Scalable Computing:
Practice and Experience

Scientific International Journal
for Parallel and Distributed Computing

ISSN: 1895-1767

Universitatea de Vest din Timişoara

Volume 25(1) January 2024

SUBSCRIPTION INFORMATION: please visit `http://www.scpe.org`

Scalable Computing: Practice and Experience

Volume 25, Number 1, January 2024

TABLE OF CONTENTS

Papers in the special issue on Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions:

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 1–10, DOI 10.12694/scpe.v25i1.2391

SENSOR NETWORK SOLUTIONS FOR AIRCRAFT ROUTE SCHEDULING AND PARKING ALLOCATION WITH LOCALIZATION AND SYNCHRONIZATION

CHUNXIN HUANG*, YAN YAO†, AND LINA WEI‡

Abstract. The Wireless Sensor Network (WSN) is the modernized version of the sensor networking, earlier the concerned networking system used to be wired. The wireless and modified features have been able to increase the efficiencies of the networking in routing scheduling of aircraft and allocating the parking. This new approach to the WSN system is not much different from the typical sensor networks framework that contains sensors, a communication system, and a controller. Instead of a communication system, a wireless protocol is applied within the sensor network. The smart system of parking has effectively implied due to its hugely innovative as well as in-ground sensors. This can monitor individual spaces of parking and able to relay the status of occupancy to the smart-sport gateways. Then, this can send the live status data to the platform of the smart cloud. This entire process enables the real-time parking data to be accessed and observed on multiple devices.

Key words: Sensor Network, Wireless Sensor Network (WSN), Aircraft Route Scheduling, Parking Allocation, Localization and Synchronization

1. Introduction. The sensor network constitutes a set of small and powered devices along with the infrastructure of both wired and wireless networks. These groups are able to record conditions following any number of environments such as farms, hospitals, and industrial facilities [21]. For vehicles and transportation systems, the use of sensor networks has been seen in the route scheduling of aircraft and in parking allocation systems for localizing and synchronizing.

From the above-presented diagrammatic representation 1.1, the versatile uses of Wireless Sensor Networking (WSN) are reflected [15]. The mentioned security and surveillance and tracking vehicles are widely used for traffic control and monitoring for land and air vehicles as well. In terms of aircraft scheduling, earlier approximately 80 percent of flight accidents were caused by machines, and nearly 20 percent of accidents happened due to human errors. The implementation of WSN in aircraft is one of the reasons behind the improved technological aspects to reduce system and machine failures [12]. Hence, today the scenario has changed entirely, nearly 80 percent of aero plane accidents occur due to human errors such as mechanics, controllers, pilots, and so on whereas machine or equipment failures related accidents are 20 percent. The major contribution of the work is as follows,

1. **Transition to Wireless Sensor Networks (WSN):** The research addresses the transition from traditional wired sensor networks to Wireless Sensor Networks (WSN). This shift reflects the advancement in technology, offering increased flexibility and efficiency in various applications.
2. **Efficient Routing Scheduling for Aircraft:** The research introduces the concept of using WSN for routing scheduling of aircraft. This innovation demonstrates the potential to enhance the management and efficiency of aircraft movements within airport premises.
3. **Parking Allocation Optimization:** The study explores the application of WSN in allocating parking spaces, particularly in the context of aircraft. This novel approach can lead to more efficient parking allocation, reducing congestion and optimizing resource utilization.
4. **In-Ground Sensor Implementation:** The research presents the innovative use of in-ground sensors for monitoring individual parking spaces. This implementation has the potential to revolutionize parking management by providing accurate real-time occupancy status.

*College of Civil Aviation, Shenyang Aerospace University,Shenyang,China,110136, email: chunxinhuang1@outlook.com
†Shenyang Aircraft Design and Research Institute, Shenyang, China, 110035
‡College of Civil Aviation, Shenyang Aerospace University,Shenyang,China,110136

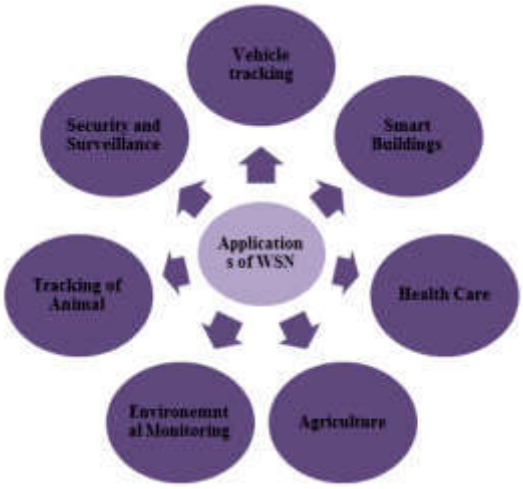


Fig. 1.1: Applications of Wireless Sensor Networks (WSN)

5. **Smart Sport Gateways Integration:** The research integrates the concept of smart sport gateways, acting as intermediaries between the in-ground sensors and the cloud platform. This integration demonstrates a holistic approach to data relay and connectivity.
6. **Real-time Parking Data:** By utilizing WSN and in-ground sensors, the research achieves the capability to gather real-time parking occupancy data. This data can be accessed by various devices, enabling users to make informed parking decisions on the go.
7. **Cloud-based Data Platform:** The study introduces the usage of a smart cloud platform to aggregate and manage the real-time parking data. This cloud-based approach enhances data accessibility, analysis, and utilization.
8. **Multi-Device Observability:** The research offers multi-device observability of real-time parking data, allowing users to access this information through different devices such as smartphones, tablets, and computers.
9. **Integration of Multiple Technologies:** The study showcases the integration of various technologies, including WSN, in-ground sensors, smart sport gateways, and cloud computing. This multidisciplinary approach demonstrates the research's innovative nature.
10. **Practical Implementation Potential:** The research's findings highlight the potential practical implementation of a smart parking system based on WSN. This has implications for modernizing parking management strategies in various domains.
11. **Data Accessibility and User Convenience:** The research contributes to the ease of access to real-time parking data, enhancing user convenience and potentially reducing congestion through more informed parking decisions.

The research introduces a novel approach to networking by applying Wireless Sensor Networks to optimize aircraft routing scheduling and parking allocation. The integration of in-ground sensors, smart sport gateways, and cloud-based platforms further enhances the capabilities of the proposed system. This innovative research opens up opportunities for more efficient and smart parking management systems, while also showcasing the potential of WSN in broader applications.

2. Objectives. The main objective of the research is to explore and demonstrate the potential benefits and practical implementation of using Wireless Sensor Networks (WSN) in optimizing aircraft routing scheduling and parking allocation. The research aims to showcase the innovative application of WSN technology and its integration with in-ground sensors, smart sport gateways, and cloud-based platforms to create a smart parking management system.

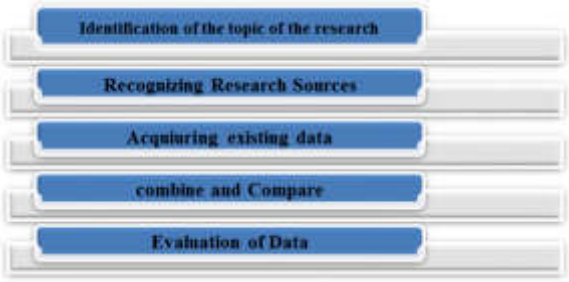


Fig. 3.1: The way of conducting Secondary research

1. To define the sensor networks for aircraft route scheduling and parking allocation with localization and synchronization
2. To find out the significance of the sensor networks
3. To detect the risk of using sensor networks in scheduling the aircraft route and localizing and synchronizing the parking route
4. To determine the potential solutions to the threats of using sensor networks
5. To analyze the architecture of the sensor networks
6. To evaluate the application of sensor networks in advanced technology

3. Methodology. A secondary qualitative research method has been applied here to find out the solutions and usage of sensor networking in terms of route scheduling of aircraft and synchronizing and localizing the parking [2]. Following this, secondary research can be described as the method of using data that already exists which can be from research articles of other researchers, books, published academic papers, statistical databases, and so on.

In the above-presented figure 3.1, the way of conducting secondary research has been presented [23]. Sensor networks and their uses in aircraft and parking have been identified as the topic. The data sources of the concerned research topic have been identified and acquired from other research articles, statistical databases, and so on. Qualitative research has been carried out on the collected secondary data that is used to explore and get deeper insights into the issues of the real-world such as sensor networking [25]. Unlike the quantitative data, the qualitative research did not collect numerical data, it helped to generate hypotheses and further research and comprehend the quantitative data research.

4. Discussion on Sensor Netwwork. Due to the technical improvements and modifications of nodes, the use of WSN has increased and is used mostly than wired sensor networks in aircraft and traffic [10]. The data rate of wireless sensor networking (WSN) ranges between 80 kb/s - 250 kb/s for operating in different areas.

The aforementioned Figure 4.1 has described that with the support of the internet, the WSN is integrated into the aircraft and traffic system in order to carry out the analysis, storage, mining, and processing. This is done through its sensors, which, lead to two sensing regions mentioned in the figure. It connects the base station with the sensors so that the data can be transcended and accessed in real time.

The aforementioned Table 4.1 described thoroughly the applications of sensor networking in aircraft and traffic [3]. In this regard, for aircraft, the use of the wireless system in sensor networking has been seen as it is more beneficial and less complicated than the wired system.

5. Impact of Sensor Networking for Aircraft Route Scheduling and Parking Allocation. Table 5.1 in the above presentation discussed several activities of sensors in the context of parking allocation [16]. The synchronization phase within two nodes is seen to be a two-way communication.

The above-presented diagrammatic representation 4 has depicted the two-way communication within nodes. Along with this, the synchronization phase continued with the sender-to-receiver communication [1]. In parking allocation, the movement can be located and synchronized by commencing with the root node and propagating

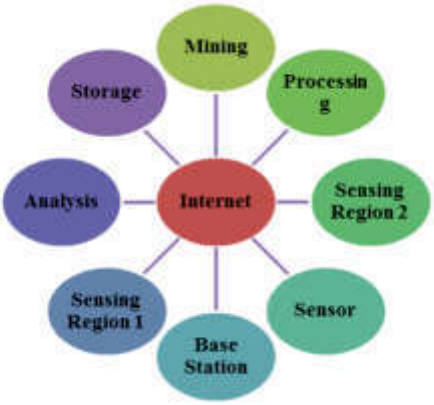


Fig. 4.1: Process and Components of WSN used for aircraft and traffic

Table 4.1: Use of WSN in aircraft and parking

Route Scheduling of Aircraft	**Localizing and Synchronizing the Parking**
Recently, the airplane monitoring system (AMS) has integrated the airborne wireless sensor networks (AWSN) to leverage the advantage, especially flexibility.	It is seen that the activity of detecting sensor nodes' physical coordinates within WSNs is known as positioning or localizing. This is considered as the key aspect in the current system of communication so that the estimation and measures can be taken regarding the place of origin of incidents.
The easy deployment system and the low-cost nature of the concerned system have made it beneficial to implement in the aircraft.	The inductive detector of the loop (ILD) sensor is acknowledged as the most commonly used sensor in terms of traffic management.
The new approach of monitoring the aircraft through the AWSN resolves many issues of wire-based tools like the efficiency of fuel, emission of carbon, and mass of flight.	Sensor networking is used for acquiring the flow of traffic, occupancy of vehicles, length, and speed.

through the network [5]. In this regard, it can be stated that the parking sensors are acknowledged as proximity sensors [9]. This is being used for designing road vehicles so that the drivers can stay alert about the obstacles of parking.

On the other hand, in terms of Aircraft Route Scheduling, the travel route scheduling schema with the mobile collector (TRP-MC) can be possible [6]. It considers a short route which can take as many sensors of AWSNs as possible [14]. In this regard, the communication system of the concerned AWSN can be used.

In the above-presented diagrammatic representation, it is reflected that the concerned communication of AWSN consists of four components, which are beyond AWSN, smart sensors, inter-AWSN, and remote servers [13]. Hence, the smart sensors within the AWSN networking can be deployed on the airplane so that the connection can be made through the AWSN. It can be stated that among all the sensor nodes, the AWSN is dependent on wireless transceivers [20]. It is because the beyond AWSN component along with the access point nodes and the gateway forms a bridge to other networks in airplanes such as portable devices, displayers in the cockpit, and system of control [17]. In this regard, it has been seen that the higher-level implementation of data is entirely based on the concerned sensory networking. The networking of satellites, ground stations, the centre of air traffic control, and the system of management all are involved in the higher level of applying data.

Table 5.1: Traffic activities by sensors

Synchronization of Traffic and Parking Allocation	**Use of sensors**
Traffic light sensors	In the traffic sensor context, the traffic light or intersection needs to be placed once, then the sensor can be able to determine the vehicle in different areas which are predefined. This would help in allocating the parking as well. After that, this is able to activate one or many relays. These relays are responsible for triggering the traffic light to be red or green.
Vehicle movement sensors	For the detection of vehicle movement, a radar antenna is equipped with a sensor of the traffic light. The movement of vehicles can determine the allocation in the smart parking system.
Localization of Nodes	Localization of the concerned sensor nodes within the WSNs is seen to be serving a significant role in monitoring the traffic. In this regard, the main aim of the localization process is to seek the coordinates of all targeted nodes through their connection with the anchor nodes. In this way, the vehicle movement can be detected by the sensor networking in traffic and verifying particular vehicles if required.

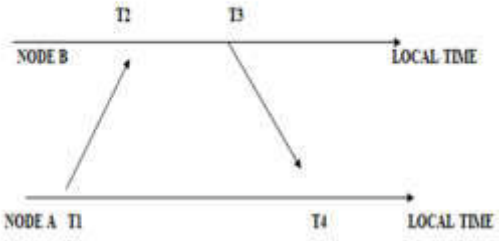


Fig. 5.1: The two-way communication between nodes in the synchronization phase

6. Potential Risks of using Sensor Networking. There are certain disadvantages to using sensor networking that poses threats in the fields the WSNs are being used. Generally, the range of WSNs is acknowledged as limited according to their design [19]. The concerned sensors are designed to work over a few hundred meters at most. Hence, if routing and scheduling for aircraft or localization and synchronizing the traffic are required to get coverage of a larger area, the WSN would fail. The use of multiple sensors can temporarily solve the issue; however, it is extremely expensive and too complicated to manage [27]. In addition, the dependence on wireless communication, the WSN is considered to be highly susceptible to interference with other devices. This can turn out to be extremely dangerous for any transportation from aircraft to road traffic. The loss of data or data corruption can happen, and even the performance of the network can get affected entirely.

The above-presented Table 6.1 discussed the issues caused by WSN to aircraft routing [7]. It is seen to be difficult to plan a reasonable route of travelling in terms of acquiring data efficiently [28]. This concerned issue is affecting people to plan a travel route without any inconvenience of scheduling other places.

7. Possible Solutions to the Threats. There might be certain disadvantages of WSN, however, the issues can be solved and the advantages of the WSN can be able to create balance to reduce the extremeness of these issues.

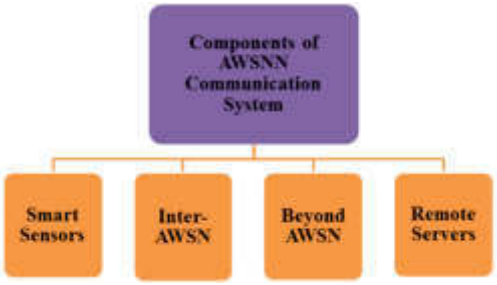


Fig. 5.2: The components of AWSN communication system

Table 6.1: Risks in Routing due to sensor networking and its design

Challenges of routing and issues in the design of WSN	**Description**
Deployment of nodes	The deployment of nodes in the WSNs is seen to be dependent on the application. This application-centric nature can potentially affect the routing protocol performance.
Considerations of energy	The sensor nodes might consume all the limited supply of energy, which is allocated for performance consumption, and transmission of information in the wireless environment.
The delivery model of data	The data delivery model is driven by events. Hence, it is prone to react immediately due to any sudden and drastic changes.

Table 7.1 has reflected the potential solutions to the rising threats from the WSNs application. In this regard, it is seen that the system is much more cost-effective than the previous wired system [11]. However, the multiple expansions can cost more in WSN, which is still effective for the budget as the initial cost was low. In addition, it would require only an operator to handle the delivered data and a technical expert to manage the entire network [29]. Therefore, the low maintenance of the concerned system makes it more effective. In terms of parking allocation, the use of electromagnetic or ultrasonic sensors has been observed in recent times. This is an extremely cost-effective and scalable system than the previously used wired networking.

8. Architecture of Sensor Networks. The sensor network is seen to contain numerous sensor nodes that are detection stations. These nodes are small, portable, and lightweight, and each is equipped with a transducer, transceiver, microcomputer, and power source [2]. This aforementioned transducer produces electrical signals, which rely on sensed physical effects and phenomena.

The block diagram of a sensor node has been depicted in the above-presented Figure 8.1 [18]. From the diagram, it can be assessed that a modular design approach of each sensor node offers much-needed flexibility as well as versatility in the concerned platforms such as aircraft and parking allocation. Through this, the requirements of a large variety of implementations can be possible [9]. The parking allocation of sensor networking is capable of handling the traffic, generating the signals according to the vehicle movement, allocating the appropriate parking to vehicles, and localizing, and synchronizing the allocation to enhance the smart parking and traffic system entirely.

9. Sensor Networking and its Use in Advanced Technology. The above-presented Table 9.1 has shown the integration of different advanced technologies with the sensor nodes of WSNs to enhance the smart parking and aircraft system [11].

Table 7.1: Potential effectiveness of WSN

Factors	Description
Energy efficiency	In terms of energy consumption, it is seen that the WSN is a new approach, which apparently consumes less energy than the traditional wired system. The battery-operative WSN system and lack of physical connections in the modernized WSN networking trigger the lesser consumption, which can be modified further targeting the zero consumption of energy.
Cost-effectiveness	It is seen to be less expensive to implement the WSN networking than the previous wired networking system. Installing the wired system would cost a lot more than inserting the WSN system. Through this concerned cost-effective process, the flexibility has increased a lot.
Scalability	The scalability of the concerned WSN system is huge and it can be expanded by adding more sensors within the structure of the network. Therefore, it is the enabler of investigating a larger area more than its limited expansion range and more events can be detected through this.

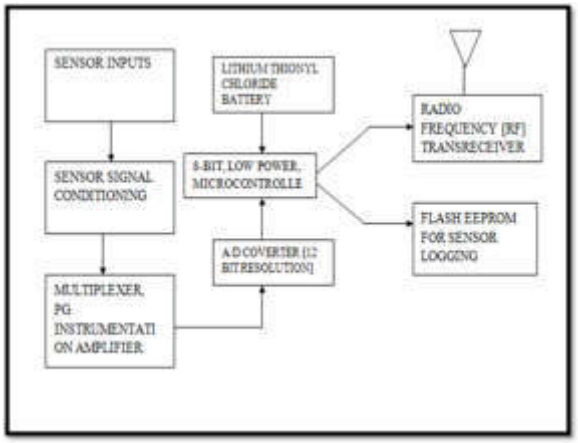


Fig. 8.1: A sensor node's functional block diagram

Table 9.1: Integration with advanced technologies

Technologies	Integration with sensory network
AI	It is seen that the combination of multiple sensors enables an AI-driven robot to detect the size, recognize an object, and locate its distance. On the other hand, AI technology can be potentially applied to detect the hidden risks of WSNs networking. The prevention of certain undetected security threats of WSNs can be resolved through the integration of AI technology.
IoT	The integration of WSNs in IoT can be able to create an infrastructure-less wireless network. This can be utilized in the deployment of a large number of wireless sensors, which can monitor the system to appropriately carry out the parking allocation as well as an aircraft routing system. Certain physical and environmental situations are also deployed through this.

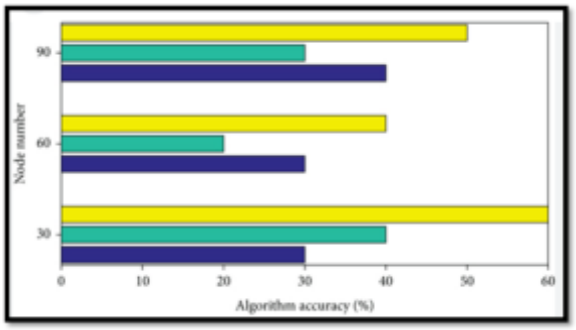


Fig. 10.1: Node number and algorithm accuracy

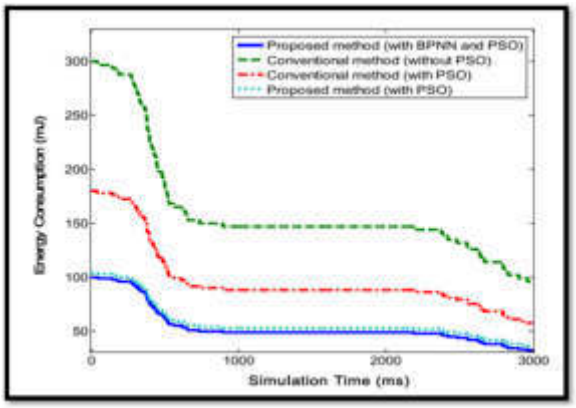


Fig. 10.2: Consumption of energy and simulation of time

10. Results. Sensor networking technology has an immense effect on controlling the routes of traffic for aircraft. In terms of efficiency and productivity increase the wireless networking system has been proven useful in sectors like parking of cars, buildings made with smart technology, health care sectors, agricultural field, monitoring of environmental issues and animal tracking.

From the figure 10.1 it can be seen that different networking signals have been colour-coded here as blue for DCCS, green for DMC-CS and yellow for EDACP-CS. These signals have been compared for their accuracy of algorithms in percentage and number of nodes [4]. For node numbers 0 to 30 or 40 the accuracy of signals is found to be 30% for DCCS, 40% for DMC-CS and 60% for EDACP-CS. The percentage values of algorithm accuracy range from 20%, 30% and 50% for node numbers varying between 45 and 75. 60 is the average node number in between and EDCAP-CS shows the maximum accuracy value and DMC-CS shows the lowest accuracy value [8]. For several nodes varying between 75 to 105 accuracy values of the algorithm varies in between 50%, 30% and 40% for three types of signals. Here also it can be seen that the highest accuracy value is seen for EDACP-CS and the lowest value of accuracy is seen for DMC-CS. From this analysis, it can be effectively concluded that for traffic signal controlling EDACP-CS signal is the most efficient as it has the most accuracy of the algorithm and DMC-CS is the least efficient network signalling method because it has the lowest number for accuracy of algorithm per number of nodes [15]. These signals are frequently used in controlling network traffic for air routes and other vehicles. In the case of air traffic controls the placed methods are highly effective when the accuracy of the algorithm is higher.

The figure 10.2 depicted above represents a stimulation of the consumption of energy and time required for the stimulation. The experiment is done to analyze and compare energy consumption in a particular

period. The results of the simulation experiments are generally verified over data collected from over 1000 stimulations [22]. The obtained data from the stimulation experiments have been analyzed and transformed into graphical forms. The data is observed and analyzed by keeping about the mathematical deviations and optimal calculations. The figure illustrates the relationship between energy expenditure in a given fixed period. As evident from the graph, it can be clearly stated that with the increase in the total number of stimulations, the consumption of energy has decreased for all the involved systems [24]. It can be said that in general terms equally distanced nodes with conventional usage with PSO use more energy as compared to BPNN and PSO which consumes 100 mJ in over 3000 ms this is because of the back propagation phenomenon. Conventional methods of the depiction of models of stimulation show more usage and consumption of energy in values of 300 mJ in 3000 ms. The same conventional methods use very less amount of energy when used in conjunction with PSO. The accuracy of energy models is summed up by the more efficient energy usage and efficiency [26]. The model of the hybrid approach has been phenomenal and most efficient in terms of energy and productivity. More energy consumption would indicate more wastage of resources which will not be beneficial in the long run. That is why it is more efficient and productive than BPNN and PSO methods used to modify the proposed methods.

11. Problem Statement. The sensor network is a significant solution to managing the aircraft route schedule as well as the allocation of parking, however, recently it is facing certain problems. This should be acknowledged as early as possible to modify the system and improve it more than earlier to be more effective. Some nodes of sensors are seen to be either failed or blocked because of damage physically, lack of power, or interference from environmental factors [7]. Only rerouting or actively adjusting powers of transmission are the only ways to handle this fault tolerance issue. However, this is decreasing the performance and productivity for routing the aircraft or allocating the parking.

12. Conclusion. The primary significance of using the wireless sensor network lies in generating authentic real-time data in each field. The aviation system and smart parking allocation matter are two important fields where the inclusion of the WSN has immensely improved the field. The data collected through concerned sensor networking can be evaluated and utilized by connecting it to the internet or the computer network. Considering the advancement of technologies in this modern era, the importance of sensors in WSNs has increased everywhere, especially in environments, phones, workplaces, and vehicles. One of the primary takeaways from this study is the remarkable role that WSN plays in ensuring the generation of accurate and real-time data. In the aviation sector, for instance, the implementation of WSN facilitates more precise and efficient aircraft routing scheduling. Real-time data collected from sensors placed at strategic points can be harnessed to monitor the movement and status of aircraft, enabling optimization of routes and schedules. This not only enhances operational efficiency but also contributes to overall safety and better resource utilization within airport premises.

Similarly, the realm of smart parking allocation has been revolutionized by the integration of WSN technology. The utilization of in-ground sensors to monitor individual parking spaces enables the continuous tracking of occupancy status. By relaying this information to smart sport gateways and ultimately to cloud platforms, users gain access to real-time parking availability updates. This not only streamlines the parking experience for users but also offers opportunities for effective parking space management, reducing congestion, and optimizing space utilization.

REFERENCES

[1] R. Ahmad, R. Wazirali, and T. Abu-Ain, *Machine learning for wireless sensor networks security: An overview of challenges and issues*, Sensors, 22 (2022), p. 4730.

[2] U. A. Bukar and M. Othman, *Architectural design, improvement, and challenges of distributed software-defined wireless sensor networks*, Wireless Personal Communications, 122 (2022), pp. 2395–2439.

[3] A. Bulashenko, S. Piltyay, Y. Kalinichenko, and O. Bulashenko, *Mathematical modeling of iris-post sections for waveguide filters, phase shifters and polarizers*, in 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), IEEE, 2020, pp. 330–336.

[4] A. Bulashenko, S. Piltyay, A. Polishchuk, and O. Bulashenko, *New traffic model of m2m technology in 5g wireless sensor networks*, in 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), IEEE, 2020, pp. 125–131.

[5] Y. Cao, C. Chen, D. St-Onge, and G. Beltrame, *Distributed tdma for mobile uwb network localization*, IEEE Internet of Things Journal, 8 (2021), pp. 13449–13464.
[6] J. Chen, C. W. Yu, and W. Ouyang, *Efficient wireless charging pad deployment in wireless rechargeable sensor networks*, IEEE Access, 8 (2020), pp. 39056–39077.
[7] Z. Deng, S. Tang, X. Deng, L. Yin, and J. Liu, *A novel location source optimization algorithm for low anchor node density wireless sensor networks*, Sensors, 21 (2021), p. 1890.
[8] A. Di Graziano, V. Marchetta, and S. Cafiso, *Structural health monitoring of asphalt pavements using smart sensor networks: A comprehensive review*, Journal of Traffic and Transportation Engineering (English Edition), 7 (2020), pp. 639–651.
[9] J. Diez-Gonzalez, R. Alvarez, N. Prieto-Fernandez, and H. Perez, *Local wireless sensor networks positioning reliability under sensor failure*, Sensors, 20 (2020), p. 1426.
[10] H. El Alami and A. Najid, *Ech: An enhanced clustering hierarchy approach to maximize lifetime of wireless sensor networks*, Ieee Access, 7 (2019), pp. 107142–107153.
[11] K. Gulati, R. S. K. Boddu, D. Kapila, S. L. Bangare, N. Chandnani, and G. Saravanan, *A review paper on wireless sensor network techniques in internet of things (iot)*, Materials Today: Proceedings, 51 (2022), pp. 161–165.
[12] D. S. Ibrahim, A. F. Mahdi, and Q. M. Yas, *Challenges and issues for wireless sensor networks: A survey*, J. Glob. Sci. Res, 6 (2021), pp. 1079–1097.
[13] Y. Jin, J. Xu, S. Wu, L. Xu, D. Yang, and K. Xia, *Bus network assisted drone scheduling for sustainable charging of wireless rechargeable sensor network*, Journal of Systems Architecture, 116 (2021), p. 102059.
[14] A. Jung, P. Schwarzbach, and O. Michler, *Future parking applications: Wireless sensor network positioning for highly automated in-house parking.*, in ICINCO, 2020, pp. 710–717.
[15] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, *Applications of wireless sensor networks: an up-to-date survey*, Applied system innovation, 3 (2020), p. 14.
[16] O. I. Khalaf and B. M. Sabbar, *An overview on wireless sensor networks and finding optimal location of nodes*, Periodicals of Engineering and Natural Sciences, 7 (2019), pp. 1096–1101.
[17] M. N. R. Khan, H. Haque, K. Labeeb, M. Aktar, R. K. Datta, and M. Z. Abedin, *Internet of things and wireless sensor network solution in smart environmental monitoring*, in 2021 6th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2021, pp. 1–5.
[18] H. Landaluce, L. Arjona, A. Perallos, F. Falcone, I. Angulo, and F. Muralter, *A review of iot sensing applications and challenges using rfid and wireless sensor networks*, Sensors, 20 (2020), p. 2495.
[19] U. K. Lilhore, O. I. Khalaf, S. Simaiya, C. A. Tavera Romero, G. M. Abdulsahib, and D. Kumar, *A depth-controlled and energy-efficient routing protocol for underwater wireless sensor networks*, International Journal of Distributed Sensor Networks, 18 (2022), p. 15501329221117118.
[20] D. Liu, Y. Xu, Y. Xu, Y. Sun, A. Anpalagan, Q. Wu, and Y. Luo, *Opportunistic data collection in cognitive wireless sensor networks: Air–ground collaborative online planning*, IEEE Internet of Things Journal, 7 (2020), pp. 8837–8851.
[21] J. Liu, P. Tong, X. Wang, B. Bai, and H. Dai, *Uav-aided data collection for information freshness in wireless sensor networks*, IEEE Transactions on Wireless Communications, 20 (2020), pp. 2368–2382.
[22] J. Lu, L. Feng, J. Yang, M. M. Hassan, A. Alelaiwi, and I. Humar, *Artificial agent: The fusion of artificial intelligence and a mobile agent for energy-efficient traffic control in wireless sensor networks*, Future generation computer systems, 95 (2019), pp. 45–51.
[23] R. S. Mphahlele and L. Mbati, *Revised research methodology guiding tool for m & d proposals.*
[24] A. Mukherjee, D. K. Jain, P. Goswami, Q. Xin, L. Yang, and J. J. Rodrigues, *Back propagation neural network based cluster head identification in mimo sensor networks for intelligent transportation systems*, IEEE Access, 8 (2020), pp. 28524–28532.
[25] M. Newman and D. Gough, *Systematic reviews in educational research: Methodology, perspectives and application*, Systematic reviews in educational research: Methodology, perspectives and application, (2020), pp. 3–22.
[26] S. Pundir, M. Wazid, D. P. Singh, A. K. Das, J. J. Rodrigues, and Y. Park, *Intrusion detection protocols in wireless sensor networks integrated to internet of things deployment: Survey and future challenges*, IEEE Access, 8 (2019), pp. 3343–3363.
[27] G. Santana Sosa, J. Santana Abril, J. Sosa, J.-A. Montiel-Nelson, and T. Bautista, *Design of a practical underwater sensor network for offshore fish farm cages*, Sensors, 20 (2020), p. 4459.
[28] A. P. Singh, A. K. Luhach, X.-Z. Gao, S. Kumar, and D. S. Roy, *Evolution of wireless sensor network design from technology centric to user centric: An architectural perspective*, International Journal of Distributed Sensor Networks, 16 (2020), p. 1550147720949138.
[29] N. Temene, C. Sergiou, C. Georgiou, and V. Vassiliou, *A survey on mobility in wireless sensor networks*, Ad Hoc Networks, 125 (2022), p. 102726.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 3, 2023
Accepted: Oct 4, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 11–24, DOI 10.12694/scpe.v25i1.2432

COMPREHENSIVE EVALUATION MODEL FOR COMPETITIVENESS OF MASS MEDIA COMPANIES IN THE IOT SENSOR NETWORKS

MENGYING XI*

Abstract. Mass media companies can be elaborated as organizations that manipulate technological components for their various departments such as movie studios, publishing houses, radio and television station management teams that impact a large range of audience via vast communication strategies. The companies have also been referred to as media conglomerates, media groups, or media houses further illustrating their grasp over the global markets and their revenue structures. The largest media companies such as Apple, Disney, and Comcast among others, offer products and services to users that are diverse individuals as well as large organizations leading to significant revenues as well as challenges that have been further explored in a comprehensive manner in the study. IoT utilizes wireless networks that are without infrastructure to install a huge number of wireless sensors that track system, physical, and environmental conditions. If you're wanting to integrate WSN into your business, our highly driven and experienced engineers can give you an all-encompassing solution.

Key words: Mass media, IOT sensors, Models, Communication strategies

1. Introduction. This study addresses the application of technological elements such as Internet technology for media-based establishments that fulfill several duties in society. Some of those responsibilities include focusing on entertainment, education, information-sharing, development of a public forum for discussion, and acting as a watchdog for governments, business, and other institutions in a vigilant manner. However, the operators associated with the mass media industry can be assessed to possess personal agendas because of political inclinations, demand for advertisement funds, differences in ideological bias, that have relatively constrained their competitive abilities. The development of technological privileges has also revolutionized traditional revenue streams, such as print advertising, that have necessitated greater strategic countermeasures to prevail in global markets. The analytical discussion on the performances of the various establishments and their impact in the prolonged period of time has further explored to identify the barriers that are required to be overcome to maintain a recurrent nature of profitable revenue stream across global digital platforms.

The rapid evolution of technology has ushered in an era of unprecedented change within the media industry. Mass media companies are navigating an intricate web of new platforms, channels, and consumer behaviors. The motivation to understand how these conglomerates harness technology to adapt and thrive in this dynamic landscape fuels our research. Exploring the financial underpinnings of media giants like Apple, Disney, and Comcast unveils a fascinating narrative of revenue diversification. At the same time, this exploration shines a light on the complex challenges they face in an era of changing consumption patterns, digital disruption, and evolving regulatory frameworks.

The essence of contemporary media hinges on immediacy and relevance. IoT sensor networks enable real-time data collection from a multitude of sources, granting mass media companies the ability to capture up-to-the-minute information, trends, and user behaviors. This real-time insight empowers agile decision-making, ensuring content delivery aligns with current audience preferences. Understanding the audience has always been central to media success. IoT sensor networks provide a panoramic view of user interactions, preferences, and consumption patterns across platforms. This nuanced understanding enables media companies to tailor content to individual preferences, thus bolstering user engagement and loyalty. IoT sensors facilitate the seamless monitoring of user experiences across diverse media platforms. By tracking user behavior and response, media companies can refine interfaces, optimize content delivery, and create personalized experiences that resonate

*Business Administration major, Beijing University of Posts and Telecommunications Zhengzhou City 450003 China (mengyingxiresc@outlook.com)

with each user.

Leveraging IoT sensor data, media conglomerates can optimize content delivery mechanisms. Insights into user engagement levels, content preferences, and viewing habits allow for targeted content recommendations and scheduling adjustments, enhancing viewer satisfaction and engagement. IoT sensors extend their influence beyond content delivery. They enable the monitoring of equipment performance, infrastructure utilization, and energy consumption. This data-driven oversight contributes to operational efficiency, reduced downtime, and informed maintenance strategies. Informed decision-making is the hallmark of successful media strategies. IoT sensor data guides strategic choices by offering empirical evidence of content performance, user engagement trends, and emerging patterns. This data-driven approach enhances the likelihood of producing content that resonates with audiences. The media landscape evolves rapidly. IoT sensor networks facilitate the monitoring of industry trends and audience preferences. This agility equips media companies to innovate and adapt swiftly, ensuring they remain relevant in the face of disruptive forces.

1. **Objectives**
2. To understand the various major mass media conglomerates that are active
3. To identify the barriers for mass media establishments
4. To explore areas of market opportunities for the development of the institutions
5. To assess the future success rate of the industry amidst globally changing trends.

The research contributes by delving into the qualitative aspects of the impact of media technologies on mass media companies. By exploring digital datasets and qualitative elements, the study uncovers nuanced areas of improvement influenced by media technologies, providing a comprehensive understanding of their influence.

2. Methodology. The development of this study has been made possible by a vast array of digital datasets and insights. The use of qualitative elements in the study has also been a crucial factor in terms of identifying the areas of improvement that have been impacted by the media technologies. The use of Internet as a medium for amassing greater range of viewership has propelled several mass media establishments to introduce drastic approaches to gain better prominence that have also been elaborated by the use of secondary datasets that have been reviewed across scholastic platforms.

Incorporating detailed examples of how prominent media companies have harnessed IoT sensor networks to transform their operations can provide concrete evidence of the technology's influence. For instance, illustrating how Disney utilized IoT devices to enhance visitor experiences in its theme parks or how news organizations integrated IoT sensors for real-time data collection during major events could vividly depict the benefits and possibilities of IoT integration. These examples not only demonstrate the applicability of IoT in the media industry but also lend credibility to the research findings.

3. Discussion on IoT sensor networks . The internet of things technology can be elaborated to have resulted in a significant level of mass media conglomerates that has provided them with a diverse manner of challenges and opportunities [28]. The comparative discussion of the approaches in which internet technology is important for mass media companies have been further illustrated below.

The growth of new revenue streams that have replaced the regressive traditional revenue stream accumulation tactics such as print advertising by media companies. In comparison, the identification of new ways to generate capital from their content, such as through online advertising, paywalls, and subscriptions have also been valuable for their revenue performances [1]. On the other hand, the internet has made it easier for new communities to engage the market, resulting in an incline in competition for established media companies.

Another element of IoT and its application also implies that media companies can collaborate with other media institutions. The linkage of various social media platforms such as Facebook, Instagram, LinkedIn as well as their influencer communities can enable greater interaction with new audiences by developing more immersive forms of content [25]. The above figure shows a glimpse of the connections of IOT connections for mass communication

It can be elaborated, however, that the internet has fragmented audiences, making it relatively difficult for media companies to expand their outreach in terms of a large, unified audience. On the other hand, the ability to conduct investments in mass media companies that diversify their monetary assets into research and development for testing out new platforms can be more successful for investing parties as well. The internet

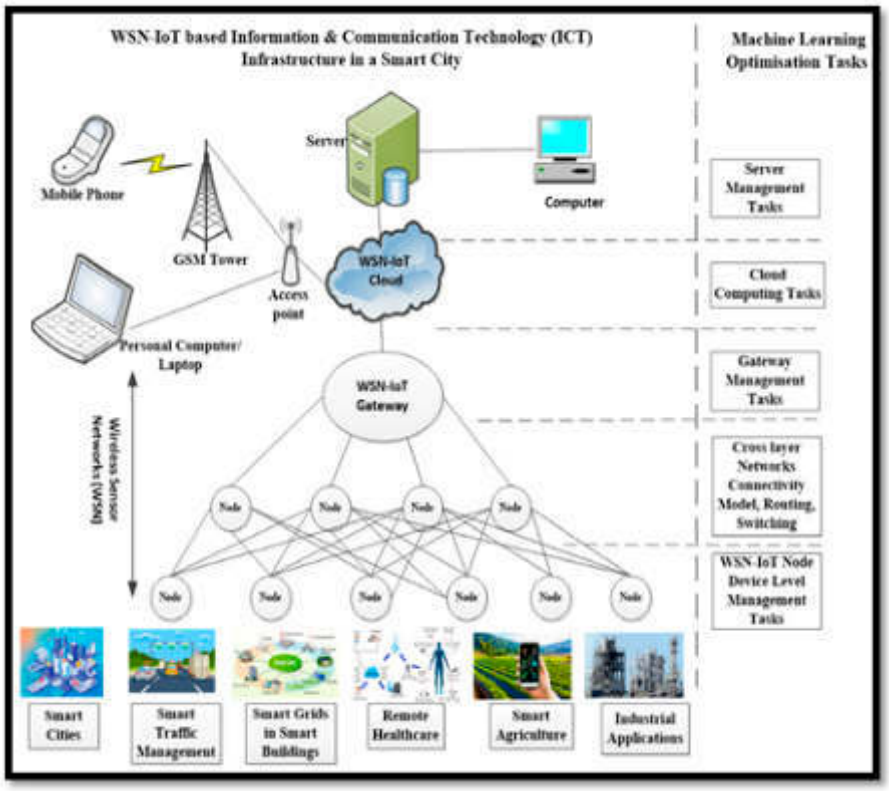


Fig. 3.1: IOT based information and communication technology [6]

technology that has revolutionized over the past decade has also been expressed to unleash a major proportion of attention-deficit-disorder among viewers leading to greater saturation in traditional and social media [14].

One of the demerits of a fractured audience relay implies that the internet has been partially able to immerse itself in the media environment to reach a more diverse and global audience, as well as to target specific demographics through social media and other online platforms. However, it can be argued that media conglomerates aspire immensely for development of policies that facilitate their control of the markets around the world and their associated digital privileges [11, 5].

The internet has yielded beneficial situations for establishments such as Google, Amazon and others that are reliant on mass media conglomerates. However, it has created new threats and setbacks that has also enabled better management of new opportunities for revenue, collaboration, and audience engagement [26, 18]. Media companies that are willing to invest in research and development and adapt to new technologies will likely be more capable of predicting performance across global platforms.

4. Performance of mass media entertainment industry. The statistical information provided above informs of the various establishments that are active across the globe owing to their prominent revenues [2]. Enterprises such as the Comcast, Meta-verse and others that have been further addressed in the study to develop more comprehensive insights. The various means of improving revenues by the global conglomerates have been further elaborated below. Mass media companies can upgrade their revenue streams by diversifying their business structures and adapting to the advent of new technologies. By implementing strategic discourses, media companies can generate revenue from diverse sources and lower their reliance on traditional revenue streams [24]. The performance of establishments such as Amazon, Apple, Comcast have been significant in terms of revenue collection as denoted by their revenue streams that have been approximated to be ranging across 250-300 billion euros.

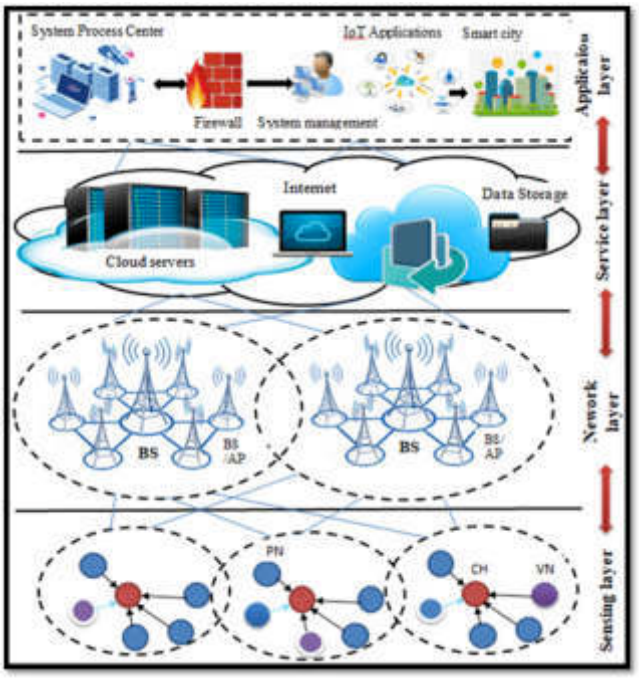


Fig. 3.2: The mass communication tech connecting Iot Applications [19]

The use of statistical data as relayed in the image above informs that establishments such as Apple, Microsoft and others control the majority of the media markets. Some of the most common forms of revenue expansion include induction of digital advertising on company websites and email subscriptions for newsletters to generate revenue. Digital media establishments also charge users for access to premium content, such as exclusive publications, videos, and podcasts that leads to better revenue accumulation [17].

One of the most lucrative approaches that are engaged by competing digital media companies is by earning commissions by promoting products or services through affiliate links. Additionally, the hosting of digital meetings to commercialize their own products or merchandise, such as books, clothing, and accessories can also be beneficial [16]. The hosting of live events can also lead to a growing number of publishers willing to generate revenue by hosting live events that leads to better market performance.

5. Barriers of mass media companies in markets. A vast number of threats and setbacks can be assessed to be poised at the integration of Internet facilities to mass media conglomerates. Barriers to internet technology adoption in mass media companies can also lower their brand image among shareholders [8]. The detrimental setbacks to Internet Technology have been further explored in the following sections. In general, the term "media" to refer to a platform via which content is disseminated from the creator to the audience. The media sector is made up of the businesses and people who produce, handle, distribute, and use this content. The Indian media business is one of the economy's fastest-growing and is expanding significantly. The sector is on the verge of initiating an even greater phase of expansion, supported by increasing consumer demand and improving revenue, showcasing its elasticity to the world. The digitalisation and internet usage in the previous ten years have led to an exponential growth in the business.

Lack of awareness: A prominent community of people can be observed to be ignorant to the use of the internet and its resultant impacts that include lowering of cost and labor [29]. The lack of awareness can act as a setback for technologically less-adept communities.

Lack of relevant areas of application: The lack of acknowledgment for the relevance of the internet in daily lives can act as a barrier to adoption.

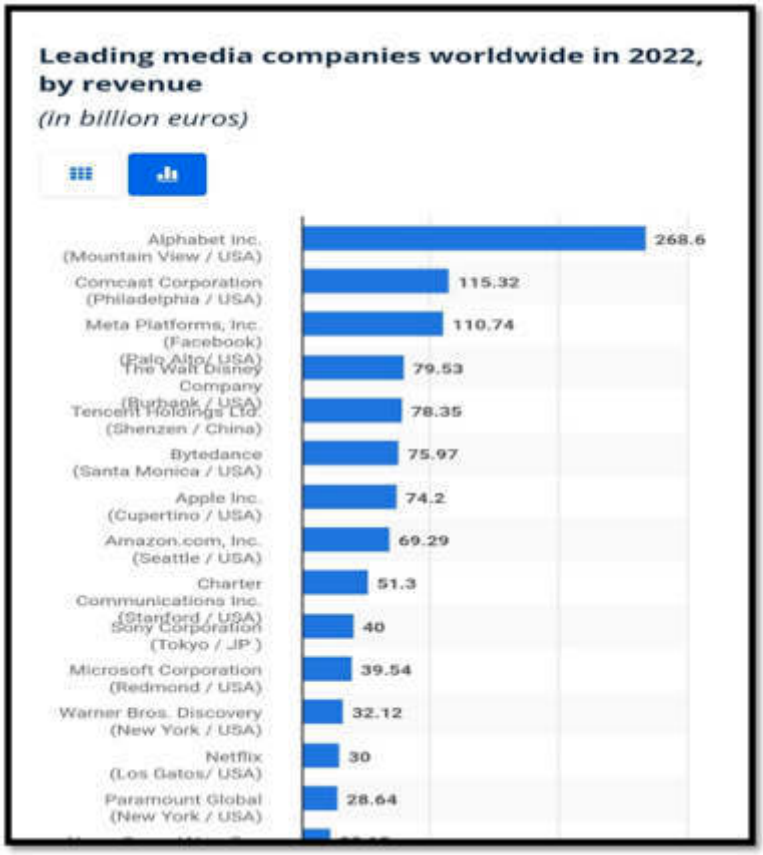


Fig. 4.1: Performance of global mass media enterprises [29])

Mismanaged infrastructure development: In some countries that are suffering in terms of economic stability, the infrastructure necessary to support internet adoption can be scarce in nature, which can act as a barrier to adoption of digital technologies [9].

Physical challenges: One of the most prominent challenges to digital application is the need for its simplification to the point where it can be used seamlessly by elders. The older community members may face physical challenges that complicates its ability to use electronic devices acting as a barrier to digital engagement.

Lack of mental confidence: It has been observed that several communities can suffer from ineffective identity issues leading to the lesser desire to operate digital services [4]. The communities that lack confidence in their ability to accrue knowledge and properly apply it while using electronic devices can further act as a barrier to adoption.

Competitive forces: The presence of competing nature of relationships across global conglomerates have been a significant barrier for several establishments in terms of introducing more lucrative deals than others. Furthermore, the internet has made it more accessible for new players to enter the market, increasing competition for pre-existing media companies [13].

Fragmentation: One of the most prominent issues in terms of mass media usages include a fragmented nature of audience with diverse trends. Additionally, this has also led to Attention deficit disorder issues for social media application users on a global scale. Since the beginning of the industry's recognition, this has been a problem. The media houses have always been concerned about issues relating to the complexity of contracts, advertising, handling (and settlement) of finances, acquisition and retention of employees and material, and ambiguity in having clients and producers on board. Additionally, while the model for advertising has historically been

The 100 largest companies in the world by market capitalization in 2023

(in billion U.S. dollars)

Ranking of the companies from 1 to 100	Market capitalization in billion U.S. dollars
Apple (U.S.)	2,746.21
Microsoft (U.S.)	2,309.84
Saudi Arabian Oil Company (Saudi Aramco, Saudi Arabia)	2,055.22
Aplhabet (U.S.)	1,340.53
Amazon (U.S.)	1,084.06
NVIDIA (U.S.)	708.4
Meta Platforms (U.S.)	599.82
Tesla (U.S.)	539
LVMH Moet, Hennessey, Louis Vuitton (France)	482.45

Fig. 4.2: Largest mass media conglomerates (Influenced by [4])

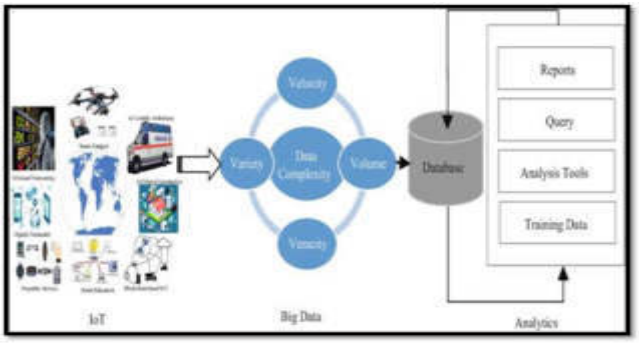


Fig. 5.1: Elements of big data analytics (Influenced by [21])

straightforward, there are still a number of components that must be put together, such as the type of media being used and whether direct or indirect advertising is being done. The result is a completely distinct need for accounting, ranging from sales through financial planning, its analysis, and financial management. This raises the possibility of errors that could have an impact on the analysis as a whole. As more and more individuals move towards digital technology, attacks like social media accounts malware, phishing attacks, and other frauds have also gotten easier and quicker to transmit as well as news, facts, and data. Currently, social media accounts are one of the most valuable assets in the media sector. Hacks can simply gain access to these accounts and spread fake information, endangering the feelings of many people and bringing negative attention to the media outlet.

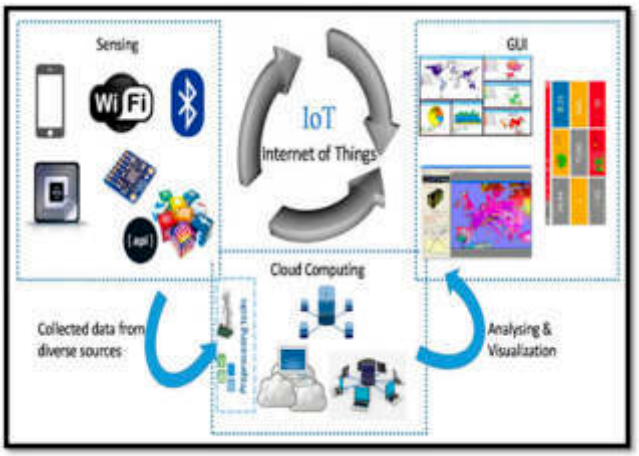


Fig. 5.2: Nature of Internet technology(Influenced by [18])

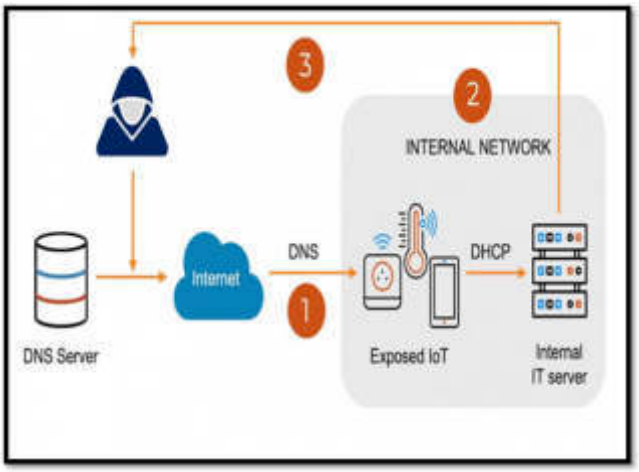


Fig. 6.1: Types of digital connections (Influenced by [9])

6. Comparative assessment of mass media institutions. The diagrammatic expression provided above informs of the various elements of digital activities necessary for maintaining a coherent nature of development of Internet Technology. The global mass media industry is dominated by a small array of vast establishments that control a significant portion of the market in a perennial nature. A comprehensive discussion as provided below illustrates some of the major global mass media companies and their performances [30].

Apple: The enterprise maintains a market capitalization rate that approximates $2.74 trillion signifying major financial benefits. Apple Corporation offers a wide selection of products and services, including smartphones, tablets, computers, and streaming services that have undergone a relatively ever-increasing form of growth.

Disney: The establishment manipulates a market capitalization of $238.21 billion elevating it to the stature of one of the largest media conglomerates in the world. It has distributed operations between four segments: media networks, parks and hospitality, studio entertainment, and consumer product management [21].

Comcast: Comcast is one of the largest global media, entertainment, and communications companies that have prevailed owing to its ability to enable collaboration across multiple platforms. The establishment

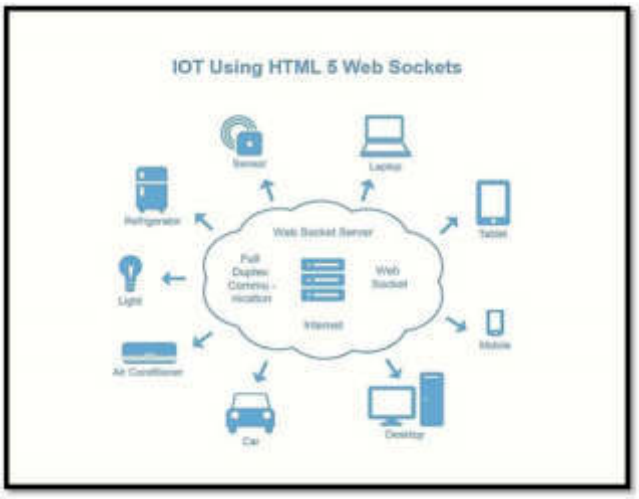


Fig. 6.2: Application of Internet services (Influenced by [7])

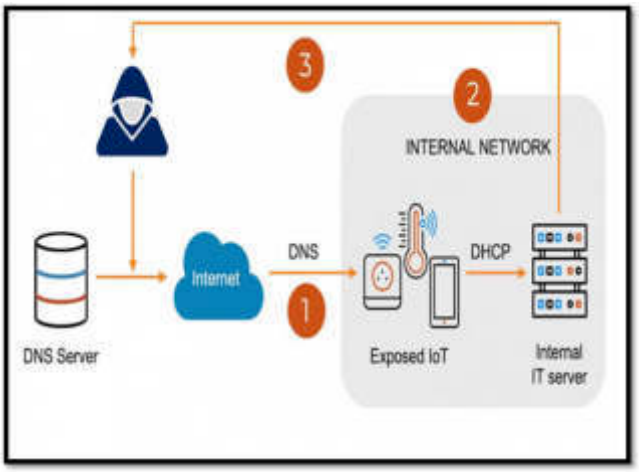


Fig. 7.1: Major mass media companies (Influenced by [27])

operates across five segments namely, cable communications, cable networks, broadcast television, motion-picture entertainment, and theme parks.

The pictographic expression provided above informs the various uses of IoT that are engaged by mass media establishments to render a greater range of competition to their institutions [20]. Some of the services can also be denoted to yield synergy of content that can have profound impacts. The internet has permitted media companies to synchronize their content by broadcasting the same insights and products across multiple platforms, which significantly aids in reducing relative first edition costs owing to the marginal internet costs [3]. The reduction in the costing of internet privileges can also be considered a motivating factor. The synchronization of IoT devices have also expanded to be able to be used within vehicles that have further led to a higher range of integration.

7. Opportunities for mass media establishments. The diagrammatic expression provided above informs about the major establishments that are currently functional on a globally immense scale and are capable of impacting the global economy in a diverse manner [22]. The adoption of internet technology has enabled

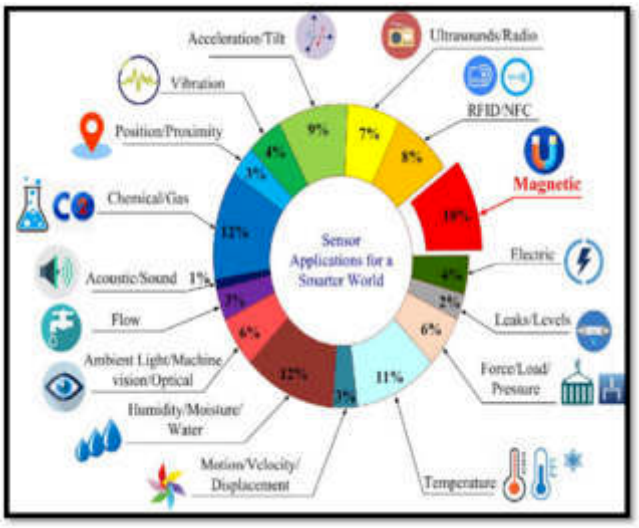


Fig. 7.2: Application of sensor technology (Influenced by [28])

media companies to interact directly with their audiences through social media and other online platforms that have further enabled establishments such as Microsoft, Apple and others to identify and target probable patterns among consumer demographics.

The diagrammatic expression provided above informs the use of sensor based data to identify digital commands that are crucial for improving digital integration. Sensor-based technology has become an intrinsically important element in the mass media industry owing to its ability to simplify human intervention to its basic stages [7]. The following are some types of sensor-based technology used in mass media that are effective today as well.

Sensor journalism: Sensor journalism implies the application of sensors to generate or collect data. The following steps include analyzing, visualizing, or using the data to support journalistic inquiry. The aforementioned approach to journalism involves the innovation of data with sensor equipment that is relatively disparate from data journalism [12].

Drones: Drones are unmanned aerial vehicles that can be stubbed with sensors to capture images and videos from a relative aerial distance. The use of it is becoming increasingly common in the media industry for aerial photography and videography services.

Location-based sensors: The use of Location-based sensors can be valuable in identifying trends in specific locales in terms of content consumption. Privileges such as GPS, used to collect data on the location of individuals and objects have also proven to be functional. The datasets expressed above can be used by media companies to create personalized content and advertising in a more subjective format [5]. The discussion of the various components used in technological developments further informs the areas of IoT that are in deft need of undergoing changes.

8. Relevance of technology in mass entertainment platforms. The diagrammatic expression provided above informs of the major challenges in mass media communications that can hamper the competitive forces active between different companies.

The diagrammatic expression provided above informs of the various establishments functional on a global scale and their ability to accumulate vast revenue structures. However, it can be expressed that there are a variety of susceptibilities that need to be addressed efficiently as further explored. The lack of transparency in the media industry, which can lead to a lack of faith from the public that hampers their brand image [6].

Media companies such as Meta, Microsoft and others are also necessitated to comply with various laws and regulations, which can be daunting due to the constant nature of unpredictability prevalent in the industry.

Fig. 8.1: Challenges in mass media industry (Influenced by [2])

The growth of new media channels, such as social media sites has led to greater threat in terms of traditional media channels.

9. Future scope of mass media industries. The diagrammatic expression provided above informs of the various sources of information that are crucial to media houses for developing a better strand of services for communities in need. One of the most valuable elements of mass media is its engagement of social media that has led to a wide range of impacts.

Social media and its adoption has led to a variety of significant impacts on mass media communication as well [23]. One of the earliest impacts include incline in audience participation that have also led to better expansion of mass media industries.

Expansion of digital reach: Social media has surpassed the reach that traditional broadcast and print media can cover as evident from its persistence in countries that relocated across ecologically diverse planes. The encapsulation of all types of audiences and communities on a global basis can be observed to contribute towards their future performance as well.

Instant communication: Social media enables instantaneous connection that permits individuals to exchange data in the form of information, and engage in conversations in an ever-present phase [19].

Formulation of public opinion: Social media can be denoted to be one of the most prominent tools for the shaping of personal and subjective outlooks. The full range of its application also has the ability to redirect public opinion and influence perceptions, since individuals share and discuss and share insights and opinions on news, and events.

The diagrammatic expression provided above informs various sources of information that are employed by mass media authorities to better manage their services across diverse platforms [26]. Some of the sources have been further discussed below.

Newspaper sources and editorial magazines: They can be considered as traditional sources of mass media information that have been prevalent for several decades. These data sources are capable of providing in-depth coverage of new developments, current events, and other topics of interest.

Television and radio broadcast channels: These can be considered as one of the oldest sources for imparting mass media information that have been active for multiple generations [10]. The aforementioned sources provide

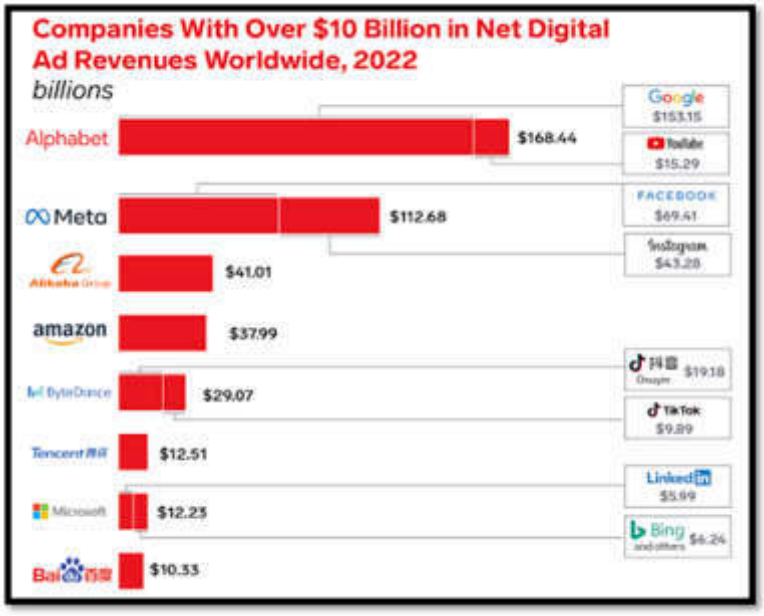


Fig. 8.2: Global mass media companies (Influenced by [25])

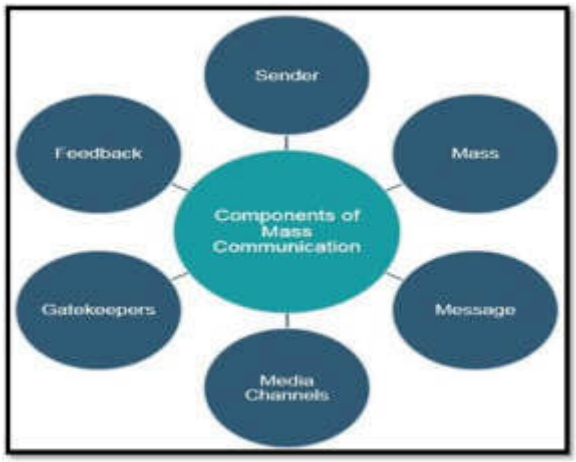


Fig. 9.1: Components of mass communication(Source: [16])

news, entertainment, and other programs that are demanded for viewership by a wide audience.

Internet: The advent of Internet of Things (IoT) technology has become a major source for accumulating mass media information in recent years. It provides unrestricted access to news, entertainment, and other content on an international scale.

Social media: Social media platforms that have developed owing to technologies such as Facebook, Twitter, and Instagram have become increasingly trending sources of mass media information [27]. They have also been known to allow users to share news, opinions, and other content with a wide audience leading to a relatively better coherence in sharing insights.

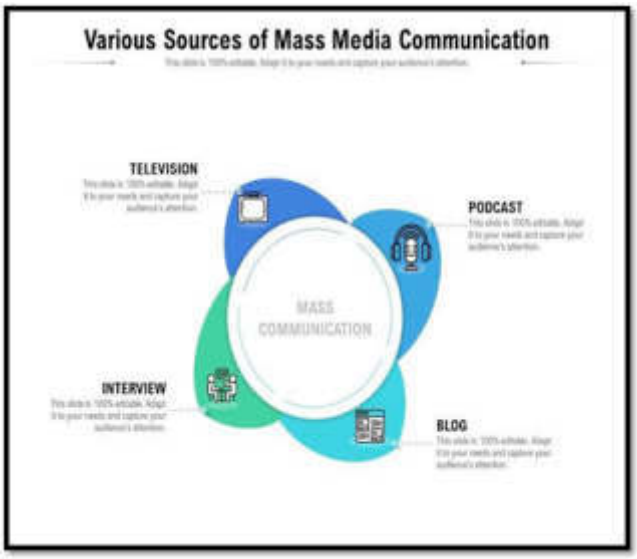


Fig. 9.2: Sources of mass media communication(Influenced by [13])

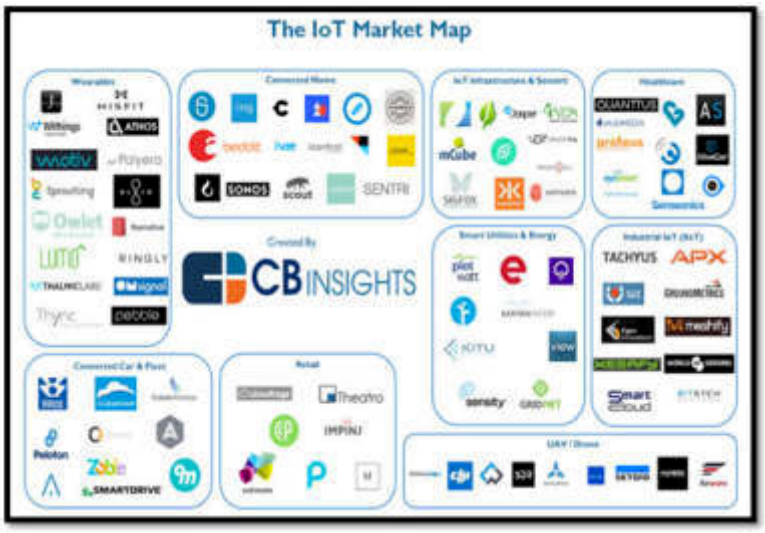


Fig. 10.1: Faculties registered to IoT (Influenced by [23])

10. Results. The diagrammatic expression provided above informs of the IoT market that is prevalent on a global scale. Establishments such as Amazon, Microsoft, and Apple among others are conglomerates that have been able to foster stakeholder trust by maintaining an annual nature of incline in their revenue generation.

The diagrammatic expression provided above informs of the basic privileges that are received from the application of digital technologies in the mass media industry [15]. Its compatibility with digital components such as Machine learning by the use of algorithms and models have also elevated the range of its services among users.

11. Conclusion. This research sheds light on the competitive dynamics that characterize the global mass media industry landscape. Through an analytical exploration of industry giants such as Google, Microsoft, and

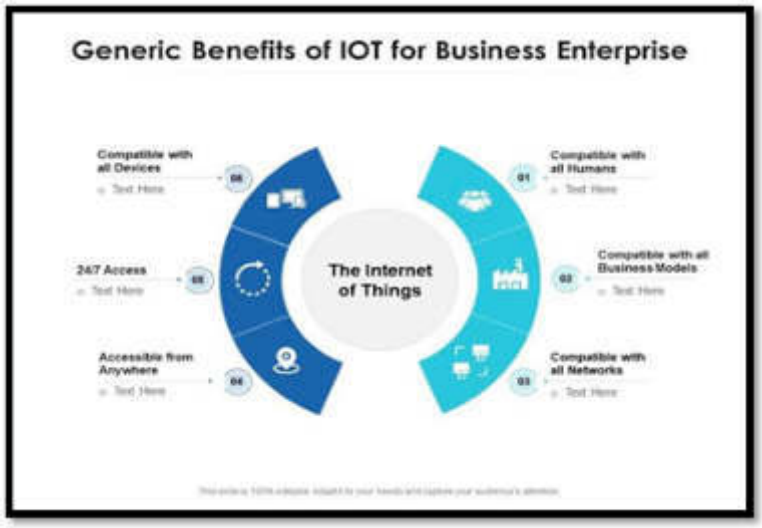


Fig. 10.2: Benefits of IoT technology (Influenced by [1])

others, a deeper comprehension of mass media technologies has emerged. Furthermore, the study delves into the role of social media applications as integral components of mass media manipulation strategies. However, it's important to acknowledge the limitations associated with Internet of Things (IoT) technology. The study recognizes that certain barriers can hinder its effectiveness in various aspects. Factors such as data privacy concerns, interoperability challenges, and potential security vulnerabilities need to be addressed to fully unlock the potential of IoT in the mass media sector.

Looking ahead, there are promising avenues for future research. Exploring innovative approaches to mitigate the limitations of IoT in mass media, such as developing robust data protection frameworks and enhancing device compatibility, holds significant potential. Additionally, investigating the evolving landscape of media consumption patterns and audience behaviors in the context of IoT-driven media experiences could yield valuable insights for industry stakeholders. This study offers a multifaceted perspective on the mass media industry's competitive landscape and the role of IoT technology within it. While acknowledging limitations, the research also highlights the prospects for overcoming challenges and shaping the future of mass media through strategic adaptation and technological advancements.

REFERENCES

[1] S. H. H. Al-Taai, H. A. Kanber, and W. A. M. al Dulaimi, *The importance of using the internet of things in education*, International Journal of Emerging Technologies in Learning (Online), 18 (2023), p. 19.

[2] L. J. Albert, S. Rodan, N. Aggarwal, and T. R. Hill, *Gender and generational differences in consumers' perceptions of internet of things (iot) devices.*, E-Journal of Social & Behavioural Research in Business, 10 (2019).

[3] L. Amodu, O. Omojola, N. Okorie, B. Adeyeye, and E. Adesina, *Potentials of internet of things for effective public relations activities: Are professionals ready?*, Cogent Business & Management, 6 (2019), p. 1683951.

[4] S. O. Amran, M. D. Hendra, A. C. Triyandra, A. S. Putera, and A. Adriyani, *Adoption of mass media technology on industry 4.0 perspective*, Jurnal Ranah Komunikasi (JRK), 7 (2023), pp. 25–32.

[5] V. B, M. S, P. N, J. L, N. V, and K. S, *Artificial conversational entity with regional language*, in 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022, pp. 1–6.

[6] M. Basingab, *Investigating the adoption of internet of things technology using agent-based simulation*, in Proceedings of the International Conference on Industrial Engineering & Operations Management, Bangkok, Thailand, 2019, pp. 5–7.

[7] A. Benis, *Social media and the internet of things for emergency and disaster medicine management.*, 2022.

[8] A. Hidayat, V. A. Wardhany, A. S. Nugroho, S. Hakim, M. Jhoswanda, I. N. Syamsiana, N. A. Agustina, et al., *Designing iot-based independent pulse oximetry kit as an early detection tool for covid-19 symptoms*, in 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), IEEE, 2020, pp. 443–448.

[9] I. Irwandi, I. Sari, R. Oktavia, and M. Syukri, *Mems and iot applications in isle-based stem physics learning media*

for mechanics topic with labview integration, in Journal of Physics: Conference Series, vol. 1462, IOP Publishing, 2020, p. 012066.
[10] J. I. Khan, J. Khan, F. Ali, F. Ullah, J. Bacha, and S. Lee, *Artificial intelligence and internet of things (ai-iot) technologies in response to covid-19 pandemic: A systematic review*, Ieee Access, 10 (2022), pp. 62613–62660.
[11] W.-S. Kim, W.-S. Lee, and Y.-J. Kim, *A review of the applications of the internet of things (iot) for agricultural automation*, Journal of Biosystems Engineering, 45 (2020), pp. 385–400.
[12] E. Korneeva, N. Olinder, and W. Strielkowski, *Consumer attitudes to the smart home technologies and the internet of things (iot)*, Energies, 14 (2021), p. 7913.
[13] A. Korte, V. Tiberius, and A. Brem, *Internet of things (iot) technology research in business and management literature: results from a co-citation analysis*, Journal of Theoretical and Applied Electronic Commerce Research, 16 (2021), pp. 2073–2090.
[14] L. H. Larsen and J. Mensa-Annan, *A critical techno-anthropological view on the iot in danish media*, (2020).
[15] M.-H. Maras and A. S. Wandt, *Enabling mass surveillance: data aggregation in the age of big data and the internet of things*, Journal of Cyber Policy, 4 (2019), pp. 160–177.
[16] M. Maroufi, R. Abdolee, and B. M. Tazekand, *On the convergence of blockchain and internet of things (iot) technologies*, arXiv preprint arXiv:1904.01936, (2019).
[17] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, and A. Zanella, *Iot: Internet of threats? a survey of practical security vulnerabilities in real iot devices*, IEEE Internet of Things Journal, 6 (2019), pp. 8182–8201.
[18] N. Misra, Y. Dixit, A. Al-Mallahi, M. S. Bhullar, R. Upadhyay, and A. Martynenko, *Iot, big data, and artificial intelligence in agriculture and food industry*, IEEE Internet of things Journal, 9 (2020), pp. 6305–6324.
[19] D. O. Okocha and D. O. Monday, *Public relations in the digital age: Implications for nigerian public relations practitioners*, The Social and Management Scientist, 14 (2023), pp. 12–22.
[20] P. Qian, B. Feng, D. Zhang, X. Tian, and Y. Si, *Iot-based approach to condition monitoring of the wave power generation system*, IET Renewable Power Generation, 13 (2019), pp. 2207–2214.
[21] A. Rozalena, M. Sulaeman, S. Mulyati, and H. Gunawan, *Business communication skill model based on internet of thing (iot)*, in Journal of Physics: Conference Series, vol. 1477, IOP Publishing, 2020, p. 072010.
[22] K. P. Seng, L. M. Ang, and E. Ngharamike, *Artificial intelligence internet of things: A new paradigm of distributed sensor networks*, International Journal of Distributed Sensor Networks, 18 (2022), p. 15501477211062835.
[23] H. Sequeiros, T. Oliveira, and M. A. Thomas, *The impact of iot smart home services on psychological well-being*, Information Systems Frontiers, (2021), pp. 1–18.
[24] Y. Shi, A. B. Siddik, M. Masukujjaman, G. Zheng, M. Hamayun, and A. M. Ibrahim, *The antecedents of willingness to adopt and pay for the iot in the agricultural industry: An application of the utaut 2 theory*, Sustainability, 14 (2022), p. 6640.
[25] K. Vaigandla, N. Azmi, and R. Karne, *Investigation on intrusion detection systems (idss) in iot*, International Journal of Emerging Trends in Engineering Research, 10 (2022).
[26] B. Vivek, A. Arulmurugan, S. Maheswaran, S. Dhamodharan, A. S. Dharunash, and N. Gowtham, *Design and implementation of physical unclonable function in field programmable gate array*, in 2023 8th International Conference on Communication and Electronics Systems (ICCES), 2023, pp. 152–158.
[27] M. G. S. Wicaksono, E. Suryani, and R. A. Hendrawan, *Increasing productivity of rice plants based on iot (internet of things) to realize smart agriculture using system thinking approach*, Procedia Computer Science, 197 (2022), pp. 607–616.
[28] A. R. Yanes, P. Martinez, and R. Ahmad, *Towards automated aquaponics: A review on monitoring, iot, and smart systems*, Journal of Cleaner Production, 263 (2020), p. 121571.
[29] M. Yousif, C. Hewage, and L. Nawaf, *Iot technologies during and beyond covid-19: A comprehensive review*, Future Internet, 13 (2021), p. 105.
[30] X. Yu, *Research on the strategy of multi ethnic culture blending in school field based on internet of things*, in Journal of Physics: Conference Series, vol. 1744, IOP Publishing, 2021, p. 042022.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 21, 2023
Accepted: Sep 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 25–33, DOI 10.12694/scpe.v25i1.2152

A SECURE METHOD OF COMMUNICATION THROUGH BB84 PROTOCOL IN QUANTUM KEY DISTRIBUTION

CHUNDURU ANILKUMAR *, SWATHI LENKA †, N. NEELIMA ‡, AND SATHISHKUMAR V E §

Abstract. Security awareness is one of the most pressing topics in today's globe. The idea of cryptography is introduced when the subject is information security. Conventional cryptography-based security techniques rely on the presumption that keys are shared before secure connections. The most crucial factor to consider when integrating cryptographic operations into account when integrating cryptographic operations in with any system is the safe key management strategy required for sending and transferring a secret key between two entities The systems will be vulnerable to bugs and possibly fatal external assaults if the fundamental management methods are poor A method for securely encrypting data sent between parties is quantum cryptography. and spotting eavesdroppers trying to overhear the conversation. Quantum cryptography may be the solution to these issues a quantum cryptography application, Quantum Key Distribution (QKD), refers to the production of a cryptographic key with unconditional security assured by physical rules. Quantum cryptography is a kind of encryption. We examine the quantum key exchange protocol (BB84 protocol) in this study and the way that it significantly improves data transfer security when compared to standard encryption techniques. The main objective of quantum cryptography is to offer a trustworthy way to provide a secure method of communication between the intended peers only and to detect the Eavesdropper presence.

Key words: Security, Cryptography, Quantum Cryptography, Rivert Shamir Adleman Algorithm, Shor's Algorithm, Quantum Key Distribution, BB84 protocol.

1. Introduction.

The emergence of the implementation of quantum computers brings significant risks to current encryption methods; for example, the implementing the Shor algorithm can obsolete in a very short period. This has probably led us to seek alternative methods of encrypting data with a greater degree of safety. To encrypt messages, provable secure cryptosystems (for example, OTP) rely on the exchange of a secret key between sender and receiver. Quantum cryptography, sometimes referred to as quantum encryption, uses quantum physics to encrypt communications so that only the end user can decipher them.

Photons and their inherent quantum characteristics are used in quantum cryptography to create a secure cryptosystem. While the quantum state of any entity cannot be determined without destroying it, quantum cryptography relies on the usage of photons and their intrinsic quantum features to create an unbreakable cryptosystem. These are the optical fiber cable data transmitters, a trustworthy channel for communications with extremely high bandwidth. The fundamentals of quantum physics indicate that noticing a quantum state causes disruption. Because of the various QKD techniques, any possible listener intending to track the delivered photons will interfere with the communication. This interference will cause transmission issues, which authorized users would be capable of identifying. That has been done to guarantee the safety of the given keys.

1.1. Quantum Cryptography vs. Traditional Cryptography.

A classical bit is the fundamental element of traditional computation and information systems. Similarly, the basic element of quantum information as well as quantum computation systems is the qubit, a term invented by Benjamin Schumacher.

In a traditional system, a bit can be either 0 or 1. A qubit has two basic states in quantum systems, which are expressed as $|0\rangle$ or $|1\rangle$, where $|\ \rangle$ is Dirac notation.

*Department of Information Technology, GMR Institute of Technology, Rajam, Andhra Pradesh, 532127, India, ORCID ID:0000-0002-3537-127X (anilkumar.ch@gmrit.edu.in)

†Department of Information Technology, GMR Institute of Technology, Rajam, Andhra Pradesh, 532127, India, ORCID ID: 0009-0004-8240-2560 (swathi.l@gmrit.edu.in)

‡Department of Information Technology, RVR & JC College of Engineering, Guntur, Chowdavaram, Andhra Pradesh, India (neelimanalla1979@gmail.com)

§Department of Computing and Information Systems, Sunway University, 47500, Petaling Jaya, Selangor Darul Ehsan, Malaysia (sathishv@sunway.edu.my)

$$|\Psi 00 = |0\ ,$$
$$|\Psi 10\ = |1\ ,$$
$$|\Psi 01 = |+ = 1/\sqrt{2}|0\ + 1/\sqrt{2}|1\ ,$$
$$|\Psi 11 = |- = 1/\sqrt{2}|0\ - 1/\sqrt{2}|1\ .$$

The traditional form of cryptography is the method of mathematically encrypting the message so that only one person with the correct key can read it. There are two types of key distribution in traditional cryptography: symmetric key and asymmetric key [4]. Asymmetric cryptography encrypts communications using a public key and decodes them employing a private key, in contrast to symmetric key algorithms that decrypt and encrypt data with a unique secret key. Traditional cryptography techniques have been trusted because it would take classical computers an impractical amount of time to factor the required large numbers, that are required to make up both private and public keys [18].

Unlike conventional encryption, quantum cryptography is based on the ideas of quantum physics. And, unlike traditional cryptography, which is rooted in mathematical equations and calculations quantum cryptography is considerably more difficult to decrypt because perceiving the involved photons alters the expected outcome, alerting both the sender as well as the recipient to the involvement of an eavesdropper [15]. Because the process requires fiber optic cables, as well as repeaters, as well as repeaters distributed out to boost the signal, quantum cryptography usually has a distance or field of view associated with it.

Existing encryption systems, on the other hand, are threatened by quantum algorithms. Another quantum approach capable of defeating symmetric encryption is the Grover algorithm. For instance, the popular Shor method can decipher asymmetric encryption schemes like Elliptic Curve and RSA. Whereas key exchanges are protected by quantum physics in quantum cryptography [8]. Moreover, the security of ordinary encryption is threatened by insecure random key generators, increases in CPU power, innovative attack strategies, and the development of quantum computers. Such encrypted information has no significance in the case of quantum computers. In the future, quantum computers will be able to intercept and preserve encrypted data for decryption [6]. Quantum cryptography has the advantages of "unconditional security" and Eavesdropper detection. These qualities may be useful in addressing cyberspace security issues for the next-generation internet and associated applications like the internet of things as well as smart cities.

- RSA algorithm is implemented as an example of a Conventional Key Exchange algorithm.
- The Shor's algorithm is implemented to show how Quantum algorithms (Shor's) breaks Conventional algorithm.
- Quantum Key Distribution in conventional cryptography provides a secure method of communication.
- Data transmission security is elevated to a greater level via QKD.
- Python tools like Qiskit are used to implement QKD using the BB84 protocol.

The remaining sections of this paper are demonstrated as follows: Introduction to Quantum Cryptography, Section 2: Using the BB84 protocol for quantum key distribution and then the comparison of Quantum cryptography and traditional cryptography, Section 3: The related work needed to the paper, Section 4: The process for distributing quantum keys is described, Section 5: The results are analysed and explained, section 6: It encloses the conclusion.

1.2. The BB84 Protocol for Quantum Key Distribution (QKD). The four-state BB84 protocol is integrated with the quantum key distribution (QKD) technique, a quantum cryptography technique, by assuming an ideal quantum channel atmosphere in which the eavesdropper is the only factor contributing to QBER greater than zero. N binary bits are first generated by Alice and must be transferred to Bob. Alice randomly selects a polarization basis from the diagonal () or rectangle (+) to encrypt a binary bit into a qubit.

Binary data 0 and 1 might be represented on the rectangular basis, for instance, by a qubit having polarizations. As a result, a qubit having polarizations can diagonally denote 1 and 0, respectively. The no-cloning theorem, which states that any arbitrarily defined unknown quantum state cannot be flawlessly copied, ensures this.

The data is encoded as non-orthogonal qubits, which is essential for detecting eavesdropping. Naturally, Eve may try to capture those quantum carriers and measure them. She is unaware of the precise group of carriers Alice pre-selected for each important component, just like Bob. She could be unable to distinguish

Table 1.1: BB84 protocol polarization scheme

Basis	Bit	Polarization of Photon
Rectangular (+)	0	→
	1	↑
Diagonal (x)	0	↖
	1	↗

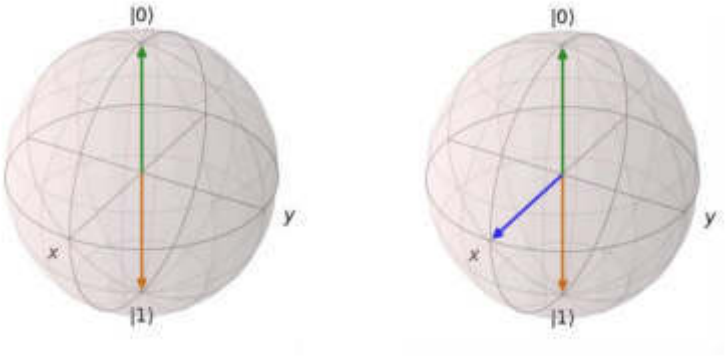


Fig. 1.1: Representation of Qubits as Bloch sphere

between $|0\rangle$ and $|1\rangle$, when Alice encrypts a bit as $|+\rangle$ or $|-\rangle$, or vice versa, just like Bob. But in contrast quantum cryptography, this usually involves Alice (A) and Bob (B) wishing to exchange confidential details while eavesdropper Eve (E) tries to intercept the message without even being detected. The main objective of quantum cryptography is to offer an effective means of detecting Eve's activity.

1.3. Qubits Representation and their Properties. The BB84 protocol is a well-known quantum key distribution (QKD) protocol that enables secure communication between two parties using the principles of quantum mechanics. It uses qubits, the fundamental units of quantum information, to encode information in a secure and tamper-evident manner. This report aims to explain the representation of qubit states through the BB84 protocol and discuss their properties[4].

In the BB84 protocol, qubit states are used to encode information. A qubit can exist in a superposition of two basis states, usually denoted as $|0\rangle$ and $|1\rangle$. These basis states correspond to the classical bit states 0 and 1, respectively. The qubit states can be represented as linear combinations of the basis states, such as $\alpha|0\rangle + \beta|1\rangle$, where α and β are complex probability amplitudes that satisfy the normalization condition $|\alpha|^2 + |\beta|^2 = 1$.

The BB84 protocol involves the following steps: Qubit Preparation: The sender (Alice) prepares a series of qubits in random states. These qubits can be in either the $|0\rangle$, $|1\rangle$, or superposition states. Qubit Transmission: Alice sends the prepared qubits to the receiver (Bob) through a quantum channel, which can be a physical medium like optical fibers. Measurement Basis Selection: Bob randomly chooses a measurement basis for each received qubit from two possible options, denoted as the computational basis ($|0\rangle$, $|1\rangle$) and the Hadamard basis ($|+\rangle$, $|-\rangle$). d. Measurement: Bob measures each qubit in the chosen basis and obtains classical measurement outcomes [5]. Basis Announcement: Alice and Bob publicly communicate the bases they used for each qubit transmission but not the measurement outcomes. Key Generation: Alice and Bob retain the bits for which their measurement bases matched, forming a shared secret key for secure communication.

In the BB84 protocol, different qubit states are used to encode information. These states have specific properties that contribute to the security of the protocol: $|0\rangle$ and $|1\rangle$ States: These are the computational basis states and represent the classical bit states. They are orthogonal and form the basis for secure key generation in the protocol.

B $|+\rangle$ and $|-\rangle$ States: These are the superposition states in the Hadamard basis. They are also orthogonal and provide a second basis for key generation. The $|+$ state represents an equal superposition of $|0\rangle$ and $|1\rangle$, while the $|-\rangle$ state represents their difference [6].

Randomness and Security: The security of the BB84 protocol relies on the randomness of the qubit states chosen by Alice and the measurement bases chosen by Bob. The random choice of states ensures the security of the key against eavesdropping attempts.

2. Related Work. The application of a unique quantum key distribution (BB84 protocol) and the way it may be utilized with conventional encryption methods to increase the security of data transmission. Moreover, it compares the encryption, decryption, avalanche impact, and performance of both QKD free versions - and QKD of these operations to assess the performance of various cryptographic techniques for a variety of file sizes. This work explores quantum cryptography is possible uses in secure communication systems, building on earlier research into the subject [5, 9, 25]. Solid evidence that uses a communicative architectural model and execution to mimic the concepts of quantum physics. It employs both the presence and absence of an eavesdropper in the quantum key distribution (QKD), implemented with BB84 protocol. Heisenberg's uncertainty principle and no-cloning principle can be utilized to find an eavesdropper, according to simulation findings. according to simulation results, although the chances of them accurately guessing which polarization state to listen in on is quite tiny [1, 13]. Quantum computing's current status of development and its uses in cryptography. It looks at the resistance of current encryption techniques to quantum computing and how the quantum computer can be employed to predict secret keys for communication decryption. The development of an application that enables users to utilize this technique to decode encrypted communications is also covered [17, 12]. Quantum computing algorithms, particularly Shor's algorithm, will be examined in this session to see how they may be used instead of conventional techniques to break encryption systems. In order to evaluate the efficacy of various quantum computing techniques, it will also examine the topics of storage capability, computation time precision, correctness, integrity, availability, and efficiency [14]. Factorial quantum technique for RSA cracking is presented in this paper without specifically calculating the modulus of n. Its foundation is the phase estimation and quantum inverse Fourier transform. The Shanks' SQUARE Form Factorization method, the Lehman methodology, and there have been several investigations on the RSA Quantum Polynomial-Time Fixed-Point Attack and compared to this strategy as approaches to the Integer Factorization Problem (IFP) [23]. Extensive overviews of cutting-edge QKD-protected optical networks that will have an impact on communication networks in the coming decades. The fundamental setup technique is described, as well as the procedures and methods used in QKD-protected optical networks. It contains a full explanation and comparison of the many ways proposed in the literature to manage networking-related difficulties [24]. The application of wireless body sensor networks (WBSN) for remote medical surveillance during the COVID-19 pandemic. Following an examination of the most recent security vulnerabilities to WBSN data, a unique upgraded BB84 Quantum Cryptography Protocol (EBB84QCP) is proposed as an effective way for safe key distribution without the direct exchange of secret keys [10]. Current state of research in post-quantum cryptography and quantum key distribution (QKD) approaches. This work employs QKD to improve current encryption protocols such as Rivest-Shamir-Adleman (RSA) and render them more resistant to quantum computer assaults. The paper also discusses how utilizing a QKD protocol to initialize may assist avoid brute force attacks by trying to prevent Eve from learning N and breaking the protocol via a brute force technique [20].

For authentication, QKD employs the PRF (Hash, Once) MAC paradigm. Because of the variety of functionality it offers, this MAC is suited for QKD. Yet, PRF is more important than the Wegman-Carter paradigm, one of most popular MAC approach in QKD (Hash, Nonce). It ensures eternal security, which implies that even with unlimited computational power, the attacker cannot learn any additional knowledge about the generated keys as far as authentication is not interrupted during QKD execution [19, 16, 11]. The Bennett-Brassard-84 (BB84) quantum key distribution (QKD) protocol's upper bounds on false-negative and false-positive ratios for eavesdropping detection are examined in this study. In order to deal with the constantly shifting quantum channel circumstances, it additionally offers a clustered BB84 protocol and a combinatorial eavesdropping detection method. The authors conducted a detailed simulation analysis to evaluate their proposed methodologies. The results showed that they can detect eavesdroppers with a minimum of 99.92% accuracy in such situations [22, 21].

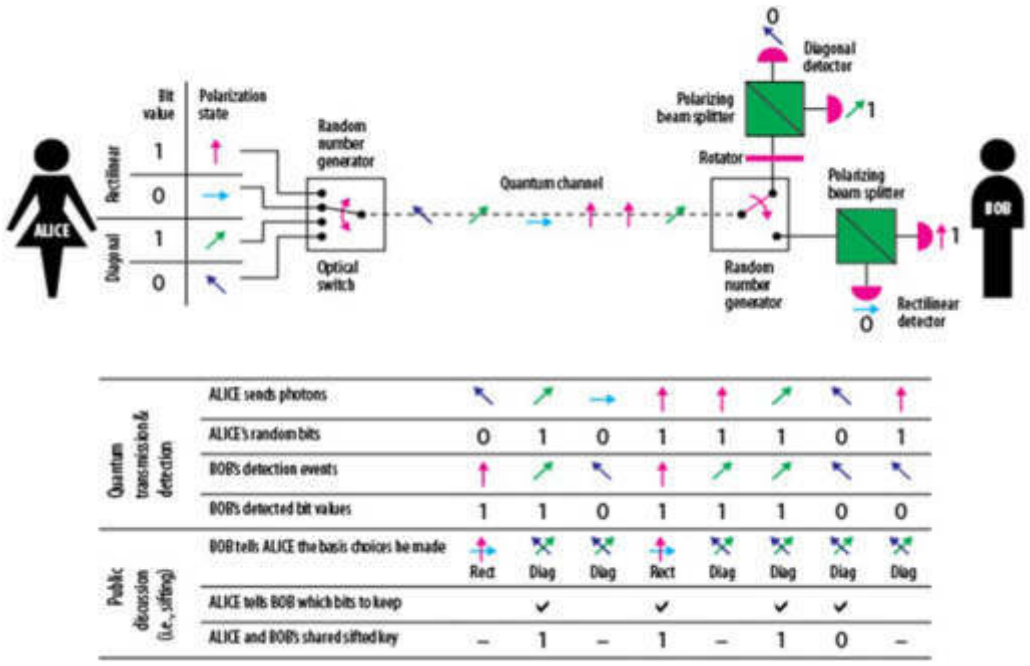


Fig. 2.1: Working process of the bb84 protocol

3. Proposed Methodology. The Quantum Key Distribution is integrated with the BB84 protocol to focus more on security proof. To generate the key, two people, Alice and Bob, employ quantum signals known as quantum bits, or simply qubits. Each attempt by an eavesdropper (say, Eve) to obtain the key causes a disturbance in the quantum signal, which eventually leads to Eve's discovery. Our project's major goal is to offer a method of secure communication only between the two intended communicating peers namely, Alice and Bob. This communication achieves security with the secure transmission of a secret key only. The major steps involved in this methodology are:

1. Key Generation
2. Key Sifting
3. Key distillation

Key Generation.

- The emitter transmits a photon whose polarization is chosen at random among the four states for each bit. He keeps track of the orientation in a list.
- The photon is sent across the quantum channel.
- The receiver sets the direction horizontal or diagonal of a filter that allows it to differentiate between two polarization states at random for each incoming photon. He keeps track of these orientations as well as the results of the detections — photons deflected to the right or left.

Key Sifting. This information is used by the emitter to compare the orientation of the photons he has delivered with the matching filter orientation. He informs the recipient in which circumstances the orientations are compatible and which are not. As seen in the BB84 protocol diagram, following sifting, the two parties have a sequence of bits known as the sifted key, that are identical in the absence of an eavesdropper. and can function as a secret key.

Key Distillation. If no eavesdropper was present during the transmission and the apparatus utilized was optimal, the key should be error-free after Key Sifting. To avoid risking the key's security, these mistakes are all attributed to the eavesdropper. Following that, a post-processing procedure called as Key Distillation is carried out.

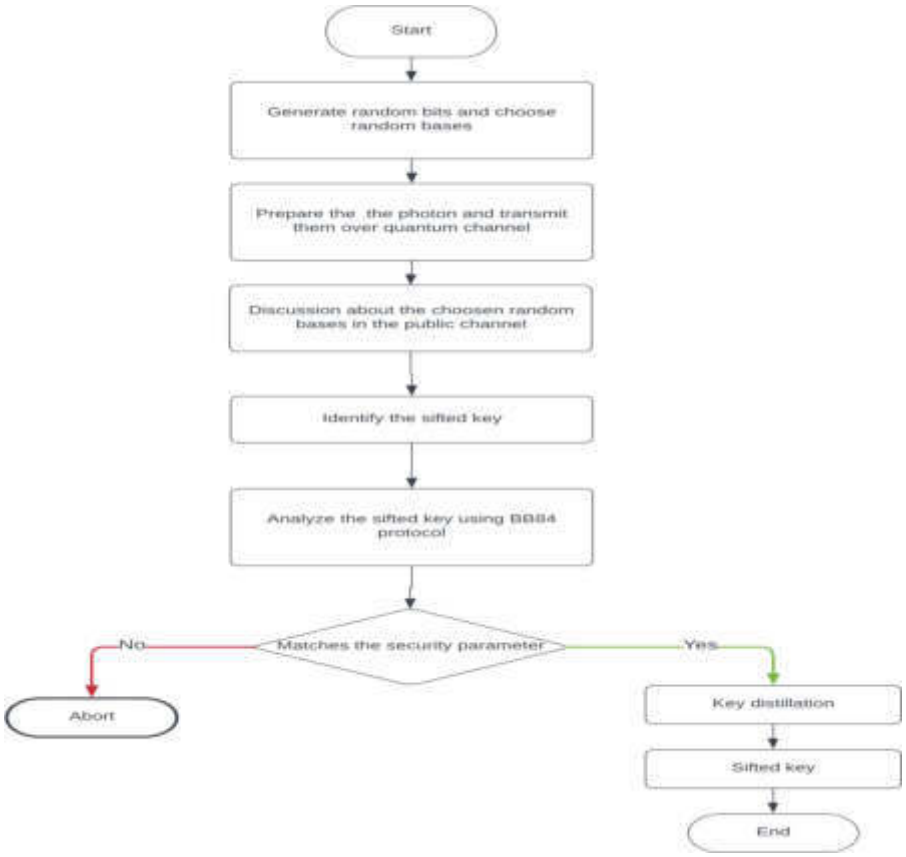


Fig. 3.1: Flow chart of the QKD Methodology using BB84 Protocol

4. Results and Discussions.

4.1. Shor's Algorithm Breaking the RSA Encryption. Shor's algorithm has successfully deciphered the RSA algorithm, by finding the factors of the given N value and also finding the private key. The below Fig.4.1 is the results shown after the implementation of Shor's algorithm. Here, the time taken to crack the factors of N value is increasing with the value of N linearly.

4.2. Analysis of the generation of Quantum keys' Time Complexity. There are three steps in Quantum Key Distribution: key creation, key filtering, and key distillation. The below fig.4.2 describes the time taken for key generation through the Quantum key distribution. The below results discuss that the time that quantum key generation takes is very small with respect to the number of bits and less when compared with the other conventional key generation algorithms such as RSA.

4.3. Quantum Key Distribution without Eve's Presence:. At the first step of Quantum key distribution Alice generates the random bits and chooses the random bases to transmit them to Bob. This key transmission process is n happens through the quantum channel. And then Bob will choose the random bases(filters) to receive the bits sent by Alice. The below Fig.4 discusses the random bits, basis, and polarised photons sent by Alice and the corresponding bits and basis received by Bob. Here in Table 4.1: $h = \rightarrow$; $v = \uparrow$; $r = \nearrow$; $l = \nwarrow$

In Table 4.1 shows the sifted key after the key generation and key sifting process. The actual secret key will be generated after the key distillation process. Below is the secret key of 16 bits that is distributed between Alice and Bob with the restriction that an eavesdropper does not exist between the communicating peers

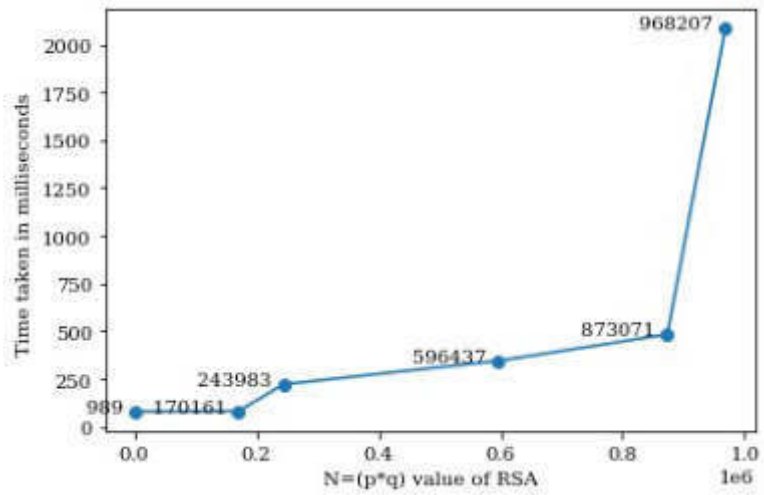


Fig. 4.1: Time Required for Shor's Algorithm to Defeat the RSA Scheme.

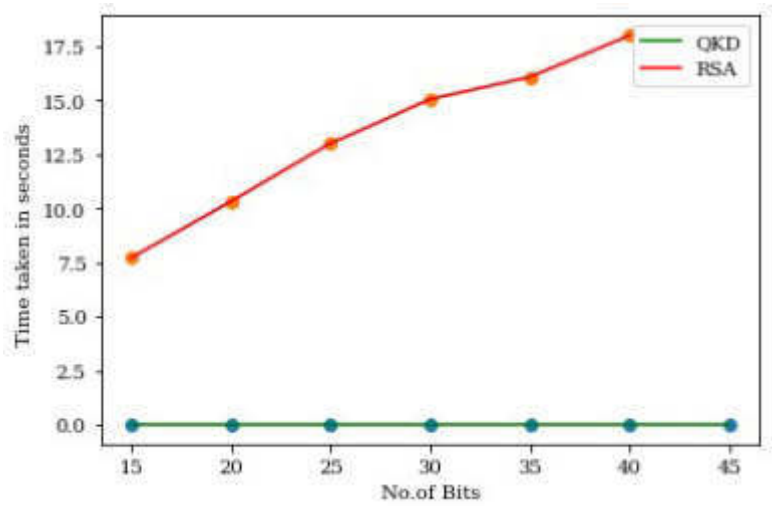


Fig. 4.2: Analysis of the Time Complexity of Quantum Key Generation

Table 4.1: Alice and Bob's Bitstream along with bases and identifying the sifted key

A_bits	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	0	0	1	1	1	0	1	0	0	1	1	1	1	0	0	0
A_basis	+	x	x	x	+	+	x	x	x	x	+	+	+	+	+	+	x	x	x	x	x	x	x	+	+	x	x	x	x	+	x	x
A_photons	h	l	l	l	h	h	l	r	l	r	h	v	h	h	h	h	r	r	l	l	l	r	l	v	v	l	l	l	l	v	r	r
B_basis	x	+	+	x	x	x	+	+	x	x	+	+	+	x	+	x	x	x	+	x	+	x	+	x	x	x	+	+	x	+	+	+
B_bits	1	0	1	1	0	1	1	1	1	0	1	0	1	0	1	1	0	0	1	1	0	0	0	1	1	1	0	0	1	0	0	
Sifted key				1					1	0	1	0	1		1		0	0		1						1			1	0		

Table 4.2: Secret key of 16 bits that is Distributed between Alice and Bob

Secret Key	0	0	0	1	0	0	0	0	1	0	1	0	1	0	1	0

Table 4.3: Alice and Bob's bitstream in the presence of Eavesdropper

A_bits	1	1	0	1	0	1	0	1	1	1	1	0	0	1	1	1	0	0	1	0	1	1	1	1	0	1	0	1	0	1	1	0
A_basis	+	+	x	x	x	x	+	+	x	+	x	+	x	+	+	x	+	x	+	+	+	+	x	+	+	x	x	+	x	+	+	+
A_photons	h	h	r	l	r	l	h	h	r	v	l	v	r	h	h	r	h	l	h	v	v	v	l	h	h	l	l	v	l	h	h	h
E_photons	r	l	r	l	h	l	h	l	h	v	v	v	v	h	h	h	l	l	h	v	v	v	l	h	r	l	l	v	l	h	h	h
E_bits	1	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1	0
B_basis	x	+	+	x	+	x	x	+	x	+	+	+	x	x	+	x	+	x	+	x	+	x	x	x	+	+	+	+	+	+	+	x
B_bits	0	1	1	0	1	0	1	0	1	0	1	1	0	1	0	0	1	1	0	0	1	1	0	0	0	1	1	1	0	1	0	1
Sifted key		1							1				0		0													1		1		

4.4. Quantum Key Distribution During Eavesdropper Presence:. Here is the other case, which involves the distribution of quantum keys while being monitored by an Eavesdropper. To get the value of the bit, he must witness the photon, which will disrupt the conversation and betray his existence. A more sophisticated technique would be for the eavesdropper to detect the photon, register the bit value, and prepare a new photon based on the result to broadcast to the receiver. The two legitimate parties collaborate in quantum cryptography to prevent the eavesdropper from doing so by compelling him to create mistakes. Eve tries to intercept the quantum channel's qubit using an intercept resend attack. Bob's modified and measured qubits alter the initial shared key. Here in Table 3; $h = \rightarrow$; $v = \uparrow$; $r = \nearrow$; $l = \nwarrow$

In Table 4.3 represents the bitstream of both Alice and Bob in the Eavesdropper and demonstrates how the existence of the Eavesdropper causes Quantum Bit Error Rate to rise. Thus, the provided need to be discarded and the whole process need to be started again.

5. Conclusion. In this work, the proposed method and the Quantum Key distribution Using protocol were implemented. Here, the primary exploit in the traditional cryptographic key exchange technique has been exposed, namely the ability to defeat the RSA algorithm's fundamental building blocks using Shor's algorithm. The time required to produce the key using a quantum cryptography method and a conventional cryptographic algorithm were compared and analyzed. Quantum Key Distribution takes less time to generate the key for the encryption of the data. As computer power grows, cyber security becomes more complex. This project provides a comprehensive model of QKD communication. This is a huge and encouraging step towards a day when we may feel more confident in our interactions. As a result, we may anticipate that QKD will have a profound influence on basic physics, altering our understanding of how quantum mechanics evolved. For the time being, our technology provides a reasonable solution for two-way encrypted communication. Yet, someone may soon be able to utilize sophisticated tools to breach this system, jeopardizing security. As a result, security policies and procedures must be updated on a regular basis.

REFERENCES

[1] Dariush Abbasinezhad-Mood and Morteza Nikooghadam. An anonymous ecc-based self-certified key distribution scheme for the smart grid. *IEEE Transactions on Industrial Electronics*, 65(10):7996–8004, 2018.

[2] Akwasi Adu-Kyere, Ethiopia Nigussie, and Jouni Isoaho. Quantum key distribution: Modeling and simulation through bb84 protocol using python3. *Sensors*, 22(16):6284, 2022.

[3] A Ahilan and A Jeyam. Breaking barriers in conventional cryptography by integrating with quantum key distribution. *Wireless Personal Communications*, pages 1–19, 2022.

[4] Vaishali Bhatia and KR Ramkumar. An efficient quantum computing technique for cracking rsa using shor's algorithm. In *2020 IEEE 5th international conference on computing communication and automation (ICCCA)*, pages 89–94. IEEE, 2020.

[5] Khodakhast Bibak and Robert Ritchie. Quantum key distribution with prf (hash, nonce) achieves everlasting security. *Quantum Information Processing*, 20(7):228, 2021.

[6] Ivan B Djordjevic. Qkd-enhanced cybersecurity protocols. *IEEE Photonics Journal*, 13(2), 2021.

[7] Aayush Joshi, Rutuja Kumbhar, Akshat Mehta, Vaishali Kosamkar, and Harshith Shetty. Breaking rsa encryption using quantum computer. 2022.

[8] V Kalaivani et al. Enhanced bb84 quantum cryptography protocol for secure communication in wireless body sensor networks for medical applications. *Personal and Ubiquitous Computing*, page 1, 2021.

[9] Chankyun Lee, Ilkwon Sohn, and Wonhyuk Lee. Eavesdropping detection in bb84 quantum key distribution protocols. *IEEE Transactions on Network and Service Management*, 19(3):2689–2701, 2022.

[10] Yonghong Ma, Xiuyu Wang, and Dandan Cui. Secure communication mechanism for smart distribution network integrated with subcarrier multiplexed quantum key distribution. *Power Syst. Technol.*, 11:036, 2013.
[11] Mosayeb Naseri. Revisiting quantum authentication scheme based on entanglement swapping. *International Journal of Theoretical Physics*, 55:2428–2435, 2016.
[12] Vanga Odelu, Ashok Kumar Das, Mohammad Wazid, and Mauro Conti. Provably secure authenticated key agreement scheme for smart grid. *IEEE Transactions on Smart Grid*, 9(3):1900–1910, 2016.
[13] Amritha Puliadi Premnath, Ju-Yeon Jo, and Yoohwan Kim. Application of ntru cryptographic algorithm for scada security. In *2014 11th international conference on information technology: new generations*, pages 341–346. IEEE, 2014.
[14] Neetesh Saxena and Santiago Grijalva. Dynamic secrets and secret keys based scheme for securing last mile smart grid wireless communication. *IEEE Transactions on Industrial Informatics*, 13(3):1482–1491, 2016.
[15] Purva Sharma, Anuj Agrawal, Vimal Bhatia, Shashi Prakash, and Amit Kumar Mishra. Quantum key distribution secured optical networks: A survey. *IEEE Open Journal of the Communications Society*, 2:2049–2083, 2021.
[16] Vishal Sharma, Kishore Thapliyal, Anirban Pathak, and Subhashish Banerjee. A comparative study of protocols for secure quantum communication under noisy environment: single-qubit-based protocols versus entangled-state-based protocols. *Quantum Information Processing*, 15:4681–4710, 2016.
[17] Jia-Lun Tsai and Nai-Wei Lo. Secure anonymous key distribution scheme for smart grid. *IEEE transactions on smart grid*, 7(2):906–914, 2015.
[18] Yahui Wang, Huanguo Zhang, and Houzhen Wang. Quantum polynomial-time fixed-point attack for rsa. *China Communications*, 15(2):25–32, 2018.
[19] M Xin, Z Liang, P Ma, N Jin, and M Zhu. Optical fiber transmission solution of measurement and control signal between substations based on quantum key distribution and one-time pad. *Automation of Electric Power Systems*, 41(12):212–217, 2017.
[20] MIAO Xin and CHEN Xi. Quantum logic circuit of quantum bit error correction coding and decoding for quantum communication in smart grid substation. In *Zhongguo Dianji Gongcheng Xuebao/Proc. Chin. Soc. Electr. Eng.*, volume 34, pages 4359–4363, 2014.
[21] Hao Yuan, Yi-min Liu, Guo-zhu Pan, Gang Zhang, Jun Zhou, and Zhan-jun Zhang. Quantum identity authentication based on ping-pong technique without entanglements. *Quantum information processing*, 13:2535–2549, 2014.
[22] Piotr Zawadzki, Zbigniew Puchała, and Jarosław Adam Miszczak. Increasing the security of the ping–pong protocol by using many mutually unbiased bases. *Quantum information processing*, 12:569–576, 2013.
[23] Baokang Zhao, Bo Liu, Chunqing Wu, Wanrong Yu, Jinshu Su, Ilsun You, and Francesco Palmieri. A novel ntt-based authentication scheme for 10-ghz quantum key distribution systems. *IEEE Transactions on Industrial Electronics*, 63(8):5101–5108, 2016.
[24] J Zhou, L Lu, Y Lei, and X Chen. Research on improving security of protection for power system secondary system by quantum key technology. *Power Syst. Technol*, 38(6):1518–1522, 2014.
[25] Tianqi Zhou, Jian Shen, Xiong Li, Chen Wang, and Jun Shen. Quantum cryptography for the future internet and the security analysis. *Security and Communication Networks*, 2018:1–7, 2018.

Edited by: Achyut Shankar
Special issue on: Machine Learning for Smart Systems: Smart Building, Smart Campus, and Smart City
Received: Mar 16, 2023
Accepted: Nov 11, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 35–44, DOI 10.12694/scpe.v25i1.2172

NETWORK SECURITY WITH VIRTUAL REALITY BASED ANTIVIRUS PROTECTION AND REDUCED DETECTION DELAYS

CHUNNA SONG*, JINFANG CHENG† AND GUOQIU ZHANG‡

Abstract. Addressing the persistent delay problem in traditional network security antivirus protection systems, this paper introduces an innovative approach utilizing virtual reality (VR) technology. The primary objective is to significantly reduce detection delays and enhance the efficiency of network security measures. An enhanced decision-making algorithm is proposed to identify relevant features associated with network security. These features are then weighted and optimized to improve the overall detection process. An injection list is generated through web crawling techniques to strengthen security measures. A virtual protection block is also developed to serve as a barrier against potential threats. The proposed method claims a detection delay of only 75.33 milliseconds, significantly outperforming two traditional methods that recorded 290.11 milliseconds and 337.30 milliseconds, respectively. This substantial decrease in detection delay emphasizes the effectiveness of automatic detection within the context of VR technology. Practical implementation and empirical evidence further validate the success of this approach. The automatic detection of network security vulnerabilities within the VR technology framework is efficient and exhibits considerable progress. As such, this research offers a promising solution to the delay problem in network security antivirus protection. Embracing VR technology achieves shorter detection delays, ultimately improving the security posture of network systems.

Key words: Network Security, Antivirus Protection, Virtual Reality Technology, Detection Delay, Automatic Detection

AMS subject classifications. 15A15, 15A09, 15A23

1. Introduction.

Virtual reality technology has widespread applications across diverse domains, including entertainment, education, and healthcare. In computer network security, leveraging virtual reality technology for vulnerability detection represents a novel and emerging approach. Computer network systems are naturally vulnerable to security threats, which put system components and data integrity at serious risk. Malicious viruses use these gaps, frequently relying on shortcomings in security rules, protocols, hardware, and software implementations, to obtain access to the system without authorization. For instance, vulnerabilities such as flaws in the Network File System (NFS) protocol's authentication method, logic errors in microchips, or misconfigurations by Unix system administrators setting up anonymous File Transfer Protocol (FTP) services can all be exploited as security vulnerabilities within a system [11].

The Morris worm virus, which emerged during the early stages of computer networks, occupies a pivotal position in the records of computer viruses. Its profound impact was a catalyst for the subsequent proliferation of computer viruses. In the wake of the Morris worm, the computer virus has experienced an explosive expansion, characterized by a sheer increase in numbers and a marked elevation in sophistication. Notably, this propagation has been accompanied by a distinct shift in the motivations driving the creation and dissemination of viruses. Initially conceived as experiments, viruses have metamorphosed into instruments wielded for profit-driven endeavours, including cybercrime and data theft. This evolutionary shift has been accompanied by a substantial augmentation of viruses' capabilities, enabling them to escape detection and inflict significant harm upon computer systems and networks. Consequently, computer security has had to adapt continuously, developing strategies and technologies to counter these ever-evolving threats and safeguard digital environments from an expanding array of viruses and malware [4, 15].

The escalating frequency of attacks and threats to computer networks can be attributed to multiple factors. On the one hand, diverse security vulnerabilities and their growth rates continue to rise, often without

*Information Engineering College, Shijiazhuang Vocational College of Finance & Economics, Shijiazhuang, 050000, China
†Department of Student Affairs, Shijiazhuang Preschool Teachers College, China, 050228 (Corresponding author: jinfangcheng9@126.com)
‡Department of Aeronautical Engineering, Shijiazhuang Engineering Vocational College, China, 050061

proportionate attention. On the other hand, the increasing complexity of network systems inherently escalates security risks, making it imperative to explore innovative methods, such as those grounded in virtual reality technology, for robust vulnerability detection and mitigation [13].

Currently, computer network vulnerabilities often appear in the following dimensions [17, 3]:

- Inherent System Vulnerabilities: Present-day operating systems, including Unix and Windows, exhibit various security risks. Virtually every operating system contains known security vulnerabilities that have been identified, addressed, and remain potential threats.
- Unauthorized User Access: Unauthorized users often manage to breach network security measures, gaining entry to networks to which they should not have access.
- Unauthorized Escalation of User Privileges: In some instances, legitimate users may exploit vulnerabilities to increase their access privileges without proper authorization, potentially compromising system security.
- Multi-Vector Attack Vulnerabilities: Network systems are susceptible to attacks from multiple angles and approaches, leaving them vulnerable to various security threats.

These typical vulnerabilities can be traced back to various underlying factors, encompassing security weaknesses within network protocols, vulnerabilities inherent to operating systems, and flaws within different applications and software components.

2. Literature Review. Computer network security frequently encounters challenges, with a particular emphasis on addressing numerous vulnerabilities. These vulnerabilities may result from poor usage, design defects in software or hardware components, or both.

The vulnerabilities in computer network security are predominantly evident in three key dimensions [10]:

(a) Operating System Vulnerabilities: These vulnerabilities inherent to operating systems are the foundation for network infrastructure. Security weaknesses within operating systems can provide entry points for attacks and unauthorized access.

(b) Computer Software Vulnerabilities: Vulnerabilities in various software applications, utilities, and programs utilized within the network environment present another significant risk. These software vulnerabilities can be exploited to compromise network security.

(c) Network Hardware Facility Weaknesses: Hardware components within the network infrastructure may also contain vulnerabilities. These weaknesses can expose critical network assets to threats, making them susceptible to attacks and breaches.

Addressing these vulnerabilities is dominant to strengthening computer network security and mitigating the risks associated with cyberattacks and unauthorized access. Failure to address vulnerabilities within the operating system can have severe repercussions on computer network security. The operating system is a critical component of computer networks, and any vulnerabilities can pose a significant threat. The most substantial security risk within the network environment lies in remote attacks that exploit operating system vulnerabilities. Consequently, ensuring a secure operational environment for computer network operating systems is imperative. Most operating systems exhibit various security vulnerabilities, which may only become apparent during routine usage. Some security risks remain covered within these system vulnerabilities, making them particularly challenging to identify and mitigate [18].

Computer software vulnerabilities constitute a substantial impediment to preserving computer network security. Allowing such vulnerabilities to persist may ultimately result in these flaws evolving into software defects susceptible to external attacks. In instances where software security vulnerabilities occur frequently, particularly if they predominantly consist of high-risk vulnerabilities, they pose a grave threat to the overall security of computer networks. Failing to rectify these vulnerabilities promptly leaves networks susceptible to exploitation and attacks by malicious hackers. It's crucial to acknowledge that software vulnerabilities are the underlying source of security incidents. This is especially significant when hackers gain access to sensitive information, as it can lead to severe cases of fraudulent activities and other serious security breaches [8].

Numerous security vulnerabilities within computer networks can be attributed to weaknesses in network hardware facilities. For instance, using removable storage media, such as USB flash drives, can potentially lead to the inadvertent disclosure of sensitive information when these drives are borrowed or shared. In conventional methods of detecting computer network security vulnerabilities, inspectors typically must interact

Table 2.1: Comparison of network security vulnerability and detection approaches

Technology	Advantages	Disadvantages
Operating System Vulnerabilities	Critical for network infrastructure security. Understanding and addressing these vulnerabilities are essential.	Difficult to identify and mitigate. Vulnerabilities may not be apparent in routine usage.
Computer Software Vulnerabilities	Significant for overall network security. Prompt remediation is crucial.	Can evolve into software defects susceptible to attacks High-risk vulnerabilities pose a serious threat
Network Hardware Facility Weaknesses	Addresses potential weaknesses in network hardware. Recognizes risks of removable storage media.	Limited by conventional detection methods. Challenges with accuracy, speed, and cost.
Network Vulnerability Detection Methods	Involves simulation experiments and dataset modelling. Offers automated updates in vulnerability detection.	Complexity in acquiring network flow data tables Reduced detection efficiency
Automatic Detection Method	Optimizes weaker variables efficiently. Creates injection point lists and simulates attacks. Utilizes virtual reality technology.	Mainly applied in controlled simulated experiments. Uncertain real-world applicability Requires further empirical testing.

with computer terminals and rely on various vulnerability detection tools. However, these tools often exhibit significant limitations in practical use, including issues related to accuracy, speed, and the associated high costs of vulnerability detection [2].

Network vulnerability detection methods typically involve modelling a sample dataset and extracting network variables for subsequent simulation experiments. For instance, some literature employs the hidden Markov model to represent the sample dataset of network information, aiming to achieve automated updates in network security vulnerability detection. However, this approach introduces complexity in acquiring network flow data tables, reducing detection efficiency and efficiency-related issues [1].

The automatic detection method involves extracting dynamic numerical variables and analyzing static attack processes. It's crucial to highlight that this approach has mainly been employed in controlled simulated experiments, often utilizing open-source software. Although these experiments have provided valuable insights, they come with certain limitations. One notable limitation pertains to the constrained scope and specificity of the experimental outcomes. Since these simulations are artificial and controlled, they may not have adequately explored the full range of real-world scenarios and vulnerabilities. Consequently, this method's practical applicability and real-world effectiveness remain uncertain, as its performance may vary when challenged with the intricacies and diverse landscapes of actual network environments. Further research and empirical testing are warranted to establish the method's reliability and suitability for addressing the intricate and evolving challenges of network security vulnerability detection [14].

An optimization strategy is employed to fine-tune the weighting of weaker variables. This optimization process involves the identification and analysis of vulnerability behaviours by deploying data entry points, as well as the creation of injection point lists. Additionally, virtual reality technology is harnessed to simulate malicious attacks effectively. Through the comprehensive documentation and analysis of fundamental attributes related to virtual protection, research endeavours to implement automatic network security detection within virtual reality technology have been completed [12]. Table 2.1 provides a comparison of network security and vulnerability detection techniques.

3. Research on Automated Network Security Vulnerability Detection using VR Technology. The network security and vulnerability detection technology involves performing source code analysis of the relevant program to identify and address vulnerabilities. In the context of virtual reality technology, automatic network security vulnerability detection relies on simulating and capturing malicious attacks using virtual reality

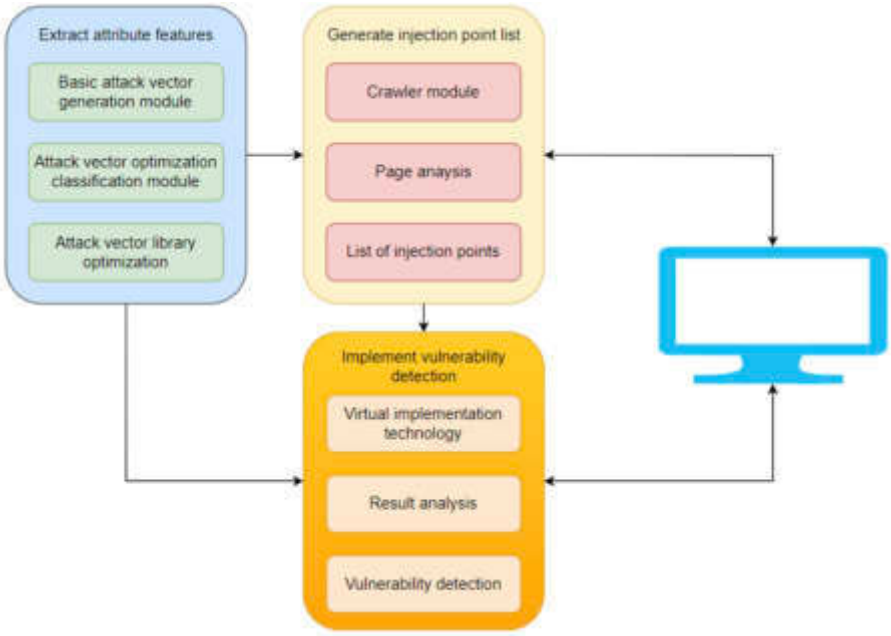


Fig. 3.1: Proposed network security and vulnerability detection process

technology. This process involves the automated generation of attack graphs through detection mechanisms. The comprehensive architectural depiction of this method is illustrated in Figure 3.1.

During the development of the attack graph, the primary focus centres on the automated construction of this graph. It involves the creation of an attack vector repository through optimization calculations. Additionally, analysis of attack injection points aims to identify potential entry points within the network. It is achieved by employing web crawlers to retrieve web pages and utilizing crawler tools for web page analysis. The attack and analysis process entails simulating the attack sequence using virtual reality technology. Subsequently, the outcomes generated by this simulation are analyzed to arrive at a final assessment of vulnerabilities.

3.1. Extracting Network Security Vulnerability Characteristic Attributes. Leveraging virtual reality technology for vulnerability detection enables the identification of diverse vulnerabilities within a virtual network environment created for this purpose. Using VR technology, vulnerability testers can interact with vulnerabilities within virtual environments, engaging in activities such as analyzing network traffic, conducting penetration tests on operating systems, and evaluating firewall defences. These operations closely mimic real-world attack scenarios, enhancing vulnerability detection accuracy.

In the automated network security and vulnerability detection process, the network system treats vulnerabilities on nodes as if they were part of its codebase. Employing optimization calculations, the system adjusts the weights assigned to various attributes associated with these vulnerabilities, generating an attack graph. Attackers simulate network attacks through different vulnerabilities over time, with the disparities in weights between vulnerabilities reflected in the attack graph through the aggregation of vulnerability groups [6]. At time t, a cluster with m vulnerabilities is calculated. Suppose there are k individual genes in each vulnerability, the vulnerability gene studied by the author adopts the difficulty gene, exposure gene, repair gene, exposure gene and repair gene, therefore, the value of h: is 3, and the location of a vulnerability i can be expressed as $q_i = (q_{i1}, q_{i2}, q_{i3})$, where $i = 1, 2, \cdots, m$. In the process of generating the attack graph, the starting position of the vulnerability can be regarded as a point of the spatial coordinates, so the position of vulnerability i can be digitized to represent it during coordinate iteration. Set the loophole movement speed, $v_i = (v_{i1}, v_{i2}, v_{i3}), v_{i1}, v_{i2}, v_{i3}$ as the loophole movement speed in the three coordinate axes, and its attributes

can be expressed as:

$$V_{i+1} = \omega\vartheta_{ir} + \varepsilon_1 (z_{ir} - p_{ir}) + \varepsilon_2 (z_{ir} - p_{ir}) \tag{3.1}$$

In the above Equation, ε_1 is the learning coefficient, ε_2 is the network learning coefficient, ω is the inertia coefficient, ϑ_{ir} is the initial speed of the vulnerability, z_{ir} is the optimal location of the vulnerability, and p_{ir} is the probability of the vulnerability being applied in the attack, so p_{ir} is given by the Equation

$$\sum_{i=1}^{m} p_{ir} = 1, 0 < p_{ir} < 1 \tag{3.2}$$

When ω is between 0 and 1, the vulnerabilities in motion can converge; therefore, when the ω value is determined, the number of malicious attacks can be predicted. When the vulnerability exposure in the network is large, the degree of repair will also increase, and the vulnerability will be difficult to apply. Therefore, the three attributes of vulnerabilities will increase with time, so p_{ir} will decrease accordingly, and the extraction of vulnerability characteristic attributes is completed [5].

3.2. Generate Injection Point List. Based on the extracted vulnerability feature attributes, an analysis of injection points is conducted to produce a comprehensive list of these points. This process relies on a web crawler's assistance, an automated program that captures the vulnerability feature attributes. The crawler's function involves evaluating these attributes to establish the prioritization order for the URLs to be captured while preserving the essential attributes necessary for the analysis.

During the web crawler search process, a significant portion of the URL queue may contain duplicate entries, impeding retrieval speed. We employ the BloomFilter algorithm to mitigate this issue for duplicate removal, vulnerability detection, adjustment, and data storage. As the URL queue is obtained through crawling, each entry undergoes assessment. Entries not already present in the collection are added to it. When identified with an ID of Y, it signifies inclusion in the collection. Thus, they are placed into the URL set. This operation effectively eliminates duplicates, resulting in the compilation of a list of injection points.

The system utilizes the data obtained through crawling attack injection points to perform vulnerability analysis and record the results. This process yields an attribute injection point list to conduct a comprehensive vulnerability analysis based on these records. Subsequently, the system automatically selects the most suitable method to detect the identified vulnerabilities. When implemented through virtual reality technology, vulnerability detection significantly enhances detection speed. Traditional vulnerability detection methods often necessitate manual intervention or specialized vulnerability detection tools, resulting in comparatively slower detection processes. In contrast, vulnerability detection leveraging virtual reality technology can be carried out programmatically, thereby substantially improving the speed and efficiency of the detection process.

3.3. Implementation of Vulnerability Detection under Virtual Reality Technology. Virtual reality technology employs a sense of telepresence to recreate a lifelike perception of the simulated environment. Applying this technology to network security vulnerability detection enables a comprehensive three-dimensional analysis of vulnerabilities within virtual settings. During virtual attack testing, the information in the generated injection point list is a guide. Virtual reality technology is leveraged to simulate attack behaviours and facilitate interaction with the server in line with the identified vulnerabilities [7].

Within this process, since the attacker's identity is entirely virtual, the attributes of network security vulnerabilities serve as the fundamental genetic elements for constructing virtual attacks. Information embedded within network security vulnerabilities is systematically extracted through the established virtual reality environment and virtual attack mechanisms. Security vulnerabilities are classified based on categorizing vulnerability genes, exposure genes, and repair genes inherent in attack information. This constitutes the operational structure during the virtual attack vulnerability detection process.

In the detection phase, artificial virtual vulnerability attacks are executed following the selection and cross-compilation of virtual attack genes. The outcomes of these attacks are then factored into calculating a fitness function alongside other vulnerability-related data. Ultimately, this process culminates in the successful detection of network security vulnerabilities. With this, the research into automatic detection methods for network security vulnerabilities within the virtual reality technology framework is concluded [9].

3.4. Protection Technology for Computer Network Security Vulnerabilities. The diverse types and their high complexity characterize computer network vulnerabilities widespread in many computers and significantly threaten network security. Their presence renders the networks susceptible to external attacks, disrupting regular operations. To overcome these risks, conventional practices involve the installation of vulnerability patches and continuously updating operating systems. Additionally, specialized Trojan-killing software is deployed to eliminate existing Trojan horses, and this software is frequently updated to ensure thorough system checks and eradication. Real-time monitoring functionality is enabled to safeguard against external virus intrusions and system damage proactively [16]. Computer operating systems and software frequently contain numerous vulnerabilities that hackers can exploit to compromise system security. Consequently, keeping the operating system and software up to date by promptly installing the latest patches and security updates is imperative. This practice effectively mitigates the risk of attackers exploiting known vulnerabilities and compromising the system's integrity.

(1) Security configuration switch

One commonly employed method to minimize the forwarding of unicast broadcast traffic on specific ports is the application of multicast or unicast broadcast blocking attributes. By reducing traffic volume on a per-port basis, several advantages are realized. This approach enhances network security by limiting traffic and prevents network devices from processing unneeded, directionless packets. Port security allows or denies traffic based on the host Media Access Control (MAC) address. Depending on the switch model, there may be varying maximum MAC address allowances. This functionality specifies the permissible number of hosts per port, which can then be configured to meet network requirements.

(2) Security configuration router

Routers can be configured to establish a secure perimeter and defend against external attacks within a defined range. Typically, access lists are employed to restrict the source and destination addresses of packets traversing through the router. Additionally, some routers implement the Reverse Path Forwarding (RPF) check to enhance security further. Furthermore, it's possible to configure the "no IP directed broadcast" setting on all routers potentially linked to the target subnet. To improve security measures, the router can turn source route options off using the command "no IP source route". This precautionary step serves as an effective deterrent against source route attacks.

(3) Set up the computer.

To increase security, it's essential to avoid guest accounts, conceal IP addresses, exercise caution with unfamiliar emails, install and regularly update effective security software, turn off "printing and file sharing" when not needed, promptly close unused ports, regularly update administrator information, prevent empty connections, and address program logic vulnerabilities by updating to corrected software versions. When hackers misuse legitimate program functionalities, gaining insight into their tactics is crucial. To defend against attacks, users should avoid hacker-exploited program steps and employ methods to bypass their attack vectors, thus enhancing overall cybersecurity.

(4) Securely configure the Windows server.

In managing Windows servers, it is imperative to prohibit using Remote Registry, Messenger, and Telnet. Additionally, safeguarding important files and directories can be achieved by modifying the registry to hide them. Strengthening defence involves setting and managing accounts with complex passwords. To ensure the security of all network connections within the local system, it is crucial to protect them using Windows Firewall promptly. This requires configuring the appropriate Group Policy parameters in the system environment of Windows Server 2008. Trusted access is essential for verifying user requests and establishing trust when connecting to cloud computing resources, ultimately ensuring cloud security.

Moreover, the cloud security management and defence system, consisting of a cloud security detection platform, a cloud security response platform, and a cloud security recovery strategy, plays a pivotal role in eliminating security vulnerabilities and upholding the security of hosts, networks, and application layers. Regularly updating and maintaining Windows Firewall rules is recommended to safeguard network security. The core of the cloud security management and defence system is the trusted link platform system, which constructs a comprehensive framework through user identity authentication, data encryption, and authorization management, ensuring operational security.

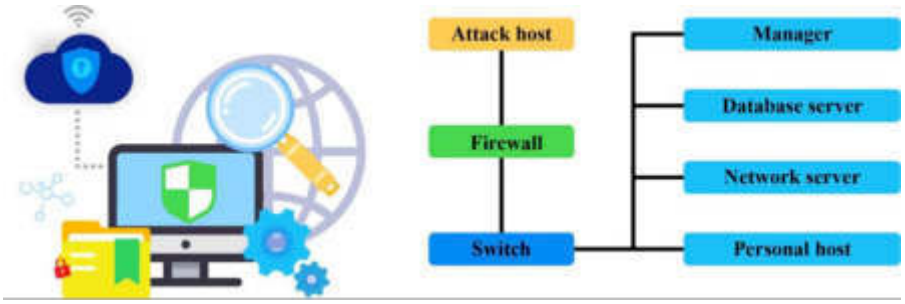


Fig. 4.1: Experimental network topology

Table 4.1: Intranet node vulnerability settings

Node	Vulnerability Number	Server or software	Permission promotion	Ease of use 1%
Network server B	Me10-062	WWW	U-A	70
Network server B	Me11-002	WWW	U-A	100
Database server C	Me11-012	Oracle	O-A	90
Database server C	Me10-065	Oracle	O-U	85
Manager D	Mc10-045	Windows	O-A	100
Personal host E	Mc10-006	Office	O-A	70

4. Simulation Results. To assess the network security vulnerabilities within the VR technology framework established previously, simulation experiments are employed to validate the reliability of the proposed automated vulnerability detection method. These experiments involve comparing the detection outcomes with those obtained using two conventional vulnerability detection technologies, thereby confirming the advantages of the proposed method.

4.1. Experimental Environment and Vulnerability Parameter Settings. In network security and vulnerability detection research, establishing a well-structured experimental environment and carefully configuring the experimental network structure is essential, as shown in Figure 4.1.

Generally, the topology consists of five nodes: A, B, C, D, and E. Node A represents external network host access. The other four internal network nodes serve distinct roles: network services, database services, server protection management, and important data storage. Communication within or between internal network nodes does not necessitate firewall traversal. However, when external network nodes seek to access internal network nodes, they must pass through the firewall, which grants access solely to network servers. However, certain inherent vulnerabilities within the intranet expose internal network nodes to real risks. Table 4.1 outlines these weaknesses in the intranet nodes.

4.2. Experimental Results and Analysis. After constructing the experimental environment, the traditional modelling, numerical, and automatic detection methods of network security vulnerabilities under the virtual reality technology designed by the author are used for experiments.

The key step to reduce the detection time is to improve the network interaction frequency during detection. By optimizing the structure of the neural network, the number of network layers and parameters can be reduced, and the computational efficiency and response speed of the network can be improved. Reducing input data size can reduce the network's processing time and computational complexity, thereby improving the network's response speed. To obtain better vulnerability detection results, 70 detection samples are set to view the change in interaction frequency of the two methods in the detection process. Record the time spent by each method to view requirements, as shown in Figure 4.2.

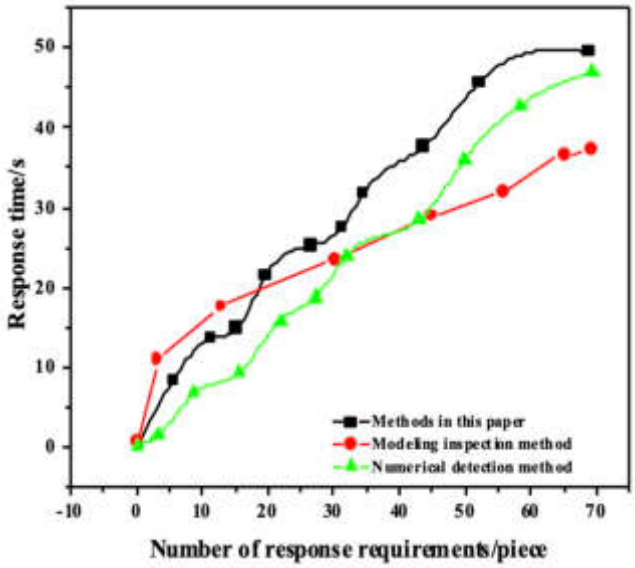


Fig. 4.2: Comparison of interaction frequency detection

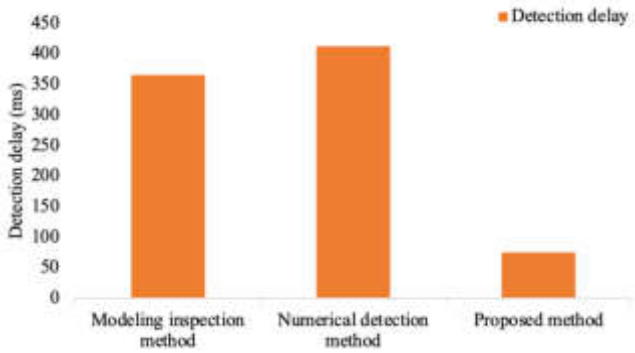


Fig. 4.3: Test and detection delay of the three methods

It can be seen from the above results that with the increase in the number of test samples, the number of methods proposed by the authors to respond to requests is increasing. The first 20 seconds of detection has strong advantages because the modelling detection method covers part of the sample information. When the test time reaches 20 seconds, the method proposed by the author is compared with the traditional two methods. The number of response requirements simultaneously is higher than that of the other two methods, the interaction frequency is higher, and the number of vulnerabilities detected is more, which has obvious advantages.

The simulation network vulnerability detection experiment is carried out for the four target nodes B, C, D and E. Record the time delay of the two methods to detect vulnerabilities and compare the detection results, as shown in Table 4.2 and Figure 4.3.

From the simulation results, the proposed detection method exhibits a detection delay of 75.33ms. Compared to the two traditional methods, this represents a reduction of 290.11ms and 337.30ms. This significant reduction in detection delay serves as compelling evidence for the efficiency of the automatic network security vulnerability detection method within the framework of virtual reality technology.

Table 4.2: Comparison of delay detection for the target nodes

Method	Target Node	Detected vulnerability number	Detection delay (ms)
Modelling inspection method	B	Me10-062	365.44
	C	Mc10-065	
	D	Me10-045	
	E	-	
Numerical detection method	B	Mc10-062 Me11-002	412.63
	C	-	
	D	Me10-045	
	E	Mc10-006	
Proposed method	B	Mc10-062 Mc11-002	75.33
	C	Mc11-012 Mc10-065	
	D	Me10-045	
	E	Mc10-006	

5. Conclusion. Traditional network security antivirus protection suffers from prolonged detection delays. An automatic network security and antivirus protection system based on virtual reality technology is proposed. The system employs optimization techniques to fine-tune vulnerability weights, extracts network security vulnerability features, utilizes web crawlers to capture and identify vulnerability characteristics and employs virtual reality technology for testing malicious attacks. Through extensive experimentation, this novel approach is compared with two traditional methods, resulting in significant reductions in detection delays of 290.11ms and 337.30ms, respectively, thereby confirming the efficiency of automatic network security detection within the virtual reality technology framework. Computer network security and vulnerability detection within virtual reality technology is an emerging and promising field with several potential avenues for future development. These include enhancing interactivity to create more realistic detection environments, integrating deep learning for improved accuracy and speed, and ensuring multi-platform support to provide various operating systems.

REFERENCES

[1] A. Ahmadian Ramaki, A. Rasoolzadegan, and A. Javan Jafari, *A systematic review on intrusion detection based on the hidden markov model*, Statistical Analysis and Data Mining: The ASA Data Science Journal, 11 (2018), pp. 111–134.
[2] R. Amankwah, P. K. Kudjo, and S. Y. Antwi, *Evaluation of software vulnerability detection methods and tools: a review*, International Journal of Computer Applications, 169 (2017), pp. 22–27.
[3] Z. Chen, X. Zuo, N. Dong, and B. Hou, *Application of network security penetration technology in power internet of things security vulnerability detection*, Transactions on Emerging Telecommunications Technologies, 33 (2022), p. e3859.
[4] R. J. Cole, *Computer worms, detection, and defense*, in Encyclopedia of Information Ethics and Security, IGI Global, 2007, pp. 89–95.
[5] X. He, *Analysis of network intrusion detection technology based on computer information security technology*, Journal of Physics: Conference Series, 1744 (2021), p. 042038.
[6] J. Hu, J. Chen, L. Zhang, Y. Liu, Q. Bao, H. Ackah-Arthur, and C. Zhang, *A memory-related vulnerability detection approach based on vulnerability features*, Tsinghua Science and Technology, 25 (2020), pp. 604–613.
[7] ———, *A memory-related vulnerability detection approach based on vulnerability features*, Tsinghua Science and Technology, 25 (2020), pp. 604–613.
[8] W. Huang, X. Li, and Z. Huo, *XSS vuinerability detection technology based on EBNF and twice crawling strategy*, Application Research of Computers, 36 (2019), pp. 2458–2463.
[9] X. Jia, *Research on college sports training based on computer virtual reality technology*, Journal of Physics: Conference Series, 1648 (2020), p. 032132.
[10] J. M. Kizza and J. M. Kizza, *Introduction to computer network vulnerabilities*, Guide to Computer Network Security, (2017), pp. 87–103.

[11] J. Li, P. Cao, and J. Yang, *Research on noc static vulnerability detection system based on big data technology*, Modern Electronics Technique, 42 (2019), pp. 77–81.
[12] R. Liu, *A computer network intrusion detection technology based on improved neural network algorithm*, Telecommunications and Radio Engineering, 79 (2020).
[13] G. Luo, *Research on network security vulnerability detection method based on artificial intelligence*, Journal of Physics: Conference Series, 1651 (2020), p. 012005.
[14] I. Medeiros, N. F. Neves, and M. Correia, *Automatic detection and correction of web application vulnerabilities using data mining to predict false positives*, in Proceedings of the 23rd International Conference on World Wide Web, Republic of Korea, 2014, pp. 63–74.
[15] H. Orman, *The morris worm: A fifteen-year perspective*, IEEE Security & Privacy, 1 (2003), pp. 35–43.
[16] C. Wang, T. Ren, Q. Li, X. Wang, G. Guo, and J. Dong, *Network computer security hidden dangers and vulnerability mining technology*, IOP Conference Series: Materials Science and Engineering, 750 (2020), p. 012155.
[17] M. Yi, X. Xu, and L. Xu, *An intelligent communication warning vulnerability detection algorithm based on iot technology*, IEEE Access, 7 (2019), pp. 164803–164814.
[18] D. Zhaokun, L. Yuliang, H. Zhao, H. Hui, and Z. Kailong, *Network program vulnerability detection technology based on program modeling*, Journal of Beijing University of Aeronautics and Astronautics, 45 (2019), pp. 796–803.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Mar 21, 2023
Accepted: Oct 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 45–54, DOI 10.12694/scpe.v25i1.2173

COMPUTER NETWORK VIRUS DEFENSE WITH DATA MINING-BASED ACTIVE PROTECTION

XIAOHONG LI*, YANG LI†, AND HONG HE‡

Abstract. A novel approach is presented in this paper to address the limitations of virtual machine technology, active kernel technology, heuristic killing technology, and behaviour killing technology in computer network virus defence. The proposed method provides data mining technology, specifically Object-Oriented Analysis (OOA) mining, to detect deformed and unknown viruses by analyzing the sequence of Win API calls in PE files. Experimental results showcase the Data Mining-based Antivirus (DMAV) system's superiority over existing virus scanning software in multiple aspects: higher accuracy in deformed virus detection, enhanced active defence capabilities against unknown viruses (with a recognition rate of 92%), improved efficiency, and a reduced false alarm rate for non-virus file detection. Furthermore, the paper introduces an OOA rule generator to optimize feature extraction, enhancing the system's intelligence and robustness. This research provides a promising solution to support virus detection accuracy, active defence mechanisms, and overall efficiency while minimizing false positives in virus scanning, thus contributing significantly to the advancement of computer network security.

Key words: Metamorphic virus, PE documents, Win API sequence, Data mining, OOA mining

1. Introduction. The use of the internet and the growth of online businesses have made network security a big concern. It's a problem because it threatens the safety of networks and the information stored on them. Hackers can get into systems differently, stealing user data and private information and causing systems to stop working correctly. This creates risks and challenges for businesses, governments, and individuals. Traditional security methods like operating system strengthening and firewalls are static, meaning they don't adapt well to new and complex attacks. Because of this, researchers are now working on intrusion detection systems and dynamic defence technology to better protect against online threats [11].

A typical cyber-attack can be broken down into three stages: information gathering, the actual attack, and exploiting unknown system vulnerabilities for illegal activities. Deliberate attacks often follow a specific sequence. Take Distributed Denial of Service (DDoS) attacks as an example: the process begins with identifying and scanning numerous hosts to locate a vulnerable target. Subsequently, a remote exploit program is employed to breach the target, gaining control over the system. Lastly, an attack daemon is installed and executed on the compromised host at the DDoS distribution point, which, in turn, deploys multiple compromised machines to scan and attack a single target. Studying the patterns of attack behaviours is significant in advancing intrusion detection technology [4].

Intrusion detection involves identifying unauthorized access or security breaches within a computer network or system. It gathers data from critical points in the network or system and then analyzes it to identify potential security policy violations and signs of network or system attacks. An Intrusion Detection System (IDS) is a combination of hardware and software designed for this purpose, serving as a tool to recognize potential intrusions or attacks on computer systems or networks. Intrusion detection plays a vital role within the broader security framework, serving as a dynamic defence technology that contributes significantly to overall system security [13].

The IDS can be categorized in various ways, primarily based on the source of detection data and the

*Department of Network and Communication Engineering, Shijiazhuang Information Engineering Vocational College, Shijiazhuang, Hebei, 050000, China
†Department of Network and Communication Engineering, Shijiazhuang Information Engineering Vocational College, Shijiazhuang, Hebei, 050000, China (Corresponding Author: yangli23@126.com)
‡Department of Software Engineering, Shijiazhuang Information Engineering Vocational College, Shijiazhuang, Hebei, 050000, China

intrusion detection methods employed. Regarding the source of detection data, IDS can be classified into two main types: Host-Based Intrusion Detection Systems (HIDS) and Network-Based Intrusion Detection Systems (NIDS). HIDS focuses on detecting events occurring within the host itself, such as file modifications, process creation, and system calls. In contrast, NIDS identifies events within network traffic, including network connections and packet content [1].

The authors used another classification technique based on the intrusion detection methods. This categorization divides IDS into two distinct approaches: misuse detection and anomaly detection. Misuse detection identifies known attack patterns through predefined rules, typically created by security experts. On the other hand, anomaly detection relies on learning algorithms to identify unknown attack behaviours, often employing machine learning and data mining techniques to discover abnormal activities [23].

IDS functions as a crucial second line of defence behind firewalls, offering real-time intrusion detection and implementing appropriate security measures such as evidence collection for tracking, network connection termination, and recovery. However, as intrusion detection technology evolves, so does intrusion technology. The demand for security products based on multi-level and comprehensive defence strategies has emerged. The evolution of intrusion detection systems has progressed through three stages: from active response within the intrusion detection system to interaction between the intrusion detection system and the firewall, and ultimately to the latest development of intrusion prevention systems [14].

The current state of intrusion prevention systems can be viewed as a fusion of firewall and intrusion detection capabilities. These systems leverage intrusion detection as their basis for identifying potential threats, but they take proactive measures to thwart attacks once detected. While introducing intrusion prevention systems has partially met the demands of interconnected network users within specific environments, they still exhibit shortcomings, including adaptability, scalability, limited intrusion event analysis capabilities, reliance on a single defence approach, and difficulties in handling complex attacks [17].

Intrusion prevention systems face challenges due to the constantly changing network environment and evolving attack techniques, which hinder their ability to stay up-to-date with emerging attack methods and malicious behaviours. This vulnerability renders them susceptible to evasion by attackers and reduces their overall effectiveness. Additionally, their rigid design and implementation constrain their adaptability and customization to diverse network setups and attack scenarios, limiting their practicality and scalability. Furthermore, these systems predominantly focus on detecting and reporting intrusion events, neglecting comprehensive analysis and traceability, which in turn hampers their ability to facilitate timely responses and efficiently manage intrusion incidents. Addressing these deficiencies in intrusion prevention systems is essential to strengthen their resilience and enhance network security in the face of evolving threats and complex attack landscapes [7].

To investigate security systems founded on distributed technology, featuring adaptable and open structures and incorporating the seamless integration of intrusion detection and defence technologies are proposed. Such research and development efforts are essential for advancing network information security technology and products towards comprehensive three-dimensional protection and bolstering national security defence initiatives [9].

The paper is structured into five sections with an overview of the research topic and outlines the study's objectives. Section 2 offers a comprehensive literature review, presenting the existing knowledge and research in computer network virus defence, including various technologies and approaches. The details of the proposed method, explaining the utilization of data mining technology for virus defence and the specific techniques, are employed in Section 3. The results of experiments and subsequent discussions and the performance of the proposed method are discussed in Section 4. Finally, Section 5 provides a conclusion summarizing the key findings and contributions.

2. Literature Review. Currently, the predominant antivirus technology still relies on signature-based methods, and its primary drawback lies in the static nature of the signature codebase, necessitating continuous updates to combat emerging viruses. This signature-based approach often results in a reactive defence stance, where antivirus solutions lag behind new threats. Researchers have introduced novel antivirus technologies to tackle this challenge, including behaviour detection, machine learning, and sandboxing. These approaches avoid reliance on fixed signature codes, instead analyzing virus samples' behaviours and characteristics to determine their malicious nature. In contrast to signature-based techniques, these innovative technologies exhibit superior adaptability and generalization capabilities, enabling them to counteract new and mutated

viruses effectively. Consequently, the evolution towards intelligent virus-active defence systems represents an inevitable virus detection and mitigation progression [21].

Prominent domestic antivirus software companies like Kingsoft, Rising, and Jiangmin primarily rely on traditional signature scanning technology for virus detection and eradication. However, they also commit to integrating advanced technologies into their antivirus strategies. For instance, Jinshan Antivirus Laboratory has harnessed the power of suspicious tool scanning and "honeypot" technology, resulting in notable successes in identifying suspicious files. Their in-memory virus detection technology effectively identifies and eliminates highly concealed viruses. Ruixing, on the other hand, has ventured into semi-virtual machine technology as a means to detect previously unknown viruses. Jiangmin, meanwhile, has implemented the "signature broad-spectrum method" to detect and thwart certain forms of deformed viruses. This diversity of approaches, in line with the "hundred schools of thought contend, each with its own merits", collectively furnishes effective security assurances for the computer systems of most users [18].

The authors introduced an innovative Android virus software detection method that combines the strengths of learning-based and signature-based approaches. This method automatically synthesizes semantic malware signatures from a limited number of instances within a virus software family. The common functionalities shared by the virus software family are captured by the pattern represented through the call graph between components of Android applications. Utilizing MaxSAT, the virus software signature is synthesized by detecting the Maximum Suspicious Common Subgraph (MSCS) within a group of ICCG [2].

Optimized network structures are explained using genetic algorithms, comparing the performance with traditional Back-Propagation (BP) and Levenberg–Marquardt (LM) algorithms. Their comparison reveals that the GA-LM model accurately predicts Dissolved Oxygen (DO) values and outperforms traditional neural networks as a rapid interpolation and extrapolation tool [16].

The authors introduced executable malware detection technology, leveraging cluster analysis techniques to categorize executable files into multiple feature regions. This approach uses cluster analysis to detect known and unknown malware and measures byte distribution similarity between malicious and normal executable files. Consequently, this technology efficiently identifies malicious software without complex command analysis, minimizing system execution overhead [15].

The contributors introduced an active virus defence approach that employs object-oriented association mining technology within the Windows platform. By conducting a static analysis of WinAPI call sequences within PE files and integrating it with OOA mining technology, authors have developed and executed a DMAV system [22].

3. Research Methods. The primary structure of the DMAV system is illustrated in Figure 3.1. Initially, when dealing with potentially suspicious PE files that have been compressed (e.g., UPX, ASPack, etc.), the PE file parser extracts all WinAPI call sequences from the imported table, following the order of code execution. Each function within the sequence is assigned a unique 32-bit integer ID through a database query based on the API calls. This research employs the same methodology to derive API functions called by all samples within the training dataset. Subsequently, the WinAPI call sequences extracted from all samples are stored as integer vectors in the feature database. To actively defend against polymorphic and unknown viruses, this study incorporates the OOA mining algorithm from data mining technology to analyze the feature database.

Through a rule generator, it identifies and stores association rules that fulfil specific criteria in a rule database. To determine whether a suspicious PE file is a virus, the system compares the WinAPI sequence it generates with the antecedents of each rule in the rule library. The suspicious PE file is categorized as a virus if the distance between vector V_u and V_s surpasses a predefined threshold [6]. This approach offers the advantage of rapid scanning and detecting numerous PE files while maintaining a high detection rate for known viruses.

3.1. PE file parser. The Portable Executable (PE) file format is the predominant executable file format within the current Windows platform. In this paper, we have developed a PE file analyzer designed to extract all WinAPI call sequences found in the PE file's import table. The implementation of this parser is structured into three key steps:

Step 1: The import table is initially located within the PE file, following the PE file structure. This table contains pointers to the serial numbers and names of all WinAPI input functions.

Fig. 3.1: Main structure of DMAV system

Table 3.1: Sample file database

ID	API sequence	Sample category
1	API1,API5	BENIGN
2	API1,API3	TROJAN
3	API1,API2,API4,API5	BENIGN
4	API1,API2,API3,API4,API5	WORM
5	API3,API5	BACKDOOR
6	API2,API4	BENIGN
7	API1,API4,API5	SPYWARE
8	API2,API5	BENIGN

Step 2: Next, all code segments are systematically scanned within the PE file, extracting the corresponding call instructions and their associated target addresses based on the sequence of code execution. Subsequently, the corresponding WinAPI input functions are identified by cross-referencing the target addresses with the input address table.

Step 3: Using the API query database, each exported WinAPI function is mapped to a globally unique 32-bit integer API ID number to streamline the process. This database stores the names and ID numbers of commonly used Win32DLLs and their API functions. By representing the WinAPI sequence as a 32-bit integer vector, we significantly enhance the efficiency of similarity measurement, saving valuable computation time [12, 8].

3.2. Rule generator. Both decision trees and Bayesian networks have been employed in virus detection, but these methods are susceptible to a common issue known as over-learning. Over-learning occurs when a model performs well on training data but poorly on test data due to excessive complexity or inadequate training data. Excessive model complexity can lead to memorization of the training data's characteristics without the ability to generalize to new test data. In contrast, insufficient training data prevents the model from acquiring enough information for effective learning.

This paper integrates the OOA mining algorithm from data mining technology to address these challenges. It analyses 5000 viruses and 2000 non-virus samples using a PE file parser, storing their distinctive attributes in a database, as shown in Table 3.1. Subsequently, it mines association rules tailored to specific objectives through a rule generator and archives them in a rule database. This paper optimizes feature extraction processes to enhance system efficiency further [5, 10].

In the context of association rule mining, the OOA mining algorithm distinguishes itself from constraint-

based association mining by focusing on extracting association rules that fulfil specific objectives and demonstrate utility. The rule extraction process employed in this paper exemplifies an application of OOA mining. As illustrated in Table 3.1, we are engaged in mining association rules that satisfy the target criteria Obj = (Group = MALICIOUS (TROJAN ∨ WORM ∨ BACKDOOR ∨ SPYWARE)), essentially aiming to differentiate Obj = (Group = ¬ BENIGN).

3.2.1. Related concepts. Set the database DB, and the item set $I = \{i_1, i_2, \ldots, i_m\}$ is composed of two parts: $I = I_{\text{obj}} \cup I_{\text{nobj}}, I_{\text{obj}} \cap I_{\text{nobj}} = \emptyset, I_{\text{obj}}$ is the item set that constitutes the target, I_{nobj} is the antecedent that constitutes the association rule to be mined, and the transaction set T has n transactions, specifying the minimum target support mos% and the minimum target confidence moc%.

Definition 3.1. *Objective (Obj) is a logical formula composed of items in* I_{obj}. $I_{\text{obj}} = \{Group\}, \text{Obj} = (Group = \neg\text{BENIGN})$ *can be set. If* Obj *is true in a transaction* $t \in T, t$ *is said to meet the goal. If there is* $X \subseteq t\,(X \subseteq I_{nobj})$ *at the same time, then* $X \rightarrow$ Obj *is satisfied in* t. *OOA mining is to mine rules like* $X \rightarrow$ Obj.

Definition 3.2. *If project set* $X = \{a_1, a_2, \cdots, a_i\}$ *is set, and* $X \subseteq I_{nobj}$, *the transaction count that meets* X *is count* (X, DB), *and the transaction count that meets* $X \rightarrow Obj$ *is count* $(X \rightarrow \{Obj\}, DB)$, *then the target support os% and target confidence oc% of* $(X \rightarrow \text{Obj})$ *are followed using Equation* (3.1)*:*

$$\begin{aligned} os\ \% &= \frac{\text{count}(X \cup \{Obj\}, DB)}{|DB|} \times 100\% \\ oc\ \% &= \frac{\text{count}(X \cup \{Obj\}, DB)}{\text{count}(X, DB)} \times 100\% \end{aligned} \tag{3.1}$$

If os% $\geq$ mos %, X *is considered as frequent* OOA.

Definition 3.3. *If the project set* X *has* os% $\geq$ mos % *and* oc% $\geq$ moc %, *then* $X \rightarrow$ Obj(os%, *oc%) is an OOA rule.*

Theorem 3.4. *If the item set* X *is OOA frequent,* $\forall Y \subset X, Y \neq \emptyset$, *then* Y *is also OOA frequent.*

Theorem 3.5. *If the item set* X *is not* OOA *frequent,* $\forall Y \supset X, Y \subseteq$, *then* Y *is also* OOA *frequent.*

3.2.2. Main algorithm implementation of rule generator. OOA mining meets the two basic properties of the Apriori algorithm, so OOA mining can be implemented with the Apriori algorithm.

Taking Table 3.1 as an example, if mos%=25%, moc%=65%, then the frequent set FP={API1}, {API2}, {API3}, {API4}, {API5}, {API1, API3}, {API2, API4}, {API2, API5}, {API3, API5}, {API4, API5}, {API2, API4, API5}. The rules obtained are: 1) {API3} → Obj (37.5%, 100%); 2) {API1,API3} → Obj(25%,100%); 3) {API3,API5} → Obj(25%,100%); 4) {API4,API5} → Obj(25%,66.7%); 5) {API2,API4,API5} → Obj(25%,66.7%).

3.3. Similarity measurement. The similarity measurement is implemented in two steps, namely, sequence rearrangement and similarity calculation.

3.3.1. Sequence rearrangement. A sequence rearrangement procedure is employed based on two feature vectors to enhance the precision of similarity measurement between the two vectors before performing the similarity calculation. This sequence rearrangement algorithm can be executed using a matrix, as depicted in Figure 3.2.

The specific implementation of sequence rearrangement is as follows: If the first row and first column of matrix $A_{(M+1)*(N+1)}$ are row 0 and column 0, then a_{ij} can be obtained using Equation (3.2),

$$\begin{aligned} a_{ij} = \begin{cases} 0, \text{ If } a_i 0 \neq a_{a_j} \\ 1, \text{ If } a_i 0 = a_{a_j} \end{cases}, (i \in [1, M], j \in [1, N], \\ M = \text{ length(Sequence1)}, N = \text{ length(Sequence2))} \end{aligned} \tag{3.2}$$

Let's explore an illustrative instance of the sequence rearrangement algorithm to provide a more concrete demonstration. Following the meticulous sequence rearrangement procedure outlined in Figure 3.3, this paper intentionally inserts the value 0 into any vacant positions within the sequence. This deliberate step gives rise to the creation of two fresh API sequences, denoted as $BV_s{}'$ and $V_u{}'$.

	W	A	N	D	D	R
W	X					
A		X				
R				X		
E					X	
R						X
S						

Fig. 3.2: Sequence rearrangement algorithm represented by matrix

	W	A	N	O	E	R
W	1	0	0	0	0	0
A						
D						
E						
R						
S						

	W	A	N	O	E	R
W	1	0	0	0	0	0
A	0	2	1	1	1	1
D						
E						
R						
S						

	W	A	N	O	E	R
W	1	0	0	0	0	0
A	0	2	1	1	1	1
D	0	1	2	3	2	2
E						
R						
S						

	W	A	N	O	E	R
W	1	0	0	0	0	0
A	0	2	1	1	1	1
D	0	1	2	3	2	2
E	0	1	2	2	4	3
R						
S						

	W	A	N	O	E	R
W	1	0	0	0	0	0
A	0	2	1	1	1	1
D	0	1	2	3	2	2
E	0	1	2	2	4	3
R	0	1	2	2	3	5
S						

	W	A	N	O	E	R
W	1	0	0	0	0	0
A	0	2	1	1	1	1
D	0	1	2	3	2	2
E	0	1	2	2	4	3
R	0	1	2	2	3	5
S	0	1	2	2	3	4

Fig. 3.3: Example of sequence rearrangement algorithm

3.3.2. Similarity calculation. The similarity is calculated by taking the mean value of Equations (3.4) to (3.6). The most commonly used method is Euclidean distance method to calculate the similarity between two vectors, as given by Equation (3.3). However, Euclidean distance cannot accurately measure the similarity between two vectors. Equation (3.4) calculates the cosine of the angle between two vectors $V_s{}'$ and $V_u{}'$ to determine their similarity, Equation (3.5) is the Jaccard extended cosine algorithm, and Equation (3.6) is the

Table 4.1: Identifying deformed viruses using various virus-scanning software

Samples	N	M	D	K	SAVE	DMAV
Beagl	✓	✓	✓	✓	✓	✓
Beagl1	✓	✓	×	✓	✓	✓
Beagle2	✓	×	×	✓	✓	✓
Beagle3	×	×	×	✓	×	✓
Beagle4	✓	✓	×	×	×	✓
Blaster	✓	✓	✓	✓	✓	✓
Blaster1	✓	✓	✓	✓	✓	✓
Blaster2	×	×	×	×	×	✓
Lovedoor	✓	✓	✓	✓	✓	✓
Lovedoor1	×	×	✓		✓	✓
Lovedoor2	×	×	×	×	×	✓
Lovedoor3	×	✓		✓	×	✓
Mydoom	✓	✓	✓	✓	✓	✓
Mydoom1	×	×	×	×	×	✓
Mydoom2	×	×	×	?	✓	✓

Pearson correlation measurement algorithm [20].

$$D\left(V_s', V_u'\right) = \sum_{i=1}^{\min\left(\left|V_s'\right|,\left|V_u'\right|\right)} \left[\left(V_{s_i}' - V_{u_i}'\right)^2\right]^{1/2} \tag{3.3}$$

$$s^{(C)}\left(V_s', V_u'\right) = \frac{V_s'^T V_u'}{\left\|V_s'\right\|_2 \cdot \left\|V_u'\right\|_2}, \|v\|_p = \left[\sum_{i=1} \mid V_i|^7\right]^{1/p} \tag{3.4}$$

$$S^{(J)}\left(V_s', V_u'\right) = \frac{V_s'^T V_u'}{\left\|V_s'\right\|_2^2 + \left\|V_u'\right\|_2^2 - V_s'^T V_u'} \tag{3.5}$$

$$s^{(P)}\left(V_s', V_u'\right) = \left[\frac{\left(V_s' - \nabla_s'\right)^T \left(V_u' - \nabla_u'\right)}{\left\|\left(V_s' - \nabla_s'\right)\right\|_2 \cdot \left\|\left(\left(V_u' - \nabla_u'\right)\right)\right\|_2} + 1\right] /2 \tag{3.6}$$

4. Result Analysis and Discussion. The DMAV system is developed within the VC++6.0 environment, coupled with a MySQL database. VC++6.0 is an integrated development environment well-suited for C++ development on the Windows platform. It offers robust tools and libraries that expedite the creation of Windows applications, including virus detection systems. Utilizing VC++6.0 streamlines the development of Windows GUI applications, and frameworks like MFC can be employed to simplify the development process further. The experiment used 5,000 virus samples and 2,000 non-virus samples for rule extraction, while 150 virus samples and 500 non-virus samples served as test specimens. All samples were sourced from the Jinshan Antivirus Laboratory. The experiment is primarily divided into two components: detecting deformed viruses and detecting unknown viruses.

4.1. Detection of deformation virus. The experiment and analysis involved Win32PE virus samples, including well-known examples such as Lovedoor, Mydoom, Blaster, and Beagle. Virus deformation technology is employed for each type of virus sample to transform these samples into various deformed versions. Subsequently, different virus scanning software and the DMAV system were used to scan and assess these deformed viruses. The experiment outcomes are summarized in Table 4.1, revealing that the DMAV system consistently exhibits superior accuracy in detecting deformed viruses compared to other conventional virus scanning software.

Table 4.2: Detection of unknown viruses through different virus-scanning software

Malware Samples	N	M	D	K	SAVE	DMAV
1	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓
3	✓	✓	×	✓	✓	✓
4	✓	×	×	×	×	×
5	×	✓	✓	×	✓	✓
6	✓	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓	✓
8	×	×	×	×	✓	✓
9	✓	✓	✓	✓	×	✓
10	?	✓	×	✓	✓	×
… …	… …	… …	… …	… …	… …	… …
150	✓	×	×	×	✓	✓
Statistics	50	68	48	75	82	138
Ratio/%	33.4	453	32	50	54.8	92

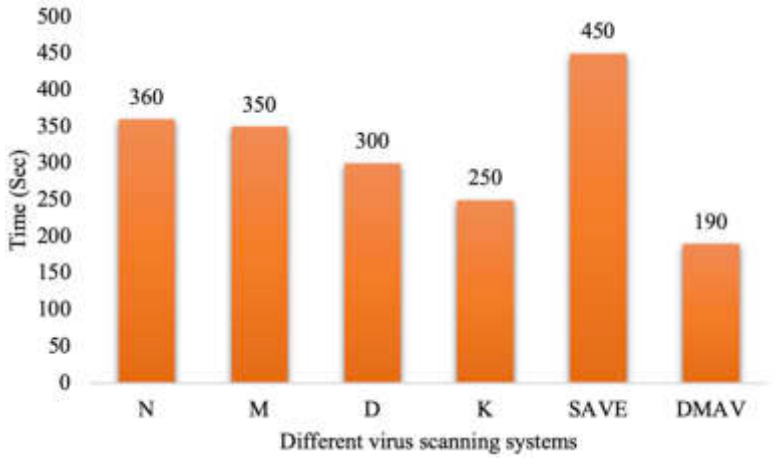


Fig. 4.1: Comparison of the efficiency of different virus scanning systems

4.2. Detection of unknown virus. To evaluate the detection capabilities of the DMAV system concerning unknown viruses, 150 virus samples with undisclosed characteristics were subjected to scanning using various virus scanning software and the DMAV system. The outcomes of this experiment are presented in Table 4.2, highlighting that the DMAV system demonstrates notably heightened efficiency in actively defending against unknown viruses compared to other conventional antivirus software. Impressively, the DMAV system achieves a recognition rate of 92%.

Remarks: ✓indicates successful detection, × indicates failed detection, and ? indicates suspicious. All scanning software adopts the latest version. N-Norton, M-McAfee, D-DF Web, K-Kaspersky, SAVE-Static Analyzer for Vicious Executable, DMAV-Data Mining-based, Antivirus system.

To assess the system's efficiency and false positive rate (FP), we conducted experiments within the same testing environment as the SAVE system. The SAVE system is an open-source network security monitoring and defence system with many functions, including intrusion detection, vulnerability scanning, and traffic monitoring. The SAVE system incorporates various technologies, such as rule-based detection, anomaly-based detection, and machine learning, to enhance detection efficiency and diminish false positives.

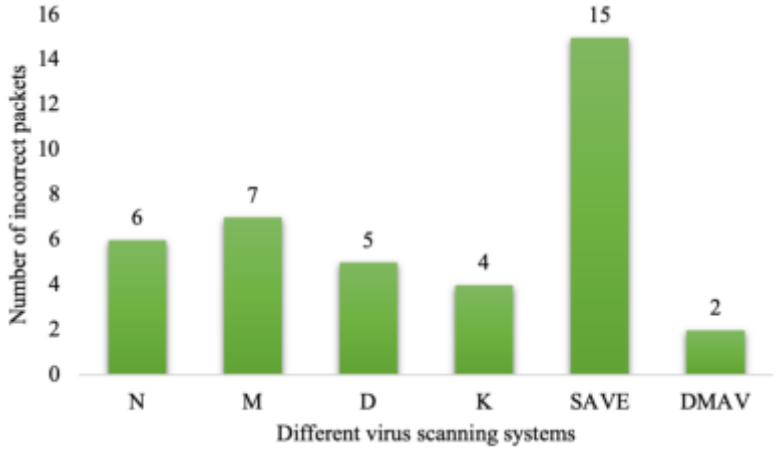


Fig. 4.2: Comparison of false alarm rates of different virus scanning systems

The proposed system is evaluated using the test environment consisting of an Intel P4 1GHz processor with 1GB of RAM, running on MS Win2000. The experimental outcomes, illustrated in Figures 4.1 and 4.2, underscore that the DMAV system exceeds other virus scanning software's efficiency and capacity to maintain a relatively lower false positive rate when detecting non-virus files than traditional virus scanning software [3, 19].

5. Conclusion. By employing static PE file analysis to extract WinAPI call sequences, a proactive virus defence system named DMAV has been proposed, operating within the Windows platform and rooted in data mining technology. DMAV performs the dual function of detecting deformed viruses and actively safeguarding against unknown ones. This system consists of three fundamental modules: a PE file parser, a rule generator, and a similarity measurement algorithm. The rule generator, empowered by OOA mining, streamlines feature extraction processes. Empirical findings demonstrate that DMAV exceeds conventional virus scanning software in terms of accuracy, efficiency, and false alarm rate when detecting unknown and deformed viruses. The computer virus defence domain, propelled by data mining, is dynamic and cutting-edge. As computer networks continue to evolve, traditional antivirus methodologies prove inadequate, prompting researchers to integrate data mining into virus defence strategies seamlessly. The forthcoming developments in data mining-driven virus defence may encompass the adoption of deep learning for more precise models and the utilization of multimodal data analysis techniques to enhance accuracy through the joint analysis of various data types, including text, images, videos, and audio, thereby stimulating network security in the face of ever-evolving virus threats.

REFERENCES

[1] A. T. Azar, E. Shehab, A. M. Mattar, I. A. Hameed, and S. A. Elsaid, *Deep learning based hybrid intrusion detection systems to protect satellite networks*, Journal of Network and Systems Management, 31 (2023), p. 82.
[2] M. Campion, M. Dalla Preda, and R. Giacobazzi, *Learning metamorphic malware signatures from samples*, Journal of Computer Virology and Hacking Techniques, 17 (2021), pp. 1–17.
[3] D. Dasgupta, Z. Akhtar, and S. Sen, *Machine learning in cybersecurity: a comprehensive survey*, The Journal of Defense Modeling and Simulation, 19 (2022), pp. 57–106.
[4] A. B. de Neira, B. Kantarci, and M. Nogueira, *Distributed denial of service attack prediction: Challenges, open issues and opportunities*, Computer Networks, 222 (2023), p. 109553.
[5] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, and A. Armando, *Functionality-preserving black-box optimization of adversarial windows malware*, IEEE Transactions on Information Forensics and Security, 16 (2021), pp. 3469–3478.
[6] M. Gharacheh, V. Derhami, S. Hashemi, and S. M. H. Fard, *Proposing an hmm-based approach to detect metamorphic malware*, in Proceedings of the 4th Iranian Joint Congress on Fuzzy and Intelligent Systems, Zahedan, Iran, 2015, IEEE, pp. 1–5.
[7] E. Gyamfi and A. Jurcut, *Intrusion detection in internet of things systems: a review on design approaches leveraging multi-access edge computing, machine learning, and datasets*, Sensors, 22 (2022), p. 3744.

[8] X. Huang, L. Ma, W. Yang, and Y. Zhong, *A method for windows malware detection based on deep learning*, Journal of Signal Processing Systems, 93 (2021), pp. 265–273.
[9] A. K. Jha, A. Vaish, and S. Patil, *A novel framework for metamorphic malware detection*, SN Computer Science, 4 (2022), p. 10.
[10] D.-y. Kim, A.-y. Jeong, and T.-j. Lee, *Analysis of malware group classification with explainable artificial intelligence*, Journal of the Korea Institute of Information Security & Cryptology, 31 (2021), pp. 559–571.
[11] Q. Li, J. Hou, S. Meng, and H. Long, *GLIDE: a game theory and data-driven mimicking linkage intrusion detection for edge computing networks*, Complexity, 2020 (2020), pp. 1–18.
[12] Y. T. Ling, N. F. M. Sani, M. T. Abdullah, and N. A. W. A. Hamid, *Metamorphic malware detection using structural features and nonnegative matrix factorization with hidden markov model*, Journal of Computer Virology and Hacking Techniques, 18 (2021), pp. 1–21.
[13] Z. Lv, D. Chen, B. Cao, H. Song, and H. Lv, *Secure deep learning in defense in deep-learning-as-a-service computing systems in digital twins*, IEEE Transactions on Computers, (2023).
[14] A. Madhuri, V. E. Jyothi, S. P. Praveen, S. Sindhura, V. S. Srinivas, and D. L. S. Kumar, *A new multi-level semi-supervised learning approach for network intrusion detection system based on the 'goa'*, Journal of Interconnection Networks, (2022), p. 2143047.
[15] R. Ognev, E. Zhukovskii, and D. P. Zegzhda, *Detection of malicious executable files based on clustering of activities*, Automatic Control and Computer Sciences, 55 (2021), pp. 1092–1098.
[16] A. A. Ojugo, C. O. Obruche, and A. O. Eboka, *Quest for convergence solution using hybrid genetic algorithm trained neural network model for metamorphic malware detection*, ARRUS Journal of Engineering and Technology, 2 (2022), pp. 12–23.
[17] T. Saba, A. Rehman, T. Sadad, H. Kolivand, and S. A. Bahaj, *Anomaly-based intrusion detection system for iot networks through deep learning model*, Computers and Electrical Engineering, 99 (2022), p. 107810.
[18] V. F. Santos, C. Albuquerque, D. Passos, S. E. Quincozes, and D. Mossé, *Assessing machine learning techniques for intrusion detection in cyber-physical systems*, Energies, 16 (2023), p. 6058.
[19] S. M. Shareef and S. H. Hashim, *Proposed hybrid classifier to improve network intrusion detection system using data mining techniques*, Engineering and Technology Journal, 38 (2020), pp. 6–14.
[20] D. Shin and J. Shim, *A systematic review on data mining for mathematics and science education*, International Journal of Science and Mathematics Education, 19 (2021), pp. 639–659.
[21] M. U. Ullah, A. Hassan, M. Asif, M. Farooq, and M. Saleem, *Intelligent intrusion detection system for apache web server empowered with machine learning approaches*, International Journal of Computational and Innovative Sciences, 1 (2022), pp. 21–27.
[22] C. Xiong, Z. Li, Y. Chen, T. Zhu, J. Wang, H. Yang, and W. Ruan, *Generic, efficient, and effective deobfuscation and semantic-aware attack detection for powershell scripts*, Frontiers of Information Technology & Electronic Engineering, 23 (2022), pp. 361–381.
[23] H. Xu, Z. Sun, Y. Cao, and H. Bilal, *A data-driven approach for intrusion and anomaly detection using automated machine learning for the internet of things*, Soft Computing, (2023), pp. 1–13.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Mar 21, 2023
Accepted: Sep 30, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 55–64, DOI 10.12694/scpe.v25i1.2227

APPLICATION OF NONLINEAR BIG DATA ANALYSIS TECHNIQUES IN COMPUTER SOFTWARE RELIABILITY PREDICTION

LI GAO*AND HAI WANG†

Abstract. This research addresses the difficulties of underfitting, overfitting, and convergence to local minima in artificial neural networks for software dependability prediction. The work specifically focuses on enhancing the performance of the conventional PSO-SVM model for software reliability prediction. The analysis of the conventional PSO-SVM model and the special features of software reliability prediction serve as the foundation. An improved PSO-SVM software reliability prediction model is developed and the PSO-SVM model and a Backpropagation (BP) prediction model are compared experimentally. The critical metrics assessed include training error, and efficiency. The experimental results reveal that the training error of the enhanced PSO-LSSVM prediction model diminishes rapidly, levelling off after approximately 200 training generations. The BP prediction model requires 1,733 generations to meet training requirements. Furthermore, the improved PSO-LSSVM prediction model demonstrates significantly higher training efficiency than the BP prediction model. The optimized prediction model exhibits superior adaptability to small sample sizes, swift training, and high prediction accuracy, making it a more suitable choice for software reliability prediction applications.

Key words: Particle swarm optimization algorithm, Support vector machine, Software reliability, Prediction, Training error, Efficiency

1. Introduction. The rapid development of the Internet industry has unleashed of innovation and transformation across various sectors. This surge in technological advancement has empowered industries to harness the benefits of the Internet, such as its wide-reaching accessibility, and the seamless sharing of information. Consequently, numerous industries have seized the opportunity to develop software systems and products tailored to their specific needs, all in pursuit of greater efficiency and competitiveness. The evolution of computer software technology has emerged as a key player towards advancing information technology. These vital sectors now largely rely on sophisticated computer software systems to run their businesses and deliver services to a global client. High levels of reliability and security are necessary in a world where these systems and goods function in an open and linked network environment [10].

The dependability and safety of these software-driven systems and products are guaranteed by multifaceted strategy. First and foremost, innovative approaches and strict standards must be incorporated into the software development process. Considerable code reviews, test-driven development techniques, and adoption of alert development practices are a few examples. By doing this, programmers can guarantee the reliability and stability of their work, lowering the possibility of unanticipated failures. Second, security issues should be taken into account during the design and implementation of the product. Access control techniques and encryption technologies are essential for protecting sensitive data inside of these systems and goods. This proactive strategy assists in reducing potential weaknesses and prevents unauthorized access to sensitive information [16].

To assess security vulnerabilities within the software regularly and it is entailed by conducting routine vulnerability scans and swiftly implementing remedies to patch any identified weaknesses. Additionally, robust measures such as data backup and recovery systems should be in place to strengthen the system's flexibility in the face of unexpected disruptions or data loss. In today's digital era, the consequences of software failures cannot be underestimated. These failures can potentially cause significant disruptions and losses, affecting not only individual lives but also the trajectory of societal progress. For example, the tragic incident during the first Gulf War on February 25, 1991, where the U.S. Patriot missile system in Saudi Arabia failed to intercept

*Faculty of Computer and Software Engineering, HuaiYin Institute of Technology, Huai'an, 223003, China
†Civil and Architectural Engineering College, Nanning University, Nanning, 530000, China (Corresponding Author: haiwang39@126.com)

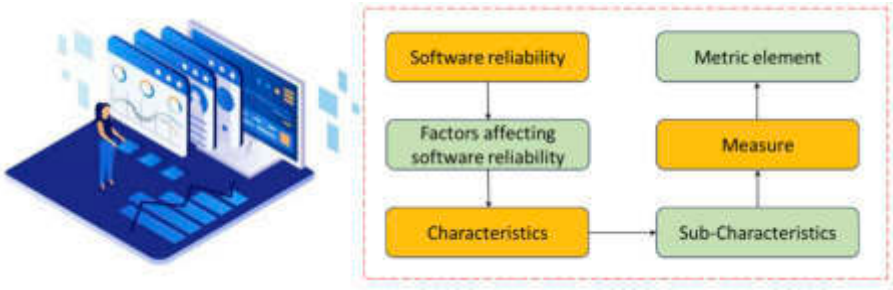


Fig. 1.1: Computation and utilization in forecasting software reliability

an Iraqi Scud missile by a mere 0.34-second error. This resulted in the tragic loss of 28 soldiers and underscored the critical importance of software reliability in military operations [1].

The aerospace industry experienced a devastating software failure on June 4, 1996, during the inaugural launch of the Ariane 5 rocket. A software malfunction caused the rocket to turn off course and explode 37 seconds after liftoff, destroying four valuable solar wind observation satellites. This incident, one of the costliest software failures in the history of space exploration, served as a stark reminder of the need for meticulous software engineering in mission-critical systems [2].

Moving closer to home, the 2011 launch of China's 12306 network ticketing system, developed for 300 million RMB, was met with high demand during the Spring Festival. However, it yielded to many issues, including website crashes, prolonged response times, payment glitches, and ticketing problems. This unfortunate episode exposed the breakability of even well-funded software systems, shedding light on architectural flaws and inadequacies in handling a massive surge in ticket transactions [12].

The vulnerabilities within the overall system architecture can be attributed to various factors. Often, system designs are not sufficiently comprehensive to account for the interactions and dependencies among different functional modules, resulting in architectural weaknesses. Additionally, programming errors, algorithmic flaws, and other implementation defects can compromise a system's normal operation or efficiency. Improper system operation and maintenance practices, such as failing to consider the impact of architectural changes during upgrades or configuration adjustments, can also lead to system failures [15].

The early iteration of the online ticketing system failed to incorporate identity verification measures, enabling scalpers to exploit the system's vulnerabilities during the Spring Festival. Even today, the system grapples with recurring challenges during peak usage, including network congestion, connection instability, and queue management issues. This ongoing struggle highlights the pressing need for sustained efforts to ensure website reliability and security. Figure 1.1 details the impact of software failures in the aviation sector. Within the aviation sector, rigorous testing protocols are followed to evaluate the software's performance. Before the first flight of an aircraft type, more than 600 software failures are meticulously scrutinized on the ground. Software-related issues have emerged as the predominant cause of aviation incidents, emphasizing the critical importance of software reliability and safety in the aviation industry [4].

The paper is organized as follows. Section 2 presents a comprehensive literature review to contextualize the research within the existing body of knowledge. The proposed PSO-SVM network is examined in detail in Section 3. The research's findings are presented in Section 4 along with an explanation. Section 5 concludes with a summary of the main findings.

2. Literature Review. In the 1970s, as information technology continued to advance, the scope of computer applications expanded gradually, leading to a sharp increase in the demand for software. However, due to the relatively primitive state of software production and management, the field remained stagnant at a level of the 1960s when computers were first introduced. This inadequacy failed to keep pace with the rapidly growing requirements of computer software development, resulting in what became known as the software crisis [3].

The software crisis established itself across several critical dimensions. First, controlling the progress of software development became a difficult challenge. Given the intricate nature of software development,

its complex designs and testing procedures, and the need for collaboration among multiple team members, maintaining a firm grip on the development timeline proved elusive, often leading to unsatisfying delays. Second, ensuring software quality proved to be equally frustrating. The absence of standardized and normalized practices within the software development process hindered the effective guarantee of software quality. Consequently, software frequently exhibited vulnerabilities and errors, sometimes resulting in system crashes and data loss [11].

The managing software costs modelled yet another tough hurdle. Owing to the inherent difficulties and problems within the software development process, cost control remained elusive, often resulting in expenditures that exceeded the established budget. Meanwhile, software systems continued to grow in scale and complexity without commensurate improvements in reliability. This mismatch led to substantial economic losses and even casualties in significant accidents [14].

A poignant illustration of the consequences of software reliability issues occurred in 2016 when the japan aerospace exploration agency (JAXA) and national aeronautics and space administration (NASA) jointly launched the X-ray astronomy satellite "Hitomi". A month after its successful launch, "Hitomi" experienced a catastrophic ground communication failure, leading to a complete loss of contact with mission control. Two months later, JAXA declared they could not regain control of the X-ray satellite "Hitomi" and were forced to abandon it [8].

Another tragic incident that underscores the gravity of software-related issues occurred on March 10, 2019, when an Ethiopian Airlines Boeing 737-MAX crashed, resulting in the tragic loss of all 157 passengers and crew members. Investigations revealed a direct link between the crash and the incorrect activation of the automatic stall prevention software known as MCAS. These real-world instances underscore the critical significance of addressing software reliability concerns. As technology advances, the imperative to enhance software development processes, elevate quality control, and mitigate vulnerabilities becomes increasingly urgent. This ensures that software can serve as a catalyst for progress rather than a source of crises. Over the decades, these painful lessons have served as cautionary stories for software practitioners, prompting the industry to grow more concerned about software reliability [9].

This heightened awareness has led to continuous development and the gradual integration of engineering principles into the software realm. In 1968 and 1969, the concept of software engineering was introduced during consecutive north atlantic treaty organization (NATO) meetings. This marked a pivotal shift as engineering principles began to guide software development practices. With software's evolution into software engineering, significant advancements were made in development technologies and management tools. Simultaneously, expectations for software reliability soared [6].

Seizing the momentum created by the burgeoning field of software engineering and drawing inspiration from traditional reliability engineering techniques, software reliability engineering emerged as a distinct discipline. This research uses SVM and PSO algorithms to investigate software reliability prediction. The inherent limitations and drawbacks of conventional PSO and SVM algorithms is investigated by examining potential optimization strategies. Consequently, an optimized PSO-SVM software reliability prediction model is established. The model's impressive predictive accuracy and applicability is demonstrated through practical examples involving limited early-stage software data samples [17, 7].

3. Research Methodology.

3.1. Analysis of traditional PSO-SVM features. The traditional PSO-SVM model is a SVM prediction model that relies on the PSO algorithm. This model possesses several noteworthy characteristics. Firstly, it is distinguished by its simplicity and ease of implementation, requiring no complicated mathematical theories or algorithmic foundations. Understanding the fundamental principles and implementation procedures of both PSO and SVM is sufficient for its deployment. Secondly, the PSO algorithm's inherent global optimization capabilities are a key feature, enabling the model to evade local optima and enhance prediction accuracy. The PSO-SVM model consistently leverages the PSO algorithm to optimize model parameters and SVM kernel parameters, improving SVM's predictive accuracy by identifying and employing the most favourable parameter combinations. Initially designed for binary classification problems, SVM later expanded into the nonlinear regression prediction. A support vector machine for nonlinear regression prediction closely resembles a classification problem as it computes a decision function based on provided data, leading to classification and prediction

outcomes. However, the regression problem retains the core characteristics of maximizing the convex function over time, with the capability to obtain nonlinear functions directly through specialized kernel functions [13].

Suppose the given data set is $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_l, y_l)\}$, where $x_i \in R^n, y_i \in R, i = 1, 2, \ldots, l$. The value of y_i for classification problems is the number of corresponding data types. For example, -1 and 1 can be used for binary classification problems, while any actual number can be used for regression problems. The original SVM problem can be expressed as Equation (3.1).

$$\begin{aligned} &\min_{\boldsymbol{\omega}\in\mathbf{R}^n, b\in\mathbf{R}} J(\boldsymbol{\omega}, \xi) = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ &\text{s.t.} \quad (\boldsymbol{\omega}\cdot\boldsymbol{\varphi}(x_i) + b) - y_i \leqslant \xi_i^* + \varepsilon; i = 1, 2, \cdots, l \\ &\quad y_i - (\boldsymbol{\omega}\cdot\boldsymbol{\varphi}(x_i) + b) \leqslant \xi_i + \varepsilon^*; i = 1, 2, \cdots, l \\ &\quad \xi_i, \xi_i^* \geqslant 0; i = 1, \cdots, l \end{aligned} \tag{3.1}$$

Given the substantial size of the eigenspace and the non-convergent nature of the objective function, we incorporate point object kernel function techniques and Wolff dual theory. These additions partition the problem into two sets of computationally manageable subproblems. One of these subproblems involves transforming the original challenge into a quadratic programming problem.

By using PSO, the model C, ε of SVM and kernel parameters σ^2 are optimized, the population continuously learns from the most available position in the current generation and the global most available position during the iterative update. Suppose the population size is m and the d^{th} dimensional space position of the ith particle is $x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$, and the velocity is $v_i = \{v_{i1}, v_{i2}, \ldots, v_{id}\}$, the optimal position of this generation is $pbest_i = \{p_{i1}, p_{i2}, \ldots, p_{id}\}$, and the global optimal position is $gbest = \{g_1, g_2, \ldots, g_d\}$, and d is determined by the dimensionality of the characteristic attributes of the target problem, and the fitness function is set according to the function for which the target problem is being optimized [18].

3.2. Analysis of model applicability and optimization counter measures. The conventional PSO-SVM model enhances final prediction results by optimizing the model and SVM kernel parameters using the PSO algorithm. This model exhibits numerous notable advantages that align seamlessly with the prediction attributes of software reliability. The benefits of this prediction model correspond to the distinctive characteristics of software reliability prediction in the following ways [5]:

1) SVM achieves the adjustment of the proportional relationship between structural risk and empirical risk via the modulation of model parameters. This mechanism effectively mitigates the issue of over-learning, subsequently enhancing the prediction model's capacity for generalization. Such an approach is particularly well-suited for addressing the challenges of acquiring software reliability samples, especially in cases with limited data.

2) Incorporating a kernel function into the prediction model, SVM efficiently transforms the multidimensional input space into a higher-dimensional space for prediction purposes. This transformation effectively addresses the challenges modelled by high dimensionality within the input space, making it highly compatible with the abundance of parameters typically associated with software reliability features.

3) The conventional PSO-SVM model conducts predictions in a high-dimensional space, initially transforming the original nonlinear problem into a prediction problem and subsequently deducing solutions for the nonlinear problem, thereby enhancing efficiency. This approach aligns well with the inherently nonlinear nature of software predictions.

However, despite the advantages of the traditional PSO-SVM prediction model, it grapples with computational limitations and inherent deficiencies in the PSO and SVM algorithms, rendering the model somewhat inadequate. The PSO-SVM model necessitates the adjustment of numerous parameters, including inertia weight, learning factor, and kernel function parameters, among others. Moreover, these parameters exhibit interdependencies, mandating extensive experimentation and adjustments, thereby compounding the model's complexity. The interaction between design deficiencies and proposed improvement strategies is explained as follows.

a) Coding rules originate from the experimental model. In the prediction model, when the kernel is not processed through SVM, it introduces significant randomness, thereby constraining the model's ability to

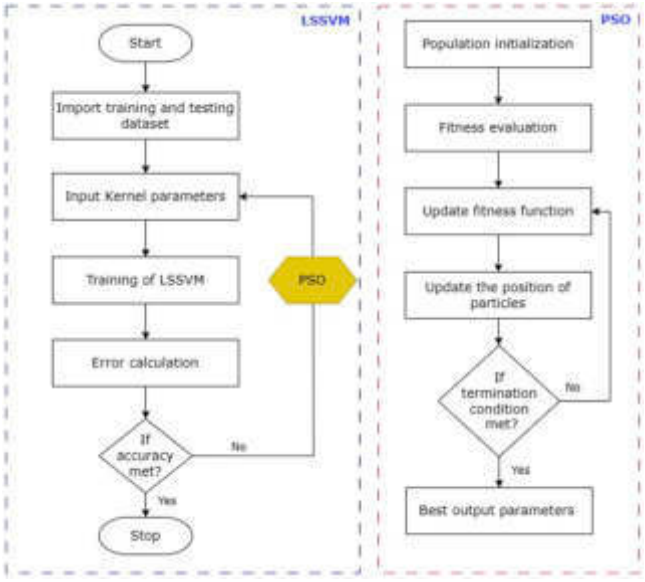


Fig. 3.1: Flowchart of improved PSO-LSSVM prediction model

undertake comprehensive global exploration in search of optimal solutions. Consequently, this can lead to a reduction in prediction accuracy and efficiency. To mitigate these limitations, a block scheme is employed to divide the initial population into several smaller subsets, facilitating the initialization of parameters. This approach enhances the diversity of the initial population, resulting in improved prediction quality and accuracy.

b) The inherent uncertainty associated with PSO within the prediction model and its constrained adaptability for conducting both global and local investigations curtails the model's capacity to achieve optimal outcomes. A variable inertia approach is employed to address this limitation, facilitating timely updates of inertia factors and optimal parameter solutions. This adjustment extends the duration of global and local surveys, enhancing the model's ability to reach optimal solutions.

c) This model transforms a low-dimensional input space into a higher-dimensional, followed by the resolution of quadratic programming problems. However, as the number of dimensions in the input space increases, computational demands grow exponentially, a significant bottleneck to the model's performance. Given the abundance of characteristic parameters in software reliability, employing the traditional PSO-SVM prediction model results in substantial computational overhead. Adapting the least squares support vector machine (LSSVM) as a modification to the SVM approach serves to streamline the SVM's details by optimizing its internal structure. Consequently, this reduces the computational burden on the prediction model and, in turn, enhances prediction efficiency.

3.3. Improved PSO-LSSVM deformation strategy. The PSO-LSSVM model represents a highly effective machine learning algorithm rooted in the SVM model's principles while optimizing model parameters through the least squares method is depicted in Figure 3.1. Additionally, it boasts robust generalization capabilities and exceptional classification performance. In the proposed model, each training sample necessitates the determination of a corresponding coefficient, ultimately forming a coefficient matrix.

The least-squares support vector machine (LSSVM) comprises two fundamental elements. Firstly, it adopts the most minor square linear system as its loss function, shifting the variable risk to the quadratic side and reformulating the support vector machine model's inequalities into equations, simplifying the constraints. Secondly, it replaces the quadratic programming approach with a system of equations to address efficiency concerns, reducing the learning complexity significantly. This transformation substantially enhances both learning efficiency and accuracy. Consequently, the original problem of the standard SVM can be simplified to the Equation (3.2)

outlined below.

$$\begin{aligned} &\min J(\boldsymbol{\omega}, \xi) = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + C\sum_{i=1}^{n}\xi_i^2 \\ &\text{s.t.} \quad y_i = \boldsymbol{\varphi}(x_i)\cdot\boldsymbol{\omega} + b + \xi_i + \varepsilon \end{aligned} \tag{3.2}$$

The block population initialization mechanism is a method employed to optimize genetic algorithms. Genetic algorithms heuristic optimization techniques simulating biological evolution rely heavily on population initialization as a pivotal step. The quality of this initialization process directly impacts the effectiveness of genetic algorithms in optimization tasks. Consequently, adopting an efficient population initialization mechanism constitutes a crucial strategy for addressing optimization challenges using genetic algorithms.

In contrast, the particle swarm algorithm (PSO) is a premier global search engine, distinguishing itself from traditional search algorithms by its exceptional global search capabilities. Therefore, when employing this algorithm, active participation in its global search process is imperative to achieve favourable global outcomes swiftly. However, during the initial stages of the algorithm, traditional particles are distributed randomly across the search space, leading to an uneven population distribution. This initial inequality can be alleviated by partitioning the search space. The main idea of using chunked population initialization is to make each particle almost uniformly distributed, assuming that the number of population particles is n, then the whole search space is divided into n regions. Each small region is randomly generated one particle is shown in Equation (3.3).

$$\begin{aligned} &X_{i,k} \in [a_k + f_k(b_k - a_k), a_k + f_{k+1}(b_k - a_k)]; k = 1, 2, \cdots, D \\ &\begin{cases} f_k = (i-1) \bmod C^{D-k} \\ f_D = (i-1) - \sum_{j=1}^{D-1} f_j C^{D-j} \\ C = \sqrt[D]{n} \end{cases} \end{aligned} \tag{3.3}$$

In Equation (3.3), a_k and b_k denote the range of values on the particle position vector in the k^{th} dimension, and mod denotes the modulus. Then the initial position of the i^{th} particle can be generated according to Equation (3.4).

$$X_{i,k} = a_k + f_k(b_k - a_k) + \frac{1}{\sqrt[D]{n}}(b_k - a_k)\cdot r \tag{3.4}$$

In Equation (3.4), r is a random number taking values within $[0, 1]$.

3.4. Inertia factor approach for adaptability. In the particle swarm algorithm, the inertia factor plays a critical role by imbuing the algorithm with historical memory during the update of particle velocities. It effectively maintains the connection between historical velocities and the global and local optima, thereby influencing the balance between global and local search capabilities. The adaptive inertia factor strategy dynamically adjusts the inertia factor's magnitude based on the current search state of the particle swarm. Consequently, it assigns a higher value to the inertia factor during the early stages of the search to stimulate global exploration. As the search progresses, the inertia factor gradually diminishes, facilitating localized exploration and ultimately improving search outcomes.

During the initial iteration, boosting the inertia weight enhances the global search capability of the PSO algorithm. This allows for exploration across a broader region, enabling the rapid identification of potential solutions in the vicinity. Subsequently, maintaining a low inertia weight throughout the iteration empowers the PSO algorithm's local search capabilities. It achieves this by decelerating particle velocities, thereby fostering effective local exploration. It is evident that the initial iterations significantly influence the algorithm's global search capabilities, while the later iterations dictate the extent of local exploration. Extending the search duration between these early and later phases can enhance the overall algorithmic performance. This is precisely where the concept of weighted arm variation, as presented in Equation (3.5).

$$w = w_{\min} + (w_{\max} - w_{\min}) \times \exp\left(-20 \times (t/t_{\max})^n\right) \tag{3.5}$$

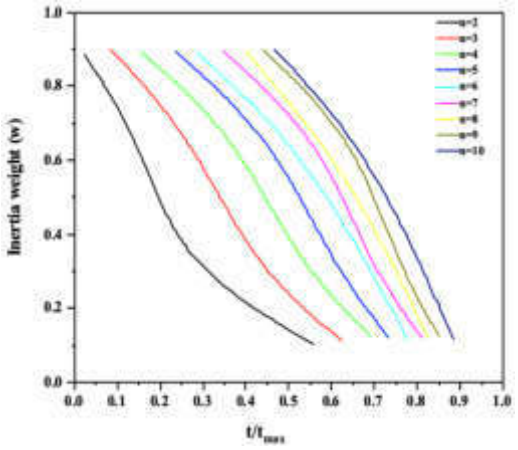


Fig. 3.2: Dynamic inertia weight adjustment curve

In Equation (3.5), $w_{\max}$ and $w_{\min}$ are taken as 0.9 and 0.1, respectively. The variation curves of the indices in Equation (3.5) are shown from the left to the right in Figure 3.1 when the indices are taken as 2 to 10 in turn. The corresponding inertia weight variation curves are shown in Figure 3.1.

As depicted in Figure 3.2, it becomes evident that when the parameter is configured as 6, both the global search duration and particle local search duration are considerably longer than other parameter values. This configuration leads to the inertia weight maintaining a high value during the initial phases of the loop and a lower value as the loop progresses. This extended duration of high inertia weight substantially strengthens the early-stage global search and late-stage local search. Consequently, this configuration achieves a well-balanced equilibrium between early-stage global search and late-stage local search, enhancing the overall global search responsiveness while concurrently augmenting local search capabilities.

4. Result Analysis and Discussions. To assess the new model's performance and compare its strengths and weaknesses against the traditional model, we conduct a series of comparative simulation experiments. In this context, we use a military software system as an illustrative example. Table 4.1 provides an overview of several metrics, including the number of module defects, for 13 modules within this military software system. Here, "SN" denotes the module number, "LOC" represents the module size in terms of lines of code, "FO" indicates module fan-out, "FI" stands for module fan-in, "PATH" reflects module control flow paths, and "FAULTS" represents the module defect count.

Evaluating the precision of an optimization model for predicting software module defect counts involved two distinct experimental trials.

In Experiment 1, all 13 data samples were partitioned into two sets. The first 10 samples were allocated for model construction and training, while the remaining 3 were reserved as test samples for assessing the model's predictive accuracy. Experiment 2 employed a different approach. The initial 6 samples were designated as training data for model development, and the subsequent 3 samples were employed as independent test data to evaluate the model's performance.

The training and prediction models under consideration encompass the BP network prediction model, PSO-SVM prediction model, and the improved PSO-LSSVM prediction model. For each of these models, normalization procedures are applied through the following steps: First, for each input feature, calculate the mean and standard deviation within both the training and testing datasets, which represent the feature's mean and variance, respectively. Next, standardize the input features by subtracting the mean of each feature in both the training and testing datasets and dividing by the standard deviation. This ensures that all input feature values are within the same order of magnitude, preventing any influence on the model's training and prediction due to differences in numerical ranges. Additionally, standardize the target variables in both the training and testing datasets to ensure they share the same order of magnitude.

Table 4.1: Summary of software metrics and defect information

Module number	Lines of code	Fan out	Fan in	Flow paths	Module defect count
1	30	3	2	3	1
2	30	3	2	3	3
3	33	1	3	1	2
4	34	2	28	3	2
5	38	6	19	15	2
6	42	6	2	13	5
7	56	1	2	11	3
8	65	5	2	13	1
9	70	2	2	7	2
10	102	3	5	11	6
11	121	2	11	21	7
12	165	13	11	220	12
13	271	8	2	79	18

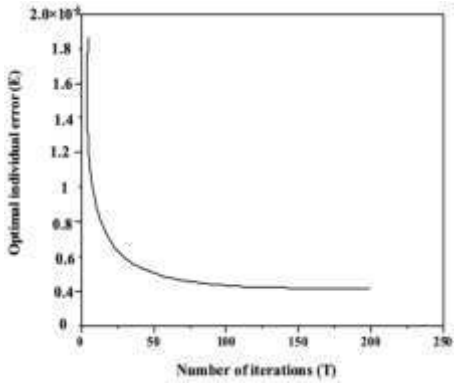


Fig. 4.1: Improved PSO-LSSVM error curve

Following data standardization, training is carried out for the BP neural network models, traditional PSO-SVM models, and improved PSO-LSSVM models. Subsequently, predictions are made on the test set using these standardized models. Specifically, the BP prediction model employs a structure with 18 hidden layers and a training target threshold of 0.00001. For the traditional PSO-SVM model, after two rounds of analysis using cross-validation and depth-first analysis, the model parameters are set to 499, and the kernel parameters are set to 5. It's important to note that both the PSO-SVM prediction model and the improved PSO-LSSVM prediction model utilize the radial basis function (RBF) as their kernel function.

During the model training process, error curves for both the BP prediction model and the improved PSO-LSSVM model are simulated. The results of these simulations are visually presented in Figure 3.2 and Figure 4.1, respectively.

Observing Figures 4.1 and 4.2, it becomes evident that the training error of the improved PSO-LSSVM prediction model exhibits rapid reduction and eventually stabilizes after approximately 200 training generations. In contrast, the BP prediction model meets the training criteria after significantly more iterations, specifically, 1,733 generations.

Following the training process, predictive models suitable for the sample data are established using the three prediction methods. Subsequently, the predictive samples are fed into their respective models for making predictions. A smoothing approach is employed to mitigate the influence of initial parameters on the BP prediction model and reduce prediction variability. This involves conducting 10 consecutive predictions using

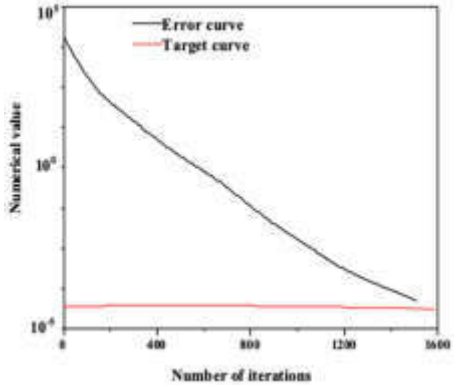


Fig. 4.2: Back propagation (BP) network error curve

Table 4.2: Comparison of prediction results of three prediction models

Experiment serial number	Test set serial number	True value	BP prediction model	Traditional PSO-SVM prediction model	Improved PSO-LSSVM prediction model
1	1	6	5. 7204	6. 2664	6. 100 9
	2	11	10. 1666	11. 4748	11. 192 5
	3	17	16. 0433	17. 6283	17. 203 7
2	1	6	5. 2234	6. 6326	6. 134 5
	3	11	9. 1999	12. 3895	11. 207 8
	4	17	14. 5860	19. 0068	17. 325 7

Table 4.3: Comparison of prediction errors of three prediction models

Experiment serial number	Test set serial number	BP prediction model	Traditional PSO-SVM prediction model	Improved PSO-LSSVM prediction model
1	1	4.66	4.44	1.68
	2	7.57	5.23	1.75
	3	5.62	3.69	1.21
	4	5.95	4.45	1.54
2	1	12.94	10.54	2.24
	2	16.37	12.62	1.89
	3	14.20	11.76	1.93
	4	14.52	11.64	2.02

the BP network and calculating the mean percentage error for the prediction outcomes. The results of these predictions are presented in Table 4.2 and Table 4.3.

In summary, enhancing and optimizing the PSO-LSSVM prediction model can be achieved by introducing several key strategies, including the adaptive inertia weight strategy, multi-objective optimization strategy, kernel function adaptive selection strategy, parallel computing strategy, and deep learning strategy. These strategies elevate the prediction model's accuracy and its generalization capabilities.

5. Conclusion. The development of the improved PSO-LSSVM prediction model through optimization techniques and non-overlapping kernels, becomes evident that the proposed model excels in prediction accuracy compared to the PSO-SVM method and the BP prediction model. Furthermore, as the number of models involved in the training process decreases, the improved PSO-LSSVM prediction model demonstrates superior overall performance compared to the PSO-SVM prediction model with fewer training samples. The predictive accuracy of the proficient PSO-LSSVM prediction model surpasses that of both the traditional PSO-SVM prediction model and the BP prediction model when trained with fewer models. Consequently, the improved PSO-LSSVM prediction model outperforms the traditional PSO-SVM and BP prediction models in terms of both training and prediction performance. This model effectively addresses the current software prediction requirements in the domestic market, especially in scenarios with limited historical data. Moreover, it can be applied to similar projects with historical models, further emphasizing its versatility and effectiveness.

REFERENCES

[1] M. K. Bhuyan, D. P. Mohapatra, and S. Sethi, *Software reliability prediction using fuzzy min-max algorithm and recurrent neural network approach.*, International Journal of Electrical & Computer Engineering, 6 (2016).
[2] G. Carrozza, R. Pietrantuono, and S. Russo, *A software quality framework for large-scale mission-critical systems engineering*, Information and Software Technology, 102 (2018), pp. 100–116.
[3] G. Coleman and R. V. O'Connor, *An investigation into software development process formation in software start-ups*, Journal of Enterprise Information Management, 21 (2008), pp. 633–648.
[4] C. Diwaker, P. Tomar, A. Solanki, A. Nayyar, N. Jhanjhi, A. Abdullah, and M. Supramaniam, *A new model for predicting component-based software reliability using soft computing*, IEEE Access, 7 (2019), pp. 147191–147203.
[5] Y. Duan, N. Chen, L. Chang, Y. Ni, S. S. Kumar, and P. Zhang, *CAPSO: Chaos adaptive particle swarm optimization algorithm*, IEEE Access, 10 (2022), pp. 29393–29405.
[6] M. P. Efthymiopoulos, *A cyber-security framework for development, defense and innovation at NATO*, Journal of Innovation and Entrepreneurship, 8 (2019), pp. 1–26.
[7] A. G. Gad, *Particle swarm optimization algorithm and its applications: a systematic review*, Archives of Computational Methods in Engineering, 29 (2022), pp. 2531–2561.
[8] J. Homola, M. Johnson, P. Kopardekar, A. Andreeva-Mori, D. Kubo, K. Kobayashi, and Y. Okuno, *UTM and D-NET: NASA and JAXA's collaborative research on integrating small UAS with disaster response efforts*, in Proceedings of the Aviation Technology, Integration, and Operations Conference, Atlanta, Georgia, 2018, p. 3987.
[9] P. Johnston and R. Harris, *The boeing 737 MAX saga: lessons for software organizations*, Software Quality Professional, 21 (2019), pp. 4–12.
[10] X. Li, H. Liu, W. Wang, Y. Zheng, H. Lv, and Z. Lv, *Big data analysis of the internet of things in the digital twins of smart city based on deep learning*, Future Generation Computer Systems, 128 (2022), pp. 167–177.
[11] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, *An empirical study of the impact of modern code review practices on software quality*, Empirical Software Engineering, 21 (2016), pp. 2146–2189.
[12] N. Medvidovic and R. N. Taylor, *Software architecture: foundations, theory, and practice*, in Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering, Cape Town, South Africa, 2010, ACM, pp. 471–472.
[13] C. Pozna, R.-E. Precup, E. Horváth, and E. M. Petriu, *Hybrid particle filter–particle swarm optimization algorithm and application to fuzzy controlled servo systems*, IEEE Transactions on Fuzzy Systems, 30 (2022), pp. 4286–4297.
[14] S. A. Ruk, M. F. Khan, S. G. Khan, and S. M. Zia, *A survey on adopting agile software development: issues & its impact on software quality*, in Proceedings of the 6th International Conference on Engineering Technologies and Applied Sciences, Kuala Lumpur, Malaysia, 2019, IEEE, pp. 1–5.
[15] T. Sommestad, M. Ekstedt, and H. Holm, *The cyber security modeling language: A tool for assessing the vulnerability of enterprise system architectures*, IEEE Systems Journal, 7 (2012), pp. 363–373.
[16] B. Wang, Q. Hua, H. Zhang, X. Tan, Y. Nan, R. Chen, and X. Shu, *Research on anomaly detection and real-time reliability evaluation with the log of cloud platform*, Alexandria Engineering Journal, 61 (2022), pp. 7183–7193.
[17] G. Wang, B. Zhao, B. Wu, C. Zhang, and W. Liu, *Intelligent prediction of slope stability based on visual exploratory data analysis of 77 in situ cases*, International Journal of Mining Science and Technology, 33 (2023), pp. 47–59.
[18] Z. Yu, Z. Si, X. Li, D. Wang, and H. Song, *A novel hybrid particle swarm optimization algorithm for path planning of UAVs*, IEEE Internet of Things Journal, 9 (2022), pp. 22547–22558.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Apr 10, 2023
Accepted: Sep 15, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 65–74, DOI 10.12694/scpe.v25i1.2254

ENHANCING INDUSTRIAL CONTROL NETWORK SECURITY THROUGH VULNERABILITY DETECTION AND ATTACK GRAPH ANALYSIS

YAN LIAO*

Abstract. Insufficient communication attack defense capabilities within industrial control networks is a serious problem that is addressed in this study. The author proposes a methodology that focuses on creating attack graphs to ease security and vulnerability studies in industrial control network systems in order to address this issue. The article provides thorough construction guidance and techniques for attack graphs, which are used for penetration testing and vulnerability analysis of networks for industrial control systems. On the created attack graph, experimental evaluations utilizing the "earthquake net" virus were carried out. The findings point to four main attack routes where the "Zhenwang" virus is most likely going to attack and cause the most damage. With a loss value of 12.2 and an attack success chance of 0.096, the first path involves cumulative attack stages. The second path consists of cumulative attack steps, with a loss value of 10.2 and an attack success probability of 0.072. The third path encompasses cumulative attack steps, with a loss value of 16.6 and an attack success probability of 0.063. The fourth path comprises cumulative attack steps, with a loss value of 18.6 and an attack success probability of 0.084.

Key words: Industrial control networks, Security vulnerability detection, Attack graph construction, Vulnerability analysis, Penetration testing, Network security

1. Introduction. The Internet has transformed people's lives significantly, facilitating global connectivity through its unique openness and data-sharing capabilities. Over nearly five decades of development, the Internet has saturated for all society, from finance, e-commerce, and industrial control to communication, transportation, healthcare, education, and beyond. It has become an essential infrastructure ensuring the stability of these critical domains. In the digital age, the Internet, with its core in cyberspace, is increasingly recognized as the "fifth space", closely intertwined with people's lives alongside land, sea, air, and sky [17].

Evolution of networking and informatization has introduced for security challenges in network security. These challenges apparent in the following ways: Explosion of Vulnerabilities: With the absence of remotely avoidable design flaws, the count of vulnerabilities halting from application software or operating system design and configuration continues to rise, maintaining elevated levels. These vulnerabilities frequently become targets for attackers, affecting an ongoing and significant threat to network security. The transformation in information carriers, transmission methods, and interconnection modes has furnished attackers with convenient avenues for launching network attacks. Furthermore, enhancing attackers' capabilities has led to more organized attack behaviours and specialized attack techniques, resulting in many new threats. Thus, vulnerabilities have evolved into a predictable security risk. According to data from the National Vulnerability Database (NVD) published by the National Institute of Standards and Technology (NIST), since 1997, the tally of known vulnerabilities has surged to 66,165. Over the years, the number of disclosed vulnerabilities has risen significantly [13].

Attack and defence represent the sides of network security. Without investigating the depths of attack theory and technology and comprehending our vulnerabilities and adversaries' tactics, we cannot effectively safeguard the security of network information systems. An integral aspect of research into network attacks centres on understanding and describing these attacks. The attack process encompasses a spectrum of distinct attack behaviours, each characterized by various stages and states. Given the complexity and diversity inherent to the attack process, deriving correlations and summarizing rules from known attack behaviours presents a formidable challenge [12].

The current theoretical foundations of attack detection technology remain incomplete. An approach rooted in attack-based analysis can furnish a structured and visual portrayal of the entire attack process. This

*Chongqing Technology and Business Institute, Chongqing, 400052 (liaoyan671@126.com)

approach is invaluable for separating and connecting the knowledge from existing research on attack behaviours. Furthermore, it enhances the usefulness of attack detection and security alerts [18].

The paper is organized into five main sections, including the introduction Section. The Literature review explores the existing body of knowledge and research in computer network security vulnerability detection and attack graph construction, as explained in Section 2. Section 3, the proposed method, outlines the novel approach and methodology proposed by the author, including the formulation of attack graphs and vulnerability detection techniques. Section 4, results and discussion, presents the empirical findings and engages in a detailed analysis and interpretation of the results obtained from experiments or simulations. Finally, in Section 5 summarizes the concluding remarks of the research.

2. Literature Review. With an increasingly intricate network security landscape and growing cyberattack threats, many organizations and institutions recognize the limitations of solely relying on detection-based defence for post-attack responses. It has become evident that a proactive and preemptive defence mechanism is needed to address security challenges fundamentally. This demand has encouraged a growing interest in risk assessment within the security [8].

Given the complexity of attacks, researchers are exploring using attack models to dissect and understand these threats. In the initial phases of assessing network system vulnerabilities, researchers primarily distilled their experiences from practical use. Then, these insights are applied to test a broader spectrum of network systems. This process effectively transitions from rule extraction to rule matching. The primary research focus revolves around generating more precise and comprehensive rulesets. Presently, mature network vulnerability scanning technologies exemplify this method. Nevertheless, rule-based vulnerability analysis methods exhibit inherent limitations. Consequently, some researchers have used formal theoretical tools like attack trees, graphs, and Petri nets to develop more holistic vulnerability analysis methodologies [19].

For instance, an information security method is proposed that quantifies the energy level associated with each attack. This assessment relies on the energy level increment of the attack and its impact on the Common Vulnerability (CV). To demonstrate the effectiveness of their proposed countermeasures, they compared CV and energy consumption across different types of attacks. These countermeasures leverage network-related security algorithms to safeguard large data communication and Distributed Critical Infrastructure Applications (DCAV), effectively mitigating CV and large data leaks during data transmission [1].

The State Grid's software, hardware, and network layers are identified as the most vulnerable points. Subsequently, the contributors investigated the threats these systems faced based on their vulnerabilities. Finally, the authors aimed to offer insights into the most productive defence solutions currently available and the imperative need for developing new defence mechanisms [11].

The researchers presented a technique for creating intelligent Production Planning and Control (PPC) systems. To improve PPC procedures, these intelligent PPC systems make use of cutting-edge technology such as the Internet of Things, big data analytic tools, and machine learning. These systems enable dynamic and near-real-time responses to changes in the production system by gaining insights from various data sources inside the production system, taking into account the expertise of production planners, and applying analytics and machine learning. The important issues and difficulties that production managers may run into when applying the suggested strategy are shown through a case study [10].

The author introduces a novel approach wherein a node's weight is defined based on three key factors: the system loss resulting from various vulnerability exploitation methods, the likelihood of attack success, and the progression of attack steps. Simultaneously, this method employs these weightings to analyze the optimal attack target within the industrial control network. Subsequently, it identifies the corresponding attack path to target this vulnerability. This approach employs an attack graph generation technology that emphasizes repairing vulnerable links. Experimental validation demonstrates the effectiveness of this technology, underscoring its substantial importance in enhancing industrial control networks' communication attack defence capabilities [3].

3. Research Methods.

3.1. Generation algorithm of attack graph for industrial control network. The security of industrial control networks is enhanced by introducing an innovative algorithm. The algorithm takes indications from the four key elements inherent to these networks: components, connections, control authority, and com-

Table 2.1: Comparison of proposed and existing approaches in industrial control network security

Aspect	Proposed Approach	Existing Methods
Methodology	Focuses on attack graph generation for analysis.	Relies on IDS, firewalls, access controls, and more.
Focus	Emphasizes visualizing attack paths and vulnerabilities.	Prioritizes preventive measures and signature-based detection.
Risk Assessment	Quantitative risk assessment with values and probabilities.	Often uses qualitative risk assessment.
Proactive vs. Reactive	Proactive, identifying vulnerabilities before exploitation.	Reactive, responding to security incidents as they occur.
Complexity	Provides a structured and visual portrayal of attacks.	May lack a comprehensive view of potential attack paths.

Table 3.1: List of utilization methods

Description	Abbreviation
Modify control parameters	ModConPa
Modify measurement parameters	ModMeasPa
Modification control procedure	ModContPr
Get permission or a password	GetPriv

munication permissions. It aims to systematically evaluate network vulnerabilities and identify potential attack routes. By conducting a thorough analysis of these elements and effectively implementing the algorithm, the research endeavors to reinforce the defense mechanisms against communication attacks within industrial control networks, thereby strengthening their overall flexibility against cyber threats.

(a) Four Elements of Industrial Control Network

The first element is the industrial control component, represented by h_i for a single component and H for a set of industrial control components. It has the following four parameters: using $host\_id$ represents the address of a single component; Use service to represent the control service provided by the component; Use vul_i to indicate the vulnerability number of components available for remote or local use; $value_i$ represents the value of the component.

The second element is the connection of the industrial control network, which is represented by C, and has three parameters in total, H_{From} is used to represent the starting component of the connection; Use Pro protocol to represent the connection protocol; Use H_{To} to represent the connected components [2].

The third element is the vulnerability of a single component, which is represented by vul_i, it has three parameters, namely $host\_id$ indicates the address of the component where the vulnerability is located; Use $Type$ to indicate the utilization way of the vulnerability; Use Att_Patt indicates the utilization mode of the vulnerability, and various utilization modes are shown in Table 3.1.

The fourth element defines user permissions on an individual component. It employs three distinct categories: "access" signifies the browsing permissions granted to regular users, "user" represents the standard operational permissions for regular users, and "root" denotes the comprehensive operational authority employed by the system administrator user over the information resources of the component.

(b) Derivation of algorithm

After a successful attack, the industrial control network changes from network state S_i to the next network state S_{i+1}, it is called state migration. If the industrial control system network needs to undergo a state transition, the following four conditions must be met simultaneously:

① When the industrial control network is in the initial state S_0, the network attacker must have sufficient authority on the attack-initiating component and can use the controlled component to attack other components

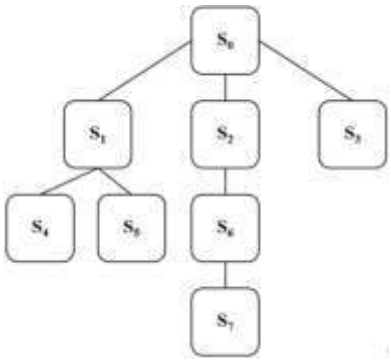


Fig. 3.1: Specific implementation process of algorithm

in the industrial control network. Use H_{victim} to indicate the components that have been successfully controlled by the attacker, and use H_{goal} to indicate the components that the attacker plans to control, namely the target components;

② There should be a relationship between H_{victim} and H_{goal} to ensure the smooth migration of network status as expressed in Equation (3.1);

$$C = (H_{victim}, Protocol, H_{goal}) \neq \emptyset \tag{3.1}$$

③ The target component H_{goal} needs to satisfy Equation (3.2), that is, it has a vulnerability so that attackers can use its vulnerability to migrate the network state;

$$Vul = (H_{goal}, Type, Att\_patts) \neq \emptyset \tag{3.2}$$

④ An attacker must obtain at least the minimum operation permission on the target component H_{goal}, and at least the minimum attack permission on the controlled component H_{victim}, to take advantage of its vulnerability to realize the migration of network state.

(c) Basic idea of algorithm

As shown in Figure 3.1, starting from the initial state S_0 of the network, use the above four rules to determine the possible state of the attacker in turn, after judgment, S_1, S_2 and S_3 meet the conditions for state migration. According to the principle of width first search, S_1, S_2 and S_3 are judged on the condition of state transition in turn, it can predict the state the attacker can reach. The attack target function H_{goal} whose S_3 state meets the above attack conditions indicates that S_3 state is a suitable attack target, so there is no need to judge its state transition conditions; Continue to judge S_1, where S_4 and S_5 are the nodes that meet the state migration conditions, and then judge S_2, and S_6 is the node that meets the migration conditions; Then continue to judge the state transition conditions of nodes S_4, S_5 and S_6 on the next layer, where S_5 state meets the state transition conditions and is a target of the attacker, therefore, it is not necessary to judge the condition of state transition; Continue to judge the state transition conditions of S_4 and S_6, when judging S_4, it is found that it can neither reach any new state node nor meet the attack target function, and it should be the last layer; When judging S_6, S_7 is the node that meets the migration conditions, and then judge the state migration conditions of node S_7, if it is found that the node can neither reach any new state node nor meet the attack target function, then S_7 is also the last layer, and the complete state migration process has been completed [4].

(d) Implementation of algorithm

Commencing from the initial state of the industrial control network, assess all potential network states based on the four state transition criteria mentioned earlier and incorporate the assessment outcomes into the state queue. The step-by-step procedure is visually depicted in Figure 3.2.

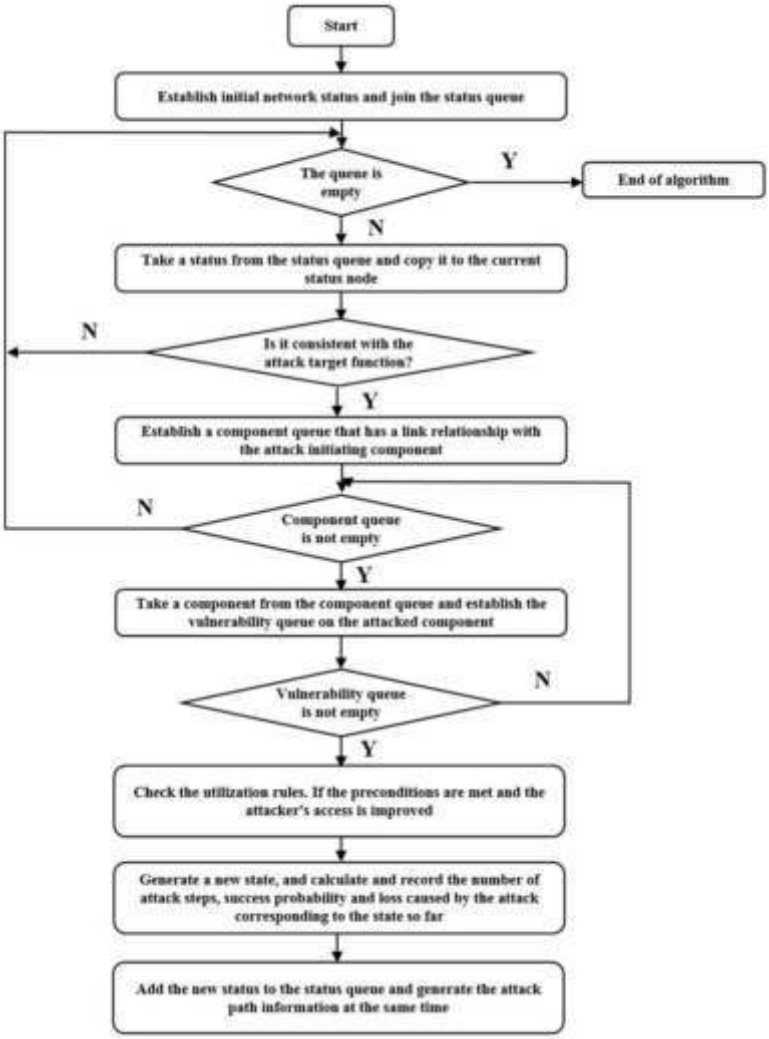


Fig. 3.2: Flow chart of the proposed vulnerability detection and attack graph algorithm

3.2. Comprehensive analysis and modeling of network vulnerabilities and attacks. A multilayered strategy for network security, encompassing three pivotal stages is introduced. Initially, it conducts a careful examination of network vulnerabilities to pinpoint potential vulnerabilities. Subsequently, the training focuses on crafting a strong attack prototype, serving as the groundwork for simulating and comprehending potential cyber threats. Finally, the research employs a systematic methodology to deduce attack sequences, providing deeper insights into prospective security breaches. The overarching objective is to fortify network defenses and strengthen against evolving cyber threats.

(1) Analyze network vulnerabilities

A data mining technique known as the association rule mining algorithm is employed to divide the attributes of vulnerability exploitation behaviour. This algorithm can uncover patterns of shared characteristics within comparable instances of vulnerable intrusion behaviours. However, the extraction process necessitates combining environmental factors with the recognized characteristics of vulnerability exploitation behaviour.

(2) Establish attack prototype

To execute an atomic attack, an attacker must fulfil specific prerequisites, including identifying vulnerabilities within industrial control components, establishing a requisite connection, and attaining a predetermined

level of control authority. Upon successfully infiltrating a component, the attacker can elevate their control authority, augmenting the network's loss value. This process forms a continuous cycle where each concluding atomic attack covers the way for the subsequent one, ultimately culminating in reaching the attack target and accomplishing the intended attack objective.

Establishing an attack prototype relies on two primary inputs: network topology and network vulnerability analysis, both of which furnish specific details about network vulnerabilities. In this context, the network topology contributes information encompassing the control network connection (C), control components (H), and control authority associated with each control component—comprising the quartet of elements characterizing the industrial control network. Concurrently, the network vulnerability analysis segment offers insights into the connection status and component vulnerabilities (Vul) within the real industrial network. Subsequently, the Attacker model, proposed in the attack graph algorithm, serves as the attack source. At the same time, the Attack_Rule rules are employed to convert all network vulnerabilities into sets of atomic attacks, thereby constituting the attack prototypes [16].

(3) Reasoning attack sequence

The process of deducing the attack sequence involves the application of pre-established attack prototypes to the real industrial control network. This application aids in ascertaining the attacker's actions and their impact on the overall network.

The term "Atom" is employed to denote an autonomous attack prototype comprising two essential parameters: the edge (referred to as "edge") and the node (designated as "point") within the attack graph. The dynamic progression of these edges and nodes within the attack graph interlinks individual attack prototypes, thereby composing the attacker's sequence of actions. To deduce the comprehensive attack sequence, inferring the transition conditions between these atomic prototypes is imperative. To establish whether two atoms are capable of transitioning, the following specific steps are undertaken:

① Determine whether the two components are connected;

② To establish the permission utilization relationship among various attack prototypes (referred to as "atoms"), the attacker's operational authorization gained from the target component in one attack prototype (atom1) must surpass the access permissions employed by the attacker on the target component of another attack prototype (atom2). Only under these conditions can the continuous transformation of atoms be sustained, ensuring the uninterrupted progression of the attack.

Attackers can be categorized into two primary types: direct and indirect. A direct attack involves the attacker directly targeting a specific component, and this type of attack occurs exclusively between two attack prototypes, referred to as "atoms". Conversely, an indirect attack occurs when the attacker targets a component through an intermediate "springboard" component. In the case of an indirect attack, the attack can transition between multiple attack prototypes of atoms.

Once the attack sequence has been deduced through the reasoning process, it becomes possible to ascertain all attack routes from the initial attack state to the ultimate attack target. These routes collectively form the basis for constructing an attack graph. In this construction, the current network state serves as the initial node, while each step in the attack sequence contributes to creating edges within the graph. These edges are defined by the associated attack behaviour, attack success probability, and component loss value at each step.

3.3. Penetration test analysis based on attack graph. The penetration test diagram is a composite representation that incorporates the four fundamental components of the traditional penetration testing process into the attack diagram. These components include test items, test objectives, test constraints, and test cases. Within this diagram, test items, test objectives, and test constraints of the penetration test are depicted as vertices, while the test cases are depicted as connecting arcs. The process of testing based on this model is as follows:

Initially, within the real industrial control network, certain security safeguards are often implemented to safeguard potential attack routes. This precautionary measure may fail to attain the intended objectives when executing corresponding test cases. Consequently, the penetration test chart obtained after a successful test may diverge from the initial penetration test chart. Therefore, adjusting the penetration test chart before initiating the test becomes imperative, ensuring that it aligns with the testing objectives and requirements.

The second step involves a comprehensive analysis of the penetration test chart after the Conclusion of the

penetration test. Initially, this analysis entails comparing the penetration test chart following the test and the vertex configuration before the test. This comparison yields the final results for the test project. Subsequently, a search is conducted based on these test results to identify the successful attack path. Finally, leveraging the weight values associated with each edge within the penetration test graph, the success probability of the attack, the resulting loss value, and the cumulative number of attack steps up to that point in the algorithm are determined. This information is then used to calculate the weight of an attack sequence, ultimately quantifying the network attack's impact on the network's security [5, 15].

3.4. Vulnerability risk assessment based on attack graph. To assess the risk associated with each vulnerability and prioritize the defence of the most difficult vulnerabilities and attack paths, a vulnerability's risk value is employed to gauge the harm it can cause. To ascertain the risk value of a vulnerability, the initial steps involve determining the vulnerability's overall probability of exploitation and the global degree of harm it can inflict. The following are the specific implementation steps for this process:

(a) Global Probability Assessment: Initially, calculate the global likelihood of the vulnerability being exploited across the entire system or network.

(b) Global Harm Assessment: Next, determine the extent of harm or damage the vulnerability can cause when exploited.

These following steps are essential in establishing the risk value attributed to each vulnerability, enabling a targeted strategy for prioritizing defence mechanisms.

(1) The breadth-first traversal algorithm is employed to compute the global utilization probability (denoted as 'P') of each node's successful utilization, calculated layer by layer starting from the initial node.

(2) Value is used to represent the value of each component, and $\emptyset$ is used to represent the independent harm degree of the vulnerability, so the loss caused by a single vulnerability to its component is $\emptyset x$ value; The letter Y represents the global hazard degree of a vulnerability, that is, the ratio of the associated hazard degree W of a single vulnerability to the sum of the value of all components in the industrial control network.

(3) 'R' signifies the risk value associated with vulnerability, calculated as the product of the global utilization probability of each node ('P') and the global hazard degree of the node ('Y') [14].

4. Result Analysis and Discussion. The "Zhennet" virus is employed as the attacker targeting the industrial control system in the experiment. The experimental process begins by establishing a network environment simulating the industrial control system. Subsequently, the attack graph is generated using the previously described method. Finally, the attack graph is the foundation for deriving the penetration test scheme and assessing vulnerability risks. The steps followed in the experimental process of the proposed method are as follows:

(1) Build network topology

It encompasses establishing an organized model for a computer or communication network. Network topology delineates the interconnections among devices and components within the network and elucidates the pathways through which data travels. It encompasses diverse configurations like star, bus, ring, mesh, and others, each with merits and drawbacks. Creating a network topology constitutes a pivotal phase in network design, facilitating streamlined data exchange and laying the groundwork for network administration and issue resolution.

(2) Generation algorithm parameter selection

Given that the "seismic network" attack gains entry into the industrial control network via a USB flash drive inserted into the operator station h_1, it designates the operator station h_1 as the initiating attack component. Table 4.1 illustrates the asset value assigned to each component within the industrial control network.

The impact coefficient of the vulnerability utilization mode is recorded as α_i, the disposable weight value of component assets is recorded as θ_i, and the component loss value is recorded as $Loss_i$, the loss value of each component after attack is shown in Table 4.2. The graphical analysis is shown in Figure 4.1.

(3) Generation of attack graph

According to the attack mentioned above graph generation method, the four paths that are most likely to attack and most harmful to the "Zhennet" virus are as follows:

① $h_1 \to h_1 \to h_3 \to h_2 \to h_5$ cumulative attack steps are 4, loss value is 12.2, and attack success probability is 0.096;

Table 4.1: Asset value table

Component No	Asset value	Component No	Asset value
h_1	2	h_4	1
h_2	3	h_5	3
h_3	4	h_6	5

Table 4.2: Calculation of loss value

Assembly	Vulnerability	$Value_i$	α_i	θ_i	$Loss_i$
h_1	vul_1	2	1	0.7	1.5
h_2	vul_2	4	1	1	4
h_3	vul_5	3	1	1	3
h_4	vul_6	1	1	1	1
h_5	vul_4	3	1.4	0.7	3.5
h_6	vul_4	5	2.4	0.7	10

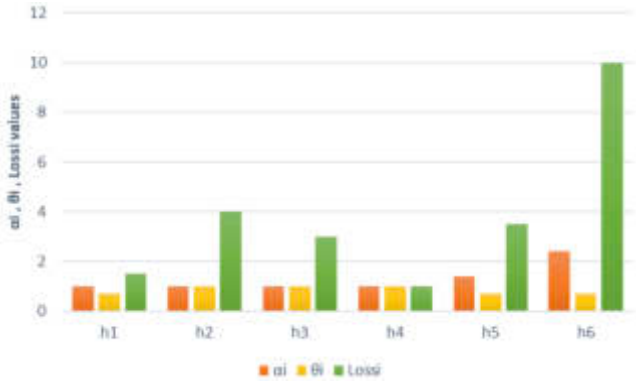


Fig. 4.1: Loss value analysis for various assembly

② $h_1 \to h_1 \to h_4 \to h_2 \to h_5$ cumulative attack steps are 4, loss value is 10.2, and attack success probability is 0.072;

③ $h_1 \to h_1 \to h_4 \to h_2 \to h_6$ cumulative attack steps are 4, loss value is 16.6, and attack success probability is 0.063;

④ $h_1 \to h_1 \to h_3 \to h_2 \to h_6$ cumulative attack steps are 4, loss value is 18.6, and attack success probability is 0.084.

(4) Penetration test based on attack graph

In penetration testing, parameters such as test objectives, test items, and test constraints are depicted as vertices, while test cases are represented as arcs within the penetration test diagram. These penetration test schemes are generated using the depth-first traversal method. Upon generating the penetration test chart, a comparison is made with the initial penetration test chart created at the outset of the test, and the results are found to be entirely consistent [6, 7].

(5) Vulnerability risk assessment based on attack graph

The risk value associated with each vulnerability is determined using a standard vulnerability scoring system and a vulnerability utilization diagram, as illustrated in Table 4.3 and Figure 4.2.

As shown in Table 4.3, the biggest vulnerability risk is vul_4, which is the DLL loading policy defect

Table 4.3: Specific risk value of each vulnerability

Vulnerability code	Component	Value quantity	Probability P	Overall hazard degree	Risk
vul_1	h_1	2	0.5	0.54	0.34
vul_2	h_2	4	0.7	0.74	0.6
vul_3	h_2	4	0.77	0.73	0.58
vul_4	h_5, h_6	8	0.90	0.8	0.82
vul_5	h_3	3	0.75	0.4	0.37
vul_6	h_4	1	0.4	0.4	0.2

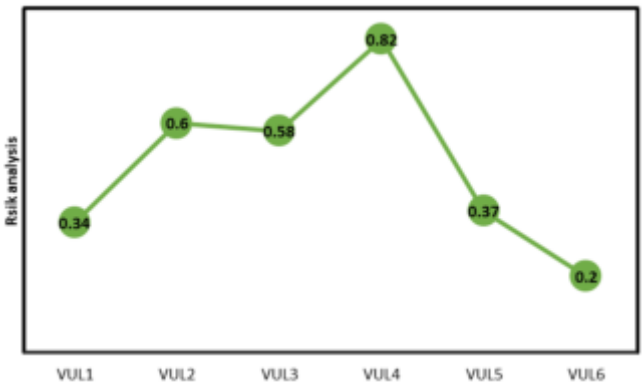


Fig. 4.2: Risk value of each vulnerability

in WINCC. It can be seen that the components installed with WINCC software are the most dangerous vulnerabilities to the system in the "seismic network" attack, which means that the attack path must be focused on defence [9]. The defence situation is the same, meaning the analysis results are correct. The penetration test analysis scheme is derived from the attack graph generated using the abovementioned method. The vulnerability risk values assessed through this penetration analysis scheme align with the actual scenario, demonstrating the feasibility and efficacy of the attack graph generation method.

5. Conclusion. The paper introduces a noteworthy approach to industrial control network security through an attack graph generation technology. The primary objective of this technology is to facilitate comprehensive security and vulnerability analyses within the industrial control network domain. The approach entails the development of a precise generation algorithm and explaining meticulous construction steps for the attack graph, which serves as a foundational tool for understanding network vulnerabilities and potential attack pathways. One key facet of this research is the empirical testing conducted using the "Zhennet" virus as the attacking agent. This practical application aims to validate the accuracy and effectiveness of the attack graph generation method proposed. The results obtained from these tests offer valuable insights into the practical utility of the approach. Notably, a degree of subjectivity is involved in calculating two critical parameters: the system loss value and the probability index of successful attacks. These calculations form pivotal components of the attack graph generation process, and their accuracy directly influences the reliability of the generated attack graph.

REFERENCES

[1] A. Algarni and V. Thayananthan, *Autonomous vehicles: The cybersecurity vulnerabilities and countermeasures for big data communication*, Symmetry, 14 (2022), p. 2494.

[2] O. Briones, R. Alarcón, A. J. Rojas, and D. Sbarbaro, *Tuning generalized predictive pi controllers for process control applications*, ISA Transactions, 119 (2022), pp. 184–195.
[3] B. Fan, C.-X. Zheng, L.-R. Tang, and R.-Z. Wu, *Critical nodes identification for vulnerability analysis of power communication networks*, IET Communications, 14 (2020), pp. 703–713.
[4] Y. Feng, G. Sun, Z. Liu, C. Wu, X. Zhu, Z. Wang, and B. Wang, *Attack graph generation and visualization for industrial control network*, in Proceedings of the 39th Chinese Control Conference, Shenyang, China, 2020, IEEE, pp. 7655–7660.
[5] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, *Badnets: Evaluating backdooring attacks on deep neural networks*, IEEE Access, 7 (2019), pp. 47230–47244.
[6] I. Kotenko and M. Stepashkin, *Attack graph based evaluation of network security*, in Proceedings of the 10th IFIP TC-6 TC-11 International Conference, Heraklion, Crete, Greece, 2006, Springer, pp. 216–227.
[7] J. Luan, J. Wang, and M. Xue, *Automated vulnerability modeling and verification for penetration testing using petri nets*, in Proceedings of the Cloud Computing and Security: Second International Conference, Nanjing, China, 2016, Springer, pp. 71–82.
[8] S. Mubarak, M. H. Habaebi, M. R. Islam, A. Balla, M. Tahir, A. Elsheikh, and F. Suliman, *Industrial datasets with ics testbed and attack detection using machine learning techniques*, Intelligent Automation & Soft Computing, 31 (2022), pp. 1345–1360.
[9] E. Normanyo, F. Husinu, and O. R. Agyare, *Developing a human machine interface (hmi) for industrial automated systems using siemens simatic wincc flexible advanced software*, Journal of Emerging Trends in Computing and Information Sciences, 5 (2014), pp. 134–144.
[10] O. E. Oluyisola, S. Bhalla, F. Sgarbossa, and J. O. Strandhagen, *Designing and developing smart production planning and control systems in the industry 4.0 era: a methodology and case study*, Journal of Intelligent Manufacturing, 33 (2022), pp. 311–332.
[11] V. D. Savin, *Cybersecurity threats and vulnerabilities in energy transition to smart electricity grids*, in Navigating Through the Crisis: Business, Technological and Ethical Considerations: The 2020 Annual Griffiths School of Management and IT Conference (GSMAC) Vol 2 11, Springer, 2022, pp. 71–83.
[12] M. T. Siponen and H. Oinas-Kukkonen, *A review of information security issues and respective research contributions*, ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 38 (2007), pp. 60–80.
[13] Y. Su, M. Zhao, C. Wei, and X. Chen, *Pt-todim method for probabilistic linguistic magdm and application to industrial control system security supplier selection*, International Journal of Fuzzy Systems, 24 (2022), pp. 1–14.
[14] R. Vishwakarma and A. K. Jain, *A survey of ddos attacking techniques and defence mechanisms in the iot network*, Telecommunication Systems, 73 (2020), pp. 3–25.
[15] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, *Neural cleanse: Identifying and mitigating backdoor attacks in neural networks*, in Proceedings of the IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 2019, IEEE, pp. 707–723.
[16] S. Wang and Y. Gong, *Adversarial example detection based on saliency map features*, Applied Intelligence, (2022), pp. 1–14.
[17] W. Xia, R. Neware, S. D. Kumar, D. A. Karras, and A. Rizwan, *An optimization technique for intrusion detection of industrial control network vulnerabilities based on bp neural network*, International Journal of System Assurance Engineering and Management, 13 (2022), pp. 576–582.
[18] J.-P. A. Yaacoub, H. N. Noura, O. Salman, and A. Chehab, *Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations*, International Journal of Information Security, (2022), pp. 1–44.
[19] P. Zeng, G. Lin, L. Pan, Y. Tai, and J. Zhang, *Software vulnerability analysis and discovery using deep learning techniques: A survey*, IEEE Access, 8 (2020), pp. 197158–197172.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 4, 2023
Accepted: Sep 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 75–84, DOI 10.12694/scpe.v25i1.2266

IMPROVING SEMANTIC ANALYSIS IN VISUALIZATION WITH META NETWORK REPRESENTATION AND PARSING ALGORITHM

CHUNMEI JI*, NING LIU†, ZANSEN WANG‡, AND YAPING ZHEN§

Abstract. This article aims to advance semantic analysis models, particularly in visualization, by proposing a novel semantic representation method utilizing the semantic Meta Network (MNet). MNet is a complex framework comprising semantic elements, internal and external relationships, and feature attributes, defined hierarchically through recursive processes, aiming to depict the comprehensive semantic space from phrase-level components to complete texts. The methodology involves the development of a general construction algorithm for MNet, encompassing meta relationships, tree structures, and network structures, and a Parsing method for specific semantic analysis problems, including a bottom-up specification-based MNet semantic dependency tree construction algorithm and a network construction algorithm tailored for natural language interface parsing. Empirical experiments confirm the effectiveness of these algorithms, particularly in parsing natural language control interface instructions in Supervisory Control and Data Acquisition (SCADA) systems, bridging specific semantic analysis problems with the general construction and parsing processes of MNet, accounting for internal semantics concerning language unit structures and foreign language meanings in the linguistic context, thereby contributing significantly to the field of natural language semantic analysis.

Key words: Natural language processing, SCADA, Natural language interface, MNet, Parsing algorithm

1. Introduction and examples. SCADA technology, also known as Computer for Remote Sensing (telemetry, remote control, remote signalling, remote adjustment), is an automated system founded on the Computer, Communication, Control, Sensor (3C+S) framework. It seamlessly integrates monitoring, control, and data acquisition functionalities. Communication technology enables data communication within cross-regional and long-distance SCADA systems. To address the distribution of SCADA systems, the complexity of regulating control objects, and the concurrent data collection and real-time demands of automatic control, advanced computer network communication technology is essential for constructing a distributed SCADA system [6].

The authors introduced a VLAN-based distributed SCADA system implemented within the overall automation system of the Yellow River Diversion Project. They explored its system architecture, distributed data acquisition, and regulation control algorithms, all based on parallel databases. Nonetheless, it is evident that China's manufacturing industry currently faces challenges in attaining a high level of sophistication, particularly when harnessing information technology for industrial production. In this regard, a considerable gap exists between China and industrialized developed nations. One prominent technology that stands out in industrial production is SCADA, a prevalent and indispensable industrial information system [15].

The full potential of SCADA systems in China's manufacturing industry is yet to be realized, leaving substantial room for growth and advancement. Closing this technological gap and fully integrating SCADA technology into the industrial production landscape is essential for China's progress in this arena and its efforts to compete globally. Currently, traditional SCADA systems predominantly rely on copper wires. However, as the industry transitions towards IP-based massive data collection SCADA systems, it confronts two significant challenges. The first challenge pertains to the adaptability of data collection protocols. This issue arises due to multiple IP-based data collection and transmission protocols, each characterized by substantial variations. Notably, there is a unified industry standard for the IP data transmission protocols used by sensors [9].

*School of Information & Security, Yancheng Polytechnic College, Yancheng, Jiangsu 224005, China (Corresponding author: chunmeiji57@163.com)
†School of Information & Security, Yancheng Polytechnic College, Yancheng, Jiangsu 224005, China
‡School of Information & Security, Yancheng Polytechnic College, Yancheng, Jiangsu 224005, China
§School of Information & Security, Yancheng Polytechnic College, Yancheng, Jiangsu 224005, China

The different IP sensors employ a range of protocols such as HTTP, FTP, SNMP, SSH, TELNET, and MODBUS, each presenting distinct formats for the transmitted data messages. This diversity poses a formidable obstacle for SCADA systems seeking to integrate many sensors into their infrastructure seamlessly. This results in substantial pressure for the widespread adoption of IP-based SCADA systems [5].

Distributed processing of massive data collection is a primary advantage of IP-based SCADA systems, as it allows for nearly unlimited system capacity, facilitating the creation of large-scale data acquisition and monitoring control systems encompassing up to 100,000 points. However, such extensive systems' scalability introduces challenges in efficiently managing SCADA systems' substantial influx of data. Failure to address this issue could potentially hinder the widespread adoption of SCADA systems in large enterprises, posing significant hurdles to their application [8].

Semantic analysis, a fundamental process in natural language processing, encompasses various tasks depending on the language unit under consideration. These tasks encompass word sense disambiguation at the word level, role labelling at the sentence level, and referential disambiguation at the discourse level. Recent years have seen semantic analysis primarily focus on two major approaches: rule-based and statistical methods. Rule-based methods rely on a series of language rules rooted in generative linguistics, often beginning with establishing "predicate-argument relationships", featuring concepts like first-order predicate calculus, semantic networks, concept dependency graphs, and frame-based representations. On the other hand, statistical methods draw insights from extensive corpora analysis, employing probabilistic and data-driven techniques. Advancements in deep semantic analysis have given rise to concepts such as semantic dependency analysis trees, dependency analysis graphs, Abstract Semantic Representation (AMR), combinatorial category logic, and knowledge graphs. Deep networks, primarily founded on distributed learning, including word embeddings and sub-embeddings, have delivered promising outcomes in shallow semantic analysis tasks like named entity recognition, word relationship extraction, text classification, and automatic term extraction. Notably, practical applications have seen a continuous integration of rule-based and empirical statistical methods, yielding favourable results [3, 4].

2. MNet Methodology. This section presents an abstract definition of the semantic analysis model, MNet. Subsequently, we investigate a comparative analysis, highlighting the similarities and distinctions between MNet and other relevant methods. Conclusively, we explain the conceptual framework for constructing MNet and underscore that the construction of MNet inherently represents the process of semantic analysis and its corresponding solution [2].

2.1. Definitions. Definition 2.1. *Meta network (MNet) is an ordered group, MNet=*$(n_1, n_2, \cdots, n_m; r, R, P)$*, among them,* n_i $(i = 1, \cdots, m)$ *is also a semantic network, which is a sub semantic network of N (this is a recursive definition, this is particularly important, and it is also an important feature to distinguish other semantic networks),* $m \geq 1$*, for example, under the framework of natural language description, ni can be any form from words, phrases, phrases to sentences and texts [14].*

R is the set of internal relationships of N, $r = r_{ij}, j = 1, \cdots, m,$ and $i \neq j$;

R is the set of external relationships of N, $R = R_{ik} = 1, \cdots, m, k = 1, \cdots, \infty$;

The set of P attributes, $P = p_i = 1, \cdots, \infty$;

Definition 2.2. *The internal relationship of the semantic MNet:* r_{ij} *is a quaternion (*η_i*,* n_j*, relation, P), among them, relationship is the name of the relationship where ni points to* n_j*, where* $\eta_i \in$ *MNet and* $n_j \in$ *MNet.*

Definition 2.3. *The external relationship of the semantic MNet:* R_{ik} *is a quaternion (*η_i*,* n_k*, relation, P), the relationship is* η_i *points to* η *the relationship name of k, among them,* $\eta_i \in$ *MNet,* $n_k \in$*MNet.*

Definition 2.4. *Attribute set P:* p_i *is a binary (AttriName, AttriValue) consisting of attribute names and values.*

Definition 2.5. *Meta relationship: If* n_i*,* n_j *are independent words, then* r_{ij} *or* R_{ik} *are meta relationships.*

Here are a few clarifications within the MNet framework: firstly, MNet suggests that the smallest semantic unit is a word, with the connections between words referred to as meta relationships. Secondly, MNet exclusively defines binary relations, even though real semantic units often involve multiple relationships; for instance, in the natural language interface parsing for SCADA instructions, ternary relationships are commonplace but can

generally be simplified into binary relations for shortness. However, the precise transformation details are not expounded here [11].

2.2. Comparative Analysis. When MNet's target object is a sentence, similar methods predominantly centre around sentence dependency analysis, represented by dependency analysis trees and diagrams. Semantic dependency analysis's primary objective is to define the genuine or logical semantic connections among words within a sentence's structure. In a broader context, these connections encompass syntactic functional relationships within sentence components, thereby encompassing syntactic dependency analysis within the purview of semantic dependency analysis [16].

From Definition 2.1, it can be seen that the semantic dependency tree is a special form of semantic meta-network. When semantic meta-network N meets the following restrictions, it is simplified into a natural language semantic dependency tree:

1) $S = n_1, n_2, \cdots, n_m$ is a complete natural language sentence composed of words;

2) R satisfies the following qualifications related to the semantic dependency tree:

The directed acyclic tree composed of r as an edge and $n_1, n_2, \cdots, n_m$ as nodes, is a single root node; $N_i(i = 1, \cdots, m)$ has only one parent node; If the word n_i depends on n_j, then all words between n_i and n_j belong to n_j, which means there are no edge intersections in the dependency tree.

3) R and P are empty sets.

The semantic dependency graph represents a significant advancement by overcoming the limitations of semantic dependency trees, particularly in terms of edge crossing and multiple parent nodes. This expansion of functionality contributes to a richer semantic description. AMR transcends traditional syntactic tree structures by abstracting a sentence into a semantically coherent, single-rooted, directed acyclic graph. Comparable methods for sentence-level semantic analysis include semantic networks, which essentially depict the conceptual relationships among words in a graphical network format. When the focus extends to sets of words as language units within a specific natural language, the functionalities of the Italian Meta Network (MNet) align with those of ConceptNet, WordNet, and HowNet, as well as knowledge base tools like Knowledge Graph. Various internet companies and organizations have introduced their knowledge base tools, including Google Knowledge Graph, BabelNet, DBPedia, DBary, and Microsoft Concept Graph. But, the widespread adoption of these tools is currently limited [12, 10, 18].

In essence, MNet unifies various expression methods, including knowledge base tools like WordNet, semantic dependency trees (graphs), and semantic networks to represent semantic relationships comprehensively. A noteworthy aspect of MNet is its incorporation of external relationship definitions and recursive structures, pivotal features of this unified model.

2.3. MNet Construction Ideas. The cognitive process is facilitated by MNet, using Figure 2.1 as an illustrative example. It provides insight into the thought process guided by common sense and perception. Hypothetical evaluation is incorporated to shed light on the natural language processing tasks involved in reading comprehension. The sentence "someone smiles and walks towards a table with apples and water cups" serves as an overarching description of a particular real-life scenario, necessitating the prediction of the subsequent actions of the individual involved. According to common-sense reasoning, these actions could encompass "eating apples" or "drinking water". Regarding the static features of the semantic analysis model, the internal and external relationships inherent to each layer of elements within the scene correspond to the internal and external relationships of the semantic units within the sentence [7].

While speech transmission typically follows a chronological sequence from left to right, creating the impression that the brain processes speech and text linearly, word by word and sentence by sentence, the actual relationships between semantic units in language are fundamentally grounded in the objective scene. Each semantic unit represents a distinct element of the dynamic natural scene, independent in time and space. Consequently, constructing and deducing these relationships can be orchestrated and interconnected through these semantic units. In other words, establishing relationships need not adhere strictly to a word-by-word sequential order as per sentence input; it can instead unfold in a bottom-up, parallel manner akin to the processing of visual imagery [17].

Modern cognitive neuroscience posits that the brain can convert language into visual information, with the cortex processing auditory signals like visual inputs from the eyes.

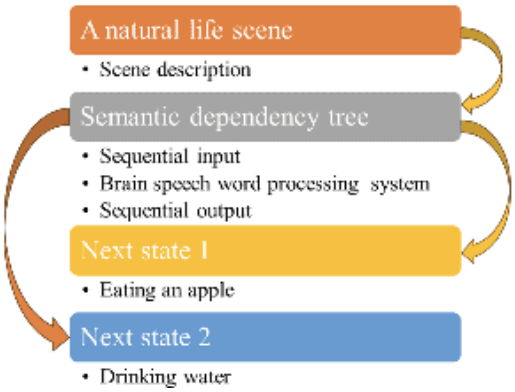


Fig. 2.1: Natural language understanding task.

In analysis and reasoning, the leverage of external relations, often manifested as contextual connections and background knowledge embedded within language and text, proves invaluable. In natural language processing, semantic dependency trees are valuable for delineating internal and external relationships among scene elements facilitating logical reasoning. However, these trees alone do not provide a comprehensive view. The interplay between the intrinsic relationships inherent to a sentence itself and the extrinsic relationships governed by contextual knowledge engenders a complex network structure within the semantic units of the sentence.

Consider the sentence "someone smiles and walks towards a table with apples and water cups". Internal relationships encompass actions such as "walking towards the table" and "smiling", intertwined with external relationships like "apples and water cups may be food" and "people can eat food and drink water". Navigating this intricate web of relationships allows us to deduce that the potential next state involves "eating apples" and "drinking water".

In light of this analysis, the author employs MNet to investigate the composition and relationships of elements that characterize objective scenes. The construction of MNet adopts a bottom-up self-organizational approach, with the ultimate goal of addressing specific practical challenges in natural language processing. The subsequent steps employed in the process of constructing MNet are as follows [1]:

1) Meta-relationship construction involves establishing a foundational semantic relationship library among words, serving as a cornerstone for resolving Met's internal relationships. A deep learning approach leveraging bidirectional GRU and an attention mechanism was employed to construct the meta-relationship library between words, accompanied by corresponding experimental investigations.

2) Tree construction primarily encompasses the development of internal relationships based on these meta-relationships. This is achieved through a parallel bottom-up self-organization method to induce and construct a comprehensive semantic dependency tree. The author introduced a bottom-up specification-based MNet semantic dependency tree construction algorithm supported by an accompanying experimental analysis.

3) Conversely, Web construction primarily deals with external relationships. Given the impracticality of exhaustively constructing all external relationships of semantic units within sentences, a selective approach is adopted. Tailored to specific natural language computing tasks, the semantic tree established in the preceding step is extended to encompass external relationships. As an illustrative example, the natural language control interface of the smart home SCADA system is considered, with the objective being to map the relationships between nodes in the MNet semantic dependency tree and target instructions. Algorithmic details and experimental outcomes are presented separately to explain this process.

Given current cognitive limitations, our exploration of the MNet method has not delved into a rigorous mathematical framework. Rather, our focus has centred on refining the MNet approach for semantic analysis within the context of solving natural language manipulation interfaces, approaching it primarily from a design perspective.

Table 2.1 provides an overview of the MNet method, evaluating it from various dimensions of semantic

Table 2.1: Assessment of the MNet semantic analysis model.

Index	Description
Visualization	The relationship between semantic units expressed in network graphics
Concurrency	Each language unit can independently and parallelly calculate its relationship with others.
Complexity	Recursive definition reduces model complexity.
Uniformity	Consistent with natural reading habits and good consistency
Variability	Support incremental semantic parsing from meta-relationship tree relationships to graph relationships with good model variability.

representation modelling. MNet exhibits strong concurrency, variability, visualization, and consistency characteristics in many respects and maintains low complexity.

3. Structure of MNet.

3.1. MNet Relationship. MNet meta relationships can leverage existing knowledge base tools such as WordNet and HowNet. However, it becomes imperative to pre-construct bidirectional GRU and word vectors through sample training when addressing domain-specific challenges. This process, in combination with sub-character level and sentence-level attention mechanisms, facilitates the training of word relationships based on sample data.

Character embedding and position embedding information vectors of word pairs involved in the prediction relationship are employed to represent the input. This approach incorporates a word-level attention mechanism and combines bidirectional GRU with a character attention mechanism to create the embedding vector representation of the sentence.

For instance, consider a set of n sentences encompassing relational pairs (word1, word2), denoted as $s_i(i = 1, \cdots, n)$. The embedded expression vector within each sentence carries information about whether it includes the relationship r. We incorporate a sentence-level attention mechanism to leverage information from all these sentences when predicting relationship r for the pairs (word1, word2). This mechanism enables us to represent these n sets of sentences using feature vectors encapsulating embedded expression information from all the sentences. Subsequently, we conduct comprehensive training, a strategy that offers the advantage of justifying the noise impact from inaccurate standard data.

3.2. Tree Construction. The analysis of semantic dependency trees typically encompasses both transfer-based and graph-based methods. To align more closely with human language thought patterns and accommodate concurrent execution, we have integrated both transfer-based and graph-based approaches. Building upon the probabilistic assessment of relationships between words within a sentence, our method employs a bottom-up approach involving neighbouring word competition and a dependency mechanism to construct a semantic dependency tree. This approach differs from traditional transfer-based methods in several key ways:

1) It is no longer constrained by the input order of the sentence, irrespective of whether left-associative or right-associative dependencies take precedence.

2) When dealing with words that have not yet determined their dependent objects, it considers the dependency relationships with adjacent words and those of the words upon which the adjacent words rely.

3) We have implemented optimizations to address the occurrence of multi-subtree phenomena during the construction process.

3.3. MNet Construction. The construction of the MNet network varies based on the specific semantic understanding tasks at hand. Fundamentally, it revolves around tree construction, where we re-label words and their relationships to align with the diverse requirements of downstream natural language processing tasks. This secondary annotation process can occur either after the completion of semantic dependency tree parsing or concurrently during the parsing process. Consequently, MNet can iteratively optimize natural language parsing, thereby continually deepening its grasp of semantics. This adaptability and incremental refinement distinguish it from neural network models, providing distinct advantages in the field [13].

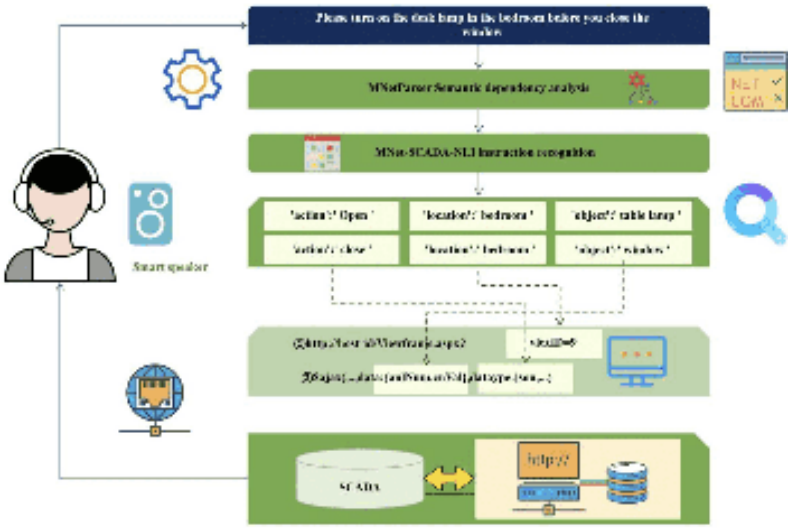


Fig. 3.1: MNet SCADA-Natural Language Interface (NLI) analysis algorithm.

To illustrate the process of constructing a specific MNet network, let's consider the application of a natural language interaction interface in a SCADA system. In the SCADA system's control mode, there are primarily two types of commands: query and control. Query commands retrieve status data from on-site processes or equipment, while control commands modify on-site equipment or process parameters. Control commands in the SCADA system typically consist of three key components: actions, objects, and parameters. Among these, the parameter often includes the position, indicating the specific location of the controlled object.

For instance, a natural language instruction like "turn on the desk lamp in the bedroom" can be transformed into an intermediate language representation with a data structure such as "{Object=desk lamp, Location=bedroom, Action=turn on}" using a natural language processing program. Subsequently, formal rules can be applied to generate the SCADA system's command instructions. Figure 3.1 illustrates the parsing principle of the SCADA system's natural language control instructions. This example highlights the process of adapting natural language input into actionable commands within the SCADA system.

To enhance the efficiency of accomplishing the intended tasks, it is possible to seamlessly integrate the construction of specific MNet networks into the MNet tree construction process. This approach allows for the modification of STEP3 within "MNetSParser" to align with specific application requirements. Subsequently, the processed data can undergo further processing using the MNet-SCADA-NLI algorithm. This integration streamlines the task execution process and enhances overall efficiency.

4. Experiments and Applications.

4.1. Building MNet Relationships with Word Relationship Knowledge Base. There is a lack of standardized definitions for word relationships, even in resources such as WordNet, HowNet, ConceptNet, and others, which are customized and incomplete. The original word relationship samples are constructed to address this issue using the "evsam05.zip" Chinese semantic dependency analysis and evaluation dataset published by the Natural Language Processing and Chinese Computing Conference (NLP & CCC 2013). The training data follows the CoNLL format for a Chinese dependency corpus, while the experimental data is sourced from the Tsinghua database. The data samples have been uniformly converted to facilitate model training, as detailed in Table 4.1. This approach establishes a more comprehensive and tailor-made understanding of word relationships.

The Tsinghua semantic dependency tree database enumerates 69 types of word dependency relationships, as presented in Table 4.2. For example, in the sentence "The Eighth Wonder of the World Appears", the relationship between "the world" and "the miracle" is "limited". At the same time, negative samples were

Table 4.1: Semantic relationships in sample sentences.

Word 1	Word 2	Relationship Type	Sentence
World	Miracle	Limit	The eighth wonder of the world appears
Within	Hall	Locative word dependency	The hall is carpeted with red carpet
Red	Carpet	Describe	The hall is carpeted with red carpet
World	Appear	Null	The eighth wonder of the world appears

Table 4.2: Serial numbers and their associated types.

Serial Number	Type	Serial Number	Type
0	The word 'dependent' on '	9	Connection Dependency
1	Original state	10	quantity
2	content	11	degree
3	Process Period	12	result
4	Relationship subject	13	possessor
5	source	14	objective
6	comment	15	Final state
7	reason	16	Tone dependency
8	result event	17	Comparative quantity

added, and the relationship type was NULL, indicating that there was no dependency relationship between the group of words, the data was divided into training and testing sets.

Two cases were tested separately, one without negative samples and the other with negative samples, among them, 80% of the training set with negative samples included the addition of negative samples, which significantly improved the efficiency and accuracy of training. Figure 4.1 and Figure 4.2 show the changes in loss function and accuracy rate during the training process of adding negative samples using the TensorboardX tool. Finally, select the model with the training fitting accuracy acc of 0.98 and the accuracy of this model in predicting word relationships on the test set is 89.9

4.2. MNets Parser Semantic Dependency Analysis for Tree Construction. The MNet implies utilizing the MNet parser to perform semantic dependency analysis in tree construction. The proposed method suggests that the MNet parser plays a pivotal role in dissecting semantic relationships, which are the meaningful connections between words, and this analysis is crucial for creating tree-like structures.

In the data source and analysis process, the evaluation indicators test set uses the Chinese semantic dependency analysis and evaluation data package published by the NLP & CCC 2013 to evaluate the tested system using three indicators, namely: Labeled Attachment Score (LAS); Unlabeled Attachment Score (UAS); Labeled Accuracy (LA). Assuming that the total Number of words in the entire test corpus is N, the dependency of any word is represented by a triplet<wordi, wordj; Depreij>represents. Among them, 'wordi' is the word itself, and 'wordi' is dependent on 'wordj' with a relationship of 'deprej', the correct Number of words' wordi 'for all wordj is' Nuas', the correct word data for all deprij is 'Nla', and the correct Number of words for all wordj and deprij is 'Nlus'. So, the calculation method for testing indicators using Equations (4.1) to (4.3):

$$LAS = \frac{N_{\text{las}}}{N} \tag{4.1}$$

$$UAS = \frac{N_{\text{uas}}}{N} \tag{4.2}$$

$$LA = \frac{N_{\text{la}}}{N} \tag{4.3}$$

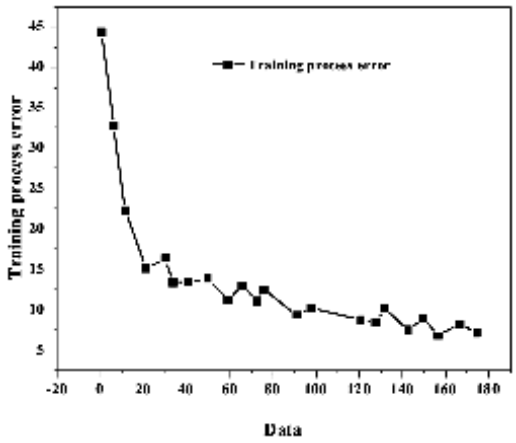


Fig. 4.1: Evolution of training process errors.

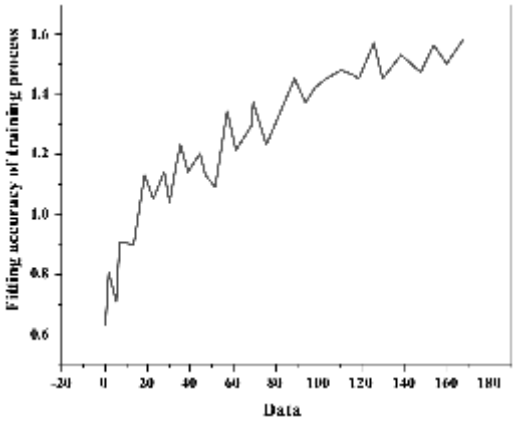


Fig. 4.2: Training process fitting accuracy.

Regarding computational complexity analysis, "MNet Parser" can efficiently construct semantic dependency trees in O(n log n) time, similar to the straightforward edge-first approach. Evaluating accuracy using the same word relationship calculation model and comparing it with the effective transfer-based edge-first algorithm reveals certain advantages of MNet Parser, with an improvement of approximately 2-3 percentage points, as demonstrated in Table 4.3. However, the overall accuracy remains somewhat modest, primarily attributed to a relatively small sample size and up to 70 dependency categories. It is anticipated that enhancing accuracy can be achieved by reducing the Number of dependency categories and increasing the sample size. Factors such as part-of-speech and dependency category prioritization introduced during the parsing process can improve accuracy.

4.3. MNet-SCADA-NLI Methodology. In the context of natural language interaction within smart homes, a limited-scale questionnaire survey was conducted to compile a dataset comprising roughly 100 frequently utilized language manipulation commands. TF-IDF51 and the proposed MNet-SCADA-NLI were employed for intermediate language recognition within the SCADA system.

The algorithm's performance is assessed by measuring accuracy (P), recall (R), and F-values, which are defined by Equations (4.4) to (4.6):

Accuracy (P) quantifies the fraction of accurately parsed parameters within the predicted intermediate

Table 4.3: Comparative analysis of test results for MNet parser.

Method	LAS	UAS	LA	Time complexity
MNet Parser	59%	72%	79%	o(n log n)
Edge First Algorithm	57%	69%	77%	o(n log n)

Table 4.4: Comparison of natural language manipulation instructions in SCADA system.

Method	Precision (P)	Recall (R)	F value
TF-IDF	0.71	0.52	0.6
Proposed MNet-SCADA-NLI	0.90	0.90	0.90

language results.

$$P = \frac{N}{Total_P} \tag{4.4}$$

Recall rate (R): It reflects the proportion of correctly parsed parameters in natural manipulation language samples.

$$R = \frac{N}{Total_R} \tag{4.5}$$

F-value: It is calculated as the harmonic mean of precision and recall and is defined as:

$$F = \frac{2 \times P \times R}{P + R} \tag{4.6}$$

where N = Number of correctly parsed parameter indicators; $Total_P$: Number of parameter indicators in the algorithm prediction results; $Total_R$: The Number of parameter indicators in the expected results of the original natural manipulation language sample.

Table 4.4 displays the comparative outcomes between the TF-IDF method and the proposed MNet-SCADA-NLI algorithm in terms of instruction recognition within the intermediate language of the SCADA system. The experimental findings show that the proposed MNet-SCADA-NLI holds notable advantages, particularly when dealing with small sample training sets. TF-IDF primarily relies on keyword extraction for sorting and differentiation based on part of speech and category. While TF-IDF performs well in straightforward natural language instruction parsing, its performance notably diminishes when tackling complex natural language instructions. In some cases, it may fail to parse certain instructions altogether.

For instance, when presented with a sentence like 'Please turn on the desk lamp in the bedroom before closing the window,' TF-IDF often struggles to parse a comprehensive set of commands. Alternative approaches like Seq2Seq are attempted but yield less-than-ideal results due to the limited sample size. Seq2Seq typically demands substantial-high-quality training data to achieve favourable outcomes.

In contrast, the proposed MNet-SCADA-NLI method excels in handling complex natural language instructions. It yields high accuracy when coupled with domain-specific rules and dependency analysis results. However, it's important to acknowledge that this method has certain limitations, particularly regarding question formulation and openness. Furthermore, there is a prerequisite to enhance the accuracy of semantic dependency analysis to achieve superior results for intricate language sequences.

5. Conclusion. A comprehensive semantic analysis methodology called the MNet is introduced from the Semantic Web, deep web, and dependency analysis. MNet is designed to encompass various semantic elements, internal and external relationships, and feature attributes, all structured hierarchically to capture semantic

nuances from individual phrases and sentences to entire texts. Developing a general MNet construction algorithm involves three pivotal processes: Meta relationship, tree structure, and network structure. A novel bottom-up specification-based MNet semantic dependency tree construction algorithm, demonstrating its effectiveness through experiments, is introduced in resolving challenges related to semantic dependency analysis and natural language control interfaces, particularly within the context of SCADA system interfaces. The proposed approach effectively translates the semantic analysis procedure used in SCADA system natural language manipulation interfaces into the broader construction framework of MNet, presenting a promising path for advancing natural language semantic analysis. Potential areas for exploration include utilizing established knowledge bases like WordNet and HowNet to extract word vector features, positions, and parts of speech for MNet meta-relationship construction integrating deep reinforcement learning into the dependency selection process of the MNetParser algorithm.

REFERENCES

[1] A. A. AL-Banna and A. K. AL-Mashhadany, *Natural language processing for automatic text summarization [datasets]-survey*, Wasit Journal of Computer and Mathematics Science, 1 (2022), pp. 156–170.
[2] U. Ashraf, A. Ahmed, M. Al-Naeem, and U. Masood, *Reliable and qos aware routing metrics for wireless neighborhood area networking in smart grids*, Computer Networks, 192 (2021), p. 108051.
[3] S. Conia and R. Navigli, *Probing for predicate argument structures in pretrained language models*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022, Association for Computational Linguistics, pp. 4622–4632.
[4] D. C. Dima, J. Culham, and Y. Mohsenzadeh, *Semantic representations of human actions across vision and language*, Journal of Vision, 23 (2023), pp. 5511–5511.
[5] C. Fachkha, *Cyber threat investigation of SCADA modbus activities*, in Proceedings of the 10th IFIP International Conference on New Technologies, Mobility and Security, Canary Islands, Spain, 2019, IEEE, pp. 1–7.
[6] K. Ghonasgi, S. N. Yousaf, P. Esmatloo, and A. D. Deshpande, *A modular design for distributed measurement of human–robot interaction forces in wearable devices*, Sensors, 21 (2021), p. 1445.
[7] K. He, R. Mao, T. Gong, C. Li, and E. Cambria, *Meta-based self-training and re-weighting for aspect-based sentiment analysis*, IEEE Transactions on Affective Computing, (2022).
[8] K.-C. Kao, W.-H. Chieng, and S.-L. Jeng, *Design and development of an iot-based web application for an intelligent remote scada system*, 323 (2018), p. 012025.
[9] H. Kim, *Security and vulnerability of scada systems over ip-based wireless sensor networks*, International Journal of Distributed Sensor Networks, 8 (2012), p. 268478.
[10] M. Lyu and S. Mo, *Hsrg-wsd: A novel unsupervised chinese word sense disambiguation method based on heterogeneous sememe-relation graph*, in Proceedings of the International Conference on Intelligent Computing, vol. 14089, Singapore, 2023, Springer, pp. 623–633.
[11] D.-D. Nguyen, M.-T. Le, and T.-L. Cung, *Improving intrusion detection in scada systems using stacking ensemble of tree-based models*, Bulletin of Electrical Engineering and Informatics, 11 (2022), pp. 119–127.
[12] Y. Qin, Z. Liu, Y. Lin, and M. Sun, *Sememe-based lexical knowledge representation learning*, in Representation Learning for Natural Language Processing, Springer Nature Singapore Singapore, 2023, pp. 351–400.
[13] R. M. Samant, M. R. Bachute, S. Gite, and K. Kotecha, *Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions*, IEEE Access, 10 (2022), pp. 17078–17097.
[14] B. Sousa, M. Arieiro, V. Pereira, J. Correia, N. Lourenço, and T. Cruz, *Elegant: Security of critical infrastructures with digital twins*, IEEE Access, 9 (2021), pp. 107574–107588.
[15] N. O. Tippenhauer, B. Chen, D. Mashima, and D. M. Nicol, *vbump: Securing ethernet-based industrial control system networks with vlan-based traffic aggregation*, in Proceedings of the 2th Workshop on CPS&IoT Security and Privacy, Online, 2021, pp. 3–14.
[16] S. Wang, L. Pan, and Y. Wu, *Meta-information fusion of hierarchical semantics dependency and graph structure for structured text classification*, ACM Transactions on Knowledge Discovery from Data, 17 (2023), pp. 1–18.
[17] Y. Xiao, C. Li, M. Thürer, Y. Liu, and T. Qu, *User preference mining based on fine-grained sentiment analysis*, Journal of Retailing and Consumer Services, 68 (2022), p. 103013.
[18] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, and N. J. Yuan, *Multi-modal knowledge graph construction and application: A survey*, IEEE Transactions on Knowledge and Data Engineering, (2022).

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 11, 2023
Accepted: Oct 16, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 85–93, DOI 10.12694/scpe.v25i1.2267

HYBRID OPTIMIZATION FOR HIGH ASPECT RATIO WINGS WITH CONVOLUTIONAL NEURAL NETWORKS AND SQUIRREL OPTIMIZATION ALGORITHM

PENGFEI LI*

Abstract. An efficient hybrid optimization algorithm is introduced in this paper to optimize the lightweight design of high-aspect-ratio wings, tackling the complexities associated with the mixed optimization design of layout and size variables within these wing structures. A hybrid binary unified coding description facilitates the optimization process for layout and size variables. The study influences one-dimensional convolutional neural networks to establish an aeroelastic surrogate model, primarily chosen for their exceptional performance in handling multi-parameter aeroelastic regression problems. Additionally, the squirrel optimization algorithm is chosen over the genetic algorithm for the mixed optimization problem, leading to notable savings in computational costs. The research demonstrates that the proposed hybrid optimization method, integrating the one-dimensional convolutional neural network and the squirrel optimization algorithm, offers superior performance in optimizing high aspect ratio wings. Specifically, it results in a reduction of 4.1% in the weight of the wing structure. Moreover, the study highlights the necessity of this hybrid approach due to the observed coupling between the layout variables of the wing ribs and the size variables of the wing beams.

Key words: High aspect ratio wings; Structural optimization design; Hybrid optimization; Convolutional neural network; Squirrel optimization algorithm

1. Introduction. Rotating machinery is critical in contemporary industrial production, constituting approximately 80

The evolution of fault diagnosis technology has progressed with mechanical advancements, unfolding across four distinct stages: During the 19th century, mechanical equipment possessed relative simplicity, warranting post-maintenance measures. Subsequently, from the early 20th century to the 1950s, the increasing complexity of machinery began to disrupt the smooth flow of production and daily life, prompting the adoption of regular maintenance as a diagnostic approach. The 1960s and 1970s witnessed a turning point with computer technology's maturation, leading to more sophisticated diagnostic methodologies that involved observing data patterns during machine operations and implementing targeted maintenance strategies [2].

The trajectory of fault diagnosis technology, foreseeing its alignment with the progression of intelligent decision-making capabilities in industrial big data, is grounded in the theoretical principles of data science. The Convolutional Neural Network (CNN) is a prominent model within deep neural networks, drawing significant attention since its initial exploration in the 1980s. Inspired by the biological visual perception mechanism, it boasts a remarkable capacity for representational learning. Notably, CNN retains its efficacy in challenging scenarios characterized by complex environmental information, ambiguous background knowledge, obscure inference rules, and samples with substantial impairments [3].

The rise of data-driven methodologies has recently witnessed widespread adoption across various domains. Its application aims to decipher the operational state of systems through data analysis, facilitating decision-making and control of equipment and production processes. Consequently, it has emerged as the prevailing technology for fault diagnosis. This approach relies on diverse datasets, delving into the intrinsic data patterns via machine learning and statistical analysis techniques, thus enabling the construction of fault diagnosis models. The author's study explores data-driven fault diagnosis methods, which can be categorized into three key stages: data collection, feature extraction, and fault classification [4].

The advancement of industrial big data is propelling the adoption of data-driven fault diagnosis methods, integrating an ever-growing range of data sources. Commonly employed data types encompass current, acoustic emission (AE), and vibration data. Particularly, fault diagnosis techniques reliant on vibration data have

* Zhengzhou Railway Vocational and Technical College, Zhengzhou, 451460, China (pengfeili69@163.com).

garnered substantial traction within the industrial sphere. Vibration data derived from machinery can effectively capture diverse indicators related to multiple components and structures, including gear meshing frequency, bearings, structural resonance, and electrical anomalies [5].

The direct installation of sensors on the casing of the monitored component presents a distinct advantage, as it minimizes the potential interference of received signals. Since the inception of the first solar-powered drone, Sunrise, more than 40 years ago, the development of High Altitude Long Endurance (HALE) solar-powered drones has been ongoing. To fulfil the demanding prerequisites of high lift-to-drag ratios and lightweight designs, the wings of these high-altitude, long-endurance aircraft are predominantly engineered with high aspect ratios and notable flexibility. This particular design paradigm, however, exhibits a heightened sensitivity to the aircraft's overall weight. Therefore, the primary goal of the structural optimization design for solar-powered unmanned aerial vehicles remains the achievement of lightweight design while ensuring the fulfilment of structural strength and stiffness requirements [6].

Two distinct strategies have emerged in addressing such complex multi-parameter structural optimization problems: the variable-by-variable and comprehensive optimization methods. While the former approach overlooks the interdependencies between multiple variables, the latter involves the mixed optimization of multiple variables, thereby considering the intricate relationships among various parameters to obtain an ultimate global optimal solution. However, this holistic approach is challenging due to its high computational complexity [7].

The authors proposed a multi-objective optimization approach for three-dimensional truss tower structures, employing enhanced Pareto evolutionary algorithms, population-based incremental learning algorithms, and archived simulated annealing algorithms to overcome the general challenges. Simultaneously, a hybrid optimization strategy for the layout and topology of reinforcing ribs in plate and shell structures using the Solid Isotropic Material with Penalization (SIMP) model and the basic structure method successfully verified the reliability of this method [8].

The structural design of solar-powered drone wings and the intricate aeroelastic characteristics stemming from high-flexibility wings under aerodynamic loads underscore the critical importance of establishing a highly reliable aeroelastic model. As computational technology has advanced, numerical calculations, notably Computational Fluid Dynamics/Computational Structural Dynamics (CFD/CSD) fluid-structure coupling technology, have assumed a pivotal role in aircraft optimization design. Traditional surrogate models, including the Polynomial Response Surface (PRS) model and the Kriging model, have been complemented by the emerging prominence of deep learning-based surrogate models, particularly adept at handling high-dimensional nonlinear problems [9].

The design of the hybrid optimization model encompasses a strategic fusion of multiple methodologies to tackle the intricacies of optimizing high aspect ratio wing structures. A pivotal element within this framework is implementing a unified coding scheme meticulously tailored to encapsulate layout and size variables. This coding methodology ensures a coherent representation of the intricate structural parameters, forming the cornerstone of the optimization process.

2. Analytical Methods. A suitable one-dimensional convolutional neural network structure is designed using a specific optimization algorithm and loss function. Using CFD/CSD technology to calculate the aeroelastic data of solar unmanned aerial vehicle wings with a high aspect ratio and using reasonable coding methods to encode the structural and aeroelastic data according to the characteristics of the wings with high aspect ratio, the aeroelastic surrogate model is obtained by training the one-dimensional convolutional neural network model. Based on the surrogate model, the squirrel optimization algorithm is used for optimization [10].

2.1. Aeroelastic surrogate model based on one-dimensional convolutional neural network.

(1) A one-dimensional convolutional neural network model The CNN functioning as a multilevel feedforward neural network is structured with distinct layers, including an input layer, a convolutional layer, an activation layer, a pooling layer, a fully connected layer, and an output layer. Its prevalence in computer vision tasks stems from its capability to extract features from localized input blocks and subsequently modularize them, optimizing data utilization. In one-dimensional sequence recognition, the distinct attributes of one-dimensional convolutional neural networks, such as pattern learning, translation invariance, and spatial hierarchy, can be effectively leveraged for multi-parameter regression analysis [11].

In deep learning, the optimization algorithm fine-tuns weight values to identify the optimal parameter combination that minimizes the loss function value. Commonly employed optimization algorithms for CNN include Stochastic Gradient Descent (SGD), SGD with Momentum, RMSProp, and Adam [12]. The expression for gradient weight is given by

$$g_t = \frac{1}{n}\nabla(\theta_{t-1})\sum_{i=1}^{n} L(f(x_i, \theta_{t-1}), y_i) \tag{2.1}$$

In Equation 2.1, g_t is the weight gradient, n is the number of small batch samples, θ is the weight, l is the time step, $f(x, \theta)$ is the forward inference result of the convolutional neural network. Y is the real label, $L(\hat{y}, y)$ is the loss function, the author trains the data set by using different loss functions and selects the loss function most suitable for determining the Mean Squared Error (MSE).

(2) Structural parameter coding

For regression analysis, the convolutional neural network necessitates a training dataset. The training dataset comprises structural parameters and aeroelastic result data within the aeroelastic surrogate model. The structural data is encoded using suitable encoding methods, while the dimensional parameters are binary. Ultimately, the two are combined into a sequence, taking the following form:

$$T_{1,1}, T_{1,2}, ..., T_{2,1}, ..., T_{n,m} | S_{1,1}, S_{1,2}, ..., S_{2,1}, ..., S_{n,m} \tag{2.2}$$

where

$$T_{i,j} \epsilon \{0, 1\}, S_{i,j} = \{a_{i,j,1}, a_{i,j,2}, ..., a_{i,j,l}\}, a_{i,j,k} \epsilon \{0, 1\} \tag{2.3}$$

In the Formula, $T_{i,j}$ represents whether the i^{th} structural member exists at position j, θ indicates nonexistence, 1 indicates existence; N represents the total number of structures, and m represents the total number of structural distribution positions; $S_{i,j}$ represents the size of the i^{th}-type structural component at position j, composed of binary code $a_{i,j,k}$; L represents the binary encoding tension.

(3) Aeroelastic surrogate model

The acquisition of sample data involves aeroelastic simulation calculations, with subsequent binary encoding of the structure. This data is then inputted into the one-dimensional convolutional neural network for training, resulting in the generation of the surrogate model, as illustrated in Figure 2.1. The aeroelastic simulation employs CFD/CSD bidirectional coupling technology, wherein the fluid domain and structure are solved independently, and data in the time domain are staggered to achieve progressive advancements, ultimately deriving the aeroelastic data of the coupled system. During the training of sample data using one-dimensional convolutional neural networks, the elastic axis displacement of the wing tip is designated as the prediction target. The average absolute error value (MAE) serves as the evaluation metric, prompting continuous adjustments to the network structure and ensuring improved prediction accuracy of the surrogate model [13,14].

2.2. Squirrel Optimization Algorithm. The Squirrel Search Algorithm (SSA) is an intelligent swarm-based optimization algorithm inspired by squirrels' foraging behaviour and aerodynamics for its rapid convergence and robust optimization capabilities [15,16]. The algorithm operates according to the following process:

(1) Random initialization

$$FS = \begin{bmatrix} FS_{1,1}FS_{1,2}...FS_{1,d} \\ FS_{2,1}FS_{2,2}...FS_{2,d} \\ ... \\ FS_{n,1}FS_{n,2}...FS_{n,d} \end{bmatrix} \tag{2.4}$$

The Equation $FS_{i,j}$ represents the value of the i^{th} squirrel in the j^{th} dimension, which can be evenly distributed using Equation 2.5.

$$FS_i = FS_L + U(0, 1) \times (FS_U - FS_L) \tag{2.5}$$

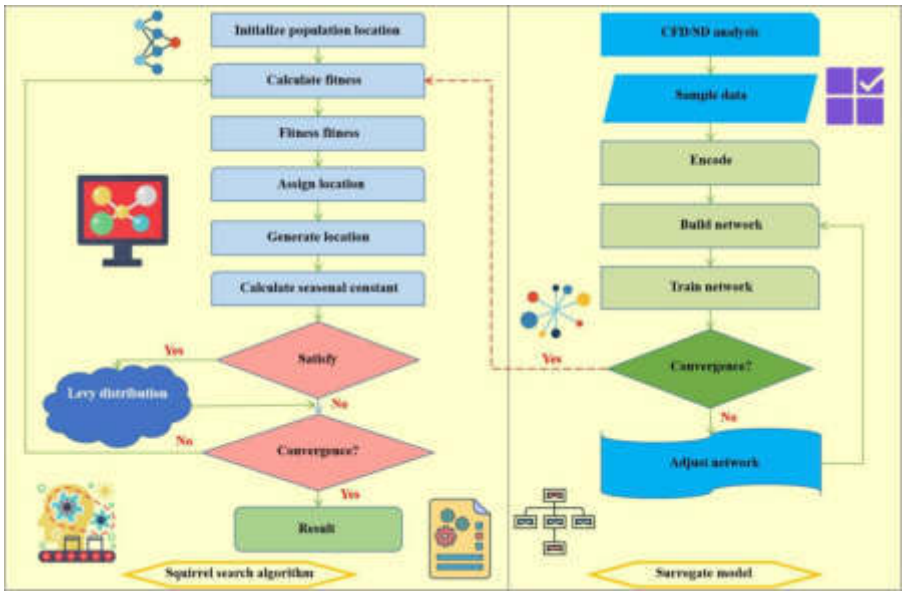


Fig. 2.1: Proposed Hybrid Optimization for High Aspect Ratio Wings with CNN and SSA

In Equation 2.5, FS_L and FS_U are the lower and upper bounds of the position of the ith squirrel, and U (0,1) is a uniformly distributed random number within the range of [0,1][17].

(2) Fitness evaluation

Applying a user-defined fitness function involves computing the positional fitness for each squirrel. Subsequently, the fitness values are arranged in ascending order. Squirrels with the minimum fitness are allocated to the hickory tree, while the subsequent three squirrels reside on oak trees, leaving the remaining to settle on ordinary trees. The foraging behaviour of squirrels is also influenced by predators, with a probability of P_{dp} =0.1.

(3) Generate new location

Squirrels encounter three scenarios during their active foraging process: The initial situation involves transitioning from an oak tree to a hickory tree:

$$FS_{at}^{t-1} = \begin{cases} FS_{at}^{t} + d_g \times G_c \times (FS_{ht}^{t} - FS_{at}^{t}), R_l \geqslant P_{dp} \\ Random, location, otherwise \end{cases} \tag{2.6}$$

In the Formula, d_g is the random gliding distance, R_1 is a random number of [0,1], FSat is the position of the squirrel on the oak tree, FS_{ht} is the position where the squirrel reaches the hickory tree, and G_c is the sliding coefficient. To balance global and local search, G_c is generally taken as 1.9. In the second case, flying from a regular tree to an oak tree:

$$FS_{nt}^{t+1} = \begin{cases} FS_{nt}^{t} + d_g \times G_c \times (FS_{at}^{t} - FS_{nt}^{t}), R_2 \geqslant P_{dp} \\ Random, location, otherwise \end{cases} \tag{2.7}$$

In Equation 2.7, R_2 is a random number in [0,1], and FS_{nt} is the squirrel's position on a regular tree.

The third scenario: Flying from a regular tree to a hickory tree:

$$FS_{nt}^{t+1} = \begin{cases} FS_{nt}^{t} + d_g \times G_c \times (FS_{ht}^{t} - FS_{nt}^{t}), R_3 \geqslant P_{dp} \\ Random, location, otherwise \end{cases} \tag{2.8}$$

where, R_3 is a random number in [0,1].

(4) Seasonal variation conditions

Seasonal monitoring conditions are incorporated into SSA to prevent the algorithm from becoming entangled in local optimal solutions. When these seasonal conditions are met, Levy Distribution introduces random positional changes for squirrels on ordinary trees [18].

2.3. Multi-parameter mixed optimization mathematical model for wing structure. According to the proposed structural coding method, the mathematical model of hybrid optimization can be expressed as:

$$\begin{aligned} &find, T, S \\ &min, m(T, S) \\ &s.t, g_i \leqslant 0.i = 1, ..., n \\ &T_{i,j}\epsilon\{0, 1\}.i = 1, ..., n.j = 1, ..., m \\ &S_i\epsilon[LB_i, UB_i].i = 1, ..., n \end{aligned} \tag{2.9}$$

The Formula incorporates T as the layout variable for the structural component, assuming values of 0 or 1. S represents the dimensional variable of the structural component, whereas M(T, S) signifies the structural mass. gi denotes the constraint conditions of the structure, while LB and UB denote the lower and upper bounds of structural size variables. The overall optimization algorithm process is visualized in Figure 2.1. The initial step involves acquiring sample data through aeroelastic simulation calculation, encompassing structural parameters alongside their corresponding wingtip elastic axis displacement. Subsequently, the structure undergoes binary encoding [19].

Integrating the squirrel optimization algorithm and the surrogate model, derived from training in generating a one-dimensional convolutional neural network, initiates the hybrid optimization process. Initially, the squirrel dimension is determined based on the number of variables. Subsequently, a decimal description is employed for initialization, and the corresponding squirrel position is coded, as demonstrated in Equation 2.2. The fitness value is computed using the surrogate model, enabling the sorting and allocating of squirrel positions across various tree types according to their respective fitness values.

Subsequent reallocation of squirrel positions is facilitated using Equations 2.6 to 2.8. The decision to redistribute squirrel positions utilizing the Levi distribution is contingent upon satisfying seasonal conditions. This iterative process continues until the optimization result is achieved [20].

3. Example analysis of wing structure optimization design. The ribs and beams are the primary load-bearing components of solar-powered drone wings with high aspect ratios. The layout and size parameters associated with these ribs and beams are crucial variables dictating structural performance. The wing structure parameters are derived from relevant literature in the author's case study. Based on the wing model, variations are introduced in the distribution of wing ribs and the diameter of wing beams. Select models with distinct parameters are chosen as samples for aeroelastic numerical analysis. The selection criterion for these samples ensures that each parameter encompasses diverse values, albeit not covering all potential values.

3.1. Numerical and Optimization Models for Wings. The wing structure is characterized by a flat and straight design, featuring an Eppler387 airfoil, a chord length of 0.41m, and a span length of 2.5m. The materials utilized adhere to isotropic models. The composite material constituting the wing rib and wing beam boasts an $1800kg/m^3$ density, a Poisson's ratio of 0.307, and an elastic modulus of 100GPa. Conversely, areas apart from the wing ribs and wing beams are filled with foam materials, showcasing a density of $20kg/m^3$, a Poisson's ratio of 0.08, and an elastic modulus of 5.4MPa. The wing structure model comprises 11 evenly distributed wing ribs, each with a thickness of 2mm. Considering the dimensions of the solar panel, a maximum of 21 wing ribs can be uniformly distributed across all feasible positions on the structure. Regarding the wing profile shape, the inner diameters of the three wing beams are 8mm, 22mm, and 14mm, while the corresponding outer diameters are 10mm, 26mm, and 16mm, respectively. The variables subject to the author's optimization efforts are the distribution of wing ribs across 21 positions and the inner diameter of the three wing beams.

During the aeroelastic numerical analysis process, special attention is given to maintaining the grid height of the fluid domain grid within the first layer of the wing boundary layer at $y+ \approx 1$. This meticulous approach

Table 3.1: Parameters of convolutional neural networks

Layer	Type	Neuron	Activation
1	Input	32	-
2	Convolution	64	Relu
3	Max pooling	-	-
4	Convolution	128	Relu
5	Global average pooling	-	-
6	Dense	16	-
7	Output	1	-

ensures that the total number of grids is 3 million. Employing a standard $k-\varepsilon$ A turbulence model, the analysis considers an incoming flow velocity of $15m/s$ and a wing angle of attack 5°. The solid domain grid is set at 500,000, with a fixed constraint at one wing beam's end. The original model's calculated elastic axis displacement of the wing tip measures 143.64mm, serving as a critical structural constraint condition.

3.2. Prediction accuracy verification of surrogate model. Utilizing the Adam algorithm, the author constructs a one-dimensional convolutional neural network to establish a hybrid optimization surrogate model. The neural network's architecture is detailed in Table 3.1.

The sample dataset comprises 180 groups, segregated into three groups using K-fold cross-validation. The AdaGrad, RMSProp, and Adam algorithms are applied to validate both the one-dimensional convolutional neural network and the fully connected network. The fully connected network comprises three hidden layers, with a network structure and the number of neurons mirroring those of the one-dimensional convolutional neural network. The evaluation metric employed is the Mean Absolute Error (MAE) value. The outcomes following 10,000 training iterations are illustrated in Figure 3.1, where Dense denotes the fully connected network and Convnet represents the one-dimensional convolutional network.

Figure 3.1 demonstrates the superior performance of the one-dimensional convolutional neural network optimized by the Adam algorithm. After 500 training steps, the MAE stabilizes without displaying signs of "overfitting." Post-convergence, the MAE mean settles around 0.8. Relative to the sample data featuring a displacement mean of 150mm, the relative error approximates 0.5%. This outcome underscores the efficacy of the surrogate model in addressing multi-parameter aeroelastic regression issues [21].

3.3. Feasibility verification of hybrid optimization algorithm. The preceding analysis demonstrates the viability of establishing an aeroelastic surrogate model using the one-dimensional convolutional neural network optimized via the Adam algorithm. Building upon this surrogate model, this section validates the practicality of the hybrid layout and size optimization algorithm within the wing structure optimization design process. To assess its efficacy, the results of the squirrel optimization algorithm were juxtaposed with those of the genetic algorithm, maintaining uniformity in key parameters such as the probability of random population changes.

The optimization capabilities are compared for the two algorithms with smaller population sizes. Fifty calculations were executed at populations of 50 and 100, respectively. The changes in average wingtip displacement and structural mass were recorded across iteration steps, illustrated in Figure 3.2 (with GA_50 and GA_100 representing genetic algorithms with populations of 50 and 100, respectively, and SSA_50 and SSA_100 representing squirrel optimization algorithms with populations of 50 and 100, respectively). Furthermore, the time taken for 1000 iterations of the two algorithms on the Inteli7-3770 processor was documented, as presented in Table 3.2 (where GA and SSA denote genetic algorithms and squirrel optimization algorithms, respectively) [22].

The observations from Figure 3.2 underscore that throughout the optimization procedure, the squirrel optimization algorithm outperforms the genetic algorithm in locating optimal values, showcasing superior stability in terms of wingtip displacement and structural mass. Examination of Table 3.2 reveals that the squirrel optimization algorithm demonstrates a more efficient time utilization when addressing this specific mixed opti-

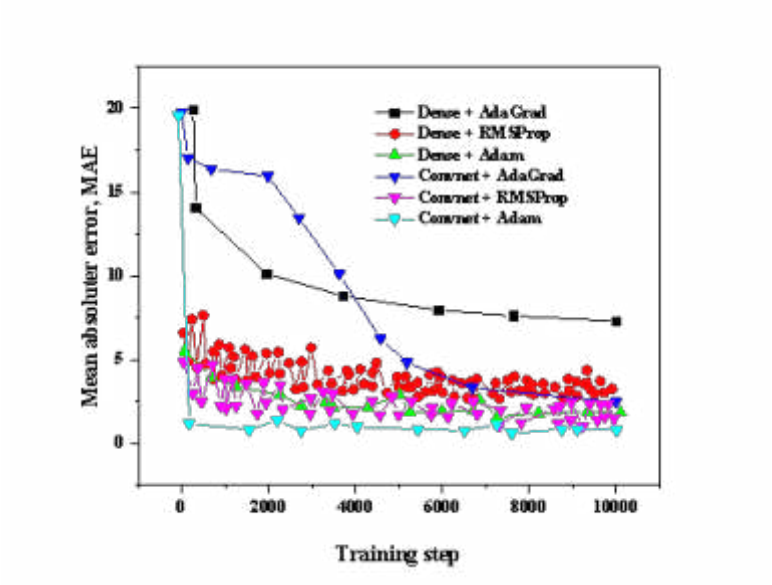


Fig. 3.1: Average Absolute Error Variation of Different Neural Networks

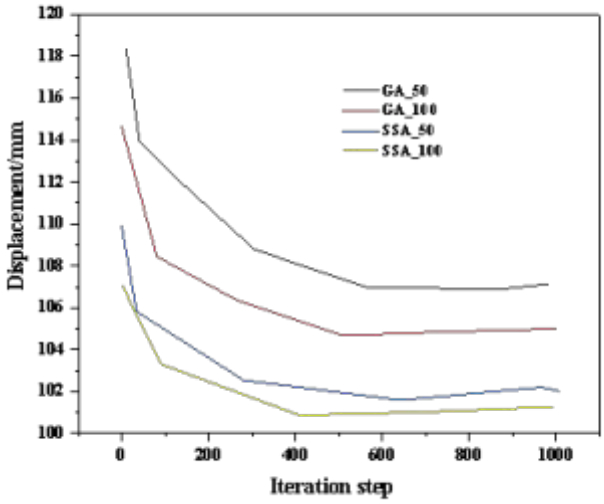


Fig. 3.2: Iteration History of Wing Tip Displacement and Structural Mass

mization problem. Notably, the time consumption is reduced by 45.51% with a population size of 50 and by 35.80% when the population is 100.

Employing the squirrel optimization algorithm with a population size of 100, the individual layout variables of wing ribs and the mixed optimization problem of wing rib layout variables and wing beam size variables were optimized, resulting in an optimized wing structure with a refined wing rib layout and a hybrid optimization of wing rib layout and wing beam size.

The optimized wing layout features five fewer ribs than the original structure, resulting in a weight of

Table 3.2: Time consumption of two optimization algorithms

Algorithm	Population	Time(s)
Genetic Algorithm(GA)	50	72.90
	100	83.81
Squirrel Search Algorithm(SSA)	50	40.09
	100	53.81

1.6588kg, marking a 3.1% reduction compared to the original design. As depicted in Figure 3.2, the mixed optimization of layout and size variables results in the wing comprising five wing ribs. Additionally, the inner diameter of the wing beams now varies from front to back, measuring 9mm, 21mm, and 15mm, respectively.

The comparison above highlights the superior performance of the aeroelastic surrogate model, established through the one-dimensional convolutional neural network, effectively handling mixed regression problems related to structure distribution and size. Comparatively, when addressing mixed optimization problems, the squirrel optimization algorithm outperforms the genetic algorithm, yielding better solutions at reduced computational costs.

4. Conclusion. The author introduces an effective hybrid optimization algorithm to target the mixed optimization design challenge inherent in the layout and size of high aspect ratio wings. Leveraging the unified coding of multiple variables, the study implements a one-dimensional convolutional neural network to establish an aeroelastic surrogate model. This is complemented by a solution workflow structured around the squirrel optimization algorithm for comprehensive search and optimization. The study highlights the efficacy of the proposed hybrid layout and size optimization method for achieving lightweight design objectives in high aspect ratio wings. Based on comprehensive CFD/CSD aeroelastic calculations and uniform coding of diverse structural variables, the surrogate model demonstrates the significant capabilities of the one-dimensional convolutional neural network. Applying the squirrel optimization algorithm effectively reduces computational costs by 35% to 45% compared to the genetic algorithm. Furthermore, the hybrid optimized structure showcases a notable 4.1

REFERENCES

[1] Wang, Z. , Lyu, Z. , Duan, D. , & Li, J. . (2021). A novel system identification algorithm for quad tilt-rotor based on neural network with foraging strategy: Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 235(11), 1474-1487.

[2] Kan, Z. , & Liu, X. . (2021). The study on void fraction prediction of gas-liquid two phase flow based on convolutional neural network. Journal of Physics: Conference Series, 2121(1), 012029.

[3] Liu, C. , He, J. , Wang, P. , Xing, D. , Li, J. , & Liu, Y. , et al. (2023). Characteristic extraction of soliton dynamics based on convolutional autoencoder neural network. Chinese Optics Letters, 21(3), 031901.

[4] Hu, Z. X., Wang, Y., Ge, M. F., & Liu, J. (2019). Data-driven fault diagnosis method based on compressed sensing and improved multiscale network. IEEE Transactions on Industrial Electronics, 67(4), 3216-3225.

[5] Yan, J., Meng, Y., Lu, L., & Li, L. (2017). Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. IEEE access, 5, 23484-23491.

[6] Wang, X., Yang, Y., Wu, D., Zhang, Z., & Ma, X. (2020). Mission-oriented 3D path planning for high-altitude long-endurance solar-powered UAVs with optimal energy management. IEEE Access, 8, 227629-227641.

[7] Liao, T., Socha, K., de Oca, M. A. M., Stützle, T., & Dorigo, M. (2013). Ant colony optimization for mixed-variable optimization problems. IEEE Transactions on evolutionary computation, 18(4), 503-518.

[8] Deb, K. (2003). Multi-objective evolutionary algorithms: Introducing bias among Pareto-optimal solutions. In Advances in evolutionary computing: theory and applications (pp. 263-292). Berlin, Heidelberg: Springer Berlin Heidelberg.

[9] Huang, C., Huang, J., Song, X., Zheng, G., & Yang, G. (2021). Three dimensional aeroelastic analyses considering free-play nonlinearity using computational fluid dynamics/computational structural dynamics coupling. Journal of Sound and Vibration, 494, 115896.

[10] Ahmad, S. N. , & Prakash, O. . (2021). Optimization of ground heat exchanger of the ground source heat pump system based on exergetic analysis using taguchi technique:. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 235(21), 5892-5901.

[11] Yang, Y., Nie, Z., Huang, S., Lin, P., & Wu, J. (2019). Multilevel features convolutional neural network for multifocus image fusion. IEEE Transactions on Computational Imaging, 5(2), 262-273.

[12] Newton, D., Yousefian, F., & Pasupathy, R. (2018). Stochastic gradient descent: Recent trends. Recent advances in optimization and modeling of contemporary problems, 193-220.
[13] Dolezel, P. , Holik, F. , Merta, J. , & Stursa, D. . (2021). Optimization of a depiction procedure for an artificial intelligence-based network protection system using a genetic algorithm. Applied Sciences, 11(5), 2012.
[14] Li, Z. , Wang, Y. , & Ma, J. . (2021). Fault diagnosis of motor bearings based on a convolutional long short-term memory network of bayesian optimization. IEEE Access, PP(99), 1-1.
[15] Li, C. , Yin, C. , & Xu, X. . (2021). Hybrid optimization assisted deep convolutional neural network for hardening prediction in steel. Journal of King Saud University - Science, 33(6), 101453.
[16] Alshaikhli, O. A. , & Al-Araji, A. . (2021). Path planning and control strategy design for mobile robot based on hybrid swarm optimization algorithm. International Journal of Intelligent Engineering and Systems, 14(3), 2021.
[17] He, Y. . (2021). Design and implementation of convolutional neural network accelerator based on riscv. Journal of Physics: Conference Series, 1871(1), 012073 (5pp).
[18] Fares, D., Fathi, M., Shams, I., & Mekhilef, S. (2021). A novel global MPPT technique based on squirrel search algorithm for PV module under partial shading conditions. Energy Conversion and Management, 230, 113773.
[19] Tanveer, A., & Ahmad, S. M. (2023). Mathematical Modelling and Fluidic Thrust Vectoring Control of a Delta Wing UAV. Aerospace, 10(6), 563.
[20] Zhao, X., Tang, Z., Cao, F., Zhu, C., & Periaux, J. (2022). An Efficient Hybrid Evolutionary Optimization Method Coupling Cultural Algorithm with Genetic Algorithms and Its Application to Aerodynamic Shape Design. Applied Sciences, 12(7), 3482.
[21] Li, W., Gao, X., & Liu, H. (2021). Efficient prediction of transonic flutter boundaries for varying Mach number and angle of attack via LSTM network. Aerospace Science and Technology, 110, 106451.
[22] Vidushi, Agarwal, M. , Rajak, A. , & Shrivastava, A. K. . (2021). Assessment of optimizers impact on image recognition with convolutional neural network to adversarial datasets. Journal of Physics: Conference Series, 1998(1), 012008.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 11, 2023
Accepted: Oct 25, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 95–102, DOI 10.12694/scpe.v25i1.2268

COMPUTER SOFTWARE MAINTENANCE AND OPTIMIZATION BASED ON IMPROVED GENETIC ALGORITHM

MING LU*

Abstract. Optimizing computer software maintenance is the key goal, which also ensures dependable and consistent network performance. In order to increase genetic operations and evaluate the satisfaction and fitness index functions, this article employs an improved genetic algorithm. Utilizing the network's performance and controlling restrictions through controlled data iterations, the architecture is refined. The study also finds a link between the number of iterations and the rate of network optimization, supporting the results of the genetic algorithm. The results show that the reliability of the network system decreases as the number of genetic operation repeats increases. If a critical point is reached, the enhancement in network reliability tends to level off due to hardware constraints or other relevant factors. Notably, the study identifies the maximum attainable value of network reliability at 0.894, precisely at 100 iterations. These conclusions offer an essential framework for optimizing the design of computer network reliability, emphasizing the necessity of a well-balanced approach to genetic algorithm-based optimization.

Key words: Computer Networks; Genetic Algorithm; Network Optimization; Network Reliability; Iterative Analysis

1. Introduction. The robustness of computer networks, fundamental to their operation, stems from their innate capacity to endure and withstand potential damage and disruptions. This flexibility empowers networks to sustain a strong and smooth operational cycle, enabling the efficient and timely implementation of updates and ensuring a consistent, stable performance during real-time operations. The ability to withstand various threats and disturbances is crucial for upholding the network's integrity and dependability, guaranteeing uninterrupted functionality even in the face of challenging circumstances. By nurturing this flexibility, computer networks can proficiently meet the demands of contemporary digital settings, maintaining an uninterrupted flow of data and information without compromising performance or stability [1]. Mismanagement of network stability can result in equipment malfunctions, severely threatening data preservation and potentially leading to data loss and system immobilization. Establishing reliable and secure computer networks necessitates integrating hardware and software functionalities. Particularly, network hubs serve as the primary stronghold of security, playing an irreplaceable role in the network infrastructure. Any failure within network hubs can directly impact user accessibility and significantly disrupt the overall user experience. Consequently, there is a pressing need for in-depth research and empirical analysis focused on optimizing computer technology, enhancing technical capabilities, and implementing refined solutions within the fundamental optimization framework [2].

Ensuring the resilience of computer networks demands a comprehensive approach that involves fortifying the system's defence mechanisms against potential threats, both internal and external. Strengthening the network's ability to withstand cyber-attacks, system failures, and data breaches is crucial to maintaining the integrity and functionality of the network infrastructure. Furthermore, prioritizing the operational cycle and optimizing the efficiency of network updates is imperative for ensuring whole and uninterrupted digital operations, allowing for quick adaptation to evolving technological demands [3].

The continuous integration of hardware and software components lays the foundation for a sturdy and dependable computer network. Harnessing the complementary features of these components in tandem is pivotal for unleashing the network infrastructure's complete capabilities. A cohesive and well-aligned hardware and software system not only boosts the network's efficiency but also reinforces its overall dependability and security, nurturing a unified environment capable of adapting to the dynamic requirements of the digital environment [4].

The foundation of evaluating the effectiveness of optimization strategies in enhancing computer network

* College of Mechanical Electronic and Information Engineering, Wuxi Vocational Institute of Arts & Technology, Wuxi 214206, China (minglu7@126.com)

reliability lies in empirical analysis. Through comprehensive empirical studies, potential areas for enhancement can be identified, leading to targeted solutions that further strengthen the network infrastructure. Continuous monitoring and evaluation facilitate the identification of potential vulnerabilities, providing invaluable insights for refining existing network optimization strategies and ensuring consistent and robust network performance over time [5].

Establishing a strong and reliable computer network demands a comprehensive approach that addresses the diverse facets of network reliability. By emphasizing the integration of hardware and software, reinforcing defence mechanisms, and conducting empirical analyses, it becomes feasible to construct a robust and secure network architecture capable of withstanding the challenges posed by the contemporary digital landscape. Through a proactive and strategic approach, it is possible to enhance the reliability and security of computer networks, fostering a seamless and efficient digital environment for diverse user needs and demands [6].

The Committee, in its recent evaluation, expresses its sincere appreciation for the State party's continued dedication to reinforcing measures that ensure the unwavering commitment of States parties to the effective implementation of the Convention on the Privileges and Immunities for the Electronic Sphere (CPIES). This acknowledgement encompasses the concerted efforts made by the State party in bolstering the reliability of computer systems, which includes. Still, it is not limited to enhancing the flexibility of the external environment, stimulating the dependability of software and hardware components, and fostering the reliability of the personnel involved [7]. It is imperative to emphasize the need for continual advancements in software algorithms, fortifying the network's dependability, and refining the overall design while considering the optimal state of the hardware system. This concerted approach is a pivotal avenue for future development [8].

The ongoing research endeavours in China concerning the stability of computer networks primarily revolve around the intricate dynamics of telecommunications signal networks. By analyzing network signal failures, the transmission of telecommunications signals is strategically harnessed to facilitate dynamic capacity changes, enabling seamless transmission through the exchange of signals. A key role of this research involves the establishment of comprehensive parameters for network security authentication, laying the groundwork for formulating a robust evaluation system. This systematic approach is instrumental in facilitating the optimization design of the network, ensuring its performance and resilience under varying conditions [9,10].

2. Literature review. Employing a strategic approach rooted in topology planning has emerged as a pivotal strategy for circumventing network impediments. By carefully delineating network topologies, researchers aim to identify and overcome potential obstacles, thus fostering a more streamlined and efficient network infrastructure. As the landscape of intelligent optimization solutions continues to expand, the quest for viable and dependable network technology solutions remains an ongoing pursuit. Notably, optimization methods have played a pivotal role in advancing this endeavour. One notable technique involves the strategic deployment of genetic algorithms, which have been effectively tailored to achieve comprehensive network optimization, particularly regarding granularity considerations. Consequently, this concerted multidimensional approach, integrating multi-objective methodologies, has continuously refined network optimization strategies, thereby establishing a robust foundation for continual evolution and enhancement of network stability [11].

The significance of reliable computer networks in modern society forms a central theme, encompassing the multifaceted implications of network failures across diverse sectors. It emphasizes the fundamental role of robust network infrastructures in supporting the smooth functioning of contemporary industries and societal operations. Examining the historical development of network reliability underscores the evolution of key concepts and breakthroughs, charting the progression of strategies to enhance network stability. This historical perspective sheds light on the trajectory of advancements in hardware, software, communication protocols, and network architectures, all of which have shaped the current landscape of network reliability [12].

The authors examine several solutions, such as fault-tolerant architectures, neural networks, genetic algorithms, and redundancy protocols, to optimize network reliability. This research provides insights into the applicability and success of each strategy in real-world circumstances by critically evaluating its strengths and limits without case studies and empirical research, which offer real-world examples of how to apply particular techniques to increase network reliability. An essential tool for understanding the state of computer network reliability research from a broad perspective is explained. Its thorough analysis identifies any shortcomings or gaps in the current research, identifying particular areas that require more study and inquiry [13].

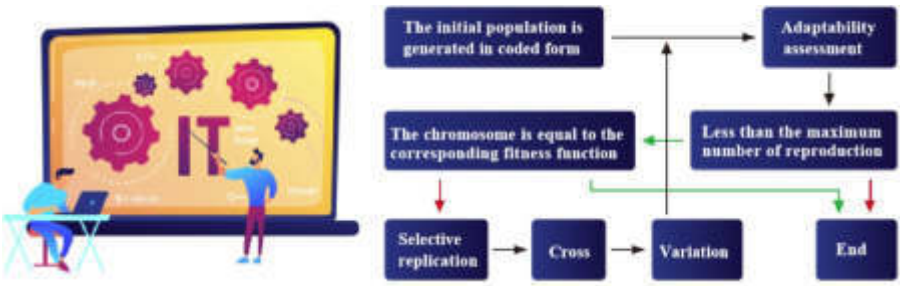


Fig. 3.1: The Proposed Genetic Algorithm Process

Providing insight into future possibilities for computer network dependability research and development, the assessment offers significant insights into the prospective trajectory of developments and the changing obstacles within the area by clarifying emerging trends, technological advancements, and creative techniques. The authors adopt interdisciplinary collaboration and exchanging ideas among researchers, practitioners, and stakeholders as a fundamental platform for ongoing exploration. Through fostering a thorough comprehension of the complex interactions among different factors that impact network reliability, the review hopes to encourage the creation of novel frameworks, approaches, and solutions that can further the development of resilient and strong computer networks [14].

3. Analysis of Computer Network Reliability Principles. The reliability of computer networks depends on the independence of individual systems. Seamless interaction among various network protocols is vital for tailored network operations. Ensuring the autonomy of each computer is fundamental, enabling the utilization of information resources across diverse network environments. This framework facilitates smooth information exchange across network terminals. Facilitating resource sharing among subnets and resource networks is critical for efficient asset utilization under varying network conditions. Data processing within computer networks is meticulously managed by implementing subnets, ensuring stable data utilization across the network infrastructure [15].

Given the expansive reach of computer networks, establishing stable communication channels while preserving network integrity is paramount. Algorithms strategically prioritize resource allocation between users and servers, enhancing the network's ability to manage and distribute resources effectively. This approach guarantees seamless data transmission, fostering a reliable and robust computer network infrastructure [16].

The innovative technique of adaptive stabilization, twisting biological mutation with computer science, has led to the conception of the genetic algorithm, represented in Figure 3.1. This algorithm serves as a pioneering solution at the intersection of these disciplines. Drawing inspiration from the cognitive processes observed in the biological realm, computer-based individuals operate autonomously, ensuring the continuous detection of algorithms across the entire system. Initially depicted as individual entities, these computer individuals combine into groups, forming an interconnected data matrix that is spatially organized. Leveraging diverse evaluation methodologies, genetic operators are crafted under specific crossing and mutation conditions, adhering to the principles of genetic computation.

Matrix computation systematically optimizes the genetic equation, enabling the seamless execution of coding and genetic operations. Within the intersection realm, the entire system's computational potency reinforces local search capabilities, supplementing the overarching search process. In optimizing genetic algorithms, the object encoded by parameters can be dynamically restructured, evading the constraints associated with data limitations. This adaptive approach significantly amplifies the breadth and depth of the search, leading to an enriched exploration of potential solutions [17,18].

3.1. Content of Computer Network Reliability Optimization. The optimization of computer network reliability involves a thorough analysis and calculation of diverse objectives. Emphasizing the performance characteristics within specific geographic regions, the focus remains on bolstering network stability and main-

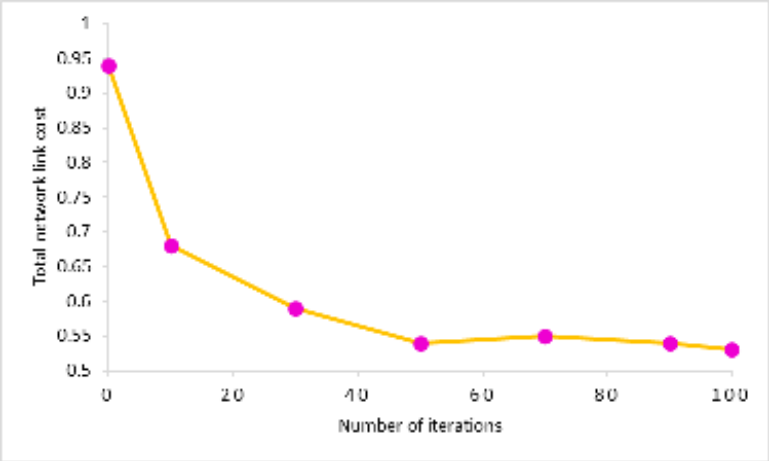


Fig. 3.2: Network Stability Data Iteration Curve

tainability. This holistic approach addresses the intricate challenges of optimizing multiple objectives, fostering a resilient and robust network structure.

The formulation of coding data schemes assumes a crucial role in enhancing computer network reliability. This process involves arranging genes into sequential chromosome patterns, generating initial entities that are subsequently grouped for streamlined operations. Computation of the fitness analysis for individual chromosomes facilitates the determination of inheritance probabilities for specific traits during subsequent stages. During the cross-operation phase, meticulous records of chromosome pairings guide the identification of an optimal pairing method. This intricate process ensures the seamless integration of diverse genetic elements, promoting the development of a sturdy and high-performing network infrastructure.

Throughout the optimization process, the satisfaction function serves as a critical metric, offering a comprehensive evaluation framework to assess the efficacy of the employed optimization strategies. This meticulous evaluation guarantees the alignment of the optimization process with predefined performance standards, thereby ensuring the stability and reliability of the computer network infrastructure is expressed as

$$sat\_cost(x) = \begin{cases} 1 \\ \frac{(1-b)(cost_{min}-x)}{cost_{opt}} + 1 \\ 0 \end{cases} \tag{3.1}$$

Among the critical factors, cost-min signifies the minimal expense linked to network optimization, while cos topt represents a slightly higher minimum value. Initial computations form the foundation for determining optimal weight values, adhering to reliability constraints, and ensuring a reliable network operation. Data failing to meet stringent constraints undergoes meticulous processing to guarantee the satisfaction of reliability benchmarks.

Utilizing the robust capabilities of the genetic algorithm, the iterative process steers the continuous evolution of the parent data through intricate algorithmic computations. This iterative process facilitates a comprehensive simulation, culminating in a stable data stability curve. This curve represents the culmination of multiple generational iterations, as evidence of the robustness and reliability of the applied optimization strategy, as depicted in Figure 3.2.

The methodology for optimizing the network aims to balance minimizing costs and maximizing reliability through this intricate process. The strategic alignment of these elements plays a crucial role in fostering an optimized network framework that is both cost-effective and resilient, meeting the demanding standards of contemporary network infrastructure.

The optimization direction is precisely determined in the initial calculation phase, ensuring the exclusion of unconstrained solutions during the genetic iteration process. With a focus on stability, various factors, including network information costs, are carefully considered, leading to iterative adjustments at different network nodes. These adjustments are guided by established satisfaction measurement standards, refining the weight calculation through systematic testing of diverse parameter values. This meticulous iterative approach guarantees the creation of a novel structural diagram that adheres to the network's topological constraints.

Significant variations in the optimized structure arise from changes in the parameter W. At the central point, each numerical variable bifurcates into three distinct trajectories, contributing to the formation of diverse structural trees. Leveraging these structural adaptations, the data iteration process is enhanced, resulting in the development of highly satisfactory genetic data following comprehensive reliability analyses. Consequently, addressing the practical challenge of increasing optimization costs, the optimization of computer network communication is successfully achieved within the parameters of the topology structure, thereby establishing a more robust and efficient network architecture [19].

4. 4.Optimization process for computer network reliability.

4.1. Optimization Criteria and Models for Computer Network Reliability. When designing the reliability of a computer, full consideration should be given to establishing a robust and efficient network infrastructure that relies heavily on carefully selecting appropriate and objective network topology models. This selection process, guided by thorough analysis and consideration, holds significant sway over the overall adaptability and performance of the network. By thoroughly assessing the specific requirements and objectives of the network environment, the chosen topology model ensures smooth data transmission, optimal resource allocation, and improved network management capabilities. Furthermore, this model forms the fundamental framework dictating the configuration and interconnection of network components, playing a crucial role in determining the network's fault tolerance, scalability, and overall operational efficiency.

Simultaneously, integrating fault tolerance and redundancy mechanisms into the network architecture is crucial to enhance its resilience and reliability. Incorporating robust redundancy protocols and fault-tolerant design elements allows the network to mitigate the impact of potential disruptions and failures effectively. This proactive design strategy minimizes downtime, data loss, and service interruptions, guaranteeing uninterrupted accessibility and consistent performance. Furthermore, integrating fault tolerance and redundancy measures fosters the development of a robust network infrastructure capable of swiftly adapting to unforeseen challenges, ensuring uninterrupted operations even during adverse events or system failures.

4.2. Network Optimization Design Based on Improved Genetic Algorithm. The optimization of genetic algorithm networks primarily depends on the effective utilization of the genetic algorithm and the critical coding scheme applied within the algorithm. The optimization process of genetic algorithm networks relies heavily on the strategic deployment of genetic algorithms, which excel in navigating complex problem spaces and deriving optimal solutions by emulating the fundamental principles of natural selection and evolution. These algorithms continuously generate increasingly refined solutions over iterative cycles by simulating natural genetic variations, selections, and reproductions. Notably emphasizing adaptability and robustness, genetic algorithms are powerful tools for addressing intricate optimization challenges, especially within computer networks.

Additionally, the efficacy of genetic algorithm network design is intricately linked to the specific coding scheme integrated into the algorithm. This coding scheme dictates how solutions are represented within the algorithm, significantly influencing the efficiency and precision of the optimization process. A delicate balance between solution intricacy and computational efficiency is achieved by meticulously defining the encoding strategy for various components and parameters within the genetic algorithm. The meticulous optimization of the coding scheme ensures that the genetic algorithm effectively explores the solution space and converges toward the most optimal network design configurations. Through meticulous optimization of the coding scheme, genetic algorithm networks can adeptly address complex network optimization challenges, thereby fostering the development of robust and efficient network architectures.

5. Simulation Results. To elaborate on the verification process and the comparison of the algorithm's effectiveness in optimizing computer network reliability design, it is crucial to emphasize the comprehensive

Table 5.1: Node Representation in Binary Gene Format

Node	Binary gene representation
N_1	$g_{11}, g_{12}, ..., g_{1n}$
N_2	$g_{21}, g_{22}, ..., g_{2n}$
...	...
N_m	$g_{m1}, g_{m2}, ..., g_{mn}$

Table 5.2: Total Cost of Network Media Comparison Across Different Algorithms

Techniques	Total cost of network media										
Neural network algorithm [20]	69.1	69.3	70.0	69.7	69.2	69.4	69.9	69.7	69.2	69.2	69.3
Inclusion and Exclusion principle algorithm [21]	68.6	68.9	68.3	68.6	68.4	68.5	68.4	68.3	68.3	68.4	68.0
Fuzzy Neural network algorithm [22]	67.7	67.9	68.2	68.1	68.6	68.6	68.6	68.7	68.6	68.6	69.4
Algorithm of this paper	67.2	67.4	67.4	67.1	67.2	67.3	67.4	67.6	67.3	67.1	67.0

approach adopted in the study. The evaluation is conducted within a standardized computer network reliability model alongside a detailed network link cost model. This standardized approach facilitates a systematic comparison between the improved genetic algorithm and other established algorithms, including the inclusion-exclusion principle algorithm, the fuzzy neural network algorithm, and the neural network algorithm. The study assesses the algorithm's effectiveness through a rigorous comparative analysis and delves into its progressive and practical implications in computer network optimization. Moreover, the experimentation uses a computer system with 32GB of memory, an Intel i7 processor, and the Windows 7 operating system. These specific configurations ensure the consistency and reliability of the experimental setup, enabling an equitable and precise comparison between the various algorithms. The insights derived from the study's findings, as outlined in Table 5.1, are invaluable in understanding the algorithm's performance and potential applicability in real-world network optimization scenarios. Such a comprehensive analysis is instrumental in evaluating the algorithm's competitiveness and prospective contribution to advancing computer network reliability optimization.

The results presented in Figure 5.1 from the simulations highlight a prominent trend: with an increasing number of genetic operations, the overall reliability of the network consistently improves. However, it becomes apparent that as the number of iterations for genetic operations reaches a certain threshold, the rate of improvement in network reliability gradually slows down. Several factors, such as hardware limitations and underlying constraints, contribute to this phenomenon. Eventually, the network reliability culminates at a peak value of 0.894 when the iteration count reaches 100%. Notably, fluctuations in the iteration count correlate with corresponding fluctuations in the resulting value, providing a comprehensive overview of the intricacies impacting network optimization and the challenges within the system.

The Table 5.2 illustrates a comparison of the total cost of network media for several techniques, including the Neural Network Algorithm, the Inclusion and Exclusion Principle Algorithm, the Fuzzy Neural Network Algorithm, and the Algorithm proposed in the current study. The table includes total cost values for multiple instances, highlighting the relative performance of each technique under various scenarios.

This critical stage underscores the necessity of alternative means for network optimization, specifically focusing on continuously reducing the overall cost linked to network chain testing. Analysis of the experimental results reveals an inversely proportional relationship between the fitness and satisfaction functions of the genetic improvement algorithm. Strengthening the efficiency of the fitness function and ensuring diverse cost regression under convergence conditions enables the derivation of the optimal function solution. This approach, centred on convergence and regression, significantly enhances network stability.

Moreover, integrating inclusive and exclusive principles in network calculations is crucial in stimulating network stability. The deliberate application of these principles facilitates the development of a more resilient network structure capable of withstanding diverse operational challenges and maximizing overall reliability.

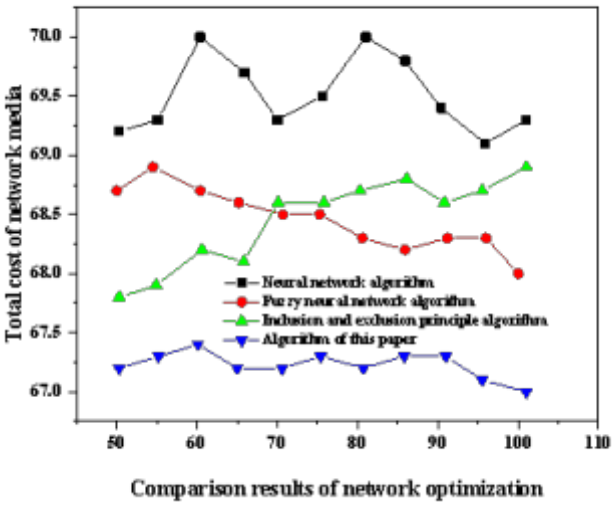


Fig. 5.1: Results of Network Optimization Comparative Analysis

6. Conclusion. The investigation of computer reliability presents a complex problem within computer development. As computers and networks increasingly converge and advance, the development of information network technology has become a fundamental requirement for the progression of network infrastructure as a whole. With an increasing reliance on information technology, computers have assumed a critical role in vital sectors such as the economy and society. Consequently, ensuring computer system reliability has become vital in guaranteeing the continuous operation of crucial sectors. Given the growing dependence on computer technology, it has become imperative to establish a robust development strategy aimed at enhancing computer reliability. By implementing a well-structured development plan, the reliability of computer systems can be substantially strengthened, ensuring the safety and stability of the entire computer infrastructure. As the reliability of computers directly impacts the overall performance and efficiency of the entire system, adopting a comprehensive approach to improving reliability for maintaining operational integrity. Acknowledging the complex nature of the challenge, researchers have incorporated sophisticated methodologies such as genetic algorithms. Researchers strive to optimize network costs by connecting the unique capabilities of genetic algorithms, thereby enhancing the overall efficiency and resilience of the computer network infrastructure. Researchers aim to establish a robust and dependable computer network framework capable of meeting the ever-evolving demands of the contemporary digital landscape by adopting an approach that simultaneously addresses multiple objectives.

REFERENCES

[1] Zhang, Y. , Song, Z. , Yuan, J. , Deng, Z. , & Li, L. . (2021). Path optimization of gluing robot based on improved genetic algorithm. IEEE Access, PP(99), 1-1.
[2] Ai, H. , Zhang, J. , Fan, Y. , & Ghafoor, K. Z. . (2022). Topology optimization of computer communication network based on improved genetic algorithm. Journal of Intelligent Systems, 31(1), 651-659.
[3] Chen, Y. . (2021). Location and path optimization of green cold chain logistics based on improved genetic algorithm from the perspective of low carbon and environmental protection. Fresenius Environmental Bulletin, 30(6), 5961-5973.
[4] Zhang, Q., Yu, H., Li, Z., Zhang, G., & Ma, D. T. (2020). Assessing potential likelihood and impacts of landslides on transportation network vulnerability. Transportation research part D: transport and environment, 82, 102304.
[5] Lu, H., Arora, N., Zhang, H., Lumezanu, C., Rhee, J., & Jiang, G. (2013, December). Hybnet: Network manager for a hybrid network infrastructure. In Proceedings of the Industrial Track of the 13th ACM/IFIP/USENIX International Middleware Conference (pp. 1-6).

[6] Mahmood, M. A., Seah, W. K., & Welch, I. (2015). Reliability in wireless sensor networks: A survey and challenges ahead. Computer networks, 79, 166-187.
[7] Sobb, T., Turnbull, B., & Moustafa, N. (2020). Supply chain 4.0: A survey of cyber security challenges, solutions and future directions. Electronics, 9(11), 1864.
[8] Akhunzada, A., Gani, A., Anuar, N. B., Abdelaziz, A., Khan, M. K., Hayat, A., & Khan, S. U. (2016). Secure and dependable software defined networks. Journal of Network and Computer Applications, 61, 199-221.
[9] De Cicco, L., Mascolo, S., & Niculescu, S. I. (2011). Robust stability analysis of Smith predictor-based congestion control algorithms for computer networks. Automatica, 47(8), 1685-1692.
[10] Wang, Z., Fan, X., & Han, Q. (2013). Global stability of deterministic and stochastic multigroup SEIQR models in computer network. Applied Mathematical Modelling, 37(20-21), 8673-8686.
[11] Tang, C. , Xue, B. , & Wang, L. X. . (2021). Optimization design of shaped charge based on improved genetic algorithm. IOP Conference Series Materials Science and Engineering, 1043(4), 042034.
[12] Alderson, D. L., & Doyle, J. C. (2010). Contrasting views of complexity and their implications for network-centric infrastructures. IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans, 40(4), 839-852.
[13] Liu, T., Wen, W., Jiang, L., Wang, Y., Yang, C., & Quan, G. (2019, June). A fault-tolerant neural network architecture. In Proceedings of the 56th Annual Design Automation Conference 2019 (pp. 1-6).
[14] Piuri, V. (2001). Analysis of fault tolerance in artificial neural networks. Journal of Parallel and Distributed Computing, 61(1), 18-48.
[15] Elyasi-Komari, I., Gorbenko, A., Kharchenko, V. S., & Mamalis, A. (2011). Analysis of Computer Network Reliability and Criticality: Technique and Features. Int. J. Commun. Netw. Syst. Sci., 4(11), 720-726.
[16] Albalawi, F. O. , & Maashi, M. S. . (2021). Selection and optimization of software development life cycles using a genetic algorithm. Intelligent Automation and Soft Computing, 28(1), 39-52.
[17] Zhao, J. . (2021). Meso-model optimization of composite propellant based on hybrid genetic algorithm and mass spring system. Journal of Physics: Conference Series, 2025(1), 012036-.
[18] Zhang, J. , Chen, C. , Si, W. , Chai, X. , Hong, Y. , & Yang, X. , et al. (2021). Research of airfoil optimization based on cst method and genetic algorithm. Journal of Physics: Conference Series, 2006(1), 012062.
[19] Xia, X. , & Wan, D. . (2021). Optimization of one-dimensional wire cutting with variable length based on genetic ant colony algorithm. MATEC Web of Conferences, 336(9), 02011.
[20] Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. Cognitive Systems Research, 52, 317-324.
[21] Lin, K. C., Liao, I. E., Chang, T. P., & Lin, S. F. (2014). A frequent itemset mining algorithm based on the Principle of Inclusion–Exclusion and transaction mapping. Information Sciences, 276, 278-289.
[22] Kuo, R. J., & Zulvia, F. E. (2021). The application of gradient evolution algorithm to an intuitionistic fuzzy neural network for forecasting medical cost of acute hepatitis treatment in Taiwan. Applied Soft Computing, 111, 107711.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 11, 2023
Accepted: Oct 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 103–114, DOI 10.12694/scpe.v25i1.2277

RESEARCH ON INTELLIGENT TRANSFORMATION PLATFORM OF SCIENTIFIC AND TECHNOLOGICAL ACHIEVEMENTS BASED ON TOPIC MODEL ALGORITHM AND ITS APPLICATION

JING WANG*, KAI WANG†, AND YANFEI CHANG‡

Abstract. Being an integral component of the national scientific and technological innovation structure, the conversion of scientific and technological breakthroughs shifts these discoveries from mere theoretical concepts to practical applications. However, this transformation process encounters issues concerning suboptimal efficiency and standards. This research investigates into the LDA theme model to extract pivotal terms and thematic phrases representing scientific and technological advancements. The aim is to ensure precise representation and endorsement, consequently enhancing the efficiency and quality of the conversion process for academic institutions and businesses. Subsequently, this paper establishes a platform for converting scientific and technological breakthroughs while examining the subsystems of information management, retrieval, and recommendations about these breakthroughs. This platform also includes aspects such as transactions and the delivery of achievements. Additionally, an analysis is conducted on the application effects of the scientific and technological achievement (STA) transformation platform, considering the transformation results and the operational evaluation of these scientific and technological innovations.

Key words: STA, LDA Theme Model, Transformation Platform, Information Management, Operational Evaluation

1. Introduction and examples. The capacity for STA innovation represents a essential measure in evaluating a nation's comprehensive strength. The transformation of STA is fundamental in navigating the complex trajectory from theoretical scientific innovations to physical applications in practical production and everyday use. It stands as the bedrock, propelling the engine of progress and paving the way for fostering a society marked by high standards and quality development. This critical process has, over time, become a subject of unwavering attention and inquiry both within national boundaries and across international frontiers, indicating the universal significance attached to the efficient transformation of knowledge into tangible outcomes [8].

The global competition strengthens, nations increasingly recognize the authoritative of development of a robust STA ecosystem, underlining the significance of advancing research and development endeavours, fostering innovation, and seamlessly integrating these advancements into practical applications. The relentless pursuit of efficient conversion mechanisms has sparked a series of comprehensive studies aiming to streamline the transformation of scientific breakthroughs into practical, real-world applications [7].

The studies conducted within the limits of academic institutions or through collaborative international efforts, underscore the critical need to bridge the gap between theoretical knowledge and tangible applications, thus highlighting the pivotal role played by the transformation of STA in encouraging a nation's overall progress and development. By systematically exploring and analyzing the details of this conversion process, researchers aim to undo the original challenges and impairments, the way for formulating comprehensive strategies and robust frameworks designed to strengthen the efficiency and impact of these transformative endeavours. As a result, these groundbreaking advancements' full potential and impact within the academic sphere often remain under-realized, posing a significant setback in the broader quest for seamless knowledge translation and practical application [9].

Recognizing the significance of bridging the gap, it it is necessary to design and implement a robust, agile, and contemporary transformation platform that is equipped to handle the dynamic nature of STA and capable of providing a comprehensive and systematic evaluation framework. This integrated approach would ensure

*Shandong Institute of Innovation and Development, Jinan, Shandong, China (jingwang516@163.com)
†Heze Public Resources Trading Center, He ze, Shandong, China (Corresponding author: kaiwang7@126.com)
‡Jinan Guida Real Estate Co., Ltd., Jinan, Shandong, 250000, China (yanfeichang@126.com)

that the transformative potential of these achievements is fully harnessed, effectively contributing to the broader landscape of STA and, subsequently, catalyzing overall societal progress and development [12].

A pioneering STA management system takes centre stage, built upon the foundation of a cutting-edge cloud table PAAS code-free development platform. Furthermore, implementing this advanced management system not only streamlines the process of organizing and tracking diverse scientific achievements but also empowers researchers and innovators to navigate the complexities of the research and development landscape more effectively. By providing a robust framework for the effective documentation, assessment, and dissemination of STA breakthroughs, this system catalyzes nurturing a culture of innovation and fostering an environment conducive to the accelerated pace of scientific progress [6].

The findings of shed light on the complexities inherent in the agricultural domain, underlining the critical importance of implementing well-informed and strategic transformational frameworks that are not only rooted in technical feasibility but also align with the practical demands and challenges of the agricultural sector. Against these insights, the current landscape of scientific and technological transformations between academic institutions and enterprises reveals a notable predicament characterized by a lack of efficiency and quality. To address this pressing concern, the research initiative culminates developing a sophisticated platform dedicated to the seamless transformation of STA breakthroughs. Leveraging the advanced capabilities of the LDA topic model algorithm, this platform represents a pioneering effort to streamline and optimize the conversion process. Through a meticulous analysis of the platform's transformative efficiency and a comprehensive evaluation of its overall operational performance, the study paves the way for formulating strategic interventions to enhance the efficacy and impact of STA transformations within the academic and industrial fields [2, 4, 13].

2. LDA Topic Model Algorithm. The Latent Dirichlet Allocation (LDA) topic model, also known as LDA, enables the representation of an article's topics as a probability distribution. It facilitates text analysis and topic clustering based on these distributions. The application of the LDA topic model lies in capturing implicit themes within STA accomplishments. Initially, it dissects the content of such achievements into individual words. Subsequently, it determines the number of topics embedded in these achievements and the proportion of each topic. It is achieved through the extraction and induction of topics and the computation of topic intensity. Ultimately, it employs this information to perform tasks such as classification, clustering, recommendation, retrieval, and other pertinent operations related to STA advancements [10].

The LDA topic model is an unsupervised machine-learning technique rooted in the bag-of-words model. It treats every document describing a scientific or technological advancement as a word vector, wherein the occurrence of each word in these achievements is quantified as a mathematical statistical representation. This approach simplifies the description of STA accomplishments into easily quantifiable data. The LDA topic model operates under the assumption that there exist STA intertwined with K topics. Moreover, it posits that all topics adhere to various distributions following the principles of the Dirichlet distribution [1]. The LDA topic model is shown in Figure 2.1 where α and β respectively represent the Dirichlet prior parameters of the multinomial distribution of each STA topic. θ_m represents the topic distribution of the m^{th} STA, ϕ_k represents the word distribution under the k^{th} topic, $Z_{m,n}$ represents the topic of the n^{th} word in m^{th} STA, and $W_{m,n}$ represents the n^{th} word in m^{th} STA.

Directly using STA as input for the LDA topic model is not feasible. It is imperative to preprocess these achievements by conducting tasks such as word segmentation, aiding in identifying and eliminating redundant words and stop words. The word segmentation procedure recognizes multiple words within these achievements and consolidates them into phrases that effectively convey sentence meaning. Following this operation, only the most frequent top 10 words are preserved, while the rest are discarded. Figure 2.2 illustrates the LDA topic model's process to underly semantics of documents pertaining to STA advancements [11].

Among them, the STA document d and the word w are the samples in the STA set M, and the word segmentation set N, and z is the unknown STA topic. Since d and w are observable variables, random STA $P(w_j \mid d_i)$ can be considered to be known, according to a large number of STA document-word segmentation information $P(w_j \mid d_i)$, we can train the STA documenttopic $P(z_k \mid d_i)$, and topic-word segmentation $P(w_j \mid d_k)$, which satisfies Equation (2.1):

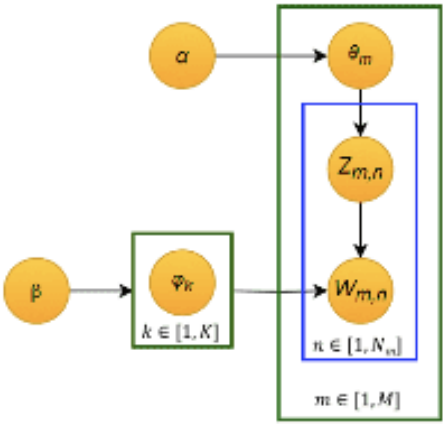


Fig. 2.1: Structure of Latent Dirichlet Allocation (LDA) topic model.

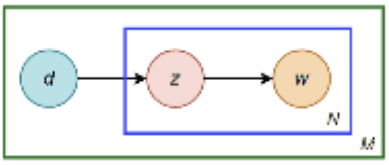


Fig. 2.2: Latent semantic analysis of STA based on LDA model.

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \tag{2.1}$$

The generation probability of each word in the STA document can be expressed as

$$\begin{aligned} P(d_i, w_j) &= \sum_{k=1}^{K} P(d_j) P(w_j \mid d_i) \\ &= P(d_i) \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \end{aligned} \tag{2.2}$$

It can calculate $P(d_i)$ from all the STA, but can't calculate $P(w_j \mid z_k)$ and, $P(z_k \mid d_i)$ so the topic classification of STA can be realized by the expected parameter θ

$$\theta = \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \tag{2.3}$$

The LDA topic model can merge the semantic data of STA, resulting in a proficient word segmentation outcome. It computes the topics of these accomplishments based on the distribution of prominent words and evaluates the significance of each topic across various samples of STA advancements [5].

3. Intelligent Transformation Platform of STA based on LDA Topic Model Algorithm. The exchange of STA innovations between academic institutions and businesses faces numerous challenges. Existing platforms designed to transform these advancements struggle to offer precise recommendations, leading to

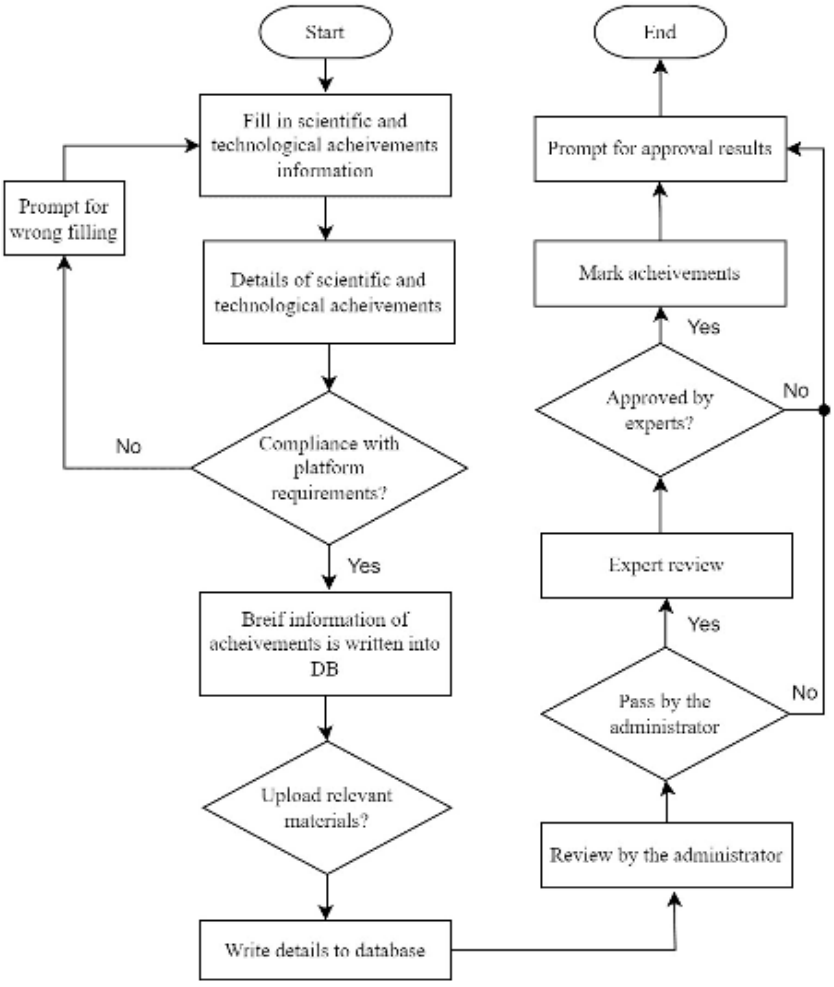


Fig. 3.1: Business process of information management subsystem.

difficulties for enterprises in identifying suitable innovations from a large pool of options. Consequently, the efficiency of this transformation process remains relatively low. To address these concerns, a platform based on the LDA topic model has been developed to facilitate the exchange of STA. The key objectives of this platform, as outlined are as follows [9]:

1. Management of information about the transformation of STA, encompassing tasks such as uploading achievements, data completion, and auditing, ensures comprehensive oversight of the achievement transformation process.
2. Implementing an intelligent retrieval and recommendation system for STA ensures the swift identification of the most compatible options.
3. Facilitating the trade of STA, including processing transaction applications, audits, and inquiries, streamlines the complex process of commercializing these innovations.
4. Facilitation of the transfer and delivery of STA, encompassing the completion of transaction reviews and the payment and delivery processes.

3.1. Information Management Subsystem of STA. The information management subsystem for STA oversees tasks such as uploading, addition, information enhancement, and review of these achievements, with patent achievements being the default. On the new STA page, essential details like the patent name, patent

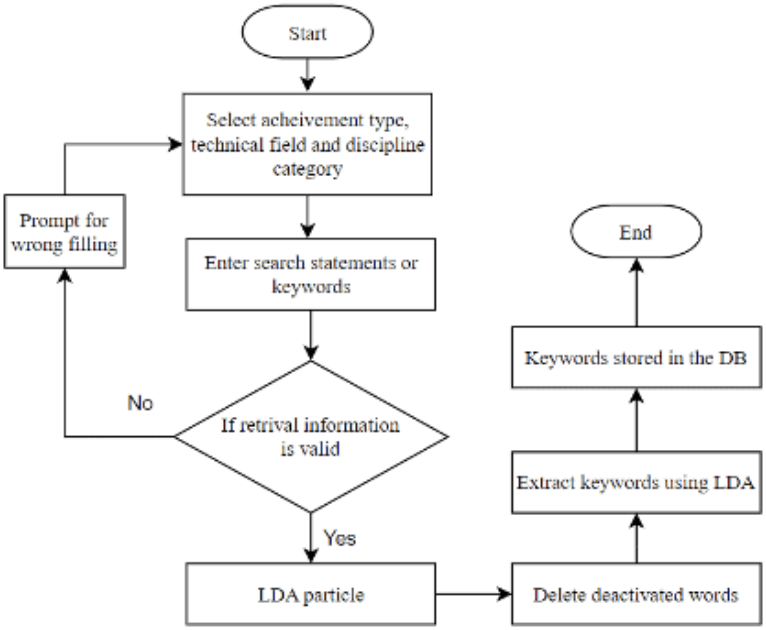


Fig. 3.2: Keyword extraction process of STA.

number, patent authorization date, and effective date must be filled out. Additionally, selecting the STA type and the subject category direction is required. Once the STA information is submitted, the format is assessed to ensure compliance with platform requirements. Subsequently, the information is forwarded to the platform system administrator for an initial audit. Upon approval during the initial audit, it is then directed to experts in relevant professional categories for a comprehensive review. Following the endorsement from the preliminary and substantive reviews, the LDA topic model algorithm is employed to categorize and label the STA. Ultimately, the platform notifies the responsible individual regarding the status of the STA. The business process flow of the STA information management subsystem is illustrated in Figure 3.1.

3.2. Intelligent Retrieval of STA and Recommendation Subsystem. The retrieval and recommendation subsystem within the STA transformation platform allows the seeker to acquire the desired transformation of STA swiftly. Unlike traditional platforms, which solely rely on keyword matching for STA searches and recommendations, this system efficiently retrieves achievements with identical semantics but varying keywords.

The intelligent retrieval process is segmented into two phases to address the challenge in retrieving STA: keyword extraction and keyword matching. During the keyword extraction phase, utilizing the LDA topic model enables the extraction of search keywords. Subsequently, in the keyword matching stage, the system matches these search keywords with those associated with STA. To ensure the usability of the extracted keywords for subsequent keyword matching, it is essential to store the obtained keywords from the extraction phase in the database. Figure 3.2 illustrates the keyword extraction process in retrieving STA.

During the keyword matching stage, the search keywords stored in the database are cross-referenced with the keywords and subject terms relevant to the STA. When a match is found, the search results display the recommendations of STA associated with the theme related to the keywords. Figure 3.3 illustrates the keyword-matching process in retrieving STA.

3.3. Transaction Management Subsystem of STA. The management subsystem for STA transactions oversees the processing and verification of transaction applications. The designated personnel responsible for STA provide essential transaction information on the platform's transformation page, including the name and method of transformation for the STA, the recipient's name for the achievements, and relevant account

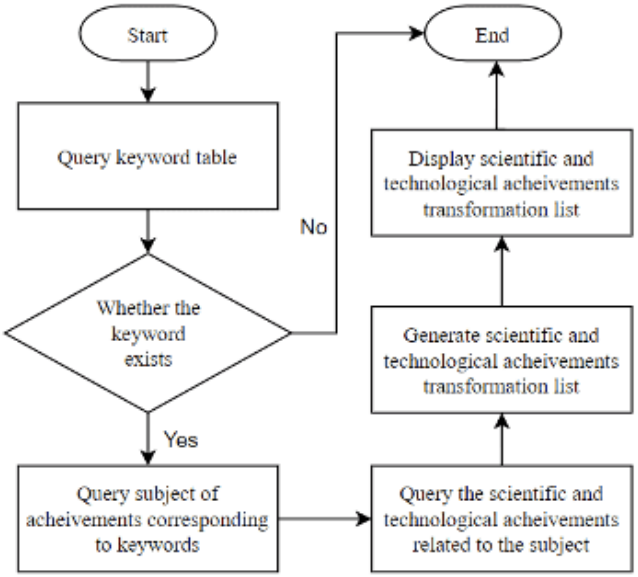


Fig. 3.3: Keyword matching process of retrieval of STA.

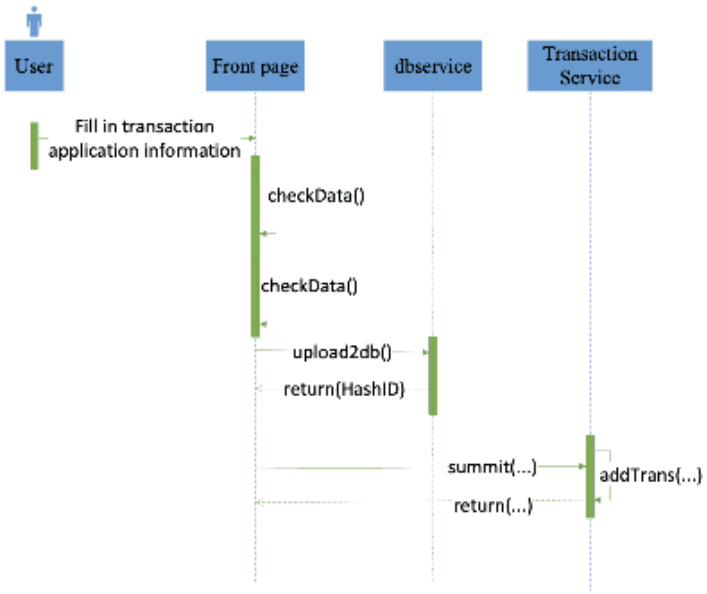


Fig. 3.4: Time sequence transactions of STA.

details. To prevent any instances of fraudulent conversions [15], comprehensive transaction contract materials must be uploaded to the conversion platform. The platform's front end ensures the logical integrity of the transaction information, validating and uploading accurate transaction details and materials to the database. Eventually, the completed parameters are transmitted to the transformation transaction service by the platform, culminating in the successful application process for STA transactions. The time sequence diagram depicting the application of STA is presented in Figure 3.4.

Table 3.1: Main fields of STA transformation.

Name of field	Type of field	Length of field	Description
Id	nvarchar	32	Number of STA
fieldId	int	8	Number of technical areas
title	nvarchar	255	Transformation title of STA
content	text	65535	Main content of STA transformation
postTime	Date	24	Publication date of STA transformation
dealObject	nvarchar	255	Transaction object of STA transformation

During the implementation of the transaction subsystem in the transformation platform, the individual overseeing STA fills out the transaction application information on the interface and submits it. Subsequently, the interface triggers the check_Information() method to validate the format and coherence of the transaction details. Furthermore, it invokes the before_Upload() function to manage the transaction materials before using the upload2db() method to upload the transaction data and materials to the database. The transaction service then uses the add_Trans() function to record the transaction information in the transformation table of STA. Ultimately, the processed data is returned to the interface. A comprehensive overview of the primary fields in the transformation table of STA is provided in Table 3.1.

3.4. Transaction Transfer Delivery Subsystem of STA. The transaction transfer delivery subsystem for STA finalizes the payment and delivery procedures associated with the transformation transactions. The transaction outcomes become visible on the information confirmation page within the STA transformation interface after completing the audit. It provides comprehensive details, including information on the STA, transaction amount, transaction contract, transformation method, and recipient information [14]. Upon reviewing and confirming the information, the transformation platform automatically updates the STA transformation status and secures the transaction amount from the recipient's platform account. Subsequently, following the transfer of ownership of the transactional achievements to the recipient, the locked transaction amount is released into the recipient's account, marking the culmination of the entire STA transformation process. The time sequence diagram depicting the transfer delivery of STA is presented in Figure 3.5.

During the implementation of the transaction transfer delivery subsystem within the transformation platform, the user accesses the STA transaction information through the front-end interface. The front-end triggers the loadFile() method to download the file about the STA transformation from the database and then displays it on the front-end interface. Following the user's confirmation, the front-end interface updates the ownership of the STA and secures the transaction amount in the recipient's account via the confirmTrans() method. Upon receiving and confirming the prompt for the ownership alteration of the returned STA, the front-end interface forwards the information to the backend purchase service. It initiates the transfer of the transaction amount to the recipient of the STA.

4. Transformation Results of STA. The metrics for assessing the transformation of STA encompass the status of the application, transformation revenue, input-output ratio, promotional progress, and funding for promotional projects. An achievement can be classified as a transformed STA accomplishment if it fulfils at least two criteria.

The application status is categorized into five groups [1]: industrial application, small-scale application, trial use, discontinuation after application, and non-application. "Industrial application" denotes the formal integration of the achievements into an industrial field, maintaining a stable application state. "Small-scale application" refers to intermittent, small-scale use post-production. "Trial use" describes the preliminary experimental application of the achievements before full implementation. "Discontinuation after application" signifies the cessation of use due to technical, financial, or strategic reasons. "Non-application" indicates that the achievements have not been practically utilized despite completing the initial transformation stages.

STA achievements in "industrial application" and "small-scale application" are considered completed transformations. Among the achievements on the platform, 303 meet the transformation criteria, accounting for

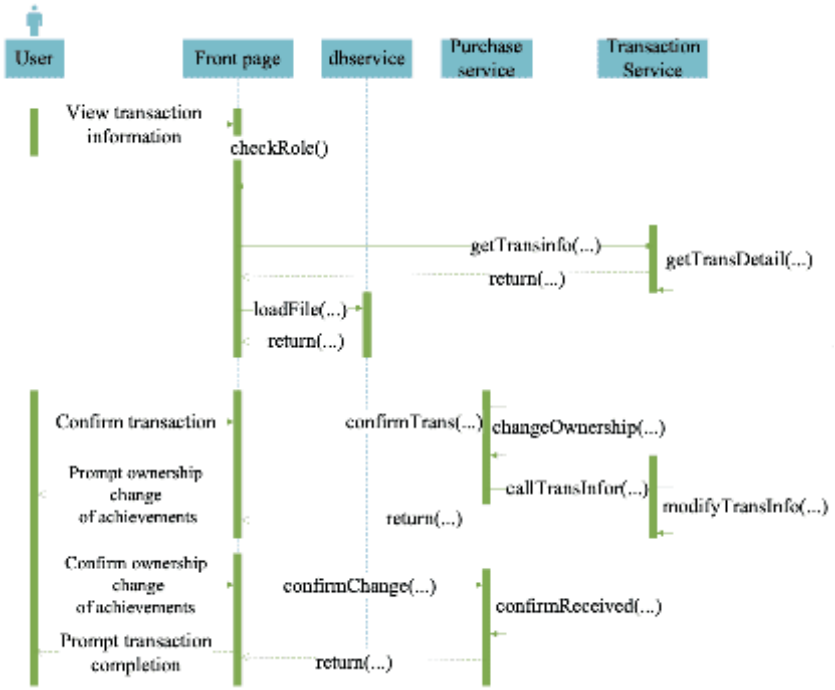


Fig. 3.5: Time sequence transfer delivery of STA.

74.81% of all STA. Specifically, there are 137 achievements in the state of "industrial application," accounting for 45.21% of the transformed achievements and 33.83% of the total achievements. Additionally, 166 achievements fall into the "small-scale application" category, comprising 54.79% of the transformed achievements and 40.99% of the total achievements. Furthermore, there are 48 achievements in the state of "trial use," constituting 15.84% of the transformed achievements and 11.85% of the total achievements. Moreover, 29 achievements are marked as "discontinuation after application," representing 9.57% of the transformed achievements and 7.16% of the total achievements. Finally, 25 achievements are classified as "non-application," making up 8.25% of the transformed achievements and 6.17% of the total achievements.

The promotion index entails a certain level of complexity, as the application of STA varies across different fields. Taking forestry as an example, its promotion involves [3]: (1) establishing a demonstration forest covering an area exceeding 300 acres; (2) cultivating over 5000 seedlings; (3) conducting at least ten training sessions, with a total of over 200 participants; (4) having more than one science and technology demonstration base; (5) operating at least one production line with an annual output value exceeding 1 million yuan.

Among the STA featured on the platform for STA transformation, a total of 196 achievements meet the transformation criteria. The percentage of achievements satisfying the conditions above stands at 44.39%, 28.57%, 14.29%, 7.65%, and 5.10%, respectively. The application and promotion of the transformation of STA are depicted in Figure 4.1.

Based on data from the STA transformation platform, analysis of various R&D subjects revealed that the enterprise sector had the highest transformation rate, reaching 86.25%. Technology promotion agencies followed closely, with a transformation rate of 70.81%. Additionally, research institutes and universities displayed transformation rates of 61.45% and 59.82%, respectively, while other institutions exhibited the lowest rate at 57.29%.

Regarding different fields, the statistics on the transformation of STA in the forest sector indicated that the field of forest-improved varieties boasted the highest transformation rate, reaching 70.24%. Subsequently, the transformation rates for forest management, pest control, forest product pricing, and ecological restoration

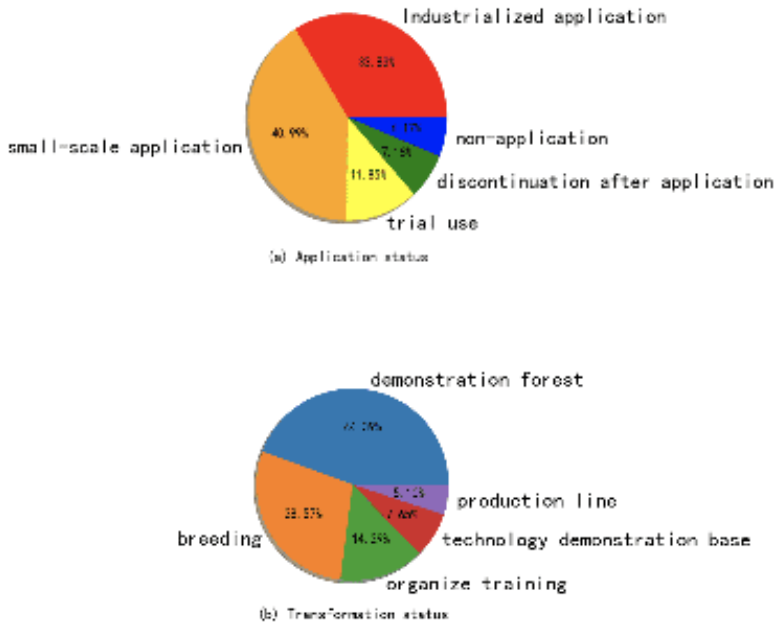


Fig. 4.1: Application and promotion of the transformation of STA.

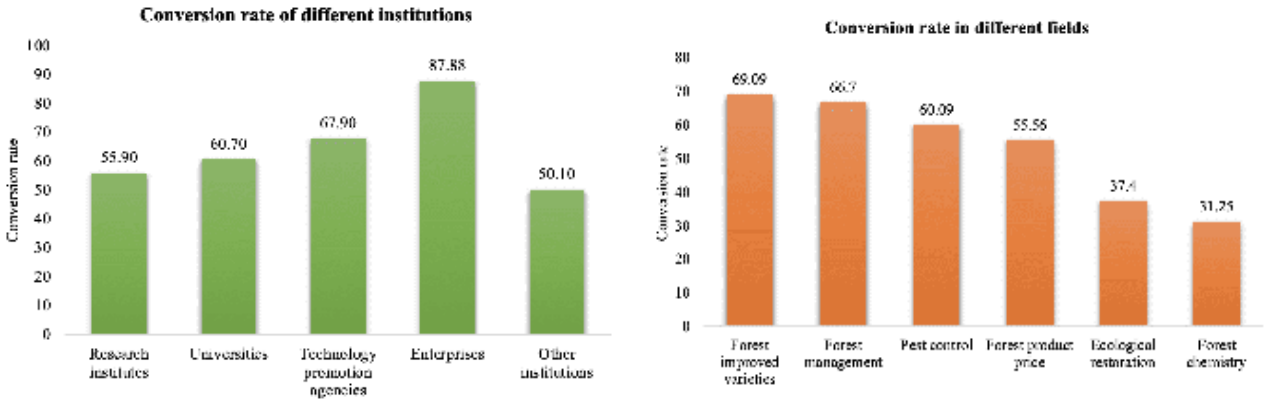


Fig. 4.2: Transformation rate of achievements in different R&D subjects and fields.

stood at 68.35%, 59.27%, 53.46%, and 37.4%, respectively. Notably, the field of forest chemistry exhibited the lowest transformation rate, at a mere 31.25%. Figure 4.2 showcases the transformation rates of achievements across various R&D subjects and fields.

The obstacles preventing the successful transformation of STA can be attributed to internal and external factors. External factors account for 73.53% of the total non-transformed achievements, with 75 STA affected. These primarily include challenges related to limited financing channels and management issues. Insufficient follow-up funding during the transformation process often hinders the progress of achievement transformation.

Additionally, some research and development sectors lack incentivizing mechanisms for converting scientific research achievements. Consequently, the efforts of achievement transformation personnel fail to yield commensurate returns, resulting in a decreased drive to convert research accomplishments.

Internal factors account for the remaining 26.47% of the non-transformed achievements, with 27 STA

Table 4.1: Weights of each index after standardization.

Index	2017	2018	2019	2020	2021
Income of achievements transfer	0.0259	0.0552	0.2156	0.2356	0.4677
Revenue from the sales of new products by the recipient enterprises	0.0031	0.0898	0.1745	0.229	0.5037
Income generated from the export sales of new products by the recipient enterprises	0.0724	0.0734	0.1043	0.2745	0.4754
Productivity of new products of transferee enterprises	0.0697	0.1063	0.14	0.2162	0.4679

impacted. This group includes instances where achievements fail to reach maturity or align with market demands, rendering them replaceable by newer innovations. Some achievements, originating from primary and pilot stages of research and development, may not meet the requirements of practical production and application. Additionally, although ahead of their time, certain innovations struggle to find suitable transformation prospects due to a misalignment with actual market demands.

4.1. Evaluation on Operation of Intelligent Transformation Platform for STA. Based on the preceding analysis of the scientific research achievements transformation platform, and guided by the principles of scientific rigor and practicality, evaluation indicators for assessing the operational performance of the platform have been identified. Drawing from pertinent research metrics [16], the chosen evaluation indicators encompass the following: achievement transfer revenue, sales revenue from the transferee enterprises' new products, export sales revenue from the transferee enterprises' new products, and the productivity associated with the transferee enterprises' new products.

First of all, it is necessary to standardize the data, which is as shown in Equation (4.1):

$$X'_{ij} = \frac{X_{ij} - \operatorname{Min}(X_{ij})}{\operatorname{Max}(X_{ij}) - \operatorname{Min}(X_{ij})} \tag{4.1}$$

wher, X_{ij} is the item j value of the i^{th} index, $\operatorname{Max}(X_{ij})$ is the maximum value of the item j value of the i^{th} index, $\operatorname{Min}(X_{ij})$ is the minimum value of the item j value of the i^{th} index, and X'_{ij} is the normalized value of X_{ij}.

Then the weight of each index is calculated by the entropy method, which is expressed in Equation (4.2):

$$H(X) = -\sum_{x \in X} P(x) \log(P(x)) \tag{4.2}$$

The weight of each index in the past five years after standardization is obtained, which is shown in Table 4.1.

The transformation ability of STA transformation platform is calculated according to Equation (4.3):

$$U = \sum W_{ij} X'_{ij} \tag{4.3}$$

where U is the comprehensive evaluation score, W_{ij} is the weight of item j of the i^{th} index, and $X'_{ij}$is the standard value of item j of the i^{th} index. The score of the platform for each year, as shown in Figure 4.3.

Figure 4.3 showcases a consistent upward trajectory in the score of the STA transformation platform. Despite a slight decline in 2019, the platform's transformative capacity has steadily grown.

5. Conclusion. The LDA topic model is conducted to address the substandard efficiency and quality during the transformation process of STA. In this paper, an intelligent transformation platform for STA, is developed. The STA information management subsystem is established to facilitate the seamless uploading, addition, refinement, and auditing of STA data. The intelligent retrieval and recommendation subsystem ensures swift access to potential users' desired STA. Moreover, the transaction management subsystem oversees the application and scrutiny of STA transactions. The transaction transfer delivery subsystem manages these

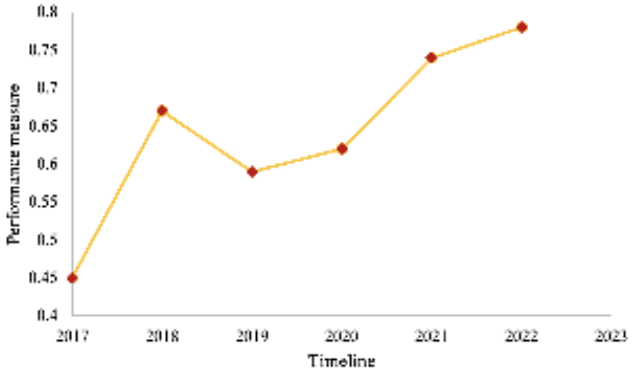


Fig. 4.3: Scores of transformation ability.

achievements' payment and delivery procedures. The analysis of the STA transformation platform is conducted from two perspectives, focusing on the transformation's outcomes and the operational evaluation of the platform itself. Taking a comprehensive view, both the transformation rate of STA and the overall rating of the transformation platform underscored the platform's capability to enhance the efficiency and quality of the transformation process significantly. The noticeable impact demonstrates the platform's effective application in streamlining the transformation of STA.

Funding Information. Key R&D Program of Shandong Province; Project Name: Evaluation of Scientific and Technological Achievements Transformation Ability of Scientific Research Institutes in Shandong Province and Research on Countermeasures to Promote the Transformation of Scientific and Technological Achievements; Project No.: 2021RKY01020.

REFERENCES

[1] S. Aziz, M. Dowling, H. Hammami, and A. Piepenbrink, *Machine learning in finance: A topic modeling approach*, European Financial Management, 28 (2022), pp. 744–770.

[2] T. Bhowmik, N. Niu, J. Savolainen, and A. Mahmoud, *Leveraging topic modeling and part-of-speech tagging to support combinational creativity in requirements engineering*, Requirements Engineering, 20 (2015), pp. 253–280.

[3] R. Congjing, S. Kai, et al., *The construction of the university transferability patent recognition model: A case analysis of artificial intelligence*, Information Studies: Theory & Application, 43 (2020), pp. 79–85.

[4] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, *Prediction of research trends using lda based topic modeling*, Global Transitions Proceedings, 3 (2022), pp. 298–304.

[5] J. Hoblos, *Experimenting with latent semantic analysis and latent dirichlet allocation on automated essay grading*, in Proceedings of the 7th International Conference on Social Networks Analysis, Management and Security, Paris, France, 2020, IEEE, pp. 1–7.

[6] M. R. Jennings, C. Turner, R. R. Bond, A. Kennedy, R. Thantilage, M. T. Kechadi, N.-A. Le-Khac, J. McLaughlin, and D. D. Finlay, *Code-free cloud computing service to facilitate rapid biomedical digital signal processing and algorithm development*, Computer Methods and Programs in Biomedicine, 211 (2021), p. 106398.

[7] E. K. Karpunina, G. K. Lapushinskaya, A. E. Arutyunova, S. V. Lupacheva, and A. A. Dubovitski, *Dialectics of sustainable development of digital economy ecosystem*, in Scientific and Technical Revolution: Yesterday, Today and Tomorrow, vol. 129, Springer, 2020, pp. 486–496.

[8] V. L. Kvint and V. V. Okrepilov, *Quality of life and values in national development strategies*, Herald of the Russian Academy of Sciences, 84 (2014), pp. 188–200.

[9] W. Li and P. Zhang, *Developing the transformation of scientific and technological achievements in colleges and universities to boost the development of low-carbon economy*, International Journal of Low-Carbon Technologies, 16 (2021), pp. 305–316.

[10] F. F. Lubis, Y. Rosmansyah, and S. H. Supangkat, *Topic discovery of online course reviews using lda with leveraging reviews helpfulness*, International Journal of Electrical and Computer Engineering, 9 (2019), p. 426.

[11] S. Qomariyah, N. Iriawan, and K. Fithriasari, *Topic modeling twitter data using latent dirichlet allocation and latent semantic analysis*, in Proceedings of the 2nd International Conference on Science, Mathematics, Environment, and Education, vol. 2194, Surakarta, Indonesia, 2019, AIP Publishing.

[12] J. Song, Y. He, X. Sun, and D. Huang, *Research on innovation evaluation of scientific and technological achievements in colleges and universities in the new era*, in Proceedings of the 22nd International Conference on Computer and Information Science, Zhuhai, China, 2022, IEEE, pp. 176–181.
[13] A. Suominen and H. Toivanen, *Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification*, Journal of the Association for Information Science and Technology, 67 (2016), pp. 2464–2476.
[14] J. Wang, Y. Li, B. Wu, and Y. Wang, *Tourism destination image based on tourism user generated content on internet*, Tourism Review, 76 (2021), pp. 125–137.
[15] M. Wu, R. Long, F. Chen, H. Chen, Y. Bai, K. Cheng, and H. Huang, *Spatio-temporal difference analysis in climate change topics and sentiment orientation: Based on lda and bilstm model*, Resources, Conservation and Recycling, 188 (2023), p. 106697.
[16] D. Yao, J. He, W. Yang, and M. Zhang, *A study on the evaluation index system of innovation and entrepreneurship education for undergraduate students majoring in interdisciplinary arts and sciences.*, Higher Education Studies, 12 (2022), pp. 135–145.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 13, 2023
Accepted: Oct 22, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 115–126, DOI 10.12694/scpe.v25i1.2279

AN EMOTIONAL ANALYSIS OF KOREAN TOPICS BASED ON SOCIAL MEDIA BIG DATA CLUSTERING

YANHONG JIN*

Abstract. An innovative approach is introduced in this paper to address the challenges in emotional topic interpretation and accuracy in emotional situation assessment. Utilizing large data from social media to improve the accuracy of emotional analysis in online debates, with a specific emphasis on Korean themes. The proposed solution, the Online Topic Emotion Recognition Model (OTSRM), builds upon the foundational Online Latent Dirichlet Allocation (OLDA) model. The OTSRM integrates the concept of emotion intensity and introduces an inventive emotion iteration framework to tackle these issues. Key innovations of the OTSRM include establishing an affective evolution channel by augmenting affective heritability using a β priori. Additionally, the model generates two critical distribution matrices: one for characteristic words and another for affective words, facilitating a deeper understanding of emotional context within topics. The relative entropy method is employed to discern emotional tones in textual content, calculating maximum emotion values for topic focus within adjacent time segments. Validation experiments using five diverse network event datasets and comparisons to mainstream models demonstrate the OTSRM's effectiveness with emotion recognition accuracy rates of 85.56% and 81.03%. The OTSRM represents significant progress in addressing challenges associated with emotional topic analysis and precise emotional dynamics assessment in Korean social media data.

Key words: Emotional Analysis, Social Media, Topic Emotion Recognition, Affective Evolution, Online Latent Dirichlet Allocation, Public Opinion Analysis

AMS subject classifications. 15A15, 15A09, 15A23

1. Introduction.

The Internet has gained widespread popularity recently, and social media has progressively assumed a significant role in people's daily lives. The real-time recording of users' online activities in the digital world has accumulated a wealth of user behaviour data within natural contexts. This growing data source has accompanied a fresh paradigm and investigative avenue for research. Social media platforms, owing to their diversity of information and complex functionalities, offer researchers many data types to explore. During the investigation, individuals often delve deeply into specific data categories for analysis and practical application. These social media data can be categorized based on their recording forms, encompassing personal account details, usage patterns, textual content, social network interactions, visual content, and other relevant cues [16].

The volume of social media data centered on Korean topics is unparalleled, boasting various informational formats. Consequently, connecting the potential of Korean topic social media big data necessitates comprehensive data collection platforms and a broad-ranging approach to data acquisition. Building upon existing social media intelligence sources, expanding the intelligence sources as extensively as possible is imperative. Concurrently, there's a pressing need to fortify the capabilities of existing data mining technologies, ensuring that recognition techniques transcend the confines of plain text and extend into the domains of images, audio, and video. This evolution aims to enable timely and exhaustive intelligence gathering about Korean topics within social media [9].

In the era of big data, the digital world is overflowing with a creative expression of interactive content driven by public sentiment, giving rise to a multifaceted tapestry of opinion evolution patterns. This torrent of digital discourse encompasses a kaleidoscope of perspectives, reactions, and dialogues, mirroring the dynamic and ever-shifting nature of public sentiment within the online realm. Particularly noteworthy are categories of public sentiment data, such as breaking news and trending events, which wield considerable influence over

*School of Foreign Studies, Lingnan Normal University, Zhanjiang, Guangdong, 524048 China (Corresponding author: yanhongjin3@163.com)

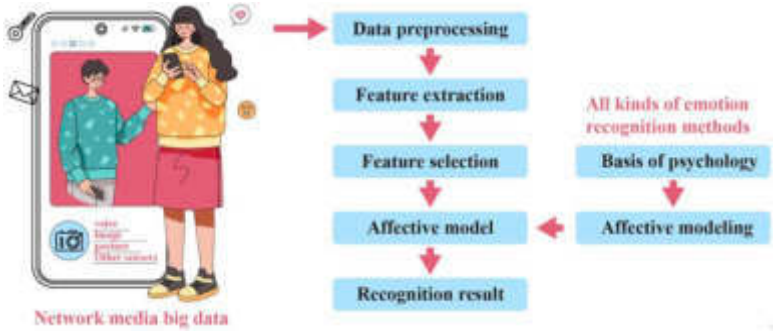


Fig. 1.1: Social media big data clustering

the trajectory of public opinion. These influential occurrences can frequently provoke polarized sentiments, inciting conspicuous shifts in the digital landscape toward either a positive or negative direction. These shifts sometimes escalate into full-fledged online public opinion crises characterized by fervent debates, misinformation proliferation, and emotionally charged exchanges [20].

The far-reaching significance of efforts transcends the confines of data analysis. They serve as vital for establishing effective early warning systems attuned to the fluctuations of public sentiment. By remaining vigilant to the evolving emotional topography of online discourse, it becomes feasible to discern the telltale signs of impending public sentiment crises and proactively implement measures to mitigate their impacts. Ultimately, the overarching goal is to nurture a harmonious and constructive online public opinion environment that fosters healthy debates, well-informed discussions, and positive interactions among netizens, as illustrated in Figure 1.1 [4].

Within this complex landscape, the imperative to mine and analyze online public opinion topics takes center stage. The process involves identifying and comprehensively understanding pivotal themes and subjects dominating online discussions. Such endeavors provide invaluable insights into the intricate dynamics of public sentiment. Equally important is assessing netizens' emotional polarity and evolution—those actively participating in online discourse. A profound comprehension of the emotional tenor of these discussions proves instrumental in accurately appraising public sentiment and prognosticating potential shifts and emerging trends.

2. Literature Review. Topic Detection and Tracking (TDT) initially emerged to uncover latent topics and monitor their developmental trajectories. The TDT model leverages mathematical statistics to condense the dimensionality of text data. However, it grapples with a limitation—its inability to harness the temporal dimension inherent in the original corpus fully. This temporal aspect is crucial for clarifying the semantic evolution of text topics over time [1].

The Online Latent Dirichlet Allocation (OLDA) model introduces a novel perspective. It posits that topic distributions exhibit both inheritance and continuity. Under this framework, the topic distribution within a historical event window serves as prior knowledge for understanding the topic's state in the current time slice. This dynamic approach allows for the online tracking of topic evolution. While the OLDA model successfully addresses the challenge of online topic modelling when new data is introduced, it grapples with a distinct issue—the redundancy stemming from intermingling old and new topics [6].

The consequence of the topic distribution redundancy is reducing the precision of topic detection and evolution analysis. It introduces a level of noise that can obfuscate the accurate identification and tracking of evolving topics within a corpus. Consequently, there is a need for more sophisticated models that can effectively disentangle the intricate interplay of old and new topics, thereby enhancing the accuracy and fidelity of topic detection and evolution analysis. A comprehensive exploration of auditory and visual environmental characteristics is embarked on a data-driven investigation encompassing 17 historical cities and towns across China. Their pioneering study harnessed posts featuring soundscapes-related keywords and street-view photographs from a prominent Chinese social media platform. The research unveiled intriguing insights into the acoustic

ambience of historic districts, revealing a symphony of artificial sounds generated by folk activities and street vendors intermingled with the soothing backdrop of natural sounds—ranging from the gentle flow of water to the melodic chorus of birdsong [19].

A novel and robust streaming media clustering algorithm is introduced, rooted in multi-edge computing, tailored explicitly for detecting streaming media traffic events. Their innovative approach incorporated domain-specific traffic-related knowledge and information. They extracted diverse elements from social media textual content to construct a heterogeneous information network (HIN) centered on traffic events. By leveraging meta-path weight calculations, the research team quantified event similarity across social media texts, offering a data-driven solution for identifying and analyzing traffic-related incidents [8].

The authors ventured into automatic emotion recognition, proposing an approach considering group types and emotional models. Their investigation delved into the facets of general datasets applicable to various emotion patterns, the overarching methodologies employed, and the reported performance metrics. The paper not only elucidated these critical properties but also analyzed the potential applications and implications of the methods discussed, charting new avenues for emotion recognition research and implementation [17].

In topic detection, the integration of emotion analysis represents a crucial dimension for unearthing hidden emotional fluctuations within topics. This interdisciplinary exploration has garnered significant attention from scholars, leading to profound advancements in two principal avenues: Firstly, researchers have sought to enhance topic detection by incorporating sentiment analysis parameters into the Latent Dirichlet Allocation (LDA) model. This approach marries the realms of topic identification and sentiment analysis, giving rise to models such as the Joint Sentiment Model (JST), Weakly Supervised Sentiment-topic Model (WSSM), and Subtopic Sentiment Combining Model (SSCM). These models share a common objective: constructing hybrid models blending topics and emotions effortlessly. By extending text-dependent topic mining to topic-dependent topic-emotion mining, these models enable the analysis of emotional evolution within topics via calculations of topic-emotion similarity. However, a drawback of these models lies in their reliance on static topic parameters, which are rooted in subjective judgments and lack dynamic adjustment capabilities, limiting their ability to track the dynamic evolution of emotions [21].

A different approach involves the integration of emotional parameters into the Online Latent Dirichlet Allocation (OLDA) model. Here, the emotional posterior of a previous time slice (t-1) is employed as the emotional prior for the subsequent time slice (t), allowing for the dynamic construction of emotional distributions across different time segments. Models such as the Time-based Subtopic and Sentiment-topic Combining Model (TSSCM) and Joint Multi-grain Topic Sentiment Model (JMTS) have emerged from this paradigm. While these models excel at dynamic topic identification, they overlook the influence of affective genetic strength in different time segments on the distribution of topic-related emotions. Furthermore, they struggle to effectively address the challenge posed by the iterative nature of emotions stemming from the interplay of old and new topics [2, 7].

In addressing the challenges above, the authors present a novel approach building upon the OLDA model, wherein emotion intensity is introduced as a pivotal element. Furthermore, it introduces the concept of emotion iteration and formulates an innovative online topic emotion recognition model. This model leverages Bayesian techniques to dynamically determine the number of topics while also harnessing the heritability of the emotion intensity operator to construct a topic-emotion distribution that unfolds over time. Through the creation of an emotion evolution channel, it adeptly discerns and tracks topic sentiment trends across diverse textual content [3, 18].

3. Research Methods.

3.1. Latent dirichlet allocation (LDA) model. LDA model constructs a three-layer Bayesian probability network by introducing Dirichlet prior parameters, an unsupervised learning model that can generate text-hidden topics. The model describes the three-layer structure relationship of document, topic and word: multiple topics have probability dependence in each document, and each topic is collected in a certain inscription according to probability. Therefore, a topic is a multinomial distribution on a thesaurus. The mixture of multiple topics constitutes the document, and the mixture of multiple feature words constitutes the topic. Assume that document d contains M documents, K topics, and V vocabulary sets. The user needs to set parameters α and β. By constructing document-topic distribution $\theta_m \left(\theta = \left\{{\theta_m}^d \mid d \in D\right\}\right)$ and topic-word distribution $\varphi_k \left(\varphi = \{\varphi_k \mid k \in [1, k]\}\right)$, the nth word $w_{m,n}$ of the mth document in the document library is obtained. The

core process of the LDA model is described as follows: (1) Set parameter α and generate document-topic distribution θ_m, namely, $\theta_m \sim \mathrm{Dir}(\alpha)$ by sampling; (2) Generate topic $Z_{m,n}$, the nth word of document m by sampling from θ_m distribution, namely, $Z_{m,n} \sim \mathrm{Mult}\ (\theta_m)$ (3). Set parameter β and generate topic $Z_{m,n}$ and its corresponding topic-word distribution φ_k by sampling, namely $\varphi_k \sim \mathrm{Dir}(\beta)$; (4) Generate the nth word $w_{m,n}$ of document m by sampling from φ_k distribution, namely, $w_{m,n} \sim \mathrm{Mult}\,(\varphi_{m,n})$.

3.2. Online latent dirichlet allocation (OLDA) model. OLDA model introduces time granularity based on the LDA model to detect the difference and continuity of online topics. By dynamically adjusting the size of the time slice, the coarse-grained requirements of the users in the time dimension of corpus analysis can be met. When processing streaming text data, the model considers that the prior and posterior probability maintains the continuity between topics, that is, the posterior weight of the word distribution φ_{t-1} in time slice $t-1$ is the prior of the word distribution φ_t in time slice t, and the Dirichlet before the word distribution φ_{t-1} in time slice t satisfies Equation (3.1) as,

$$\mathrm{Multi}\left(\varphi_k^t\right) \sim \mathrm{Dir}\left(\beta_k^t\right) \sim \mathrm{Dir}\left(\eta^\delta B_k^{t-1}\right) \tag{3.1}$$

η^δ is the heritability of the topic-word distribution on the first δ time slices. The occurrence times of words about a certain topic in the first δ time window $\{t-\delta-1, t-1\}$ are counted, and the topic-word evolution transition matrix B^{t-1}, is constructed based on this. Similarly, the document-topic distribution 0 also satisfies the similar property and the document-topic evolution transition matrix A^{t-1}, is constructed. The incremental Gibbs sampling algorithm is used in the OLDA model to approximate the over parameters α and β. Its core idea is to obtain the sample space closest to the probability distribution value by constructing a Markov chain with the probability of harvest. This algorithm considers the information of t time slice as only related to $t-1$ time slice and ignores the influence of other time slices on the corpus information. In other words, the topic-word distribution is regarded as a Markov-like chain model, and the posterior probability calculation formula of the time t slice is given by

$$P_t\left(z_i = j \mid z_{-i}, w_i\right) = \frac{\left(n_{-i,j}^{(w_i)}\right)^t + w\left(n_{-i,j}^{(w_i)}\right)^{t-1} + \beta}{\left(n_{-i,j}^{(*)}\right)^t + w\left(n_{-i,j}^{(*)}\right)^{t-1} + V\beta} \cdot \frac{\left(n_{-i,j}^{(d_i)}\right)^t + \alpha}{\left(n_{-i,*}^{(d_i)}\right)^t + K\alpha} \tag{3.2}$$

$\left(n_{-i,j}^{(d_i)}\right)^t$ represents the number of words assigned to topic j by word collection $\{V-i\}$ in the document d_i; $\left(n_{-i,*}^{(d_i)}\right)^t$ represents the total number of times the document d_i is assigned after excluding the word i; $w\left(n_{-i,j}^{(w_i)}\right)^{t-1}$ represents the number of words assigned to topic j in the $t-1$ time slice and is the same as w_i; $w\left(n_{-i,j}^{(*)}\right)^{t-1}$ represents the number of words assigned to topic j in the $t-1$ time slice. w represents heritability; V and K, represent the number of words in the corpus and the number of set topics; α and β are Dirichlet prior parameters, and the Equations give calculations.

$$\alpha_m^t = \theta_m^{t-1} \cdot A_m^{t-1} \tag{3.3}$$

$$\beta_k^t = \varphi_k^{t-1} \cdot B_k^{t-1} \tag{3.4}$$

A_m^{t-1} represents a $K \times |D^t|$-dimensional matrix arranged by θ_m^{t-1} columns of document-topic distribution on time slice $t-1$; similarly, B_k^{t-1} represents a $|V^n| \times KA_m^{t-1}$-dimensional matrix arranged by φ_k^{t-1} columns of topic-word distribution on time slice $t-1$.

3.3. Topic emotion modelling. In emotion analysis, the typical procedure involves the extraction of emotion-inducing keywords from the text, facilitated by using an emotion dictionary, followed by calculating emotion polarity. This process can generally be divided into two key stages: First, establishing a comprehensive set of emotional words is undertaken. Second, semantic proximity between emotion feature words is assessed, often employing techniques such as Similarity Calculation or Bootstrapping algorithms to derive an emotion semantic model [13, 14].

Mutual Information (MI) is widely used to calculate the semantic distance similarity between item pairs, (i_1, i_2). Generally, the emotional value of emotional feature words can be obtained by calculating its feature weight value, and the assignment of feature weight depends on the similarity distance between the combination of item vocabularies. In the index stage, all the emotion words in the text set are obtained by calculating the initial frequency value of the mixed document. The time complexity is $O\left(\left|d^2\right|\right)$ (where d is the document size) and the space complexity is $O\left(\left|T^2\right|\right)$ (where T is the number of items). The mutual information calculation is expressed as,

$$\mathrm{MI}\left(t_i, t_j\right) = \sum_{t_i \in \{0,1 | t_j \in \{0,1\}\}} \sum_{p} \log\left[\frac{p\left(t_i, t_j\right)}{p\left(t_i\right) p\left(t_j\right)}\right] \times p\left(t_i, t_j\right) \tag{3.5}$$

where $p\left(t_i, t_j\right)$ represents the probability of t_i and t_j co-appearing in the same document; $p\left(t_i\right)$ and $p\left(t_j\right)$ represent the probability that the document contains t_i and t_j, respectively. The above probability can be obtained from the initial corpus using maximum likelihood estimation.

The emotional polarity in the emotional dictionary has three kinds: positive, negative, and neutral. The classification of affective tendency generally sets an affective threshold and compares the mutual information of the calculated new item i_1 with that of the known effective item i_2. If $\mathrm{MI}\left(i_1, i_2\right)$ is less than the threshold, (i_1, i_2) is put into different affective tendencies [12, 11].

Emotional dictionaries often contain emotional, turning conjunctions and negative words. It is necessary to use the conditional clause after the turning conjunctions to replace the whole sentence before calculating the feature weight value of new words (emotional and negative words). The multi-feature linear fusion method calculates the comprehensive affective value, effectively avoiding the uncertainty of affective inclination. The calculation formula is:

$$\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) = \frac{\sum_{i_{\mathrm{pos}} \in \{0,m\} i_{\mathrm{pos}} \in I_{\mathrm{pos}}} \mathrm{MI}\left(i_{\mathrm{new}}, i_{\mathrm{pos}}\right) \cdot \mathrm{neg}}{\left|I_{\mathrm{pos}}\right|} + \frac{\sum_{i_{\mathrm{agg}} \in \{0,n\} i_{\mathrm{neg}} \in I_{\mathrm{neg}}} \mathrm{MI}\left(i_{\mathrm{new}}, i_{\mathrm{neg}}\right) \cdot \mathrm{neg}}{\left|I_{\mathrm{neg}}\right|} \tag{3.6}$$

where i_{new} represents a new word; i_{pos} and i_{neg} respectively represent the classified positive and negative emotion words in the old words. neg is the negative sign of the sentence where the nord is found. When $neg = 1$, it indicates that there is no negative word on the right of the new word, and the affective tendency is consistent with the affective feature words in front. When $neg = -1$, it indicates negative words on the right side of new words, and the affective tendency is opposite to the affective feature words in front. I_{pos} and I_{neg} represent the positive emotion word set and negative emotion word set of the emotion word list, respectively; m and n represent the number of words in positive and negative emotion word sets, respectively. When $\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) > 0$, it indicates that i_{new} has a positive tendency and is added to the positive emotion word set. When $\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) < 0$, it indicates that there is a negative tendency, and a negative emotion word set is added. When $\mathrm{SO}_{\mathrm{neg}}\left(i_{\mathrm{new}}\right) = 0$, there is no emotional inclination [10, 5].

3.4. OTSRM model description. The OTSRM model is a three-layer Bayesian network. Let the focus word N^x of t time slice contains K^t feature words and M^t emotion words. Firstly, the model obtains β priori of $t-1$ time slice according to the focus word distribution θ_m^{t-1} of $t-1$ time slice. According to the focus words obtained, the feature word $w_{m,n}^{t-1}$ is selected from feature word distribution $\emptyset_k^{t-1}$ then emotion word $S_{m,n}^{t-1}$ is extracted based on emotion word distribution μ_k^{t-1} and emotion intensity is calculated. Finally, topic emotion words with maximum emotion value are obtained through topic emotion calculation.

3.5. Calculation of emotional intensity. The OLDA model maintains the continuity between topics through topic heritability w. It is assumed that the variable distribution of t-time slice is only affected by $t-1$ time slice and has nothing to do with the text information in the previous time slice so that it can be regarded as a topic inheritance. OTSRM model refers to the property of topic inheritance and introduces the emotion iteration thought into emotion analysis. The topics-emotion distribution μ_m^t of t-time slice is regarded as the posterior of μ_m^t in $t-1$ time slice. By calculating the heritability λ of emotion intensity of $t-1$ time slice, the topic emotion intensity of t-time slice is obtained.

Emotional intensity is similar to topic intensity, which can dynamically measure the stability of topic emotion in the time dimension. Documents describing a topic have high probability distribution values on a

few topics, and sentiment words describing a topic will also show relatively high probability distribution values on one or a few topics. Similarly, suppose the probability distribution of a topic in each emotion word is relatively average. In that case, it can be determined that the emotion expressed by the document is relatively balanced and has no clear emotional tendency. In this article, Shannon's information entropy is used to represent the emotional concentration degree of the topic. The normalized topic of emotional weight w_m^t is calculated using Equations (3.7) and (3.8):

$$E\left(z_m^t\right) = -\sum_{m=1}^{M}\sum_{k=1}^{K}\mu_{m,k}^{t-1}\log_2\mu_{m,k}^{t-1} \tag{3.7}$$

$$W_m^t = 1 - \frac{E\left(z_m^t\right) - \min\left\{E\left(z_1^t\right),\cdots,E\left(z_M^t\right)\right\}}{\max\left\{E\left(z_1^t\right),\cdots,E\left(z_M^t\right)\right\} - \min\left\{E\left(z_1^t\right),\cdots,E\left(z_M^t\right)\right\}} \tag{3.8}$$

$\mu_{m,k}^{t-1}$ represents the emotion distribution of document m under the k th topic under the $t-1$ time slice. When the topic belongs to only one emotion word, the emotion information quotient under the topic of the document is 0 , and the emotion weight W_m^t is 1. In particular, when W_m^t is 0, the topic is evenly distributed on K emotions, then the corresponding affective information entropy is maximum, indicating that the affective distribution of this topic is relatively broad and does not contribute to the affective of the topic. The Formula for calculating the hyperparameter γ_m^t of topic emotion distribution in time slice t is:

$$\gamma_m^t = \lambda_m^{t-1}\cdot R_m^{t-1} \tag{3.9}$$

where R_m^{t-1} represents the topic emotion matrix in $t-1$ time slice; λ_m^{t-1} represents the heritability of topic emotion in $t-1$ time slice. The Formula for calculating the heritability λ_m^t of topic emotion in t time slice is

$$\lambda_m^t = \frac{1}{K^t}\left(K^t - \mathrm{rank}_k^{t-1}\right) \tag{3.10}$$

After calculating the emotional weight of each topic, the Formula for calculating the emotional intensity $J\left(z_k^t\right)$ of the topic is

$$J\left(z_k^t\right) = \frac{\sum_{m=1}^{M} W_m^{t-1}\mu_{m,k}^{t-1}}{M} \tag{3.11}$$

where $\mu_{m,k}^{t-1}$ represents the emotion distribution of document m under the k th topic under the $t-1$ time slice; W_m^{t-1} represents topic emotion weight under $t-1$ time slice; M is the number of documents.

3.6. Iterative model solution. OTSRM model takes dominant variable time t and word $w_{m,n}$ as initial input values, topic variable $\boldsymbol{Z}$ and emotion variable s as implicit variables, and uses the improved Gibbs sampling algorithm to solve the joint a posteriori probability function of topic distribution θ, feature word distribution φ and emotion distribution μ and the calculation formula is

$$P_t\left(z_i = k \mid z_{-i}, S_{-i}, w_i\right) = \frac{\left(n_{j,s}^{(d_i)}\right)^t + \alpha^t}{\left(n_{*,s}^{(d_i)}\right)^t + K^t\alpha^t}\cdot\frac{\left(n_{j,s}^{(w_j)}\right)^t + w\left(n_j^{(*)}\right)^{t-1} + \beta^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n^{(*)}\right)^{t-1} + V^t\beta^t}\cdot\frac{\left(n_{j,s}^{(w_i)}\right)^t + w\left(n_{j,s}^{(w_i)}\right)^{t-1} + \gamma^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n_{j,s}^{(*)}\right)^{t-1} + L^t\gamma^t} \tag{3.12}$$

where, $\left(n_{j,s}^{(d_i)}\right)^t$ represents the number of words assigned to emotion s by feature word j in the document d_i of time slice t; $\left(n_{*,s}^{(d_i)}\right)^t$ represents the total number of words that feature word j assigned to emotion s in the document d_i of time slice t; $\left(n_{j,s}^{(w_j)}\right)^t$ represents the number of words assigned by t time slice word w to feature word j and emotion word j; $\left(n_{j,s}^{(*)}\right)^t$ represents the total number of words assigned by t time slice word w to feature word j and emotion word s; $w\left(n_{j,s}{}^{(w_i)}\right)^{t-1}$ represents the heritability of emotion s assigned by $t-1$ time

slice word w to feature word j; $w\left(n_{j,s}^{(*)}\right)^{t-1}$ represents the sum of heritability of emotion s assigned by $t-1$ time slice word w to feature word j. Therefore, in each sampling, Equation (3.12) is iterated until relatively stable θ, φ, and μ distributions are obtained. Equation (3.15) of Formula (3.13) shows the corresponding probability distribution update.

$$\theta_{(z=j,s_i)}^{(d_i)} = \frac{\left(n_{j,s}^{(d_i)}\right)^t + \alpha^t}{\left(n_{*,j}^{(d_i)}\right)^t + K^t\alpha^t} \tag{3.13}$$

$$\varphi_{(z=j,s_i)}^{(d_i)} = \frac{\left(n_{j,s}^{(w_i)}\right)^t + w\left(n_j^{(*)}\right)^{t-1} + \beta^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n_j^{(*)}\right)^{t-1} + V^t\beta^t} \tag{3.14}$$

$$\mu_{(z=j,s_i)}^{(d_i)} = \frac{\left(n_{j,s}^{(w_i)}\right)^t + w\left(n_{j,s}^{(w_i)}\right)^{t-1} + \gamma^t}{\left(n_{j,s}^{(*)}\right)^t + w\left(n_{j,s}^{(*)}\right)^{t-1} + L^t\gamma^t} \tag{3.15}$$

3.7. Topic emotion calculation.

3.7.1. Topic similarity calculation. A common way to measure the similarity between two probability distributions is the KL distance. KL distance can also calculate the topic-word distribution difference in adjacent time slices. However, the asymmetry of KL distance cannot solve the symmetric topic distribution function. This article introduces relative entropy into topic similarity measurement, and a topic similarity calculation based on relative entropy is established. The specific Formula is as follows:

$$\begin{aligned}\mathrm{Sim}_{\mathrm{KL}}\left(\varphi_k^{t-1}, \varphi_k^t\right) &= -\frac{1}{2}\left[\mathrm{KL}\left(\varphi_k^{t-1}, \varphi_k^t\right) + \mathrm{KL}\left(\varphi_k^t, \varphi_k^{t-1}\right)\right] \\ &= \frac{1}{2}\left[\sum_{w\in V, k\in K} p(w)\log\frac{\varphi_k^{t-1}}{\varphi_k^t} + \sum_{w\in V, k\in K} q(w)\log\frac{\varphi_k^t}{\varphi_k^{t-1}}\right]\end{aligned} \tag{3.16}$$

$p(w)$ and $q(w)$ represent the probability of feature word w appearing in the distribution of topic Z' and Z^{t-1}, respectively.

3.7.2. Calculation of topic affective similarity. OTSRM model establishes the topic emotion model, which depends on the set of emotion words by calculating the maximum emotion value of emotion words under different topics. The specific calculation method is as follows: firstly, the position of the sentence in which the emotional word w_i of topic z is located and determined, whether there are turning conjunctions and negative words in the sentence in which it is located are determined, and the value of the negative flag neg is obtained, and the result is used to calculate the emotion $S(w_i)$, specifically, as follows:

$$S(w_i) = \mathrm{SO}_{\mathrm{neg}}(w_i) \cdot L(w_i) \tag{3.17}$$

where $L(w_i)$ represents the value of modifying degree words; $\mathrm{SO}_{\mathrm{neg}}(w_i)$ represents the comprehensive affective value of w_i. Finally, the expectation of all emotion words under topic $\boldsymbol{z}$ is taken as its final emotion value, and the calculation formula is:

$$S(z) = \frac{\sum_{i=1}^n S(w_i)}{n} \tag{3.18}$$

3.8. Experimental data. The experimental data in this article come from the news. The GooSeeker web crawler software screened five network hot events as empirical analysis cases.

Due to the space limitation, this article only takes Datal as an example to illustrate the process of topic emotion evolution analysis of OTSRM. First, the Chinese word segmentation software NLPIR cuts the text

Table 4.1: Results of feature word identification on the topic of "murder case in a mountain"

Topic	Feature words and their probabilities	Report time and probability
1	Mountain 0.013/ Traffic 0.011/ cut 0.009/ BMW male 0.012/ Dispute 0.016/ Electric vehicle 0.007/ death 0.017/ chase 0.019/ injury 0.013/ pick up	$t_1 = 0.218$, $t_2 = 0.257$, $t_3 = 0.201$, $t_4 = 0.233$, $t_5 = 0.246$, $t_6 = 0.225$
2	Procuratorate 0.021/ filing, 0.022/ conflict, 0.022/ alarm, 0.022/ alarm, 0.022/ first aid, 0.022/ drunk, 0.022/Harm 0.022/ Control 0.022/ suspected 0.020	$t_2 = 0.235$, $t_3 = 0.271$, $t_4 = 0.203$, $t_5 = 0.239$, $t_6 = 0.225$
3	Tattoo 0.021/ criminal record 0.019/ gang-related 0.021/ Good Samaritanism 0.021/ pawn shop 0.021/ burden 0.020/ illness 0.019/ Difficulties 0.019/ loans 0.019/ Cheonan Society 0.019	$t_2 = 0.243$, $t_3 = 0.296$, $t_4 = 0.233$, $t_5 = 0.277$
4	Legitimate 0.032/ safety 0.032/ excessive 0.032/ difference 0.032/ dodge 0.032/ Run 0.032/ calm 0.032/ law 0.032/ Injury 0.032/ Subjective 0.032	$t_3 = 0.312$, $t_4 = 0.217$, $t_5 = 0.296$, $t_6 = 0.211$
5	Notification 0.032/ justification 0.032/ defense 0.032/ Revocation 0.032/ Punishment 0.032/ punishment 0.032/ innocence 0.032/ immunity 0.032/ Determination 0.032/ compliance 0.032	$t_5 = 0.326$, $t_6 = 0.203$

Note: t_i indicates that the event occurred on the ith day.

into word sets and filters the stop words. Then, word sets corresponding to the report text d_c^t and its comment d_j^t belonging to the same time slice are combined into mixed word sets D_c^t. The emotion word frequency in d_j^t was measured by $(t = 1)$ day after the occurrence of the event $t = 1$ for 6 consecutive days. Finally, the word sets of t time slice is modelled successively, and an online sentiment dictionary is established according to the OTSRM model [15, 22].

4. Results and Discussion. In the OTSRM model, when $t = 1$ timepiece, the initial values of $\alpha^t, \beta^t, \gamma^t$ were set as 0.5, 0.1 and 0.1, respectively, topic feature word W_c. and emotion word W_s were set as 15, the probability generation threshold. and emotion threshold of topic features were both 0.2 and Gibbs sampling was set 2000 times.

4.1. Topic identification. The topic feature words identified by the OTSRM model and their corresponding probability distribution are shown in Table 4.1. Each topic is represented by the top 10 feature words with high probability.

From Table 4.1, a mountain murder event contains five topics. Topic 1: A mountain death caused by a traffic dispute, the duration of which is $t_1 \sim t_6$; Topic 2: The procuratorial organ files a case for investigation, the duration is $t_2 \sim t_6$; Topic 3: Background report of both parties involved, duration is $t_2 \sim t_5$; Topic 4: How to determine self-defense, duration is $t_3 \sim t_6$; Topic 5: The police decided the case. The duration is $t_5 \sim t_6$. According to Table 4.1, the evolution process of all topics in different time slices can be obtained (as shown in Figure 4.1).

From Figure 4.1, different topics coexist in the time dimension. For example, topic 2 and topic 3 overlap in the $t_2 \sim t_5$ time slice that is, "investigation by the procuratorial organ" and "background reports of both parties involved" coexist in multiple news reports, reflecting the diversity of perspectives of news reports.

4.2. Analysis of emotional evolution. According to the topic-emotion distribution, the emotional intensity of $t_1 \sim t_6$ time slice was calculated, and the topic-emotion matrix was obtained. At the same time, topic emotion words were selected, and the topic emotion values of different time slices were calculated using Equation (3.18). Taking topic 1 as an example, Table 4.2 shows the emotional outcomes identified. Due to the space limitation, other topic identification results are similar to Topic 1, which will not be repeated here. To more clearly show the emotion evolution recognition process of the OTSRM model on Data1, the emotion

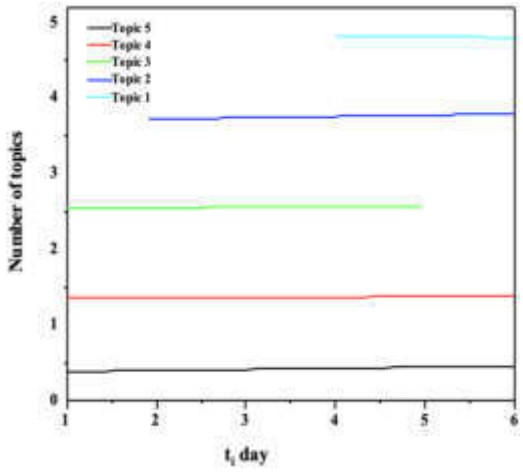


Fig. 4.1: Topic evolution process

Table 4.2: Emotion recognition results of topic 1 in $t_1 \sim t_6$ time slices

Time	Emotional words and their probabilities	Affective value
t_1	Scum 0.015/ Deserved 0.017/ innocent 0.021/ gas relief 0.018/ death 0.026/ Severe punishment 0.017/ impulse 0.026/ Courage 0.013/ show off 0.011/ brute 0.015-0.78	-0.78
t_2	Harm 0.014/ release 0.022/ Thorough investigation 0.013/ Quick 0.016/ Anger 0.017/ Compensation 0.014/ rampant 0.018/ light sentence 0.015/ Expectation 0.018/ innocence 0.026-0.42	-0.42
t_3	Hit 0.016/ disappointed 0.017/ hateful 0.021/ dead 0.016/ pathetic 0.017/ pity 0.018/ release 0.019/ Pity 0.022/ heavy judgment 0.017/ arrogance 0.018-0.26	-0.26
t_4	Bully 0.031/ Helpless 0.029/ Support 0.036/ Top 0.021/ innocent 0.032/ scum 0.024/ excessive 0.024/ sad 0.027/ harm 0.029/ sit and wait 0.033-0.34	-0.34
t_5	Kill 0.031/ legitimate 0.036/ helpless 0.027/ hateful 0.027/ poor 0.032/ Judgment 0.028/ gas relief 0.034/ sympathy 0.031/ hateful 0.026/ eradicate 0.032 0.16	0.16
t_6	Hope 0.027/ Support 0.034/ Like 0.029/ Top 0.035/ positive 0.028/ Fair 0.032/ justice 0.027/ stand 0.029/ happy 0.029/ Hold 0.027 0.43	0.43

values of five topics under different time slices were calculated respectively, and the dynamic comment emotion information evolution diagram was obtained, as shown in Figure 4.2.

From Figure 4.2, topic emotion presents a dynamic evolution over time. Topic 1 showed great emotional fluctuation, which was caused by the fact that in the initial stage of the event, netizens concentrated on expressing their condemnation and anger towards "BMW Man", which reflected strong negative emotion. When t=4, the topic of public concern gradually turned to the discussion on the judgment conditions of justifiable defence, and they expressed their sympathy for Haiming. The positive emotion they showed offset part of the negative emotion, so the overall emotion value was low-intensity negative emotion. When t=6, as the police determined the case as a justifiable defence, the positive emotion of the public reached the highest value, and strong positive emotion appeared.

The topic emotion of topic 2 and topic 3 is near the neutral emotion, which reflects the tangled and confused mentality of the public, that is, there is confusion about the defining conditions of justifiable defence and excessive defence, and the public is guessing the conclusion without clear emotional tendency. Topic 4 showed relatively stable positive emotions, and the public's sympathy for the weak party in a justifiable defence dominated. Topic 5 shows strong and stable positive sentiment, and the public's sentiment is similar to the

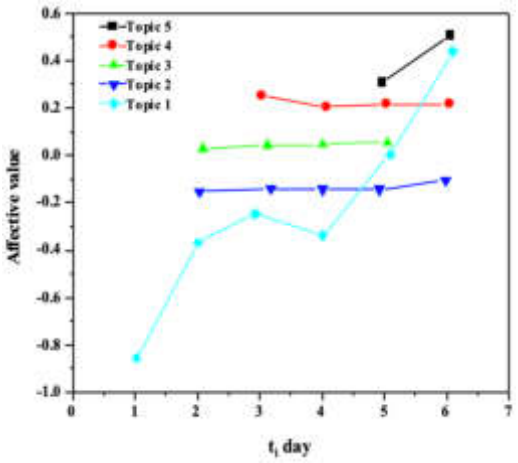


Fig. 4.2: Evolution distribution of topic emotion

Table 4.3: Comparison of emotion recognition accuracy

Model	Data 1	Data 2	Data 3	Data 4	Data 5
JST	66.16	67.09	69.81	71.36	72.95
WSSM	68.82	72.64	65.38	74.91	73.66
SSCM	74.46	75.92	72.65	75.64	74.68
TSSCM	77.68	78.84	76.31	78.88	79.35
OTSRM	78.94	80.13	79.96	81.03	80.62

sentiment tendency of topic 1 in the later stage, both of which express firm support for the judicial department to maintain social justice.

4.3. Model evaluation. In order to verify the algorithm performance of the OTSRM model, data sets of other 4 events were substituted into the model according to the above experimental process, and similar experimental results were obtained. In addition, this article uses accuracy rate, recall rate and F value as three indicators to make a comprehensive evaluation compared with the algorithms provided in the literature. Where, the initial values of the superparameter α^t, β^t and γ^t of the above model are set as 0.5,0.1 and 0.1, respectively. Topic feature word W_c and emotion word W_s were set as 15. The probability generation and emotion threshold of topic features were 0.2, and Gibbs was set to sample 1000 times. The accuracy and recall rates of the five models on the five data sets are shown in Table 4.3 and Table 4.4, respectively, and the comparison results of the F value are shown in Figure 4.3.

As can be seen from Figure 4.3, the highest emotion recognition accuracy rates of OTSRM are 85.56% and 81.03%. TSSCM and OTSRM models are significantly better than JST, WSSM and SSCM models in emotion classification because the algorithm based on the OLDA model comprehensively considers the topic relevance of different time slices. The number of topics is set dynamically according to probabilistic topics to achieve the purpose of topic temporal and spatial modelling. However, the prior topic emotion parameters of JST, WSSM and SSCM models are greatly affected by subjective experience, which easily causes skew of emotion mining, thus reducing classification accuracy. Although the TSSCM model can dynamically identify topics, it does not consider the transitivity of the topic's emotional intensity. It cannot effectively solve the problem of emotional iteration caused by mixing old and new topics. The algorithm in this article makes up for the shortcomings of the above algorithms. By introducing the emotion intensity with heritability, the topic in different time slices is acquired dynamically, and the mixed model of the coexistence of topic and emotion is established, effectively

Table 4.4: Comparison of emotion recognition recall rate

Model	Data 1	Data 2	Data 3	Data 4	Data 5
JST	76.35	74.32	77.45	78.62	76.64
WSSM	77.92	78.94	74.49	79.78	78.26
SSCM	79.81	80.18	79.46	80.64	79.98
TSSCM	81.67	83.94	82.16	83.37	82.21
OTSRM	83.33	84.62	82.77	85.56	84.33

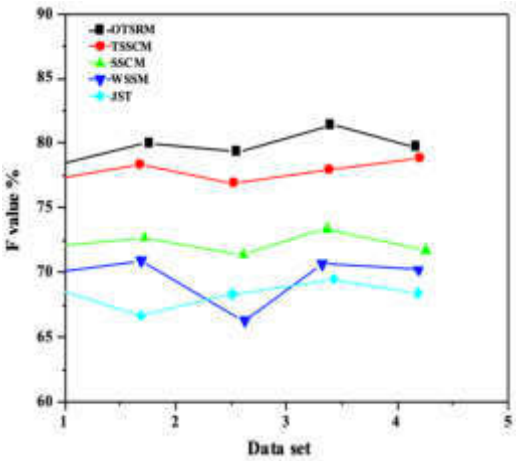


Fig. 4.3: Comparison results of F-value of emotion recognition

improving the recognition accuracy of topic and emotion.

5. Conclusion. An innovative approach for Korean topic sentiment analysis is introduced in this paper to leverage social media big data clustering. The basis of the proposed contribution is developing the Online Topic Sentiment Recognition Model (OTSRM), which introduces several key elements to enrich the topic sentiment analysis in Korean social media discussions. The OTSRM model is characterized by constructing an emotion intensity operator, establishing an emotion evolution channel based on posterior probability, and acquiring feature word and emotion word distribution matrices. These components work synergistically to enable a dynamic representation of the evolving emotional recognition within textual content. Using the relative entropy method, we calculate the maximum emotion value associated with the current topic focus, offering valuable insights into the real-time evolution of topic emotions within the text. The experimental validation has underscored the OTSRM model's effectiveness in topic emotion evolution analysis. It has demonstrated robust performance and the ability to capture the tones of emotional fluctuations within Korean social media discourse. These findings confirm the relevance and utility of the suggested model in uncovering the dynamic emotional undercurrents that shape public sentiment within online discussions.

Acknowledgements. The Key Research Project supported by Social Science research Fund of Lingnan Normal University in 2022: A Study on the Image of China in Foreign Novels in the Perspective of Powerful Cultural Country [WZ2204].

REFERENCES

[1] J. Allan, *Introduction to topic detection and tracking*, in Topic Detection and Tracking: Event-based Information Organization, Springer, 2002, pp. 1–16.

[2] Y. Alotaibi, M. N. Malik, H. H. Khan, A. Batool, A. Alsufyani, S. Alghamdi, et al., *Suggestion mining from opinionated text of big social media data.*, Computers, Materials & Continua, 68 (2021).
[3] M. Asgari-Chenaghlu, M.-R. Feizi-Derakhshi, L. Farzinvash, M.-A. Balafar, and C. Motamed, *Topic detection and tracking techniques on twitter: a systematic review*, Complexity, 2021 (2021), pp. 1–15.
[4] A. Barradas, A. Tejeda-Gil, and R.-M. Cantón-Croda, *Real-time big data architecture for processing cryptocurrency and social media data: A clustering approach based on k-means*, Algorithms, 15 (2022), p. 140.
[5] Y. Deng and D. Liu, *A multi-dimensional comparison of the effectiveness and efficiency of association measures in collocation extraction*, International Journal of Corpus Linguistics, 27 (2022), pp. 191–219.
[6] Y. Guo, F. Wang, C. Xing, and X. Lu, *Mining multi-brand characteristics from online reviews for competitive analysis: A brand joint model using latent dirichlet allocation*, Electronic Commerce Research and Applications, 53 (2022), p. 101141.
[7] D. Hyun and S. Lee, *A study of big data analysis regarding smartphone user satisfaction: Utilizing sentiment analysis based on social media data*, Korean Journal of Converging Humanities, 9 (2021), pp. 7–35.
[8] Y. Ji, J. Wang, Y. Niu, and H. Ma, *Reliable event detection via multiple edge computing on streaming traffic social data*, IEEE Access, (2021).
[9] Y. Kim, D. Sohn, and S. M. Choi, *Cultural difference in motivations for using social network sites: A comparative study of american and korean college students*, Computers in Human Behavior, 27 (2011), pp. 365–372.
[10] X. Lv and M. Li, *Application and research of the intelligent management system based on internet of things technology in the era of big data*, Mobile Information Systems, 2021 (2021), pp. 1–6.
[11] J. Meng and R. Tench, *Strategic communication and the global pandemic: Leading through unprecedented times*, International Journal of Strategic Communication, 16 (2022), pp. 357–363.
[12] M. Mosleh, G. Pennycook, and D. G. Rand, *Field experiments on social media*, Current Directions in Psychological Science, 31 (2022), pp. 69–75.
[13] K. R. Nastiti, A. F. Hidayatullah, and A. R. Pratama, *Discovering computer science research topic trends using latent dirichlet allocation*, Jurnal Online Informatika, 6 (2021), pp. 17–24.
[14] J. St John, K. St John, and B. Han, *Entrepreneurial crowdfunding backer motivations: a latent dirichlet allocation approach*, European Journal of Innovation Management, 25 (2022), pp. 223–241.
[15] J. Tong, L. Shi, L. Liu, J. Panneerselvam, and Z. Han, *A novel influence maximization algorithm for a competitive environment based on social media data analytics*, Big Data Mining and Analytics, 5 (2022), pp. 130–139.
[16] S. K. Uppada, K. Manasa, B. Vidhathri, R. Harini, and B. Sivaselvan, *Novel approaches to fake news and fake account detection in osns: user social engagement and visual content centric model*, Social Network Analysis and Mining, 12 (2022), p. 52.
[17] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, *Automatic emotion recognition for groups: a review*, IEEE Transactions on Affective Computing, 14 (2021), pp. 89–107.
[18] T. W. Wibowo, S. H. M. B. Santosa, B. Susilo, and T. H. Purwanto, *Revealing tourist hotspots in yogyakarta city based on social media data clustering*, Geo Journal of Tourism and Geosites, 34 (2021), pp. 218–225.
[19] H. Xie, Y. He, X. Wu, and Y. Lu, *Interplay between auditory and visual environments in historic districts: A big data approach based on social media*, Environment and Planning B: Urban Analytics and City Science, 49 (2022), pp. 1245–1265.
[20] C. Yang, G. Su, and J. Chen, *Using big data to enhance crisis response and disaster resilience for a smart city*, in Proceedings of the 2nd International Conference on Big Data Analysis, Beijing, China, 2017, IEEE, pp. 504–507.
[21] T. Zhou, K. Law, and D. Creighton, *A weakly-supervised graph-based joint sentiment topic model for multi-topic sentiment analysis*, Information Sciences, 609 (2022), pp. 1030–1051.
[22] Y. Zhou, L. Liao, Y. Gao, R. Wang, and H. Huang, *Topicbert: A topic-enhanced neural language model fine-tuned for sentiment classification*, IEEE Transactions on Neural Networks and Learning Systems, (2021).

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 13, 2023
Accepted: Oct 9, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 127–136, DOI 10.12694/scpe.v25i1.2280

APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGY IN ELECTROMECHANICAL INFORMATION SECURITY SITUATION AWARENESS SYSTEM

XIANGYING LIU*, ZHIQIANG LI†, ZHUWEI TANG‡, XIANG ZHANG§, AND HONGXIA WANG¶

Abstract. The information security situational awareness system is proposed in this paper to leverage big data and artificial intelligence (AI) to enhance information security situation prediction. Deep learning techniques, specifically the long short-term memory recurrent neural network (LSTM-RNN), predict security situations using complex non-linear and autocorrelation time series data from current and past system conditions. Additionally, the study incorporates the variant gated recurrent unit (GRU) within the LSTM-RNN framework. A comprehensive experimental analysis is conducted, comparing various methods, including LSTM, GRU, and others, to assess and compare their predictive performance. The experimental results reveal that LSTM-RNN demonstrates a commendable level of predictive accuracy on the test dataset, with a mean absolute percentage error (MAPE) of 8.79%, a root mean square error (RMSE) of 0.1107, and a relative root mean square error (RRMSE) of 8.47%. Both LSTM and GRU exhibit exceptional predictive accuracy, with GRU offering a slightly faster training speed due to its simplified architecture and fewer trainable parameters. Overall, this research highlights the potential of AI-based methodologies in constructing robust information security situational awareness systems.

Key words: Network security, Situational awareness system, LSTM RNN, GRU, Accuracy

1. Introduction. The beginning of the Internet era has led to significant changes in people's lives. Various industries are continually evolving due to the Internet's influence. Concurrently, there is a growing prevalence of network security threats in the Internet age. Consequently, enhancing research on information security situational awareness systems is authoritative. Current research on enterprise information security situational awareness systems and the prevailing state of information security protection highlights that enterprise information security often operates within a passive cycle of detection and remediation. Typically, enterprises install defence systems tailored to their unique operational characteristics and production nature. They proactively identify hidden vulnerabilities and risks within their network systems through risk assessments and penetration tests, followed by targeted mitigation measures [18].

When suspicious activities or attacks are detected, comprehensive investigations and analyses are conducted, encompassing the examination of security device logs and network traffic data. This process aims to determine the behaviour's specific nature and severity and resolve these issues as comprehensively as possible. Within this passive framework of information security defence in enterprises, the predominant focus often lies on the defensive aspects, with limited attention directed toward understanding and analyzing the root causes of attacks. Investment and research in system repair tend to be relatively modest, primarily relying on passive remedies in the form of patches provided by product manufacturers [6].

Simultaneously, enterprises persist in refining and enhancing their defence strategies to bolster their systems' resilience against external threats. Advances in computer technology and the growing understanding of information security have led to effective optimizations in information security defence measures. Many enterprises have implemented integrated security systems encompassing network antivirus, endpoint management, security auditing, access controls, and vulnerability discovery. These integrated systems ensure the secure and reliable

*Tangshan Sanyou Chemical Co., Ltd., Thermoelectric Branch, Tangshan, Hebei, 063305, China (Corresponding author: xiangyingliu8@126.com)
†Tangshan Sanyou Chemical Co., Ltd., Thermoelectric Branch, Tangshan, Hebei, 063305, China
‡Tangshan Sanyou Chemical Co., Ltd., Thermoelectric Branch, Tangshan, Hebei, 063305, China
§Tangshan Sanyou Chemical Co., Ltd., Thermoelectric Branch, Tangshan, Hebei, 063305, China
¶Tangshan Sanyou Chemical Co., Ltd., Thermoelectric Branch, Tangshan, Hebei, 063305, China

operation of enterprise activities, reduce information security risks, enable unified warning systems, centralize management and traceability, and mitigate the impact of information risks on routine business operations [17].

The Internet era has ushered in a surge of various network security threats, with a noticeable upward trajectory. The frequency of cyberattacks targeting countries and political entities is rising annually. Consequently, the need to swiftly address external threats affecting nations, organizations, businesses, and individuals is becoming increasingly urgent. This article primarily focuses on modern enterprises in response to the pressing network security issue. It delves into contemporary challenges, such as network attacks and data breaches, that are prevalent today. Furthermore, it underscores the importance of exploring and proposing solutions for these issues. Given this context, conducting an in-depth examination of the application of big data and artificial intelligence technology in information security situational awareness systems holds significant practical relevance [8].

In recent years, modern enterprises have increasingly relied on networks for their day-to-day operations. They depend on high-speed and secure information technology, significantly boosting work efficiency. However, a traditional approach of "defence, detection, and remediation" prevails when managing and controlling enterprise information security. Indeed, this approach proves valuable, especially in scenarios like information security penetration tests or risk assessments. However, it's noteworthy that during the optimization of information security processes, most enterprises allocate over 95% of their resources to defence, leading to a predominantly reactive approach to information security [10].

In daily operations and maintenance, enterprise information security initiatives continue to expand. Various security systems, including terminal controls, network antivirus software, and vulnerability scanning tools, are continually being implemented to safeguard business operations. However, a critical issue arises from the lack of a unified and integrated prevention and control system across these disparate security systems. This fragmentation hinders the achievement of unified management for early warning of information security threats and traceability of security issues. Data collection and perception systems rely on various network traffic and logs in enterprise information security management. During specific data collection, it's crucial to tailor the process to the unique scale and circumstances of the enterprise. This entails selecting and gathering different well-suited data types and enhancing the effectiveness and accuracy of the data collection process. Moreover, when dealing with specific traffic processing tasks, the system can leverage its technology to reconstruct data, conduct precise data analysis, and disseminate the acquired specific data. This approach facilitates other enterprises' data storage and utilization through the enterprise platform, fostering collaboration and knowledge sharing in information security [13].

After processing diverse types of network traffic data and other information, it becomes imperative to delve deeply into the data's content, proactively identifying internal issues and risks within the enterprise while promptly resolving them. This process involves two key components: Artificial Intelligence Detection of Malicious Code Technology: This technology is constructed through artificial search engines, drawing from an extensive pool of malicious and normal software samples. It seeks to identify standard information data features across different samples and build effective machine-learning models for the security scanning of unknown programs [5].

Application of Artificial Intelligence Virus Detection Technology: This application within enterprise information security situational awareness systems enables efficient identification and timely detection of viruses. It plays a pivotal role in mitigating computer system damage caused by viruses. Computer viruses have evolved and diversified in recent years, posing a significant threat to normal computer system operations. By integrating artificial intelligence virus detection technology with big data technology, enterprises can enhance their ability to detect and respond to viruses. Employing multiple virus localization methods further elevates the efficiency and accuracy of virus detection, rendering it a more precise and scientifically driven process [2].

The network security situational awareness model represents a comprehensive framework encompassing various steps and processes in achieving situational awareness within a network security context. Figure 1.1 visually illustrates this ecological framework, comprehensively representing the complex processes involved in understanding network security situations. This illustration is a valuable reference point for grasping the complexity and interconnectedness of various network security situational awareness elements. As technology and our understanding of network security continue to evolve, the development and refinement of this model

Fig. 1.1: Framework of situation awareness model

remain an ongoing and dynamic process. This continuous development reflects the ever-changing landscape of network security challenges and the need for adaptive strategies and tools to address them effectively [3].

The paper is systematically structured into five primary sections, including an introduction. Section 2 conducts an extensive literature review, providing a comprehensive understanding of previous research in information security situational awareness and the utilization of artificial intelligence within electromechanical systems. In Section 3, the paper investigates the proposed methodology, explicitly focusing on implementing the LSTM RNN with GRU within electromechanical information security. Section 4 is dedicated to the presentation and thorough discussion of the results obtained from the experimentation and analysis, explaining the efficacy of the proposed approach. Finally, Section 5 offers concluding remarks summarizing the study's primary findings and contributions.

2. Literature Review. In recent years, information security practices within enterprise units have exhibited remarkable consistency, marked by a recurring and passive pattern encompassing defence, detection, and remediation. This established procedure typically commences with the execution of penetration tests or extensive risk assessments meticulously devised to unearth vulnerabilities and latent risks lurking within information and network systems. Once these vulnerabilities are pinpointed, they trigger precise remedial actions to fortify the system's defences. A comprehensive response protocol is initiated in the unfortunate event of a cyberattack. This entails thoroughly recording and analyzing pertinent network traffic data and security device logs to unveil the intricate intricacies of attack behaviours [1].

In the framework of passive information security defence, it's worth highlighting that 95% of resources are allocated towards bolstering defensive measures. These measures are meticulously architected and implemented to thwart potential threats. However, only a meagre 5% of resources are dedicated to delving into the multifaceted realm of attack defence. This disproportionate resource distribution underscores the predominant emphasis on safeguarding against potential threats while often sidelining the comprehensive understanding and proactive management of the root causes behind these threats. It is within this context that the cybersecurity practices of the industry have predominantly unfolded in recent times [9].

The remediation process primarily revolves around applying various patches provided by the original manufacturers of the products or equipment. These patches are essential for addressing and rectifying vulnerabilities and weaknesses identified within the system. However, it's important to note that remediation efforts extend beyond patch management. They encompass an ongoing commitment to refine and fortify defensive measures, continually enhancing the overall security posture of enterprise units. Within the landscape of traditional IT network security situational awareness technology, there has been notable progress and research advancement. Assessment methods in this domain have matured considerably, offering a solid theoretical foundation for conducting security situation assessments, especially within industrial control networks. These well-established methods provide valuable guidance for assessing the security posture of networks [16].

One of the prominent drawbacks of these methods is their reliance on a single source of information, which can lead to a narrow perspective on the security landscape. Additionally, implementing these methods often entails significant time and resource investments, making them resource-intensive. Moreover, subjectivity can creep into the assessment process, introducing potential biases. Lastly, while these methods are valuable, they may exhibit limited accuracy in assessing modern network security threats' complex and dynamic nature. Many of these methods rely on manual parameter configuration or calculate overall network risk based solely on host

nodes, which can oversimplify the intricate nuances of network security [4].

This article embarks on its investigative journey by predicting the network security situations. The initial step involves scrutinizing whether it is indeed possible to forecast these situations accurately. Once the feasibility of prediction is established, the article proceeds to assess the limitations inherent in traditional prediction methodologies. A sophisticated solution called Long Short-Term Memory Recurrent Neural Networks (LSTM RNN) is introduced to address the limitations. LSTM, a deep learning model, is harnessed to predict network security situations. In parallel, the article incorporates the Gated Recurrent Unit (GRU) algorithm, renowned for its efficiency in achieving results comparable to LSTM but with significantly shorter training times [14].

Following the implementation of LSTM, GRU, and other widely used prediction techniques, the article rigorously conducts experiments and analyses the results. This comprehensive assessment reveals that LSTM and GRU outperform their counterparts in terms of prediction accuracy. These two models shine in their ability to not only account for the inherent nonlinearity of the data but also to consider the crucial aspects of autocorrelation and timing within the data. In contrast, other methods are constrained by data stationarity, the degree of data correlation, model parameter selection, noise levels, and the choice of prediction step size. Notably, LSTM and GRU excel in prediction accuracy and exhibit remarkable training speed efficiency, making them capable of real-time prediction – a highly sought-after quality in network security prediction [7].

A novel methodology in big data and artificial intelligence technology approach lies in handling complex nonlinear and autocorrelation time series data derived from current and historical values about system situations. To achieve its predictive goals, the study harnesses the power of the long short-term memory recurrent neural network (LSTM-RNN), a prominent component of deep learning known for its aptitude in handling sequential data. Furthermore, to enhance the predictive capabilities within the LSTM framework, the study incorporates the variant Gated Recurrent Unit (GRUs). This addition is noteworthy as it contributes to refining and optimizing the prediction process, particularly in scenarios where efficiency and accuracy are paramount [12].

A series of experiments are conducted to validate and benchmark the effectiveness of the proposed methodology. These experiments facilitate a comprehensive comparative analysis of the prediction outcomes obtained from LSTM, GRU, and other established prediction methods. Through this rigorous assessment, the study aims to provide insights into each method's relative strengths and weaknesses, shedding light on their applicability and potential contributions to the information security situational awareness field.

3. Investigation Methods.

3.1. Network security situation prediction based on LSTM. Network security situation prediction represents the pinnacle of situational awareness, and its precision empowers administrators to enact appropriate security measures. This prediction hinges on a foundation of situation assessment, which produces a situational value. These situational values accumulate into a time series through continual assessments over time. Given the intricate nature of network security situations and the inherent unpredictability of attacks, this situational sequence takes the form of a non-linear sequence. It is essential to acknowledge that this situational sequence is both non-linear and characterized by autocorrelation. Recognizing these features within the prediction data aids in selecting accurate prediction methods.

Traditional neural network methods like BP and RBF exhibit robust non-linear mapping capabilities. However, in their network model, layers are fully connected, but nodes within each layer lack interconnections. This structure lacks time sequence information. Consequently, traditional neural networks struggle to predict time sequences effectively. The Grey Prediction model necessitates that the function derived from the original discrete data be a smooth discrete function. Yet, when the network is under attack, the function formed by the situational sequence isn't sufficiently smooth. Despite considering the nonlinearity and timing of the situational sequence, the Grey Prediction model faces challenges in predicting time series with significant fluctuations.

Support Vector Machines are more suitable for handling small samples and are less adept at managing large-scale data. Moreover, they don't account for the timing of the situational sequence. While improvements can enhance the prediction accuracy of the mentioned methods, this article seeks a comprehensive approach that accommodates the characteristics of the situational sequence. In this pursuit, Recurrent Neural Networks (RNN) within deep learning emerge as a promising solution. RNNs can effectively address both nonlinearity and timing in the data. Moreover, RNNs suffer from the issue of gradient disappearance, prompting the development of Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) to mitigate this challenge.

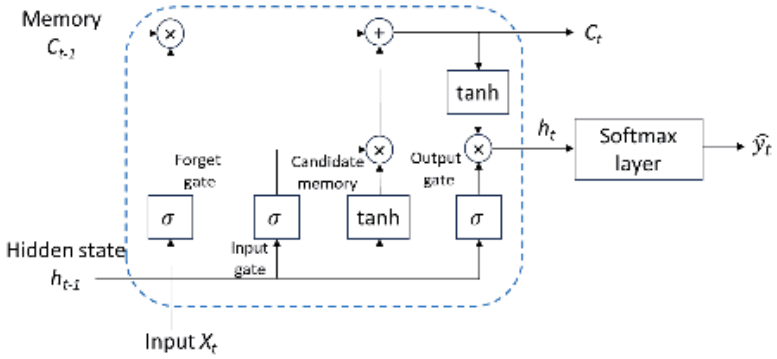


Fig. 3.1: Framework of situation awareness model

This article introduces the mathematical model of LSTM-RNN and the variant Gated Recurrent Unit (GRU) within the LSTM framework. Both LSTM-RNN and GRU are subsequently applied in predicting network security situations.

Long Short-Term Memory Recurrent Neural Networks, abbreviated as LSTM-RNN or simply LSTM, belong to a class of recurrent neural networks designed to enhance long-term and short-term memory capabilities. They excel in retaining knowledge over extended periods and have the capacity for sustained long-term learning. LSTM effectively addresses the gradient disappearance inherent in conventional RNNs, making it a deep learning technique with significant developmental potential. Its applications span various domains, including stock prediction, disease forecasting, language translation, image analysis, and more. LSTM enhances its structure by incorporating gates that regulate the flow of information. When a gate is open, the current neuron receives input from the preceding neuron, whereas a closed gate prevents this interaction. It's through these gates that LSTM achieves its exceptional long and short-term memory capabilities.

The network architecture of LSTM has introduced notable enhancements within the hidden layer of traditional RNNs. Instead of using hidden neurons, LSTM employs memory units. Each memory unit comprises one or more memory cells and incorporates three fundamental "gates", essentially non-linear summation components. These three "gates" are the Input, Output, and Forget gates. The mathematical model of LSTM is systematically introduced in this article, offering a step-by-step breakdown.

Breaking down the LSTM structure can make it appear less intricate. The primary objective of LSTM is to regulate information flow to enable long-term information retention. One straightforward approach to understanding LSTM is realizing the dot product of two matrices of equal dimensions. If the matrix values fall within the range of [0, 1], it can be interpreted as '0' indicating suppression and '1' indicating activation. With this design goal in mind, it becomes evident that the information within the Memory Cell corresponds to the horizontal line in Figure 3.1.

The following are the various gates of the LSTM RNN security awareness system [11].

(1) Forget gate

The forgetting gate controls the influence of the Memory Cell information of the previous moment on the Memory Cell information of the current moment. The oblivion gate is determined jointly by the output h_{t-1} of the previous moment and the input x_t of the current moment. The function formula of the gate f_t is as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3.1}$$

σ is the sigmoid function, which is mapped to $[0, 1]$, h_{t-1} represents the output of memory cells at the previous time, x_t represents the input at the current time, W_* is the coefficient matrix, b_* is the bias matrix.

(2) Input gate

The input gate controls the input information's contents that can affect the current memory cell. The input information includes the output h_{t-1} of the previous time and the input x_t of the current time. The function

formula of the input gate i_t is given by

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{3.2}$$

Equation (3.3) of $\widetilde{C_t}$ can be obtained from the standard of RNN:

$$\widetilde{C_t} = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{3.3}$$

After passing the above two gates, the state of memory cells at this time is as follows:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{3.4}$$

where * represents the multiplication of matrix elements.

(3) Output gate

The output gate determines the output content in memory cells and the function of the output gate. o_t is expressed as

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{3.5}$$

The output information h_t at the current time is given by:

$$h_t = o_t * \tanh\left(C_t\right) \tag{3.6}$$

The description above pertains to the conventional LSTM, but numerous LSTM variants exist. One widely adopted variant is the Gated Recurrent Unit (GRU). GRU combines the memory cell state and hidden state similar to LSTM but with fewer gates, maintaining the effectiveness of LSTM. Additionally, GRU boasts a more straightforward structure and requires less computation time than LSTM. In GRU, the forgetting and input gates are consolidated into a single update gate, responsible for governing the influence of the previous moment's state information on the current moment's state.

The mathematical model of GRU is shown in Equation (3.7):

$$\begin{aligned} z_t &= \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \\ r_t &= \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \\ \tilde{h}_t &= \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \tag{3.7}$$

where the update gate is z_t and the reset gate is r_t.

3.2. Network security prediction experiment based on LSTM. LSTM represents a neural network model well-suited for mining non-linear temporal data. In this article, LSTM is harnessed for predicting network security situations. The process commences by delineating the precise task requirements for situation prediction. This involves an analysis of the experiment-supported scenarios and the intrinsic characteristics of the situational data. Subsequently, the article provides an overview of the initial construction process for the LSTM prediction model, encompassing aspects like model architecture and strategic parameters. Lastly, the constructed LSTM model is employed in prediction experiments, and the results are subjected to comparative analysis against other models.

3.2.1. Experimental subjects and requirements. This article employs an attack simulation scenario to enhance its research methodology and practical applicability. To simulate a real-world cybersecurity incident, the researchers utilize IDS informer software to arrange a Denial of Service (DoS) attack scenario, a commonly encountered threat in information security. The attack is designed to mimic malicious performers' strategies and techniques to disrupt network services. To gather relevant data and comprehensively capture the evolving security situation, we have collected the data in discrete sets at 10-minute intervals, resulting in 1000 distinct data points over a specified period.

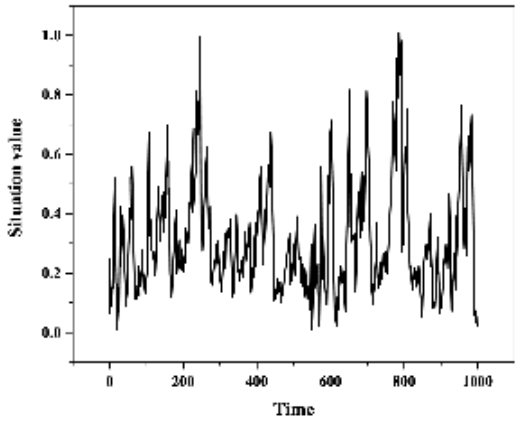


Fig. 3.2: Dynamic visualization of normalized network security systems

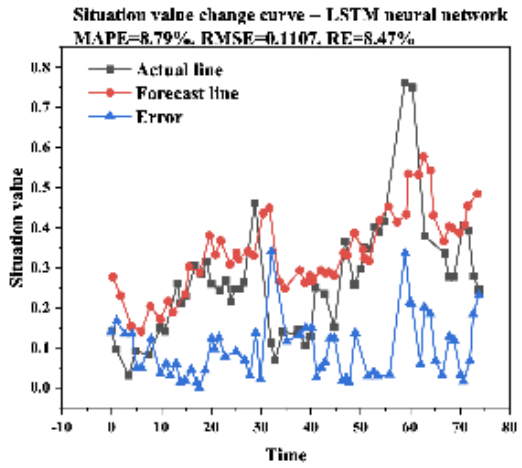


Fig. 3.3: Situation prediction results of LSTM

The situational value, a crucial metric for understanding the evolving security landscape, is meticulously computed using the established network security situational index. This index is based on a well-defined framework considering various critical factors and parameters contributing to the overall security posture. Subsequently, the calculated situational values are employed to construct a chronological change curve, visually depicting the temporal evolution of the security situation. As illustrated in Figure 3.2, the resulting curve provides a dynamic representation of how the security situation unfolds over time. This empirical approach and visualization technique give valuable insights into the nature of the attack, its impact on the network, and the effectiveness of the proposed security measures.

Figure 3.2 clearly illustrates that the ship system's normalized network security situation curve exhibits significant volatility and nonlinearity. To achieve precise situation value prediction, it's imperative to thoroughly extract timing information from existing data and comprehend its underlying mechanism via a model. In this context, the initial 925 situation values are designated as the training set, while the remaining 75 situation values constitute the test set. The input comprises historical situation values, and the output consists of the predicted situation values.

3.2.2. LSTM modeling. A three-layer LSTM with a unidirectional loop is designed to accomplish the prediction task. To optimize the applicability of the constructed LSTM, it is crucial to scrutinize its parameter configuration and the strategies implemented for learning and training. The subsequent section offers a concise exploration of these critical aspects. The LSTM-RNN represents a typical Many to One model, wherein the input consists of data from the previous moment, and the output comprises the predicted value for the subsequent moment's data. Given the enduring nature of information security threats, security situation data often exhibits prolonged temporal correlations. Hence, it is advisable to set a relatively large input dimension.

Since LSTM inherently excels at depicting long-term dependencies, increasing this value does not negatively impact model effectiveness; instead, it enhances the breadth of information extraction. Here, we set the initialization value to be equal to 30. To mitigate issues like gradient explosion or vanishing gradients and improve the efficiency of training a multi-layer model, we employ the Xavier initialization strategy for initializing the parameters of the recurrent network (LSTM-RNN) and the output nested feedforward multi-layer network (MLP) [15]. Given the limited training data, we have introduced the Dropout strategy to aid training and prevent overfitting. Additionally, we employ the Adam optimization algorithm in this experiment to ensure the efficiency of neural network training. The mean square error (MSE) is the optimization objective during training.

4. Simulation Results and Discussion. The Keras development framework is utilized efficiently to implement the designed LSTM model. Through a rigorous and exhaustive training process, the LSTM-RNN model is unlocked to generate predictions for situational data. Figure 3.3 is the culmination of the activities, providing a striking visual representation of the proposed predictive model's outcomes. This graphical representation illustrates the ability of the proposed model, as it investigates network security situations and offers valuable insights into potential threats. These predictions are a valuable asset, supporting information security and situational awareness within the context of our study.

The LSTM model has demonstrated remarkable predictive accuracy on the test dataset by its impressive performance metrics: a mean absolute percentage error (MAPE) as low as 8.79%, an exceptionally minimal root mean square error (RMSE) of 0.1107, and a mere 8.47% relative error (RE). This article conducted a comprehensive series of comparative experiments to ensure a thorough evaluation of the current model's performance. These experiments encompassed the utilization of several well-established prediction methods, including linear regression (LR), support vector regression (SVR), backpropagation (BP) neural networks, and GRU, among others. The findings and invaluable insights from these comparative experiments have been presented and visually explained in Figures 4.1 to 4.5. The comprehensiveness of these experiments offers valuable insights into the respective strengths and weaknesses of different prediction methods, ultimately reinforcing the LSTM model's resilience and effectiveness within the domain of information security situational awareness.

The discussed results indicate that LSTM and GRU exhibit higher prediction accuracy than other methods. This superiority starts from the fact that these two models account for data nonlinearity and consider data autocorrelation and timing. In contrast, other methods are constrained by data stationarity, data correlation, model parameter selection, noise levels, and the choice of prediction step size. LSTM and GRU perform exceptionally well, and their prediction accuracy is on par. GRU, in particular, benefits from its more straightforward structure and fewer training parameters, resulting in a slightly faster training time than LSTM. Specifically, the training time for LSTM is 18.9 seconds, while GRU takes 18.5 seconds. Moreover, both LSTM and GRU demonstrate rapid prediction time consumption at the millisecond level.

5. Conclusion. This article explores a method for predicting network security situations, analyzing their predictability with a focus on nonlinearity and time series characteristics. Experimental comparisons between LSTM-RNN, GRU, and existing methods reveal that LSTM exhibited robust prediction accuracy on the test data, with MAPE at 8.79%, RMSE at 0.1107, and RE at 8.47%. Both LSTM and GRU exceeded other methods, offering similar prediction accuracy. With its more straightforward structure and fewer training parameters, GRU is slightly faster training times (18.5 seconds compared to LSTM's 18.9 seconds). Notably, both LSTM and GRU achieved millisecond-level speed in predictions, satisfying real-time demands. These findings underscore the potential for improved modeling of complex network systems by addressing practical challenges and enhancing each stage. Furthermore, comprehensively representing knowledge is vital in network environments with limited local knowledge. Extending the confidence rule-based identification framework to

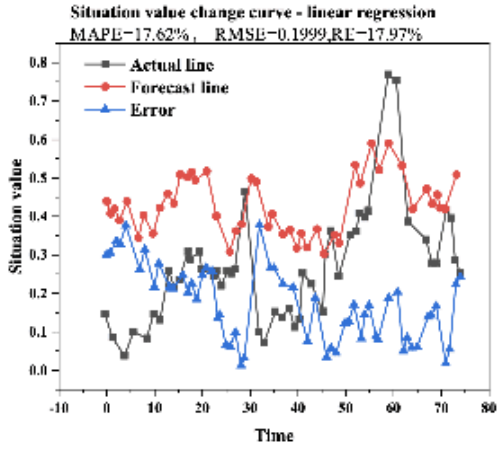


Fig. 4.1: Prediction results of linear regression method

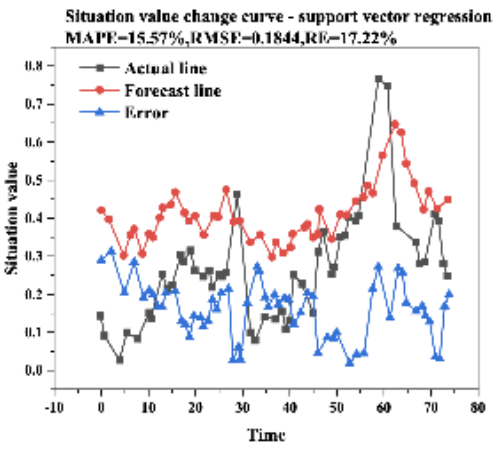


Fig. 4.2: Prediction results of support vector regression

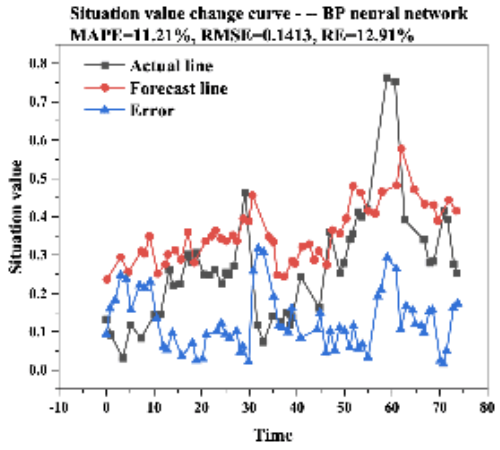


Fig. 4.3: Prediction results of back propagation (BP) neural networks

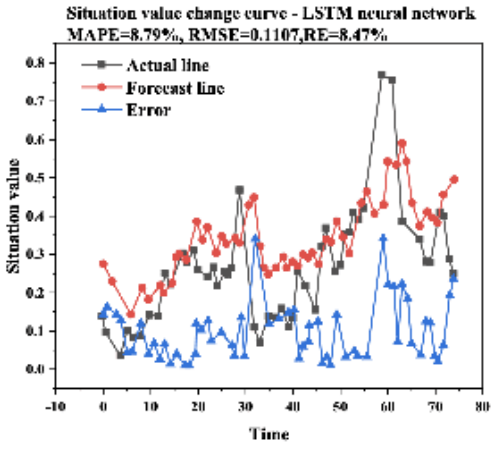


Fig. 4.4: Outcomes of the LSTM neural network's predictions

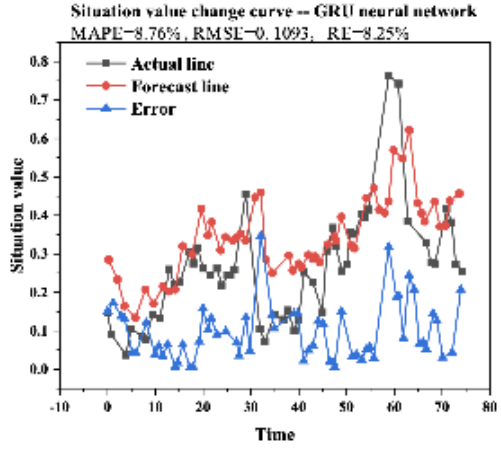


Fig. 4.5: Prediction results of gated recurrent unit

include the power set promises enhanced prediction outcomes. Future research will focus on refining observable value selection and further developing theories underpinning the implicit confidence rule base model within the power set identification framework, advancing network security situational awareness and prediction.

REFERENCES

[1] M. Abbasi, A. Shahraki, and A. Taherkordi, *Deep learning for network traffic monitoring and analysis (ntma): A survey*, Computer Communications, 170 (2021), pp. 19–41.

[2] Y. S. Afridi, K. Ahmad, and L. Hassan, *Artificial intelligence based prognostic maintenance of renewable energy systems: A review of techniques, challenges, and future research directions*, International Journal of Energy Research, 46 (2022), pp. 21619–21642.

[3] M. Alshehri, *Blockchain-assisted cyber security in medical things using artificial intelligence*, Electronic Research Archive, 31 (2023), pp. 708–728.

[4] G. K. Bharathy and B. Silverman, *Applications of social systems modeling to political risk management*, Handbook on Decision Making: Vol 2: Risk Management in Decision Making, (2012), pp. 331–371.

[5] Z. Cui, L. Du, P. Wang, X. Cai, and W. Zhang, *Malicious code detection based on cnns and multi-objective algorithm*, Journal of Parallel and Distributed Computing, 129 (2019), pp. 50–58.

[6] L. Franchina, G. Inzerilli, E. Scatto, A. Calabrese, A. Lucariello, G. Brutti, and P. Roscioli, *Passive and active training approaches for critical infrastructure protection*, International Journal of Disaster Risk Reduction, 63 (2021), p. 102461.

[7] R. Fu, Z. Zhang, and L. Li, *Using lstm and gru neural network methods for traffic flow prediction*, in Proceedings of the 31st Youth academic annual conference of Chinese Association of Automation, Wuhan, China, 2016, IEEE, pp. 324–328.

[8] Y. Gao, *Research on the application of artificial intelligence technology in the development of computer vision*, Highlights in Science, Engineering and Technology, 9 (2022), pp. 80–84.

[9] P. L. Goethals and M. E. Hunt, *A review of scientific research in defensive cyberspace operation tools and technologies*, Journal of Cyber Security Technology, 3 (2019), pp. 1–46.

[10] I. Kamwa, *Dynamic wide area situational awareness: Propelling future decentralized, decarbonized, digitized, and democratized electricity grids*, IEEE Power and Energy Magazine, 21 (2023), pp. 44–58.

[11] P. R. Kshirsagar, R. K. Yadav, N. N. Patil, et al., *Intrusion detection system attack detection and classification model with feed-forward lstm gate in conventional dataset*, Machine Learning Applications in Engineering Education and Management, 2 (2022), pp. 20–29.

[12] B. Roy and H. Cheung, *A deep learning approach for intrusion detection in internet of things using bi-directional long short-term memory recurrent neural network*, in Proceedings of the 28th international telecommunication networks and applications conference, IEEE, 2018, pp. 1–6.

[13] M. Saratchandra and A. Shrestha, *The role of cloud computing in knowledge management for small and medium enterprises: a systematic literature review*, Journal of Knowledge Management, 26 (2022), pp. 2668–2698.

[14] C. Wang, Z. Li, R. Outbib, M. Dou, and D. Zhao, *A novel long short-term memory networks-based data-driven prognostic strategy for proton exchange membrane fuel cells*, International Journal of Hydrogen Energy, 47 (2022), pp. 10395–10408.

[15] W. Wang, Y. Lei, T. Yan, N. Li, and A. Nandi, *Residual convolution long short-term memory network for machines remaining useful life prediction and uncertainty quantification*, Journal of Dynamics, Monitoring and Diagnostics, 1 (2022), pp. 2–8.

[16] D. Wu, *A network security posture assessment model based on binary semantic analysis*, Soft Computing, 26 (2022), pp. 10599–10606.

[17] I. Yaqoob, K. Salah, R. Jayaraman, and Y. Al-Hammadi, *Blockchain for healthcare data management: opportunities, challenges, and future recommendations*, Neural Computing and Applications, (2021), pp. 1–16.

[18] E. Zio, *Challenges in the vulnerability and risk analysis of critical infrastructures*, Reliability Engineering & System Safety, 152 (2016), pp. 137–150.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 13, 2023
Accepted: Sep 17, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 137–145, DOI 10.12694/scpe.v25i1.2281

ANALYSIS AND APPLICATION OF BIG DATA FEATURE EXTRACTION BASED ON IMPROVED K-MEANS ALGORITHM

WENJUAN YANG*

Abstract. This paper addresses the challenges modelled by collecting and storing large volumes of big data, focusing on mitigating data errors. The primary goal is to propose and evaluate an enhanced K-means algorithm for big data applications. This research also aims to design an extensive energy data system to demonstrate the improved algorithm's practical utility in monitoring power equipment. The research begins with an in-depth analysis of the traditional K-means algorithm, culminating in the proposal of an improved version. Subsequently, the study outlines developing a comprehensive, extensive energy data system, encompassing architectural aspects such as data storage, mechanical layers, and data access structures. The research also involves the development of a power big data analysis platform, incorporating the improved algorithm for clustering and analyzing power equipment monitoring data. Experimental results reveal that the proposed improved K-means algorithm outperforms the traditional version, with significantly improved accuracy and reduced classification errors, achieving an error rate of less than one. The improved K-means algorithm showcased remarkable enhancements, achieving a meagre misclassification rate of just 0.08% while substantially boosting accuracy levels, consistently exceeding 95% across all datasets. Moreover, the power big data system developed in this study to meet practical requirements while enhancing storage and processing efficiency effectively.

Key words: Big data, K-means algorithm, Data errors, Power equipment monitoring, Data analysis platform, Pollution detection

1. Introduction. In today's digital age, the proliferation of data generation points, primarily through data collection terminals like sensors, has led to an unprecedented surge in the sheer volume of data. Consider the well-known online giants such as Facebook, Google, Yahoo, and Baidu to put this exponential growth into perspective. These titans of the internet realm grapple with the monumental task of processing hundreds of petabytes of data each day. Likewise, global retail giants, including Wal-Mart, Carrefour, and TESCO Group, must efficiently handle millions of user requests every hour. In particle physics, exemplified by the Large Hadron Collider (LHC) since 2008, annual data production has consistently exceeded 25 petabytes. This explosion of big data presents a dual opportunity and challenge. On the one hand, it furnishes humanity with a prosperous source of information, empowering us to comprehend and exert control over the physical world to an unprecedented extent. On the other hand, this overflow of data places ever-mounting demands on server systems' processing power and efficiency [11].

Once the torrents of big data inundate the servers, a crucial step involves streamlining their processing. To optimize the efficiency of handling vast datasets, a rational approach involves categorizing big data into meaningful groups, enabling the provisioning of similar data to processing terminals with analogous functions. This systematic classification and allocation enhance the overall efficacy of data processing, ensuring that the colossal influx of information can be harnessed and transformed into actionable insights with remarkable efficiency [16].

The K-means algorithm, a cornerstone in the field of clustering techniques, was initially conceptualized by MacQueen. This classic algorithm is celebrated for its simplicity, computational efficiency, and remarkable clustering capabilities. At its core, K-means assigns each data point to the cluster whose centroid is closest in terms of Euclidean distance. However, K-means bears a notable limitation—its categorical rigidity. It operates on a strict partitioning principle, obliging every data object to be unequivocally assigned to a single cluster. The quality of its clustering outcomes hinges heavily upon the initial placement of cluster centres and the predetermined number of clusters [4].

*Shanghai Zhongqiao Vocational and Technical University, Shanghai, 201514, China (wenjuanyang5@163.com).

The Fuzzy C-means (FCM) algorithm is a more nuanced and adaptable alternative to address some of these limitations inherent in the K-means algorithm. Building upon the foundations of K-means, FCM introduces the concept of fuzzy membership. In stark contrast to K-means' rigid assignments, FCM liberates each element from the confines of strict cluster boundaries. Instead, it allows data points to exhibit degrees of membership or confidence in multiple clusters simultaneously. This intrinsic flexibility enables FCM to capture real-world data distributions' nuanced, overlapping nature. In the FCM algorithm, each data point is not restricted to a single cluster but instead conveys its affinity or level of confidence for each cluster. It is achieved by assigning continuous numbers between 0 and 1 membership values. These membership values indicate how much a data point belongs to each cluster, reflecting the inherent uncertainty or fuzziness in many practical scenarios [5].

By introducing fuzzy memberships, the FCM algorithm accommodates datasets with intricate patterns and substantial overlap and provides a more granular and nuanced representation of data relationships. This adaptability and finesse make FCM a powerful extension of the K-means algorithm, particularly well-suited for applications where data points may exhibit varying degrees of association with multiple clusters, as is often the case in complex real-world datasets [18].

The paper is organized as follows: Section 2 presents a thorough literature review, critically assessing prior work in big data and the K means algorithm. Section 3 outlines the proposed method, exploring the details of the improved K-means algorithm for feature extraction from big data. Section 4 comprehensively presents results obtained through experiments and engages in a robust discussion of these findings. Finally, Section 5 concludes the paper by summarizing key insights highlighting the contributions in big data feature extraction and analysis.

2. Literature Review. The comprehensive implementation and modernization of electricity collection systems have ushered in an era where traditional manual on-site meter readings are largely past. While this transition has undeniably boosted meter reading efficiency and dramatically curtailed labour costs, it has also brought about an unintended consequence - a reduced frequency of direct interactions between power supply authorities and consumers. Consequently, this diminished engagement has created a potential blind spot in promptly and accurately ascertaining users' actual electricity consumption behaviours, rendering them susceptible to electricity theft. Electricity theft, a clandestine practice with various modus operandi, has the unfortunate consequence of distorting real-time electricity usage data. Fortunately, the immense volume of data on residents' electricity consumption is harnessed in the age of fully integrated electricity collection systems. Leveraging advanced artificial intelligence algorithms, this wealth of big data is meticulously analyzed, thereby facilitating the effective identification of irregular electricity usage patterns among consumers who deviate from the norm [8].

In a related domain, K-means clustering and Haar wavelet transform underpin a novel optimal heart sound segmentation algorithm [17]. This innovative algorithm comprises three integral components, each contributing to a more precise and refined heart sound segmentation process. Concurrently, an advanced Orthogonal Matching Pursuit (OMP) technology significantly enhances existing methodologies. Building on this foundation, Prabhakar has replaced the K-SVD technique with K-means clustering and the Method of Optimal Direction (MOD) technology, yielding six distinctive combinations in sparse representation optimization [13].

Meanwhile, the applications of the GB-BP neural network algorithm are explored in wrestling. This research resulted in the development of a sports athlete action recognition and classification model based on the GB-BP neural network algorithm. Wang's work commenced with a comprehensive analysis of the current state of wrestling action recognition, subsequently addressing and enhancing the limitations of existing action recognition and big data analysis techniques in the domain. Through these diverse endeavours, innovative solutions and algorithmic advancements are emerging to tackle complex problems across a spectrum of domains, driven by the growing availability and utilization of big data [15].

In recent years, there has been a notable surge in research focused on big data, underscored by its profound significance in shaping the design and implementation of cutting-edge solutions across diverse applications. This surge is particularly pertinent when addressing big data's current status and challenges in various domains. In alignment with this overarching trend, the author of this study has embarked on an ambitious endeavour. The core objective of this research is to develop a robust and versatile big energy data analysis platform meticulously tailored to address the specific and evolving needs of the big energy data landscape [14].

At the heart of this initiative lies the aspiration to empower real-world analysis of big energy data, uncovering valuable insights and patterns that might otherwise remain concealed within the vast data reservoirs. By harnessing the capabilities of this platform, stakeholders can effectively pinpoint and identify crucial messages and information pertinent to energy equipment pollution. This, in turn, is a pivotal step in the larger mission to ascertain the presence and extent of equipment pollution.

Crucially, the platform leverages advanced data analysis techniques, including but not limited to the utilization of cutting-edge K-tools. These tools are instrumental in systematically collecting and analyzing data about energy equipment pollution. The platform can discern intricate patterns and anomalies within the data by employing K-means clustering and related methodologies, thereby facilitating accurate determination of equipment pollution [21].

Furthermore, the insights from this comprehensive analysis pave the way for the platform to offer tailored and customized advice. This advice is indispensable in ensuring the safety and stability of electricity usage, a paramount concern in modern energy management. In essence, this research endeavour underscores the critical role that big data plays in our contemporary world. By developing a purpose-built platform, the author contributes to advancing big energy data analytics and equips stakeholders with the tools and insights needed to navigate the complex landscape of energy equipment pollution. Ultimately, this work aligns with the broader trajectory of harnessing the power of big data to inform and optimize decision-making across many domains.

3. Proposed Improved K-means Algorithm.

3.1. Principle of improved K-means algorithm. The author introduces an upgraded version of the K-means algorithm to enhance the analysis of monitored power big data. This algorithm enhancement's essence lies in altering the conventional K-means clustering rules. Expressly, during the computation of the distance from the centroid, a novel component, represented as the particle weight proportion 'w', is incorporated. This addition enables a data point's category assignment to be determined based on the magnitude of the distance, effectively refining the clustering process.

Choose 'k' centre points from the dataset 'm', determine the cluster to which the remaining points belong, compute the mean for each cluster as the new centre point, and iterate this process until convergence. The algorithm's procedural steps can be summarized as follows:

Training sample $\{x_1, \ldots, x_m\}, x_i \in R^n$, divide it into k categories:

Step 1: Randomly select k out of m sample data: $\mu_1, \mu_2, \ldots, \mu_k \in R^n$;

Step 2: Calculate the distance between the remaining data and these k data separately;

$$C_i = \arg\min_j \|x_i - \mu_j\|^2 \tag{3.1}$$

Step 3: Redetermine the centre point of each class and recalculate the average value;

$$\mu_j = \frac{\sum_{i=1}^{m} \{C_i = j\} x_i}{\sum_{i=1}^{m} \{C_i = j\}} \tag{3.2}$$

Step 4: If the measurement function converges, terminate the program; Otherwise, continue with Step 2. The improved K-means method minimizes the evaluation function fitness $(A[1], A[2], \cdots, A[n])$.

$$\text{fitness}(A[1], A[2], \ldots, A[n]) = \sum_{i}^{k} \sum_{i}^{n} \text{Dist}\,(x_i, C_k) \tag{3.3}$$

$$\sum_{i}^{k} \sum_{i}^{n} \text{Dist}\,(x_i, C_k) = \cos\,(x_i, C_k) = \frac{\sum_1^m x_{ij} c_{kj}}{\sqrt{\sum_1^m x^2 ij \sum_1^m c^2{}_{kj}}} \tag{3.4}$$

In the formula: n is the number of data; $\text{Dist}\,(x_i, C_k)$ is the distance between x_i and the centre point C_k. The process of implementation is to generate C_k and continuously improve C_k based on the value of x_i, so that

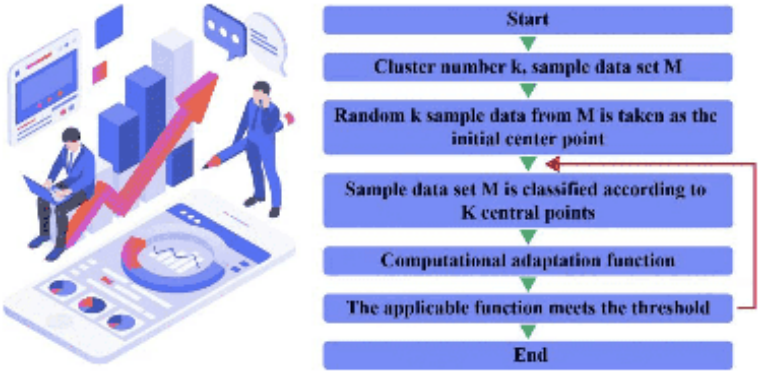


Fig. 3.1: Process flow of the improved K-means algorithm

Table 3.1: Experimental data attributes

Data set	Centre point	Number	Radius	Covariance
1	[0,0,0]	100	2	[0.300;00.350;000.3]
2	[1.25,1.25,1.25]	100	2	[0.300;00.350;000.3]
3	[-1.25,-1.25,-1.25]	100	2	[0.300;00.350;000.3]

data of the same class is more clustered and finally reaches the convergence condition.

$$C_k = \frac{\sum_{i=1}^{N} C_{ki} x_{ki}}{\sum_{i=1}^{N} C_{ki}} \tag{3.5}$$

The improved K-means algorithm introduces ω_k, and the implementation process of the same is expressed as:

$$\omega_k = \frac{1}{\sqrt{\omega_k}} \tag{3.6}$$

ω_k is the standard deviation, and when the target is C_k, the function increases by ${\Delta\varepsilon_k}^2$ after k times.

$$\Delta\varepsilon_k = \frac{1}{2m} \left(\omega_k \cdot \operatorname{dist}\left(c_k, x\right)\right)^2 \tag{3.7}$$

To assess the precision of the enhanced K-means classification method, we selected three distinct datasets for a comparative analysis of their clustering outcomes. The process flow of the improved K-means algorithm is shown in Figure 3.1.

The data attributes of the improved K-means algorithm [12] are detailed in Table 3.1, and the comparative evaluation of classification accuracy is presented in Table 3.2.

Table 3.2 reveals when applied to the classification of identical datasets, the enhanced K-means method consistently demonstrates significantly lower error rates than the traditional K-means approach, concurrent with a substantial increase in accuracy [10]. This compelling evidence underscores the efficacy of employing the improved K-means method in analyzing power monitoring equipment pollution big data, as it enables swift and precise clustering.

3.2. Design of power big data system. The following are the steps involved in data collection and processing.

Table 3.2: Comparison of data classification errors

Algorithm	Misclassification rate	Accuracy (Data 1)	Accuracy (Data 2)	Accuracy (Data 3)
K-means	13.21%	90.26%	91.94%	91.53%
Improve K-means	0.08%	95.21%	96.17%	95.82%

Fig. 3.2: Process flow in power big data system

(1) System architecture

The architecture of the power big data system encompasses various components, including data access, the mechanical floor, data storage, and data calculation. The mechanical floor comprises essential elements such as switching equipment and servers. Data access, on the other hand, encompasses multiple modules, including gateways, load balancing, and message middleware. The web management component enables diverse functionalities, including gateway management, terminal management, user management, and the parsing of original terminal messages [2]. During the data access, users can subscribe to data via message middleware, facilitating data analysis and storage.

The data import service also empowers users to import data of interest into storage, supporting various storage methods such as Hadoop, Redis, and Rdbms [20]. The platform monitoring component plays a pivotal role in overseeing the operational status of nodes and offers comprehensive monitoring across various aspects, including business, software, and systems. It also supports alarm notifications, including SMS and email.

(2) System data flow

The step-by-step procedure of the process of system data flow is represented in Figure 3.2 and explained as follows:

1) The terminal creates a long connection through LVS load balancing and gateway;

2) The gateway uses data packet decoding to encapsulate platform general data and writes it into Kafka;

3) Realize subscription of Kafka raw message data through real-time computing module and achieve data parsing; the parsed data is written in Kafka;

4) Subsequent modules can subscribe to the raw data of the terminal through Kafka or analyze the data. Thus, data can be stored and analyzed offline [9];

5) The forwarding service subscribes to data through Kafka and forwards it to other platforms;

6) The business management platform utilizes a data exchange interface to access the big data collection access module.

(3) Processing of streaming data

In the power big data system context, flow data pertains to the continuous stream of real-time data generated throughout power production, monitoring, and operational processes. The computing infrastructure employed for this purpose is the Storm system, which enhances real-time data analysis by persistently storing the computation outcomes in HBase. Stream data analysis, as a critical component, encompasses the processing, acquisition, and storage of streaming data within the framework of the power big data platform, delineating the interplay among data sources, processing stages, and computing platforms. The details of this intricate

Fig. 3.3: Module structure of power big data system

process are elucidated in Figure 3.3 to provide a comprehensive understanding of the processing workflow.

The data collection and preprocessing functions serve the crucial role of facilitating the precise, comprehensive, and instantaneous acquisition of data. These functions merge disparate datasets, encompassing incomplete data, various formats, and noisy inputs, to yield standardized data through noise reduction techniques. Subsequently, data processing applies specialized business logic to analyze standardized data thoroughly. Users have the flexibility to craft custom implementations through dedicated interfaces. Ultimately, the outcomes of these processes find their repository in HBase for storage and retrieval [19].

(4) Data collection and preprocessing

The power big data platform is a comprehensive data fusion platform, aggregating data from electricity generation and consumption domains, including sources like SCADA and energy metering systems. Notable examples of flow data encompass power equipment status monitoring data. Figure 3.4 provides a visual representation of the data collection and processing workflow. Upon gathering data from diverse facets of the power system, it is initially stored on an FTP server. Storm does not impose rigid constraints on data sources and formats, accommodating input types such as message queues, databases, and log files [6]. All that is required is implementing the corresponding interface in Spout.

(5) Result storage

Diverse data types exist within the power big data platform, each characterized by intricate structures and substantial volume. Consequently, adopting a multi-tiered storage system is imperative to cater to the varied demands of different business operations. This approach allows the platform to store its extensive big data reserves in alignment with performance and analytical requisites. For instance, data with substantial volumes and unstructured attributes, such as monitoring video data, finds its storage solution through the HDFS file system. In contrast, real-time processed data is efficiently stored in HBase, with its storage structure meticulously designed, as illustrated in Table 3.3. This strategic approach to data storage optimizes the platform's ability to handle and retrieve data following specific operational needs.

4. Experimental Results and Analysis.

4.1. System development. The power big data platform's requisites and constituent modules were meticulously examined to develop a robust power big data analysis platform. To execute system analysis and calculation tasks, a B/S architecture was employed. This framework guides users through a user-friendly interface to make crucial selections. These selections encompass picking data files, opting for intelligent algorithms, and configuring pertinent parameters. The system then seamlessly executes the chosen processes, automatically analyzing the data and presenting the results through intuitive icons for user comprehension and interpretation.

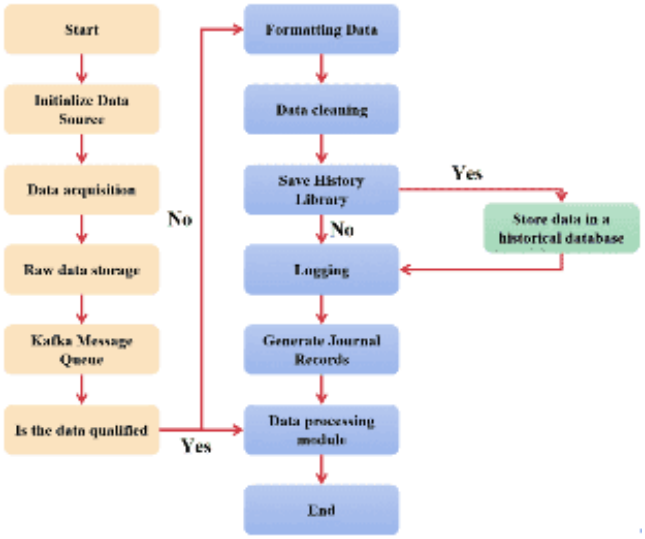


Fig. 3.4: Flow of data collection and processing

Table 3.3: HBase table storage data

Line keywords	Column cluster		Specific values of the N sampling points		
	Temperature	Dampness	V1	V2	Vn
Mac1+id	20	55%	13	12	16
Mac2+id	23	58%	14	12	15
Mac3+id	22	54%	16	14	18

4.2. System based on improved algorithms. The author monitors a specific power system daily, gathering data regularly. Whenever the environmental monitoring value surpasses the threshold of 0.1, it signifies a potential safety issue within the power grid during operation. The author harnesses an enhanced K-means algorithm to gauge the pollution status of various locations and ascertain equipment-related pollution situations.

In the initial step, the power big data analysis platform is employed to amass data about site pollution. This data is subjected to mean and variance analyses to derive valuable insights. Subsequently, the improved K-means method is applied to conduct a clustering analysis of site pollution levels. The outcomes of this analysis are vividly depicted in Figure 4.1, showcasing the clustering analysis results for power environmental monitoring stations [7].

An evaluation of the data representation within the system reveals that user identity verification is a fundamental security measure. Access to the system is granted solely upon entering the correct account and password [3, 1]. Once inside the system, users can select data files and intelligent machine learning algorithms, fine-tune parameters, and initiate data analysis. The analysis results are then elegantly presented through charts and tables, offering decision-makers precise information. This capability greatly facilitates the assessment of the power system's performance.

These findings in Table 4.1, which presents the outcomes of site classification. Following the research detailed in this article, it becomes evident that data values indicate pollution levels, with heavily polluted sites yielding the highest numerical values, moderately polluted data falling in the middle range, and lightly polluted data exhibiting the smallest numerical values. This distinct differentiation between the three data categories is accompanied by pronounced periodicity in heavily polluted data, in contrast to minimal fluctuations in lightly

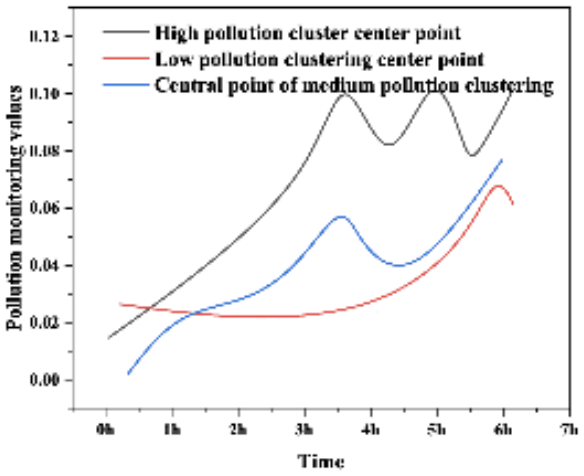


Fig. 4.1: Cluster analysis of power environmental monitoring stations

Table 4.1: Results of site classification

Classification	Pollution level	Number of sites	Processing strategy
1	Mild	353	Heavy
2	Degrees	83	Observation
3	Heavy	15	Regular cleaning

polluted sites. The extensive data mining efforts to monitor power station equipment pollution status enable timely equipment updates.

5. Conclusion. A comprehensive analysis system with both front-end and back-end components has been designed and successfully developed in response to power big data analysis requirements. The proposed system automatically performs data analysis and presents results in graphical form by guiding users through a structured process, including data file selection, intelligent algorithm choice, and parameter configuration. The improved K-means method has led to quantifiable advancements, including enhanced accuracy and a substantial reduction in misclassification rates compared to traditional K-means algorithms. When applied to classifying pollution levels in power equipment, this method significantly improves determining the true state of power equipment, thereby furnishing actionable insights for power equipment management. This comparative analysis shows that the traditional K-means algorithm yielded a relatively higher misclassification rate of 13.21% and slightly lower accuracy rates across all three datasets. These outcomes undeniably highlight the superior performance and efficacy of the improved K-means algorithm in the precise clustering and classification of data.

REFERENCES

[1] M. Aljasem, A. Irtaza, H. Malik, N. Saba, A. Javed, K. M. Malik, and M. Meharmohammadi, *Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging*, IEEE Transactions on Information Forensics and Security, 16 (2021), pp. 3524–3537.

[2] W. Bai and J. Liu, *Analysis of test scores of insurance salesman based on improved k-means algorithm*, in Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, vol. 153, Springer, 2022, pp. 1192–1201.

[3] W. Bao, K. Wang, X. Gao, J. Song, P. Zeng, D. Zhou, H. Zhu, and Q. Geng, *Verification of security measures for smart substations based on visualized simulation*, in IOP Conference Series: Earth and Environmental Science, vol. 461, IOP Publishing, 2020, p. 012002.

[4] S. Bo, *Application of k-means clustering algorithm in evaluation and statistical analysis of internet financial transaction data*, arXiv preprint arXiv:2202.03146, (2022).
[5] Z. Dong, Y. Men, Z. Li, Z. Liu, and J. Ji, *Chilling injury segmentation of tomato leaves based on fluorescence images and improved k-means++ clustering*, Transactions of the ASABE, 64 (2021), pp. 13–22.
[6] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72, Springer, 2015.
[7] J.-C. Huang, P.-C. Ko, C.-M. Fong, S.-M. Lai, H.-H. Chen, and C.-T. Hsieh, *Statistical modeling and simulation of online shopping customer loyalty based on machine learning and big data analysis*, Security and Communication Networks, 2021 (2021), pp. 1–12.
[8] W. A. Ibrahim and M. M. Morcos, *Artificial intelligence and advanced mathematical tools for power quality applications: a survey*, IEEE Transactions on Power Delivery, 17 (2002), pp. 668–673.
[9] X. Li, *Research on english teaching ability evaluation algorithm based on big data fuzzy k-means clustering*, in Proceedings of the EAI International Conference, BigIoT-EDU, vol. 467, Springer, 2022, pp. 36–46.
[10] Y. Li, R. Liu, Y. Bo, and H. Wei, *Analysis and research of students' mental health status based on k-means clustering under the background of big data*, in Proceedings of the International Conference on Cognitive based Information Processing and Applications, vol. 85, Singapore, 2022, Springer, pp. 437–444.
[11] W. Lv, W. Tang, H. Huang, and T. Chen, *Research and application of intersection clustering algorithm based on pca feature extraction and k-means*, in Proceedings of the 5th International Workshop on Advanced Algorithms and Control Engineering, vol. 1861, Zhuhai, China, 2021, IOP Publishing, p. 012001.
[12] X. Meng, J. Lv, and S. Ma, *Applying improved k-means algorithm into official service vehicle networking environment and research*, Soft Computing, 24 (2020), pp. 8355–8363.
[13] S. K. Prabhakar and S.-W. Lee, *Improved sparse representation based robust hybrid feature extraction models with transfer and deep learning for eeg classification*, Expert Systems with Applications, 78 (2022), pp. 18023–18050.
[14] F. Terroso-Saenz, A. González-Vidal, A. P. Ramallo-González, and A. F. Skarmeta, *An open iot platform for the management and analysis of energy data*, Future Generation Computer Systems, 92 (2019), pp. 1066–1079.
[15] L. Wang, K. Qiu, and W. Li, *Sports action recognition based on gb-bp neural network and big data analysis*, Computational Intelligence and Neuroscience, 15 (2021), pp. 795–798.
[16] X. Wang, C. Song, and M. Yu, *Research on power security early warning system based on improved k-means algorithm*, in Proceedings of the Big Data and Security - Third International Conference, vol. 1563 of Communications in Computer and Information Science, Shenzhen, China, 2021, Springer, pp. 73–89.
[17] X. Xu, X. Geng, Z. Gao, H. Yang, Z. Dai, and H. Zhang, *Optimal heart sound segmentation algorithm based on k-mean clustering and wavelet transform*, Applied Sciences, 13 (2023), pp. 547–552.
[18] J. Ye, J. Zou, J. Gao, G. Zhang, M. Kong, Z. Pei, and K. Cui, *A new frequency hopping signal detection of civil uav based on improved k-means clustering algorithm*, IEEE Access, 9 (2021), pp. 53190–53204.
[19] C. Zhu, Z. Liu, B. Zou, Y. Xiao, M. Zeng, H. Wang, and Z. Fan, *An hbase-based optimization model for distributed medical data storage and retrieval*, Electronics, 12 (2023), p. 987.
[20] N. Zhu and Q. Dai, *Basketball data analysis based on spark framework and k-means algorithm*, in Proceedings of the International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy:, vol. 98, Springer, 2022, pp. 853–857.
[21] J. Zhuang, C. Ren, D. Ren, Y. Li, D. Liu, L. Cui, G. Tian, J. Yang, and J. Liu, *A novel single-cell rna sequencing data feature extraction method based on gene function analysis and its applications in glioma study*, Frontiers in Oncology, 11 (2021), p. 797057.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: May 13, 2023
Accepted: Sep 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 147–155, DOI 10.12694/scpe.v25i1.2329

INTELLIGENT PREDICTION OF NETWORK SECURITY SITUATIONS BASED ON DEEP REINFORCEMENT LEARNING ALGORITHM

YAN LU*, YUNXIN KUANG†, AND QIUFEN YANG‡

Abstract. The limitations of traditional network security assessment methods characterized by manual definitions and measurements, data overload, poor performance, and non-negligible drawbacks are addressed in this research. A novel network security system employing a deep learning algorithm is proposed to overcome these challenges. The research unfolds in three key phases. First, a deep self-encoding model is developed to distinguish various network attacks effectively. Subsequently, the creation of missing measurement weights enhances pattern detection, even when dealing with a limited number of training samples. Finally, the model assesses and computes attack issues, assigns impact scores to each attack, and determines the overall network security value. Experimental results demonstrate that the deep auto encoder-based deep neural network (DAEDNN), in conjunction with the proposed unique oversampling weighting (UOSW) algorithm, significantly outperforms traditional methods such as decision trees (DT), support vector machines (SVM), and long short-term memory (LSTM) models. The F1 score of UOSW surpasses these models by approximately 2.77, 10.5, and 5.2, respectively. The deep self-encoding model employed in the proposed system offers superior accuracy and recall rates, leading to more precise and efficient measurement results.

Key words: Network Security, Deep Learning, Deep Self-Encoding Model, Security Assessment, Attack Detection, UOSW Algorithm

1. Introduction. The issue of network security has now evolved into a global concern. Managing the network security landscape and fostering effective network collaboration have become imperative topics that warrant collective discussion among nations worldwide. In the growing "Global Village era", network security assumes an increasingly critical role. The author contends that prioritizing prevention over management is paramount in controlling the network environment or establishing a robust framework of order. Consequently, this study focuses on real-time network security issue prediction. The design and optimization of the proposed model demonstrate its efficacy in timely forecasting network security problems [10].

Network security forecasting encompasses continuously monitoring the ecosystem and anticipating potential network issues, such as viruses and trojans, to safeguard computer security. Through proactive network security prediction, users can identify potential threats within the current network, maintain a historical record of related incidents, and select data patterns with specific characteristics that could impact current network security concerns. Utilizing mathematical models, this approach enables the prediction and tracking of the formation and progression of network security issues. Ultimately, this process furnishes reliable information for effective computer security management, ensuring users' online safety [16].

The term "network security problem" primarily revolves around the functionality of network equipment and the extent to which the actions of network users might jeopardize network security. In simpler terms, it encapsulates the security status of an operational network. Notably, the widespread adoption of the Internet has underscored the critical role of networks in our professional endeavours and daily lives. Consequently, numerous network security challenges have emerged, impacting a broad spectrum of users. Moreover, the patterns of network access and cyberattacks have evolved significantly, marked by diverse interactions and significant developments [11].

To anticipate the state of network security more effectively for safeguarding the network environment. Such anticipation is the foundational step toward comprehending the prevailing network security challenges. We can

*Hunan Open University, Changsha, Hunan, 410004, China
†Hunan Railway Professional Technology College, Zhuzhou Hunan, 412001, China (Corresponding author: yunxinkuang6@163.com)
‡Hunan Open University, Changsha, Hunan, 410004, China

harness this data as a foundational resource for forecasting the security landscape by extracting pertinent information regarding network security issues. This entails predicting the evolution of network security issue configurations over time within the dynamic realm of network attacks and defences, thereby enabling the proactive prevention of real-time network threats and the resolution of security concerns while enhancing prediction accuracy [3].

In practical forecasting, the landscape of attacks and vulnerabilities constantly evolves, posing a formidable challenge for security managers to address them effectively. Concurrently, monitoring the other four categories of network security content typically necessitates the application of time-quantitative analysis algorithms to forecast their security issues. Consequently, these four aspects of network protection stand as viable strategies to enhance network security measures [13].

Security personnel are crucial in furnishing accurate insights into evolving security landscapes. Alongside ensuring the consistent operation and enduring lifecycles of safety measures, short-term assessments of security measures remain equally vital. Consequently, those responsible for security management should proactively engage in the timely analysis of security systems. This proactive approach should extend to future-proofing network security configurations, accounting for application impact, prevailing work environments, and potential changes [7].

The structure of the paper is as follows: A comprehensive literature review, surveying existing knowledge and research in network security prediction, is explored in section 2. Section 3 elaborates on the proposed method, which centres around the deep reinforcement learning algorithm for intelligent prediction of network security situations. Section 4 presents the results obtained from applying the proposed method, and Section 5 summarizes the key insights and contributions of the proposed methodology.

2. Literature Review. Numerous factors exert influence over the four key facets of network security issues. Regarding asset configuration, location, quantity, and priority play a central role. Meanwhile, the bedrock of business configuration hinges on metrics such as the frequency of changes and the application of business processes. The topology model, on the other hand, is chiefly influenced by the number of switching nodes and connections. Security policies' impact typically involves parameters like the frequency of changes and access control rights within security settings. Consequently, these dynamic factors remain in constant flux, and their fluctuations carry profound implications for the future security of the network [12].

When there are alterations in the topology structure, previously feasible or infeasible attack paths can transition into feasible or unfeasible ones, consequently impacting all aspects of network security. Hence, a comprehensive analysis of the four dimensions of network security issues mentioned above should be conducted through a temporal lens to gain deeper insights into future network security challenges. In practical forecasting, utilizing the variability in distribution patterns as a basis for prediction and estimation is essential. Building upon this foundation, contrasting the natural layers of forthcoming network content should serve as the framework for budgeting and prediction. Only through this approach can we enhance the accuracy of calculated location values and establish a dependable groundwork for predicting and analyzing network security issues within site maintenance [5].

The users can discern issues within the network, dissect the root causes of these problems, identify data points indicative of network security concerns, and anticipate the trajectory of network security challenges. This process entails the development of mathematical models, fostering computer network security, and disseminating information to ensure the network environment's safety. In cases where the target under examination is extensive or the model exhibits complexity, it is often necessary to define the target through specific conditions [8].

The term "event" originally found its usage in the context of military actions. In military settings, it is commonplace to forecast the unfolding of complex, multifactorial scenarios. However, within the realm of information network construction, if the aim is to create a secure and dependable network environment, there must be shared awareness of the network's current status and a comprehensive understanding of its overall security posture. By examining the frequency, volume, and nature of network security events, we investigate the threat these events pose to the network. This amalgamation of principles about different aspects of information network security informs the broader picture. It provides a summarized narrative of the current state of network security, facilitating predictions regarding its future security trajectory [1].

The analytical focus of the model encompasses extensive and intricate subjects, and the term "condition"

frequently comes into play to characterize the inherent nature of the subject under scrutiny. Originating from a military context, "accident" refers to anticipating complex developments and situational dynamics similar to military operations. In the context of research on information network security, these terms are applied to address the challenges surrounding the creation of secure and resilient networks. This approach enhances our comprehension of network security across networked environments [9].

The swift advancement of artificial intelligence, internet technology, and automatic recognition has been accompanied by deep learning methods such as decision tree classification algorithms, gradient classification algorithms, neural networks, and convolution, rapidly finding applications in computer network security recognition. As a result, they substantially bolster the computing capacity for network security data and enhance data information efficiency, thereby expanding the domain of computer network security applications in deep learning [6].

Building upon the foundation of deep learning, the aim is to attain efficient, precise, and high-quality computer network security identification and management technology. This is achieved by harnessing the algorithmic advantages of deep learning, including feature vector extraction, recognition, information optimization, and classification. A secure, cost-effective, intelligent system is meticulously designed and developed through systematic analysis encompassing its principles, architecture, functional characteristics, and platform implementation. This system furnishes a scientifically sound reference for the design, execution, and application of the computer network security identification management system, incorporating deep learning algorithms' principles [14].

The prediction of network security issues represents a technology that accomplishes a genuine amalgamation of historical event data from diverse sources. It entails the analysis of interconnected events, comprehensive research, and informed decision-making concerning the development of network security. This domain is the primary focus within knowledge technology research on network security challenges. In tandem with the continual evolution of network attacks and defensive strategies, the landscape of network security constantly evolves, heightening the expectations and complexities associated with network attack scenarios. Consequently, the demand for predictive accuracy and real-time responsiveness in network security defences has surged significantly [15].

Renowned for its emphasis on scientific rigour, knowledge acquisition, and security management with a focus on scalability, this research endeavours to gain a deeper understanding of the intricacies and functionality of deep learning algorithms. The overarching goal is to design and implement cutting-edge technologies within computer network security management. By applying these design principles and deploying innovative techniques, this study aims to elevate the stature of network security analysis technology within the framework of deep learning. Furthermore, it seeks to enhance the utilization of deep learning methodologies alongside management control technology. The research process entails designing, predicting, and analyzing computer network security through neural networks and problem-solving training. In contrast to traditional security management processes, deep learning demonstrates remarkable improvements in prediction accuracy for security measures and overall management performance. Additionally, in terms of data augmentation, machine learning is leveraged to expand the capabilities of existing security tools, preconfigure security enhancements, and develop new security protection functionalities, thereby comprehensively fortifying the scope of computer network security management [4].

In light of the shortcomings of the methods above, the author introduces a novel approach to network security analysis rooted in deep learning. To address the challenge of assigning low values to different attack types within the dataset, a weighted data processing algorithm, referred to as UOSW, is proposed. Additionally, network attacks are disentangled by utilizing an energy-absorbing auto-encoder. This methodology involves the assessment of the impact of each attack type when deriving network attack classifications and evaluating network security. Experimental results underscore the efficacy of the author's model in conducting real-time network security assessments. The evaluation outcomes exhibit superior performance, intelligence, and overall performance metrics compared to alternative models [2].

3. Proposed Network Security Assessment Model. The network security analysis model consists of three parts: incident detection, analysis, and evaluation. The structure of the network security model is shown in Figure 3.1.

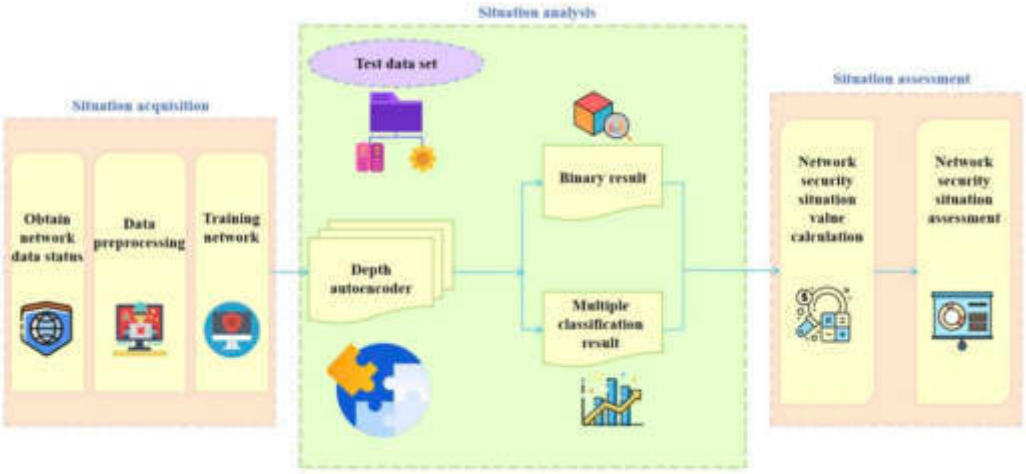


Fig. 3.1: Assessment model of network security

The merger of the three phases, namely "Nature of discovery", "Situation analysis", and "Condition assessment", constitutes what is commonly known as the "Network Security Assessment Process" or simply the "Security Assessment Process". This systematic approach is employed to assess and fortify the security of a network or information system.

(1) Nature of discovery

At this stage, traffic information on the network is received. The above NSL-KDD dataset is chosen as a road network to simulate a network processing large volumes of traffic data. After pre-planned data, the coder's deeply personal ideas in training.

(2) Situation analysis

Test datasets are integrated into training models, and binomial and multivariate distributions of test results are collected and used to calculate quantitative results for network security problems.

(3) Condition assessment

Based on the index attack distribution results, network attack probability and different network attack impact values are calculated. Also, estimate the cost of network security issues and evaluate network security issues. A detailed calculation model is presented below.

3.1. Model structure of depth self-encoder. Deep neural network (DNN) has been widely used in intrusion detection due to its accuracy and efficiency. Because DNN contains multiple hidden layers, its learning ability is significantly improved. The DAEDNN model can be applied not only to binary classification but also to multispecies. When performing a binary operation, the model function is a sigmoid function, and the model values range between 0 and 1.

$$F_{\mathrm{sgm}}(x) = (1 + e^{-x})^{-1} \tag{3.1}$$

When the model performs a multi-class task, the activation function of the model is the softmax function, which also maps the output to the range of 0 and 1:

$$F_{\mathrm{sfm}}(z_i) = e^{z_i} / \sum_{j=1}^{K} e^{z_j}, z = (z_1, z_2, \cdots, z_k) \tag{3.2}$$

K represents that the output can be divided into K classes, and Z_i represents the value obtained by each class.

3.2. Training of depth self-coder model. The training model is structured into three distinct stages:

Table 3.1: KDD - NSL data set information

Data set	Normal	DoS	Probe	R2L	U2R	Total
KDD Train+	68545	42615	10767	986	59	122972
KDD Test+	9812	7365	2530	2853	199	22759

1. The training data is initially input into the DAE network for specialized training, and the weight values are collected post-training.

2. After training the DAE model, the DAE model is integrated with the DNN model to form the DAEDNN model, which is then trained as a combined entity. To achieve the training results of the DAE model within the DAEDNN model, specific weight configurations for the DAE network are set, and the parameters of the DAE layer are designated for either learning or not learning the DNN network. During this phase, network updates are directed solely toward the DNN network.

3. Lastly, when the DAE layers encounter difficulties during training and updates for both the DAE and DNN networks are unsuccessful, adjustments to the training process are implemented. These adaptations serve the dual purpose of achieving desired learning outcomes for the DAE layer and enhancing the information message structure's visual capabilities.

3.3. Under-sampling weighted data resampling algorithm.

3.3.1. Data set description. In network security, the NSL-KDD configuration file is selected to evaluate the basic data entry for searching. The data sets used by the authors are shown in Table 3.1.

3.3.2. Under-sampling weighted data resampling algorithm. The distribution of training data among the five attack types is uneven, as illustrated in Table 3.1. Specifically, the "normal" category boasts the largest dataset, comprising 67,343 instances, while the "DoS" and "U2R" categories are comparatively limited, each consisting of only 52,995 data points. In training deep learning models, inadequate training data can hinder the model's capacity to capture all relevant data features. In contrast, an excessive amount of data may lead to overfitting. Essentially, the model learns essential features from the data, meaning that insufficient information can result in suboptimal model training, diminishing the accuracy of class recognition. In contrast, an abundance of information can lead to the opposite effect.

In data analysis, techniques like oversampling and undersampling are employed to rectify class imbalances within a dataset, commonly referred to as data resampling. Undersampling typically involves the removal of a subset of instances from an overrepresented category, while oversampling entails increasing the volume of data about a few specific data samples to achieve a more balanced dataset. To address the challenge of disparate data distribution and enhance the accuracy of minority class detection, the author introduces a weighted approach combining meticulousness, focus, and severity elements. The algorithm's steps are delineated as follows.

Set the original data set as S^1, the output data set as S^2, the data type to be resampled is $type_i$, and the original data set and sample number are S_i and x_i.

Step 1 Computing weights denoted as "A" for each data type within the dataset. When the data values for each class in the training network exhibit minimal variation, resulting in a narrow range from below the average to around the average, the network's recognition accuracy tends to be exceptionally high. Consequently, the author calculates the disparity between the observed values and the optimal model for each category, utilizing this discrepancy as a weight factor to attain a balance among all categories.

$$w_i = \sum_{i}^{n} x_i/(x_i \times n) \tag{3.3}$$

where n means that the dataset contains n categories.

Step 2 Data under-sampling: For the type with too much data, data under-sampling is performed to make the processed data sample close to the average. Use the "train_test_split" method of the sklearn library in

Table 3.2: Basic information on five types of attacks

Attack type	Description
Denial of service (DoS)	This attack renders a computer or network inoperable and unusable. A DoS attack does this by sending large amounts of traffic or data to a target.
Get permissions (User to Root, U2R)	The attack attempted to obtain the foundation's approval by illegal means.
Remote intrusion (Remote to Local, R2L)	This attack allows an attacker to access a local computer without logging in.
Detection attack (Probe)	This attack gathers network information as needed before other attacks can be launched.
Normal flow (Normal)	Normal network traffic

Table 3.3: Assessment of attack impact value

Index	Impact degree	Impact value
Confidentiality (C)	None (N) / Low (L) / High (H)	0 / 0.21 / 0.55
Integrity (I)	None (N) / Low (L) / High (H)	0 / 0.20 / 0.57
Usability (A)	None (N) / Low (L) / High (H)	0 / 0.21 / 0.58

Python to divide the data set S_i into two data sets $S_{\dot{r}\text{itrain}}$, S_{irremain}. Take S_{itrain} as the training set and add S^2, where the data volume of $S_{\dot{r}\text{itrain}}$ is $s_i = x_i \times w_i$; S_{iremain} is used for the next data oversampling operation, adding it to dataset S_{remain}.

Step 3 Data oversampling: The oversampling algorithm SMOTE is used in sampling groups with small data. The main goal of SMOTE is to create new models that are similar to existing models. The SMOTE algorithm was originally focused on two classification problems, and the following improvements were made to the algorithm due to the author's research on multi-classification problems.

(1) Merge other types of data: Combine the data set S_{remain} under-sampling processing in step 2 with a small number of data sets in the original data set, which is represented as S_{union}.

(2) Change the label. After (1), S_{union} contains data of n categories. Because the algorithm SMOTE is only for binary classification, it is necessary to distinguish the types that need oversampling from other types. Change the label of the dataset S_{union} to the same type, but different from the type.

(3) Determine the size of the data volume. To balance the data set, it is necessary to expand a small number of samples, set the expanded data size to s_i, where $s_i = x_i \times w_i$, w_i is the weight of the data type $type_i$.

(4) Data oversampling. Use the SMOTE method of the timber library in Python, combine it with other data types to generate the required data, and add it to S^2.

Repeat (1)-(4) until the oversampling operation is completed for all types whose data volume exceeds the average value.

3.4. Network attack impact value. NSL-KDD data set includes five types of network data: Normal, DoS, U2R, R2L and Probe. The basic information of the above attacks is shown in Table 3.2.

The attack impact value evaluation table is developed using the Common Vulnerability Scoring System (CVSS). The scores of confidentiality (C), integrity (I) and availability (A) are shown in Table 3.3.

The influence value (I_i) of each attack type is calculated as follows:

$$I_i = C_i + I_i + A_i \tag{3.4}$$

3.5. Quantification of the network security situation. The quantification of network security issues enables an in-depth exploration of various facets of a network. The author's approach to analyzing network

security problems typically comprises four integral components: attack analysis, estimation of attack impact, assessment of network performance security costs, and the quantitative evaluation of network security issues. The workflow for each of these sections is outlined as follows.

(1) Inspection

Randomly select several data groups from the test data set, input them into the DAEANN model, perform binary and multivariate classification, and note the distribution stops found in the second distribution.

(2) Calculate the attack impact value

The C, I, and A values of each type of attack are determined in Table 3.2 and Table 3.3, and the attack value is determined according to formula (3.5).

(3) Estimating network security issues

Network security is important to understand all the attacks on the network and the threat of each attack. Set the value of the network security problem.

$$T = \left\{ p \times \sum_{i}^{n-1} I_i \times t_i \right\} / (N - t_n) \tag{3.5}$$

Here, p is the attack probability in formula (3.1), n and N represent all n types of data and N samples, I_i represents the impact value of each type of attack, and t_i represents the number of attacks each attack, t_n represents the number of occurrences of the typical type. Because normal mode is a normal network data flow, it does not affect the network's confidentiality, integrity, or availability, so its impact score is 0, and it is only necessary to calculate the impact score of $n-1$ idle modes.

(4) Quantitative assessment of network security issues

According to the network security problem values of 0.00∼0.20, 0.21∼0.40, 0.41∼0.60, 0.61∼0.80 and 0.81∼1.00, the severity of network security problems is divided into five levels: security, low risk, medium risk, high risk and super risk.

4. Results and Discussion. The hardware environment for the experiment is Intel (R) Xeon (R) Silver processor, NVIDIAQuadroP2000 graphics card and 32GB memory. The training and test experiments were conducted on the Windows 64-bit operating system. The programming language and machine learning libraries used are Python 3.5 and TensorFlow 2.0. GPU accelerates the model's training and testing.

4.1. Evaluation indicators. The metrics used by the author are as follows.

True Positive (TP): Shows the number of models predicted to be correct when the model stalls.

False Positive (FP): A model is assumed to be normal but is a stopped model.

True Negative (TN): Refers to the number of samples that are expected to be normal and the sample to be normal.

Negative (FN): Indicates a time when the predicted sample is stopped but is a normal sample.

In the following formula, P_T, P_F, N_T and N_F represent true positive, false positive, true negative and false negative, respectively.

Precision (P): Indicates the correct attack sample frequency predicted by the model. The higher the accuracy rate, the lower the false alarm rate. It can be expressed as

$$P = P_T/(P_T + P_F) \tag{4.1}$$

$$R = P_T/(P_T + N_F) \tag{4.2}$$

F1 value: It means that the accuracy and recall of the model are comprehensively considered. It can be expressed as

$$F = 2PR/(P + R) \tag{4.3}$$

4.2. Model Vs classification results. Using the UOSW algorithm, the KDD Test data package was used to test five models: DT, SVM, LSTM, DAEDNN, and DAENDD, and precision, recall, and F1 values were selected as test parameters to compare and analyze different models. The test scores of various models

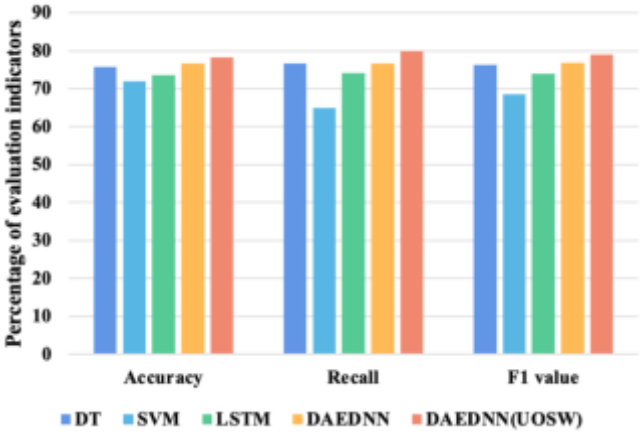


Fig. 4.1: Scores of various indicators of different models

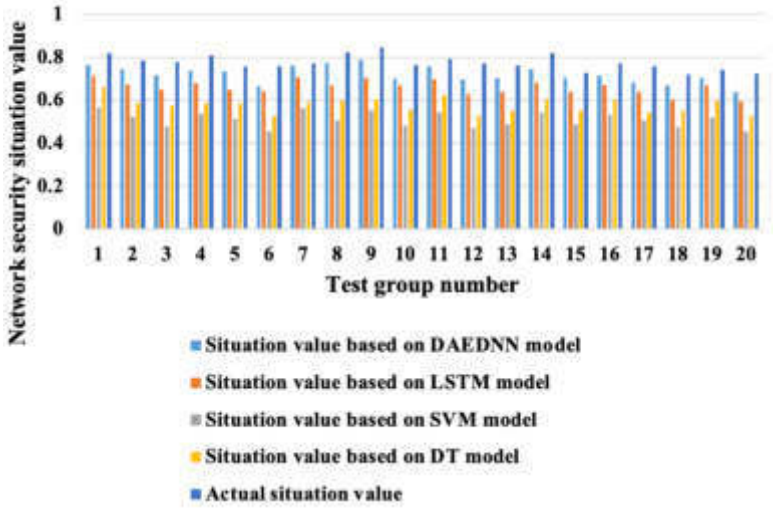


Fig. 4.2: Network security situation value of 20 groups of tests

are shown in Figure 4.1. The setting in the figure represents the percentage of test results; the higher the value, the better the model. It shows that the UOSW model outperforms the other four models regarding precision, recall, and F1 value. Experimental results show that UOSW improves the recall and accuracy of attack types on several training data but does not reduce the detection of attacks with more training examples.

It should be noted that DAEDNN has higher accuracy, recovery speed, and wider capabilities after combining with the original UOSW algorithm. Compared with DT, SVM and LSTM models, the F1 value UOSW increased by approximately 2.77, 10.5 and 5.2, respectively.

To increase the effectiveness of network security problems, the same number of test samples are selected from the test data, and different models estimate the importance of network security problems; the results of 20 groups of network security indicators.

5. Conclusion. The author introduces an innovative approach to network security, leveraging the power of deep learning. This method represents a pioneering endeavour in constructing a DAEDNN model that seamlessly integrates autoencoders and deep neural networks to identify network attacks. Drawing upon the insights

gathered from the analysis, the model calculates outage probabilities and impact values, thereby enabling the computation of network security costs. This multidimensional assessment of security issues offers a specific understanding of its potential impact on network security. The results from experimental trials underscore the efficacy of the author's proposed model, demonstrating its superiority over alternative models in both binary and multi-class classification scenarios. Moreover, the UOSW algorithm further enhances the model's ability to accurately detect attack patterns, particularly in situations with limited training samples. This augmentation empowers the model to effectively discern and evaluate diverse types of network attacks and associated security concerns, ultimately contributing to a more robust and comprehensive network security framework.

Acknowledgements. The study was supported by

1. Scientific research project of Hunan Provincial Department of Education "Research on optimization of DDPG algorithm based on Actor-Critic framework" (No.: 21C1186)

2. Hunan Vocational College Education and Teaching Reform Research Project "Research on Classroom Teaching Evaluation in Higher Vocational Education Based on Deep Learning" (No.: ZJGB2021189)

3. Hunan Natural Science Foundation project "Research on yawn detection algorithm based on AdaBoost "(No.: 2021J60038)

REFERENCES

[1] A. Akhunzada, A. Gani, N. B. Anuar, A. Abdelaziz, M. K. Khan, A. Hayat, and S. U. Khan, *Secure and dependable software defined networks*, Journal of Network and Computer Applications, 61 (2016), pp. 199–221.

[2] A. Årnes, K. Sallhammar, K. Haslum, T. Brekne, M. E. G. Moe, and S. J. Knapskog, *Real-time risk assessment with network sensors and intrusion detection systems*, in Proceedings of the Computational Intelligence and Security: International Conference, vol. 3802, Xi'an, China, 2005, Springer, pp. 388–397.

[3] M. Baykara and R. Das, *A novel honeypot based security approach for real-time intrusion detection and prevention systems*, Journal of Information Security and Applications, 41 (2018), pp. 103–116.

[4] V. Chastikova and V. Sotnikov, *Method of analyzing computer traffic based on recurrent neural networks*, Journal of Physics: Conference Series, 1353 (2019), p. 012133.

[5] M. Husák, J. Komárková, E. Bou-Harb, and P. Čeleda, *Survey of attack projection, prediction, and forecasting in cyber security*, IEEE Communications Surveys & Tutorials, 21 (2018), pp. 640–660.

[6] Y. N. Kunang, S. Nurmaini, D. Stiawan, and B. Y. Suprapto, *Attack classification of an intrusion detection system using deep learning and hyperparameter optimization*, Journal of Information Security and Applications, 58 (2021), p. 102804.

[7] N. M. Kuznetsova, T. V. Karlova, and A. Y. Bekmeshov, *Methods of timely prevention from advanced persistent threats on the enterprise automated systems*, in Proceedings of the International Conference on Quality Management, Transport and Information Security, Information Technologies, Saint Petersburg, Russian Federation, 2022, IEEE, pp. 158–161.

[8] N. Lal, S. M. Tiwari, D. Khare, and M. Saxena, *Prospects for handling 5g network security: Challenges, recommendations and future directions*, Journal of Physics: Conference Series, 1714 (2021), p. 012052.

[9] D. C. Le and N. Zincir-Heywood, *A frontier: Dependable, reliable and secure machine learning for network/system management*, Journal of Network and Systems Management, 28 (2020), pp. 827–849.

[10] H. Lin and J. Wang, *Pinning control of complex networks with time-varying inner and outer coupling*, Mathematical Biosciences and Engineering, 18 (2021), pp. 3435–3447.

[11] Z. Lin, J. Yu, and S. Liu, *The prediction of network security situation based on deep learning method*, International Journal of Information and Computer Security, 15 (2021), pp. 386–399.

[12] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, *Social network security: Issues, challenges, threats, and solutions*, Information Sciences, 421 (2017), pp. 43–69.

[13] B.-C. Seo and W. F. Krajewski, *Statewide real-time quantitative precipitation estimation using weather radar and nwp model analysis: Algorithm description and product evaluation*, Environmental Modelling & Software, 132 (2020), p. 104791.

[14] P. Wei, Y. Li, Z. Zhang, T. Hu, Z. Li, and D. Liu, *An optimization method for intrusion detection classification model based on deep belief network*, IEEE Access, 7 (2019), pp. 87593–87605.

[15] H. Wu, Q. Gao, X. Tao, N. Zhang, D. Chen, and Z. Han, *Differential game approach for attack-defense strategy analysis in internet of things networks*, IEEE Internet of Things Journal, 9 (2021), pp. 10340–10353.

[16] H. Zhang, C. Kang, and Y. Xiao, *Research on network security situation awareness based on the lstm-dt model*, Sensors, 21 (2021), p. 4788.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Jun 5, 2023
Accepted: Sep 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 157–165, DOI 10.12694/scpe.v25i1.2108

ONTOLOGICAL AUGMENTATION AND ANALYTICAL PARADIGMS FOR ELEVATING SECURITY IN HEALTHCARE WEB APPLICATIONS

NAWAF ALHARBE*

Abstract. The 4th and 5th industrial revelations provided huge access to a high volume of information in healthcare sectors. However, the vast majority of web applications are under attack constantly. Due to the huge volume of attack vectors in recent years, technical breaches to the web applications of healthcare sectors are becoming a common issue. Hence, it is essential to develop an effective framework that would help experts and healthcare practitioners in web application security management. In this research, we proposed an ontology-based technique for developing secure online web applications in healthcare sectors. This study presents a conceptual framework with 5 stages namely: idea understanding, requirement identification, design & code, threat classification, and facilitate. The proposed methodology involves several various advantage features such as providing a unified path for future professional applications. Also, the proposed solution provides a clear pathway for implementing the easy-to-use secure web application development in short term in the healthcare sector. Finally, the study used a multi-criteria decision-making methodology to undertake a performance simulation assessment. Out of the evaluated process, Fuzzy-AHP has shown better performance for scheduling each step of the final developed security system better. In the future, the proposed method will be evaluated using more databases in the healthcare sector to improve the applications of the proposed method.

Key words: Security, decision support system, healthcare web application, machine learning application

AMS subject classifications. 68T01, 68U35

1. Introduction. Virtual platforms and the Internet of Things (IoT) have been two-man pillars of the fourth and fifth revolutions [27]. Industry 4 provided easy access to the vast volume of information. New generations of wireless networks such as the fifth generation (5G) and future generation sixth generation (6G) will provide faster and more reliable to process this information [4]. Nowadays there are sensors and intelligent devices that monitor human daily activities and store information. In this environment, collected data is processed through cloud computing devices and collected data by IoT devices are uploaded to cloud servers [32]. Healthcare is one of the industries that benefited from various sources of information and the availability of datasets to provide a portfolio for each patient. By increasing the applications of IoT, the chance for adversaries to steal, and alter patient information increases as well. The healthcare sector is compacted with vital information and vast amounts of data in form of biomedical signals and patients' treatment histories. Due to the intricacies of such information, maintaining and securing the entire healthcare data is essential for any healthcare web application [2].

The main focus of these attacks is the exploitation of healthcare web apps. The prevalence of these attacks has increased from 2005 to 2022. In total, 256.65 million breach attempts have been recorded from 2005 to 2021 [16]. the total number of attacks registered and implemented with regulating authorities is 3868 from all over the world, with 627 breaches in the preceding year [23]. The number of breaches in the healthcare sector depicts the vulnerability of this industry against such an attack. IBM reported due to the swarm of attacks against the healthcare sector and the correspondent breaches caused 6.45 6.45 million dollars in loss for this industry [16]. Due to the prevalence of attacks and breaches in the healthcare section, securing the developed web application is the utmost pressing need.

Another major consequence of these breaches to the database is corresponding harm to the company's reputation. Customers don't want their online applications to be interrupted or lose data while using them [26, 25, 33]. Most of the time, the developer overlooks the possible faults and vulnerabilities since the main priority of the developer is providing an application with optimal usability over the estimated deadline [22, 31].

*Applied College, Taibah University, Medina 42353, Saudi Arabia (nrharbe@taibahu.edu.sa).

Furthermore, many firms are unable to determine the vulnerability of the proposed web application before the breaches occur.

To address the presented issues in this article, the authors proposed a solution for online web applications in the healthcare sector using a combination of AI semantics and formal ontologies. When it comes to web application security, we must consider the entire development functionality of the application. Security in web applications can only be addressed by an effective and improved security approach during the development process [23, 3]. A well-equipped team and laboratory are required to produce a secure development toward structured security of web application. From an ontology standpoint, the projected study aims to give a methodical, step-by-step approach for designing secure healthcare online applications. Ontology is a new effective concept or idea that has a lot of promise for displaying solid and successful results. The projected framework identifies the security factors that must be addressed from the beginning of the web application development process and then handled in a systematic manner [13].

The projected framework is a conceptual ontology-based approach aimed at providing efficient and secure online application development. This project creates a pathway for repositories to provide the best professional practices and processes to enable secure development in healthcare web applications. The advancement of security necessitates its application at the earliest stages of development. It is necessary to discuss and provide an artifact-based method that produces an effective result to make this practicable. This study is organized into six main sections. In Sect. 2, the history of the ontology-based method for intrusion detection and recently proposed solutions are discussed. Section 3 describes the importance of the problem and corresponding statistical characteristics related to it. In Sect. 4, the proposed model and strategy are presented. Section 5 evaluated the proposed model and verified the achieved results. Finally, Sect. 6 summarizes the research contributions, limitations, and future of the work.

2. Related Works. The fourth industrial revelations publicized the prevalence of gadgets and electronic devices in various industries. Healthcare is one of these industries with a huge volume of electronic records and contextual information [30]. However, the quality and quantity of the database have increased recently, and the number of attacks and variety of attack types has increased respectively [3]. Securing vast amounts of information and communications through the virtual platform is very important. Artificial Intelligence (AI) applications have shown huge potential for intrusion detection and improving the security of the web base applications [19]. Due to the vast majority of information, Ontology-based security frameworks and their tools can benefit from semantic analysis and its tools [24]. However, developing a proper ontology-based strategy for the healthcare section is difficult. In this section, we reviewed related research to securing healthcare web applications. The following studies have reviewed the effective solution for secure healthcare web applications using semantic ontology-based methods. Due to the recent prevalence of smart devices, the environment around the patients or even inside the patients' bodies is filled with these instruments.

The process of securing recording and transferring information for patients is vital. Any misleading information in this database might cause fatality for patients [5]. Henaien et al. [5]. Presented a method based on 4 stages of ontology to manage user profile information, medicine prescription data, senor information, and identification process. To secure the collected information, the proposed method follows the security process based on the patient's personalization information. If the patient doesn't suffer from memory, optical, or hearing impairments then they can access all of the recorded information via passwords. Otherwise, caregivers, doctors, or other third parties must help the patients to access this information.

Chen et al. [9]. Proposed an ontology-based framework in 4 layers. The presented method involves a service discovery and selection layer, a business layer, a data access layer, and a technological layer. Each of the presented layers operates to satisfy certain trends. The data access layer manages the database synchronization, business layer manages the service operation system. The proposed security layer uses the Artificial Intelligence (AI) semantic meditator to choose the user desired web service. The proposed method tried to create a secure portfolio using an integrated medical and healthcare dataset.

Sharma et al. [29]. Presented an ontology based on AI to monitories Biomedical Signals such as ECG, PPG, temperature, heart rate, and HR Variability. The authors collected the mentioned information using available gadgets such as apple watch and bio best. The authors used artificial intelligence methods such as K Nearest Neighbor (KNN) [34]to detect the COVID-19 prevalence in each cohort. The authors concluded that

the proposed ontology can detect the prevalence of COVID-19 among each cohort with 96.33% accuracy. The process of sharing information via each client is edge and server base sharing with the lowest consumption of energy. The authors concluded that the proposed method can be used for safe monitoring among COVID-19 patients via a 2-layers security transferring process.

Hady et al. [14]. Presented a combination of network and biometric metrics as features for intrusion detection. The authors used Support Vector Machine (SVM) [7], KNN, and artificial neural network [1] to detect the intrusion. The authors gathered biomedical information such as blood oxygen, temperature, and heat bit rate while recording network information such as computer identification data. The proposed program detects intrusion attacks such as spoofing attacks and data alteration. The authors created a database using 16 thousand collected records from various patients. The authors reported 92.44% accuracy using SVM for intrusion detection.

Mozzaquatro et al. [20]. Presented an ontology-based solution using the combination of semantic ontological and formal solutions. The authors presented solution monitories the reliability, operability, operability, transferability, and adaptability characteristics of each web application. The authors concluded that the proposed strategy reduces risk and ensures long-term software serviceability and security. The presented solution by the authors in this research provides the reliable automated automatic establishment of security metrics using explicit and reasoning information about situations of interest and combined knowledge from multiple security experts.

Bataityte et al. [6]. Presented a solution based on three layers of formalization to automate the formal solutions to solve the cyber security problem such as logical vulnerability, and risk assessment. The presented solution enables analysing some cyber security logical problems such as vulnerability analysis and risk assessment. The authors presented a logical theory based on situations and actions in descriptive logical theory. The authors deployed the presented solution into the ontology web language, and semantic web rule language [21]. The authors concluded that the proposed solution can provide simple analytic results related to logical vulnerability, risk assessment, and policy validation.

Shahzad et al. [28]. Presented an ontology base solution to combine information of integrated health care intelligent devices without any security problem. The proposed solution is composed of three layers of initial health system analysis, proposing ontological representation of the smart system and finally integrating every system. The proposed method is evaluated on arrhythmia detection, prostate cancer care process, and leukaemia detection. In each use case, the presented solution gathered biomedical information from each patient then detect the related disease, and finally reported the problem to caregivers or correspondent doctors. Finally, the proposed solution recorded the information about each event for each patient.

The reviewed articles attempted to fully comprehend their work before presenting a suitable framework with unique utilities and working techniques. However, the presented solution didn't consider the effect of simplicity and easy development for healthcare sectors. Adopting ontology during the development of secure applications of healthcare is critical because there is currently no specialized work accessible that discusses web-based application security from the perspective of health informatics, or slightly another web application perspective.

Another important issue is the process of validating the presented solution by ontology. Healthcare sector web applications are integrated and the concept of integration must be a crucial point to develop and test the proposed ontology solution. However, the majority of the proposed solution evaluated the ontology only in the specific domain of the presented solution, not as an integrated web application solution.

To address these issues the proposed solution by this article investigates the ontology base solution to secure the web applications in the healthcare sectors in a simple process. The proposed solution is evaluated to other solution to prove the superiority of the presented integrated ontology over similar research.

3. Problem Statement. As mentioned previously, the number of developed applications in the healthcare section has increased in the last decades. The result of the ongoing increase in the quantity and quality of the available data is the rise in the number of attacks. Table 3.1 represents the number of recorded attacks and their relation to the total volume of datasets.

The represented information in Table 3.1 is based on summarized reports from HIPPA [3]. Based on the reported information from Table 3.1, the number of reported breaches has increased 5 times over the past

Table 3.1: Number of recorded attacks based on available datasets.

Year	Amount of Data (in Millions)	Number of recorded attacks
2010	5.5300	199
2011	13.150	200
2012	2.8000	217
2013	6.9500	278
2014	17.450	314
2015	113.27	269
2016	16.400	327
2017	5.1000	359
2018	33.200	365
2019	41.200	505
2020	52.350	835
2021	65.540	987

Table 3.2: Attacks Source statistics

Year	Technical Issues and Causes	Human Error and Disclosure	Theft	Miss Handling
2010	8	8	148	10
2011	17	27	136	7
2012	16	25	138	8
2013	25	64	150	13
2014	35	76	143	12
2015	57	101	105	6
2016	113	129	78	7
2017	147	128	73	11
2018	158	143	55	9
2019	274	142	51	7
Total	850	843	1077	90

decades. Because health is inextricably linked to human life, the number of breach records is quite noteworthy and high. A data leak of this nature can put patients' lives in jeopardy. Furthermore, it is to mend its flaws after studying the attack statistics in healthcare. It is vital to initially discover the gaps in healthcare breach occurrences to identify the key difficulties. Detailed information about each category of attack has shown in Table 3.2.

Table 3.2 displays breach' counts carried out by each source, which provides a clear point-by-point explanation of healthcare flaws and aids professionals in preparing preventive actions. Based on the presented categories of attacks in Table 3.2, the maximum typical and frequent exploit generators in the healthcare sector are technical issues and human errors. Although, when comparing Table 3.1 and Table 3.2, it is clear that online application vulnerabilities generate the most troubles and exploits in healthcare. Technical flaws produce and generate troubles in the modern era, and attackers take advantage of these flaws by injecting healthcare applications on a wide scale regularly. The data and statistics presentation in Table 3.1 and Table 3.2 depict the critical issue of online medical system management, and also the necessity for functionality-centric security solutions. This research proposed an ontology-based framework to provide specialists with greater affectivity and a conceptual grasp of how to design healthcare online applications.

4. Methodology. Ontology is widely accepted and applied in various domains such as communication development as well as actual interaction between humans and machines [15]. The ontology adaption is a novel concept from the perspective of healthcare web application development. The ontology-based idea can be

Fig. 4.1: Projected framework.

used to develop intergraded web applications for an operative as well as efficient security purposes. Ontology helps the professional to translate real-world problems into machine settings and solve these problems using the machine. However, utilizing any ontology-based solution for securing healthcare virtual platforms leads to more complications in the process. Another important problem issued in this research is the vast quantity of data obtainable on many online as well as offline sites which developers accept and use without any background check to make their work easier [17].

In this article, we proposed an ontology-based framework to effectively secure the development of healthcare web applications. The proposed framework merges capable ontology security management with the traditional and commonly used development stages. The proposed methodology is created using the adopted ideology and can be easily developed in real-world web applications. Any new approach must first develop a notion before it can be fully addressed. In this work, we presented a standard development paradigm with ontology. The proposed framework has shown in Fig. 4.1.

Figure 4.1 depicts a five-step integrated systematic framework for securing applications at the early stages of the development process that combines traditional development processes with security management features. The proposed paradigm combines concept initiation with ontology to help with the development phase of the web application.

The proposed framework uses 5 phases to to secure the web application based on the developer needs. These five phases are: Requirement Identification; Idea Understanding; Classification of Potential Possible Threads; Design & Code, and facilitate. To make the process easier, no additional specific stages are there in the process that would demand extra efforts or time from the developer or the organizations. We attempted to keep things basic and straightforward because complication breeds both confusion and a lack of enthusiasm for the development process on the developer's part. framework's step-by-step working method, which is explained topic of this study is mentioned as follows:

- Idea Understanding: This stage gets people thinking about progress. In this segment, the web application's programmer, developer, as well as owner, discuss the functionality and ideas for how the web application will look after development, as well as what kinds of features the owner wants.
- Requirement Identification: This stage determines which processes, utilities, and attributes need to be handled during the program's design and development. The need identification phase integrates these demands and creates a good new framework for the application's systematic development.
- Design & Code: Designers design the cede that correlates the stated necessities and create an effective application based on them in this step.
- Threat Classification: This is the phase in the process that protects the system from exploitation. This is a process in which three viewpoints are analysed and then the existing development scenario is compared. If the scenario is compatible with the demand, it's acceptable; otherwise, the loop begins over with the requirement identification procedure.
- Facilitate: The projected framework permits designers to assist the developed web application in the

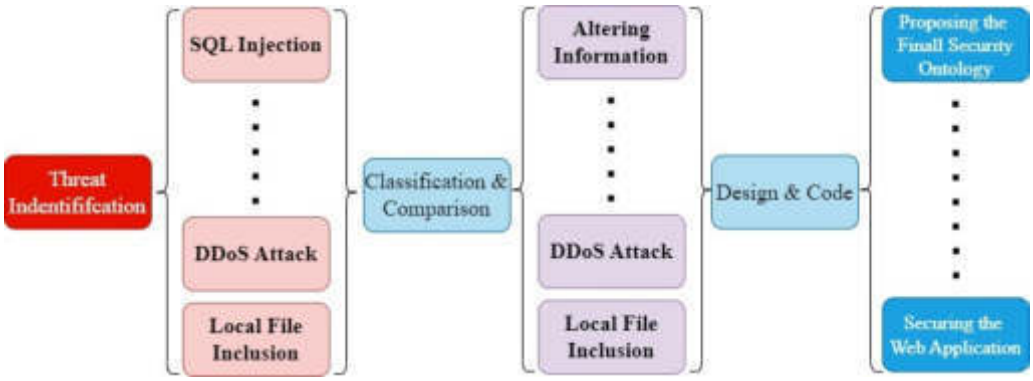


Fig. 4.2: Projected framework taxonomy.

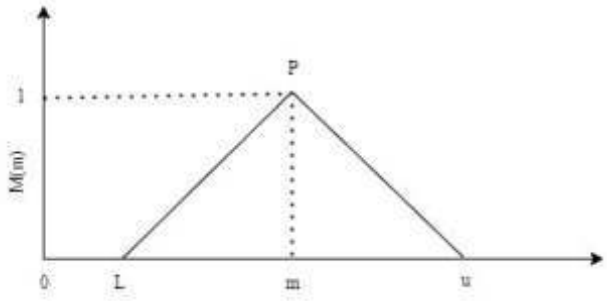


Fig. 5.1: Triangular fuzzy number (P).

healthcare industry after classifying and comparing the numerous threats and other development issues. The anticipated framework allows the designers to simplify the generated web application in the healthcare market once the full cycle becomes systematic and uninterrupted.

Various steps of identification, classifying the threats, and final design of the proposed taxonomy has shown in Fig. 4.2. Overall, after examining the framework, concept, and working flow, it is evident that the presented framework is the first stage in using an ontology-based method to assist healthcare web application development. The projected framework depicts the authors' conceptual framework and will require further mathematical testing and confirmation in the future

4.1. Experimental setting. In this research, the proposed method is compared with benchmark frameworks to determine its efficacy of it. We used six distinct frameworks namely OWASP Ontology Framework (FW1) [12], Moreira Framework (FW2) [10], Mozzaquatro Framework (FW3) [24], Maglogiannis Framework (FW4) [8], and Analytical Hierarchy Process mixed with fuzzy set theory (Fuzzy-AHP) (FW5) [18]. Fuzzy-AHP is a method that produces accurate, widely accepted, and validated outcomes. The projected framework in this research is Fuzzy-AHP.

5. Experimental Result. To compare the result, the triangular fuzzy number is used and the result has shown in Table 5.1.

We used a cut strategy to aggregate the fuzzy set into the values after detecting the pair-wise fuzzy numbers [11]. The triangular membership function has shown in Fig. 5.1.

As shown in Fig. 5.1, the Parameter L represents the minimal value, parameter m represents the most likely value, and parameter u represents the maximum value. The membership equation are shown in the

Table 5.1: Attacks Source statistics.

Model Name	**FW1**	**FW2**	**FW3**	**FW4**	**FW5**
OWASP Ontology Framework (FW1)	1.00000, 1.000000, 1.000000	0.69000, 0.886000, 1.100000	0.226000, 0.276000, 0.357000	1.000000, 1.516000, 1.933000	0.490000, 0.637000, 1.000000
Moreira Framework (FW2)	*	1.00000, 1.000000, 1.000000	0.695000, 0.950000, 1.346000	0.268000, 0.352000, 0.518000	0.166000, 0.197000, 0.253000
Mozzaquatro Framework (FW3)	*	*	1.000000, 1.000000, 1.000000	1.000000, 1.320000, 1.5520008	0.301000, 0.435000, 0.803000
Maglogiannis Framework (FW4)	*	*	*	1.000000, 1.000000, 1.000000	0.222000, 0.287000, 0.415000
Projected Framework (FW5)	*	*	*	*	1.000000, 1.000000, 1.000000

Table 5.2: Attacks Source statistics.

Model Name	**FW1**	**FW2**	**FW3**	**FW4**	**FW5**
OWASP Ontology Framework (FW1)	1.000000	1.260900	1.610200	1.002100	0.902100
Moreira Framework (FW2)	0.788000	1.000000	1.269000	0.660500	0.505000
Mozzaquatro Framework (FW3)	0.602000	0.780800	1.000000	0.650400	0.690000
Maglogiannis Framework (FW4)	0.907900	1.500400	1.530000	1.000000	0.606500
Projected Framework (FW5)	1.080700	1.810700	1.440900	1.500500	1.000000

following equations.

$$P(x|m) = \{0; x < L\} \tag{5.1}$$

$$P(x|m) = \{\frac{x-L}{m-L}; L \leq x < m\} \tag{5.2}$$

$$P(x|m) = \{\frac{u-x}{u-m}; m \leq x < u\} \tag{5.3}$$

$$P(x|m) = \{0; x > u\} \tag{5.4}$$

Table 5.2 shows the cut approach's value as well as the aggregated value of the triangular fuzzy numbers.

Based on what is described in Table 5.1, and aggregated information in Table 5.2, the projected frameworks demonstrates better option compare to other methods.

As shown in Fig. 5.2, the priority for the evaluated framework is FW5>FW4>FW1>FW3>FW2. The findings of the performance simulation mentioned in this part show that FW5 has the highest effectiveness and priority, whereas FW2 has the lowest. In general, the presented framework is one of the most appropriate

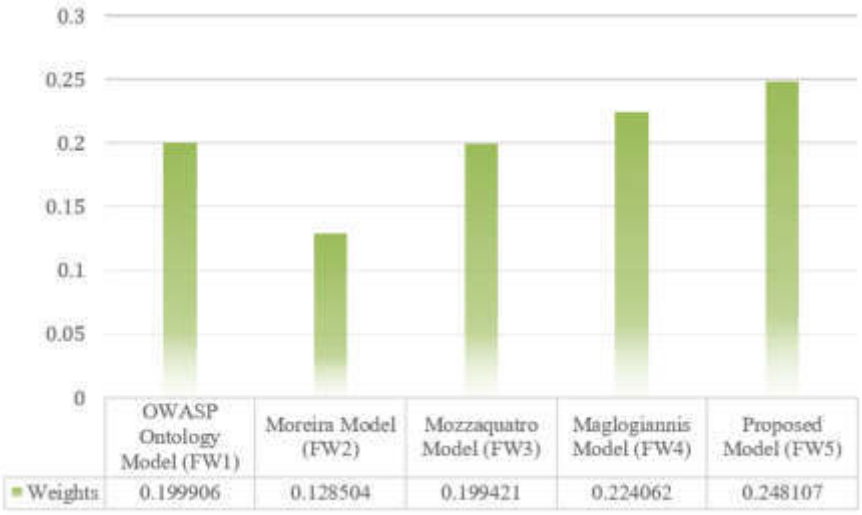


Fig. 5.2: Graphical illustration of priority.

frameworks and ideologies that academics and future practitioners in their subject might follow. The proposed framework in this section compared the performance of each section and provided a systematic workflow to effectively secure the web application in the healthcare sector. This section depicts an optimal pathway to simplify the difficulty of proposing a systematic and statistically assessed efficacy order. The projected framework has the highest effect ratio in all of the selected frameworks, with a weight of 0.248107 and a CR value of 0.001547.

6. Conclusion. Due to the development of more virtual applications and the increased volume of the dataset, the number of attacks on healthcare web applications has increased. Any breach the healthcare web applications due to the vast number of attacks lead to losing vital information and harm to the company's brands. To address these issues, we proposed an easy-to-use pre-security procedure to ensure the healthcare security section. The presented methodology is an otology-based solution to check the possible threats and develop a proper design for the proposed framework before releasing it to the public. The main focus of this projected framework is to provide an appropriate path for future researchers. The contributions of the proposed framework are: 1. Proposing the requirement of a secure web application platform in the healthcare section. 2. Providing comparison and classification path for the proposed security program. 3. Providing a novel pathway for future ontology-based security solutions in the healthcare sector. In the future, the proposed solution will be evaluated using more experimental analysis to provide a general security application in the healthcare domain.

REFERENCES

[1] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, A. M. Umar, O. U. Linus, H. Arshad, A. A. Kazaure, U. Gana, and M. U. Kiru, *Comprehensive review of artificial neural network applications to pattern recognition*, IEEE Access, 7 (2019), pp. 158820–158846.

[2] I. Abu-Elezz, A. Hassan, A. Nazeemudeen, M. Househ, and A. Abd-Alrazaq, *The benefits and threats of blockchain technology in healthcare: A scoping review*, International Journal of Medical Informatics, 142 (2020), p. 104246.

[3] M. Alenezi, *An ontological framework for healthcare web applications security*, International Journal of Advanced Computer Science and Applications, 12 (2021).

[4] R. S. Alonso, J. Prieto, F. de La Prieta, S. Rodríguez-González, and J. M. Corchado, *A review on deep reinforcement learning for the management of sdn and nfv in edge-iot*, in 2021 IEEE Globecom Workshops (GC Wkshps), IEEE, 2021, pp. 1–6.

[5] O. T. Arogundade, A. Abayomi-Alli, and S. Misra, *An ontology-based security risk management model for information systems*, Arabian Journal for Science and Engineering, 45 (2020), pp. 6183–6198.

[6] K. Bataityte, V. Vassilev, and O. J. Gill, *Ontological foundations of modelling security policies for logical analytics*, in Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16, Springer, 2020, pp. 368–380.
[7] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, *A comprehensive survey on support vector machine classification: Applications, challenges and trends*, Neurocomputing, 408 (2020), pp. 189–215.
[8] V. Charpentier, N. Slamnik-Krijestorac, and J. Marquez-Barja, *Latency-aware c-its application for improving the road safety with cam messages on the smart highway testbed*, in IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2022, pp. 1–6.
[9] S.-W. Chen, Y.-T. Tseng, and T.-Y. Lai, *The design of an ontology-based service-oriented architecture framework for traditional chinese medicine healthcare*, in 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), IEEE, 2012, pp. 353–356.
[10] J. R. Davis, *Healthcare Entities and Data Breach Threat Indicators and Deterrence: A Quantitative Study*, PhD thesis, Northcentral University, 2022.
[11] E. dos Santos Moreira, L. Andréia Fondazzi Martimiano, A. José dos Santos Brandão, and M. César Bernardes, *Ontologies for information security management and governance*, Information Management & Computer Security, 16 (2008), pp. 150–165.
[12] K. N. Durai, R. Subha, and A. Haldorai, *A novel method to detect and prevent sqlia using ontology to cloud web security*, Wireless Personal Communications, 117 (2021), pp. 2995–3014.
[13] M. Gheisari, H. E. Najafabadi, J. A. Alzubi, J. Gao, G. Wang, A. A. Abbasi, and A. Castiglione, *Obpp: An ontology-based framework for privacy-preserving in iot-based smart city*, Future Generation Computer Systems, 123 (2021), pp. 1–13.
[14] A. A. Hady, A. Ghubaish, T. Salman, D. Unal, and R. Jain, *Intrusion detection system for healthcare systems using medical and network data: A comparison study*, IEEE Access, 8 (2020), pp. 106576–106584.
[15] I. Harrow, R. Balakrishnan, E. Jimenez-Ruiz, S. Jupp, J. Lomax, J. Reed, M. Romacker, C. Senger, A. Splendiani, J. Wilson, et al., *Ontology mapping for semantically enabled applications*, Drug discovery today, 24 (2019), pp. 2068–2075.
[16] T. Hou and V. Wang, *Industrial espionage–a systematic literature review (slr)*, computers & security, 98 (2020), p. 102019.
[17] C. Islam, M. A. Babar, and S. Nepal, *An ontology-driven approach to automating the process of integrating security software systems*, in 2019 IEEE/ACM International Conference on Software and System Processes (ICSSP), IEEE, 2019, pp. 54–63.
[18] Y. Liu, C. M. Eckert, and C. Earl, *A review of fuzzy ahp methods for decision-making with subjective judgements*, Expert Systems with Applications, 161 (2020), p. 113738.
[19] B. K. Mohanta, D. Jena, U. Satapathy, and S. Patnaik, *Survey on iot security: Challenges and solution using machine learning, artificial intelligence and blockchain technology*, Internet of Things, 11 (2020), p. 100227.
[20] B. A. Mozzaquatro, C. Agostinho, D. Goncalves, J. Martins, and R. Jardim-Goncalves, *An ontology-based cybersecurity framework for the internet of things*, Sensors, 18 (2018), p. 3053.
[21] A. Patel and S. Jain, *Present and future of semantic web technologies: a research statement*, International Journal of Computers and Applications, 43 (2021), pp. 413–422.
[22] I. Rus, M. Lindvall, and S. Sinha, *Knowledge management in software engineering*, IEEE software, 19 (2002), pp. 26–38.
[23] K. Sahu, F. A. Alzahrani, R. Srivastava, and R. Kumar, *Hesitant fuzzy sets based symmetrical model of decision-making for estimating the durability of web application*, Symmetry, 12 (2020), p. 1770.
[24] K. Sahu and R. Srivastava, *Soft computing approach for prediction of software reliability*, Neural networks, 17 (2018), p. 19.
[25] ———, *Needs and importance of reliability prediction: An industrial perspective*, Information Sciences Letters, 9 (2020), pp. 33–37.
[26] M. Saito, A. Hazeyama, N. Yoshioka, T. Kobashi, H. Washizaki, H. Kaiya, and T. Ohkubo, *A case-based management system for secure software development using software security knowledge*, Procedia computer science, 60 (2015), pp. 1092–1100.
[27] N. I. Saleh, M. T. Ijab, and N. Hashim, *A review on industrial revolution 4.0 (ir4. 0) readiness among industry players*, in International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022), Atlantis Press, 2022, pp. 216–231.
[28] S. K. Shahzad, D. Ahmed, M. R. Naqvi, M. T. Mushtaq, M. W. Iqbal, and F. Munir, *Ontology driven smart health service integration*, Computer Methods and Programs in Biomedicine, 207 (2021), p. 106146.
[29] N. Sharma, M. Mangla, S. N. Mohanty, D. Gupta, P. Tiwari, M. Shorfuzzaman, and M. Rawashdeh, *A smart ontology-based iot framework for remote patient monitoring*, Biomedical Signal Processing and Control, 68 (2021), p. 102717.
[30] A. Trauth-Goik, *Repudiating the fourth industrial revolution discourse: a new episteme of technological progress*, World Futures, 77 (2021), pp. 55–78.
[31] O. I. Uwakweh, *Cybersecurity in the Retail Industry: Third Party Implications*, PhD thesis, University of Cincinnati, 2020.
[32] G. Wang and F. Xu, *Regional intelligent resource allocation in mobile edge computing based vehicular network*, IEEE Access, 8 (2020), pp. 7173–7182.
[33] S.-F. Wen and B. Katt, *An ontology-based context model for managing security knowledge in software development*, in 2018 23rd Conference of Open Innovations Association (FRUCT), IEEE, 2018, pp. 416–424.
[34] L. Xiong and Y. Yao, *Study on an adaptive thermal comfort model with k-nearest-neighbors (knn) algorithm*, Building and Environment, 202 (2021), p. 108026.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Feb 22, 2023
Accepted: Dec 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 167–176, DOI 10.12694/scpe.v25i1.2314

A HYBRID MODEL: RANDOM CLASSIFICATION AND FEATURE SELECTION APPROACH FOR DIAGNOSIS OF THE PARKINSON SYNDROME

SUMAN BHAKAR[*] MANVENDRA SHEKHAWAT[†] NIDHI KUNDU[‡] AND VIJAY SHANKAR SHARMA [§]

Abstract. Nowadays Parkinson's disease has been discovered that approximately 94% of people suffer from voice disorder problems. A neurodegenerative can identify PD patients through examination and multiple scanning tests. So, it usually takes more time to diagnose the disease at the early stage. Current work has identified that speech disorders can be a significant signal for Parkinson's disease. Therefore, this work proposed a fusion model to identify the speech disorder at the starting stage of the disease. In this process, the author has tested a model with a different pattern of feature selection method as well as classification mode and created a system with the best pattern. For the creation of pattern, three types of feature selection methods namely Chi-square, genetic algorithm and Embedded random forest method and four classifier models such as KNN, Naïve Bayes, SVM, Decision tree and Random Forest have been utilized. To analyze the performance of the system speech public dataset from the UCI repository,the authors applied the combination of the Embedded random feature selection method and random forest classification algorithm provides 97.89% of accuracy. However, this outcome is better than the recent work. The SMOTE is utilized for the balancing of the dataset.

Key words: Machine learning, Parkinson's Disease, Feature Selection, SVM Detecion

1. Introduction. Parkinson's disease is a neurodegenerative syndrome. It is the second disease after Alzheimer's that is slowly developing the neurodegenerative disorder [1]. This disease has affected 7 to 8 million people worldwide. The main source of PD is the damage to the nerve cells i.e., the portion of the brain known as substantia nigra. Nerve cells also produce a liquid called dopamine. Dopamine is the courier between the brain and nervous system. It helps to regulate the body part and body movement. Unfortunately, if nerve cells are injured due to some reason then the ratio of dopamine is minimized in the brain system. So, gradually brain movement control cannot work properly like in a normal person [3]. Both women and men are affected by Parkinson's disease. it cannot cure but can early diagnosed at the early stage, which helps the person for proper treatment and avoids the dangerous situation. Additionally, the ratio of 3:2 men and women are being affected by Parkinson's disease. This disease generally occurs at the age of sixty but it can also occur at the age of fifty [5, 35].

The starting stage of this disease is extremely gentle and unnoticeable. But these gentle symptoms go through a severe stage if predication does not occur at an early stage [6, 4]. The symptoms of PD fluctuate from patient to patient. The evaluation of the disease can be processed by the motor and nonmotor symptoms. The motor symptoms consist of rigidity, tremor, and bradykinesia and non-motor symptoms consist of sensory impairment, change in handwriting, and vocal disturbance. Many researchers have concluded that 90% of patients have shown vocal disturbance at the initial stage of Parkinson's disease.

There are several reasons for the prediction of Parkinson's disease in the initial phase. Especially the neurologists and experts can detect the Parkinson's disease after complete study and frequent scans of the patients. But both techniques are very time-consuming and difficult for patients which are aged above fifty [7, 21]. A physician with a specific degree in this Parkinson's background can diagnose the disease in the early phase through some symptoms. There is no expert doctor at the mountain area. So, there is need to develop

[*]Department of Computer and Communication Engineering, Manipal University Jaipur, India (Suman.bhakar@jaipur.manipal.edu)
[†]Department of Computer and Communication Engineering, Manipal University Jaipur, India
[‡]Department of Computer Science, SKNAU, Jobner, India
[§]Department of Computer and Communication Engineering, Manipal University Jaipur, India (Corresponding author: vijayshankar.sharma@jaipur.manipal.edu)

a model to detect the disease with high accuracy.

Many researchers have utilized many ways i.e. EMG signals, SPECT images, handwritten pictures, gait signals, and MRI images to detect Parkinson's disease [8]. The fluctuations in the speech are known as dysphonia. It is also a symptom of the early detection of Parkinson's disease. The early symptoms of Parkinson's disease can be diagnosed in the initial phase. So most researchers utilized the voice dataset to identify the disease. because it is a very low cost and simple method. Researchers have used many machine learning approaches to identify the disease at an early stage. The machine learning approach uses many pre-processing, and feature extraction approaches to identify the useful features. It also utilizes the classification model [9] to classify the disease and different validation process that is used to help the validation process. Additionally, the pattern of the disease can easily be identified by medical datasets. The pre-processing methods acquired for the balancing and normalization of the datasets. The feature engineering process involves the feature extraction and selection method. The feature selection approachs are used to enhance the accuracy and reduce the computation cost of the system [10, 2].

The major contribution of this articles are :

- The feature selection has been done by three methods such as Chi-Squared, Random forest and Genatic algorithm. The first approach and second approach select 11 features and the third approach select the 5 features out of 23 are selected. The selected feature divides into the training as well as testing datasets. Then these datasets are passed through the different classification methods to find the best detection accuracy result.
- Second, the speech dataset is imbalanced due to 130 out of 180 samples of the details of Parkinson's patients. The SMOTE is utilized to handle the imbalance datasets issues.
- At the last phase performance parameters of the classifier viz Naive Bayes, k-nearest neighbors, and andom forest are analyzed on the complete feature and reduced subsets of features. Finally, it shows the Embedded random forest .algorithm gives the better accuracy i.e. 97 results compare then the existing techniques.

The paper is prepared as tracks. Section 2 defines the overview of the literature review on Parkinson's disease. Section 3 defines the proposed methodology and datasets. The experimental results and their discussion are shown in section 4. The conclusion and future scope are given in section 6.

2. Literature Review. Many Scientists have developed many techniques to diagnose the PD through various speech signals. Little [11] developed a medical specialist model to detect Parkinson's disease.

The authors [12] developed a system to diagnose Parkinson's disease through a clinical specialist scheme. In this system, the half affected patients with Parkinson's disease voice has recorded with vocalization "a" sound for five sec. the dataset contains the recording of three different sounds. The authors utilized the 240*44 i:e row and column ratio dataset for this system. The waveform algorithm is utilized for the five feature extraction processes. The classification has been processed by the Bayesian approach. The system achieves 75.2% accuracy after the process of validation. The authors proposed a solution for dealing with small files, the concept can be used in implementing database with Hadoop and can provide a simple way to store health data [28, 30, 29].

The authors [13] proposed a decision and KNN classification method to detect Parkinson's disease. The useful features have been extracted from the cuttlefish algorithm to optimize the system. The authors utilized the HandPD, voice, and speech datasets. The 92.19% accuracy was achieved by the proposed system.

The authors [14] developed a system to predict the severity rate of Parkinson's disease by using a deep learning model. The rating scale has done by the UPDRS (Unified Parkinson's disease rating score). The performance has been measured through the motors as well as total motor signals. The authors have proved that the motors score the 81.66% of accuracy which is more than the total UPDRS signals.

The authors [15] have utilized the voice signal to detect Parkinson's disease. The feature selection has been proceeded by the eight different ranking parameters. The SVM classification method is utilized for the detection of healthy and Parkinson's disease patients. The Wilcox statistic is processed to get useful information. The system achieved the 92.21% of accuracy through the SVM method.

The authors [8] developed a model to predict PD patients through speech signals. the signals have been recorded by smartphones as well as acoustic cardioids. The Pre-processing has been performed for the identification of the voice and unvoiced signal through unique software. The feature extraction technique is also

applied to get 144 features. The classification is processed by the SVM, MLP, and KNN method to predict the PD patients. The experimental results also proved that the acoustic cardioids i.e. audio signals achieved 3% higher accuracy than smartphone signals.

The authors [16] implemented an antlion optimization algorithm to predict Parkinson's disease. The feature selection has been proceeded to get the optimized result. The filtered features are trained with the decision tree and random classification model. The maximum accuracy was achieved at 95.91%.

Researcher [17] in the article developed a model to diagnose Parkinson's disease. In this model, the data utilization is done through speech and Hand PD datasets. To enhance the optimization of the system grey wolf-based optimization feature selection method is utilized. The authors also used the three classification methods such as decision tree, KNN, and random forest. The experimental result has proven 93.87% accuracy in the speech dataset.

The authors [18] developed an adaptive-based optimization algorithm. The authors used speech signals for the detection of the disease. Authors utilized the sparse encounter method for the reduction of dimension. The classification has been processed by eight classifiers.

The authors [19] developed a model based on a spectrogram by using the feature extraction method. To implement this model PC-GITA datasets are utilized. The authors developed three methods to diagnose Parkinson's disease. In this first approach, the signal based on speech was transformed into a spectrogram and applied the ALEXNET for feature selection method to train the CNN model. In the second approach, feature extraction was done by the CNN model. In the last step, spectral and acoustic signals are passed through classifiers and achieved 99.3% accuracy by using the multilayer classification model.

The authors [20] has developed a novel method to diagnose Parkinson's disease and gender redeployment. The singular value decomposition (SVD) is applied for the features extraction process. And after the features extraction method, the feature selection process i.e. neighborhood component analysis (NCA) is applied for the selection of the features. The authors utilized the six-type based model to diagnosis the PD disease. The developed model achieved the 98.41% and 99.21% of accuracy for the detection of Parkinson's disease and gender. however, only one feature selection method is used to extract the features. and there is no criteria to select the feature for optimization of the model. Also, author applied only one machine learning approach for detection of the disease.

The authors [19] The feature selection is processed by the MRMR method to identify the useful features. The authors also applied the 8-classification method to classify the features. The experimental results proved that the hybrid of RFE and XGboost achieved 95.39% of accuracy. Still, they failed to discuss the feature selection criteria such as, at what parameter and how many features are selected for the detection model, and also did not elaborate the degree of severity for speech dataset.

The authors [23] implemented two CNN-based models to detect the disease. In this article UCI based data i.e. speech dataset was utilized. The author used the two frameworks such as feature and model level for the implementation of the model. The model level-based framework achieved 86.9% of accuracy. However, authors did not use feature extraction and processing approach for the optimisation process and also detection of disease detected by two classification models based on CNN. Also did not compare with another existing machine learning model.

The authors [20] utilised speech signals dataset . The SMOTE is utilized for the pre-processing steps. The authors applied the random forest classification. Although system achieved 89 percentage of accuracy but there is a need to develop a method to handle the unbalanced dataset.so, there is need a method to handle the imbalance dataset.

The authors [22] illustrated the work for the diagnosis of the disease. In this process , the authors used different vowels to analyze the disease. The MAMa tree is utilized for the pre-processing and the singular value and relief method is utilized for the feature selection method. the authors also applied the five-classification methods. The system achieved 92.4% of accuracy by using the KNN classifier. However, author didn't address the number of feature selected in the feature selection method and did not differentiate the experimental result optimization with and without feature selection process.

The authors [37] developed a Parkinson's diagnosis system to detect the disease. The authors used the vowels for the experiments. The relief method is utilized for the selection of Acoustic features. The authors

Table 2.1: Pros and cons of various reviewed models

DL model	Pro	Con
KNN	Simple and Intuitive. Adaptable to Changes. Effective with small datasets.	Computationally Expensive. High Memory Usage.
Naive Bayes	Efficiency with High-Dimensional Data. Works Well with Categorical Data.	Sensitive to Irrelevant Features. Limited Modeling Capacity.
Decision Tree	Handles non-Linearity and mixed data.	Overfitting. High variance
SVM	Robust to overfitting. Effective in High-Dimensional Space.	Difficulty with Large Datasets. Problem with multiclasses.
Random Forest	High Accuracy. Reduced overfitting.	Bias in Feature Selection.

applied the KNN and SVM methods for the classification. The SVM classifier achieved 91.25% of accuracy. Although ,the proposed method achieved good accuracy but still there is possibility to enhance the detection accuracy.

The authors [24] propose the three feature selection models such as Chi-Squared, Random forest and Genetic algorithm. The model is customized in such a way that it can select the 11 features from the information gain and genetic approach and 5 features from the genetic algorithm. The first approach and second approach select 11 features and the third approach selects 5 features out of 23. The table 2.1 elaborate various pros and cons of the models.

The major contribution of these articles are :

- The public dataset is collected from the UCI repository. The speech-based signal datasets comprise of 8 healthy patient's information as well as 23 Parkinson's disease patient's information.
- Three feature selection methods such as Chi-Squared, Random forest and Genatic algorithm to extract the features.
- The selected features are divided into training as well as testing datasets.
- The SMOTE is utilized to handle the imbalance datasets issues
- In the last phase Classifier's performance are analysed on the complete feature and reduced subsets of features.
- To deliver a method which provides, high accuracy, sensitivity and specificity compare than existing method.

3. Methodology. Firstly, the feature selection has been processed [24] by three methods such Chi-Squared, Random Forest and Genetic algorithm. The first approach and second approach select 11 features and the third approach selects 5 features out of 23. The selected feature is divided into training as well as testing datasets. Then these datasets are passed through different classification methods to find the best detection accuracy result.

Secondly, the speech dataset is imbalanced due to 130 out of 180 samples of the details of Parkinson's patients. The SMOTE is utilized to handle the imbalance datasets issues.

Thirdly in the last phase the performance of the classifier i.e naive Bayes, k-nearest neighbors, and random forest are analyzed on the complete feature and reduced subsets of features. Finally, it shows the Embedded Random forest algorithm gives the better accuracy i.e 97.44% of results compared to the existing techniques.

3.1. Dataset. This section describes the datasets.In this, the public dataset is collected from the UCI repository. The speech-based signal datasets comprise of 8 healthy patient's information as well as 23 Parkinson's disease patient's information. The Max Little Oxford University created the dataset. It divides into 195:31 (row: column) wherein the row defines the voice signals and the column defines the voice features. The PD patients are 147 out of 195 and leftover healthy patients' voices are available. The column also has two values 0 and 1. The value 0 defines the healthy and 1 defines the Parkinson's patients.

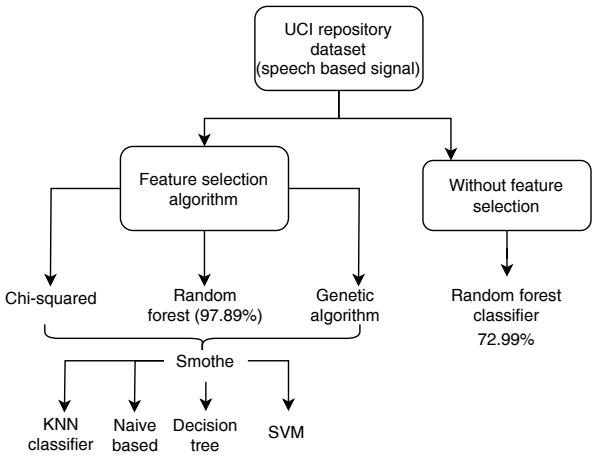


Fig. 3.1: Flowchart

3.2. Feature Selection Method. The feature selection method is the pre-processing step to recognize the vital features from the raw dataset. There are mainly two advantages to the feature selection method. The first advantage is diminishing the dimension of the datasets, and the second is optimizing the performance of the classifier model. In this proposed system, commonly three features are utilized i.e.5 as well as 11 features from 23 features.

3.3. Flowchart. The overall flow of the architecture is as shown in Figure 3.1. The authors using the UCI repository dataset applied classifier with and without feature selection method. The feature selection helps in achieving the high accuracy of 97.89%.

3.3.1. Features selection method using Chi-Squared method. It is the method to examine the freedom of the two incidents [25]. It is based on the statistics to identify the observed value as well as the expected value. It also helps to differentiate the expected value count (E) from the observed value count (O). In our dataset, there are two types of datas i.e. dependent and independent data. The Chi-square method utilizes for the retrieval of the features that are highly dependent on the response estimate as shown Equation 3.1.

$$x_c^2 = \sum sin(\frac{(O_i - E_i)^2}{E_i} \tag{3.1}$$

Here c is the degree of incidents, and 0_i, E_i is the observed as well as the expected value.

3.3.2. Features selection method using genetic algorithm . The Algorithm is also used for the feature selection method. It is the procedure to identify the predicted signal and noise. It is the process that represents the natural selection criteria. It helps to solve the complex problem that takes more time in the process. This approach is based upon the probabilistic nature to get the optimized result.

3.3.3. Features selection method using embedded random forest algorithm. It merges both methods such as the filter method and wrapper method [26]. The random forest method contains approximately 4 to 100 decision trees. It extracts the random observation from the database and extracts the different random features from the datasets.

3.4. Classification model.

3.4.1. KNN classifier. K-NN algorithm [27] is the supervised machine learning approach. It finds the relationship between new data and existing data and places the new data in the most relatable place. It is utilized for the classification and regression process. This approach is non-parametric based algorithm because it

does not contain the assumption data. Calculate the Euclidean distance between two points by the Equation 3.2.

$$Distance = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{3.2}$$

Here (X_2, X_1) and (Y_2, Y_1) are the coordinates of two-point.

3.4.2. Naïve bayes algorithm. The naïve classification algorithm [31] is based upon the Bayes algorithm. When the different features are independent then this algorithm is the best approach for the classification. The naïve Bayes algorithm provides the best result when the dataset is unbiased and dependent on features based on functionality. This approach's accuracy is not completely dependent on the scale of dependent features but also depends on the classification of the features as well as common information between features.

3.4.3. Decision Tree. Decision tree is supervised algorithm. It is utilized for both regression and classification methods [32]. In this approach, nodes represent the features, and branches/edges represent the rule of features. This process uses two types of nodes i.e., the decision node, and the leaf node. It is also known as the attribute selection method. The unnecessary nodes are deleted to find the optimal result of the classification method.

3.4.4. Support Vector Method. SVM method [33]is used for the classification process. This approach uses the best line used to help with segregation in classes. So that new data can be put into incorrect categories. The hyperplane is the best decision edge in the decision tree. The vector machine determines the vectors which help to create the hyper line. These tremendous points are known as support vectors, so this algorithm is defined as a support vector machine. In SVM multiple lines are used to help the isolate n dimension. Although it require various lines to isolate the class in the different dimension areas.

4. Performance Parameter. The classification of the Parkinson's disease datasets is measured using evaluation metrics such as F1 score, TPR, FPR, accuracy, and Kappa score [34] in Equations 4.1-4.7. These accuracy helps in the assessment of the models. Accuracy is the measure of models performance for all the classes. The Kappa score, also known as Cohen's Kappa coefficient, is a statistic that measures the inter-rater agreement between two or more raters who are assessing the same categorical items. It is particularly useful when dealing with situations where there is a need to assess agreement beyond what might be expected by chance alone as shown in equation 4.1. The True Positive Rate, also known as sensitivity or recall, measures the proportion of actual positive cases that are correctly identified as positive by a classification model as mention in equation 4.4. Similarly, the False Positive Rate measures the proportion of actual negative cases that are incorrectly identified as positive by a classification model. Mathematically, it is calculated as mentioned in 4.5.

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{4.1}$$

$$P_0 = \frac{\sum_{c=1}^{c} TP_C}{\sum_{c=1}^{c} (TP_C + FN_c)} \tag{4.2}$$

$$P_e = \frac{\sum_{c=1}^{c} TP_C * (Tp_c + FN_c)}{N^2} \tag{4.3}$$

$$TPR = \frac{TP}{TP + FN} \tag{4.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.5}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.6}$$

$$AVG = \frac{1}{k + F1 + AUC} \tag{4.7}$$

Table 5.1: KNN classifier

Feature Selection Algorithm	Accuracy(%)	F1-Score(%)	Sensitivity(%)	Specificity(%)	Precision(%)	AUC
All Features	72.11	70.11	72.11	72.11	72.11	0.66
Chi-Squared Method	79.48	85.18	88.46	61.53	82.14	0.75
Genetic Algorithm	87.17	92.06	90.62	71.42	93.54	0.81
Embedded Random Forest Method	89.74	92.59	96.15	76.92	89.28	0.86

Table 5.2: Naive Bayes' Classifier

Feature Selection Algorithm	Accuracy(%)	F1-Score(%)	Sensitivity(%)	Specificity(%)	Precision(%)	AUC
All Features	66.66	73.46	62.06	71.99	71	0.66
Chi-Squared Method	71.79	75.55	89.11	84.61	89.47	0.75
Genetic Algorithm	81.88	77.77	75.91	85.71	95.45	0.75
Embedded Random Forest Method	84.61	88.88	92.30	89.11	85.71	0.80

Table 5.3: SVM Classifier

Feature Selection Algorithm	Accuracy(%)	F1-Score(%)	Sensitivity(%)	Specificity(%)	Precision(%)	AUC
All Features	75	72	70	40	67.99	0.6
Chi-Squared Method	76.92	85.24	80.11	72	74.28	0.65
Genetic Algorithm	94.87	96.96	81.23	71.42	94.11	0.85
Embedded Random Forest Method	84.61	89.28	96.15	77	83.33	0.78

Table 5.4: Decision Tree Classifier

Feature Selection Algorithm	Accuracy(%)	F1-Score(%)	Sensitivity(%)	Specificity(%)	Precision(%)	AUC
All Features	70	72	77	70	71	0.67
Chi-Squared Method	82.05	87.27	92.30	61.53	82.75	0.76
Genetic Algorithm	87.17	92.06	90.62	71.42	93.54	0.81
Embedded Random Forest Method	87.17	90.19	88.46	84.61	92	0.86

5. Results and Discussions. This part of the section provides a compressive study of the various classifier along with the feature extraction method [36]. in this article, we applied three types of feature selection methods such as genetic, chi-square method, and decision tree. Additionally, detection of the disease is processed by the naïve Bayes algorithm, SVM, Random Forest, and decision tree. For the validation of the result, 10-k cross-validation is applied. The system performance is measured by the five parameters such as F1.

The important features are measured by the chi-square [37], genetic, and embedded random forest algorithm. The top eleven features are measured by the chi square method. In the Genetic algorithm, each feature information gain is calculated. The high values of information gain are utilized for the classification method. The top eleven features are selected by the genetic algorithm. Additionally, the last five features are calculated through a random forest algorithm [38]-[39].

The performance parameters i.e. accuracy, sensitivity and specificity are measured by classifier. In this method first features are extracted through feature selection method and then compare the optimization performance of the classifier include all the features and exclude selected features.

The KNN classifier achieves 72.11% of accuracy without feature selection method. further the features selection method named as random forest method is used to enhance the classifier performance. Then classifer optimised the 89.74% of accuracy with feature selection approach as shown in Table 5.1.

The Naïve Bayes classifier achieves 66.66667% of accuracy without feature selection method. further the features selection method named as random forest method is used to enhance the classifier performance. Then classifier optimised the 89.61% of accuracy with feature selection approach as shown in Table 5.2.

The SVM classifier achieves 75% of accuracy without feature selection method. further the features selection method named as random forest method is used to enhance the classifier performance. Then classifier achieved

Table 5.5: Random Forest Classifier

Feature Selection Algorithm	Accuracy(%)	F1-Score(%)	Sensitivity(%)	Specificity(%)	Precision(%)	AUC
All Features	72.99	74.88	34	70	67.99	0.70
Chi-Squared Method	84.61	88.88	92.30	69.23	85.71	0.80
Genetic Algorithm	93.55	95.23	93.75	85.71	96.77	0.89
Embedded Random Forest Method	97.89	96.69	92.30	92.30	96	0.92

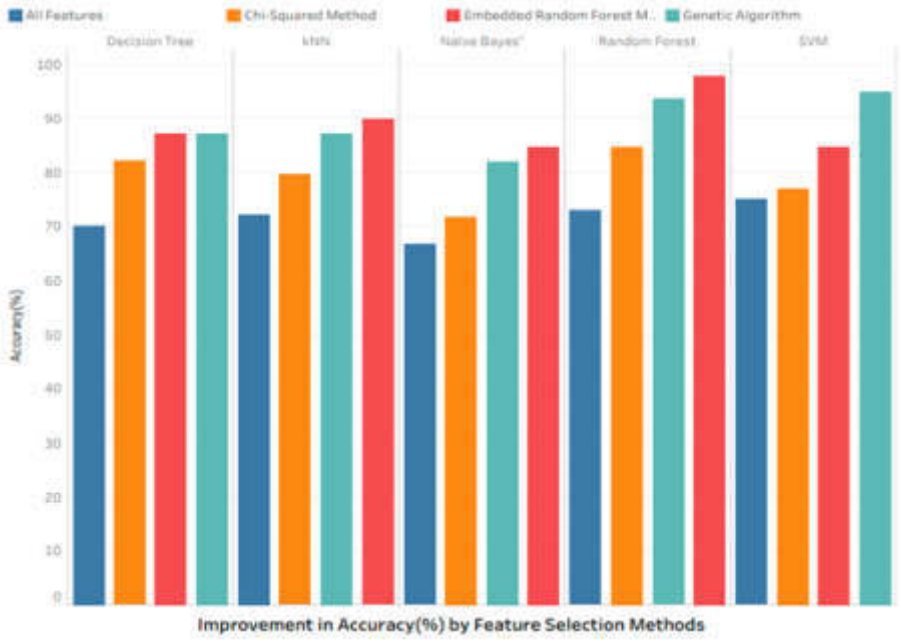


Fig. 5.1: Accuracy improvement by feature selection methods

the 84.61% of accuracy with feature selection approach as shown in Table 5.3.

The Decision classifier achieves 70% of accuracy without feature selection method. further the features selection method named as random forest method is used to enhance the classifier performance. Then classifier achieved the 87.17% of accuracy with feature selection approach as shown in Table 5.4.

The Random forest classifier achieves 72% of accuracy without feature selection method. further the features selection method named as random forest method is used to enhance the classifier performance. Then classifier achieved the 97.89% of accuracy with feature selection approach as shown in Table 5.5.

The performance improvement of the proposed methodology with the existing method is defined in Fig. 5.1. The detection accuracy of the classifier is optimized through the feature selection method. The Embedded Random Forest method provides the best result compared to the existing feature selection method. The proposed detection model of Parkinson's disease gives a better result compared to the existing model. The proposed model accuracy of 97.89% in comparison to the existing models is defined in Fig. 5.2.

6. Conclusion. Parkinson's Disease is mainly voice disorder due to neurodegenerative. Therefore, there is only one way to improve the patent's health i.e., through early detection of the disease. The proper diet plan and medication can improve the symptoms of patients with Parkinson's disease. This experimental result shows the detection of the disease at early phase. The author has utilised different combinations of feature selection as well as the classifier methods. Finally, the proposed model provides 97.89% of accuracy which is better than the existing recent work of literature.

In the future, author will plan to work on other datasets namely voice and work on slowness in handwriting

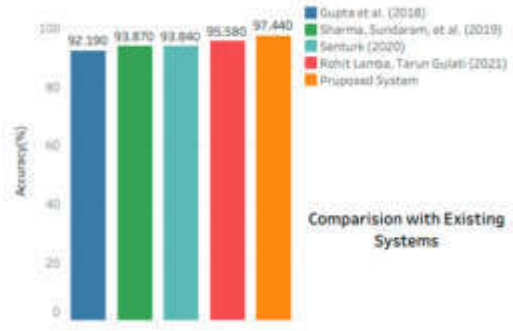


Fig. 5.2: Comparison with existing systems

skill symptoms. Developing personalized diagnostic models that consider individual variations and response to treatments is a growing area of interest. Random classification and feature selection techniques can aid in tailoring diagnoses to each patient's unique characteristics. Explaining the rationale behind a diagnosis is crucial for building trust in medical AI systems. Future research could focus on incorporating interpretability techniques alongside random classification and feature selection methods.

REFERENCES

[1] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, *Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques*, Pattern Recognition Letters, 125 (2019), pp. 55–62.

[2] S. Bhattacharya, P. K. Reddy Maddikunta, Q.-V. Pham, T. R. Gadekallu, S. R. Krishnan S, C. L. Chowdhary, M. Alazab, and M. Jalil Piran, *Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey*, Sustainable Cities and Society, 65 (2021), p. 102589.

[3] G. Chandrashekar and F. Sahin, *A survey on feature selection methods*, Computers & Electrical Engineering, 40 (2014), pp. 16–28.

[4] C. L. Chowdhary and D. Acharjya, *Segmentation and feature extraction in medical imaging: A systematic review*, Procedia Computer Science, 167 (2020), pp. 26–36. International Conference on Computational Intelligence and Data Science.

[5] F. N. Emamzadeh and A. Surguchov, *Parkinson's disease: biomarkers, treatment, and risk factors*, Frontiers in neuroscience, 12 (2018), p. 612.

[6] S. Grover, S. Bhartia, A. Yadav, K. Seeja, et al., *Predicting severity of parkinson's disease using deep learning*, Procedia computer science, 132 (2018), pp. 1788–1794.

[7] H. Gunduz, *Deep learning-based parkinson's disease classification using vocal feature sets*, IEEE Access, 7 (2019), pp. 115540–115551.

[8] S. Lahmiri and A. Shmuel, *Detection of parkinson's disease based on voice patterns ranking and optimized support vector machine*, Biomedical Signal Processing and Control, 49 (2019), pp. 427–433.

[9] R. Lamba, T. Gulati, K. A. Al-Dhlan, and A. Jain, *A systematic approach to diagnose parkinson's disease through kinematic features extracted from handwritten drawings*, Journal of Reliable Intelligent Environments, (2021), pp. 1–10.

[10] R. Lamba, T. Gulati, and A. Jain, *Comparative analysis of parkinson's disease diagnosis system*, Adv Math Sci J, 9 (2020), pp. 3399–3406.

[11] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, *Suitability of dysphonia measurements for telemonitoring of parkinson's disease*, Nature Precedings, (2008), pp. 1–1.

[12] S. A. Mostafa, A. Mustapha, M. A. Mohammed, R. I. Hamed, N. Arunkumar, M. K. Abd Ghani, M. M. Jaber, and S. H. Khaleefah, *Examining multiple feature evaluation and classification methods for improving the diagnosis of parkinson's disease*, Cognitive Systems Research, 54 (2019), pp. 90–99.

[13] Y. S. Murthy and S. G. Koolagudi, *Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (gafs)*, Expert Systems with Applications, 106 (2018), pp. 77–91.

[14] L. Naranjo, C. J. Perez, Y. Campos-Roca, and J. Martin, *Addressing voice recording replications for parkinson's disease detection*, Expert Systems with Applications, 46 (2016), pp. 286–292.

[15] I. Nissar, D. R. Rizvi, S. Masood, and A. N. Mir, *Voice-based detection of parkinson's disease through ensemble machine learning approach: A performance study*, EAI Endorsed Transactions on Pervasive Health and Technology, 5 (2019), pp. e2–e2.

[16] R. Olivares, R. Munoz, R. Soto, B. Crawford, D. Cárdenas, A. Ponce, and C. Taramasco, *An optimized brain-based algorithm for classifying parkinson's disease*, Applied Sciences, 10 (2020), p. 1827.
[17] K. Polat, *A hybrid approach to parkinson disease classification using speech signal: the combination of smote and random forests*, in 2019 scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT), Ieee, 2019, pp. 1–3.
[18] P. Rani, R. Kumar, N. Ahmed, and A. Jain, *A decision support system for heart disease prediction based upon machine learning. j reliab intell environ*, 2021.
[19] P. Rani, R. Kumar, and A. Jain, *Multistage model for accurate prediction of missing values using imputation methods in heart disease dataset*, in Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, Springer, 2021, pp. 637–653.
[20] P. Rani, R. Kumar, A. Jain, and R. Lamba, *Taxonomy of machine learning algorithms and its applications*, Journal of Computational and Theoretical Nanoscience, 17 (2020), pp. 2508–2513.
[21] G. T. Reddy, S. Bhattacharya, S. Siva Ramakrishnan, C. L. Chowdhary, S. Hakak, R. Kaluri, and M. Praveen Kumar Reddy, *An ensemble based machine learning model for diabetic retinopathy classification*, in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–6.
[22] S. G. Reich and J. M. Savitt, *Parkinson disease*, Medical Clinics of North America, (2018).
[23] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, *A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform*, Applied Soft Computing, 74 (2019), pp. 255–263.
[24] I. H. Sarker, A. Kayes, and P. Watters, *Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage*, Journal of Big Data, 6 (2019), pp. 1–28.
[25] Z. K. Senturk, *Early diagnosis of parkinson's disease using machine learning algorithms*, Medical hypotheses, 138 (2020), p. 109603.
[26] P. Sharma, R. Jain, M. Sharma, and D. Gupta, *Parkinson's diagnosis using ant-lion optimisation algorithm*, International Journal of Innovative Computing and Applications, 10 (2019), pp. 138–146.
[27] P. Sharma, S. Sundaram, M. Sharma, A. Sharma, and D. Gupta, *Diagnosis of parkinson's disease using modified grey wolf optimization*, Cognitive Systems Research, 54 (2019), pp. 100–115.
[28] V. S. Sharma, A. Afthanorhan, N. C. Barwar, S. Singh, and H. Malik, *A dynamic repository approach for small file management with fast access time on hadoop cluster: Hash based extended hadoop archive*, IEEE Access, 10 (2022), pp. 36856–36867.
[29] V. S. Sharma and N. Barwar, *An efficient approach to enhance the scalability of the hdfs: Extended hadoop archive (ehar)*, in 2021 Emerging Trends in Industry 4.0 (ETI 4.0), 2021, pp. 1–6.
[30] V. S. Sharma and N. C. Barwar, *Performance evaluation of merging techniques for handling small size files in hdfs*, in Data Analytics and Management, A. Khanna, D. Gupta, Z. Pólkowski, S. Bhattacharyya, and O. Castillo, eds., Singapore, 2021, Springer Singapore, pp. 137–150.
[31] G. Solana-Lavalle, J.-C. Galán-Hernández, and R. Rosas-Romero, *Automatic parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features*, Biocybernetics and Biomedical Engineering, 40 (2020), pp. 505–516.
[32] T. Tuncer and S. Dogan, *A novel octopus based parkinson's disease and gender recognition method using vowels*, Applied Acoustics, 155 (2019), pp. 75–83.
[33] T. Tuncer, S. Dogan, and U. R. Acharya, *Automated detection of parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels*, Biocybernetics and Biomedical Engineering, 40 (2020), pp. 211–220.
[34] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, *Comparing different supervised machine learning algorithms for disease prediction*, BMC medical informatics and decision making, 19 (2019), pp. 1–16.
[35] S. P. Vijay Kumar Gurani, Chiranji Lal Chowdhary, *Exploring breast cancer classification of histopathology images from computer vision and image processing algorithms to deep learning*, International Journal of Advanced Science and Technology, 29 (2020), pp. 43 – 48.
[36] Y. Xiong and Y. Lu, *Deep feature extraction from the vocal vectors using sparse autoencoders for parkinson's classification*, IEEE Access, 8 (2020), pp. 27821–27830.
[37] O. Yaman, F. Ertam, and T. Tuncer, *Automated parkinson's disease recognition based on statistical pooling method using acoustic features*, Medical hypotheses, 135 (2020), p. 109483.
[38] L. Zahid, M. Maqsood, M. Y. Durrani, M. Bakhtyar, J. Baber, H. Jamal, I. Mehmood, and O.-Y. Song, *A spectrogram-based deep feature assisted computer-aided diagnostic system for parkinson's disease*, IEEE Access, 8 (2020), pp. 35482–35495.
[39] T. A. Zesiewicz, Y. Bezchlibnyk, N. Dohse, and S. D. Ghanekar, *Management of early parkinson disease*, Clinics in geriatric medicine, 36 (2020), pp. 35–41.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: May 25, 2023
Accepted: Aug 24, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 177–186, DOI 10.12694/scpe.v25i1.2317

OCULAR DISEASE SEVERITY IDENTIFICATION AND PERFORMANCE OPTIMISATION USING CUSTOM NET MODEL

SUMAN BHAKAR* PARTHI VISHNAWAT † NIDHI KUNDU‡ AND VIJAY SHANKAR SHARMA §

Abstract. Early detection and timely cure of ocular disease play a vital role to avoid irreversible vision issues in daily life. The technique fundus assessment utilizes color fundus photography, which is a very effective tool though it is expensive. Since rare symptoms of the disease are detected at the initial stage of the disease, still automated and optimized models are in urgent need for the detection of the ocular disease. Additionally, existing systems focus on image-level detection for the treatment of eyes without association employing the left and right eye information. Although they concentrate only on one or two features of the ocular disease at a time. Taking into consideration severity detection and multilabel categorization plays a vital role in ocular disease detection. So, we develop a framework to detect the disease in the early phase. And then apply the classification model for the multilabel classification of the disease. our proposed experimental result proves that the proposed Custom net model provides 99.15% of accuracy compared to the existing baseline model such as Vgg16, 19, Resnet-50 and Inception V3. The performance optimization of the proposed model is evaluated on the public datasets.

Key words: Deep learning, ROP, Custom-net, Disease, Severity

1. Introduction. Retinopathy of Prematurity (ROP) is a disease occurs in premature babies having low birth weight. This causes blindness in such kids [11, 5]. The shorter the gestation or the lighter the birth weight, the more the chance of ROP disease. So, the factors viz. gestational age, birth weight and supplemental oxygen status helps in detecting the ROP [4].

The ROP screening should be done from time to time after birth of premature babies. It can be terminated once there is complete vascularization of retina without any ROP, or if the ROP has shown complete regression.The disease can be graded in terms of zone, extent in clock hour, stage of ROP(1,2,3,4a,4b,5), Aggressive Posterior (APROP) and disease status [1].

This paper proposes an elaborate design of a deep learning model to detect and classify ocular disease. The deep learning model is composed of three units: CNN model, feature extraction and classification. The CNN model is utilized for the extraction of the features from the datasets and feature extraction and image processing are utilized for the enhancement and synthesis of the features [26, 2]. The classification model creates the output of the classification. Our proposed deep learning model archives an inspirational performance on the ocular dataset. In this article, we broadly observe the results. Also, we discuss the performance parameters viz. accuracy, F1 score, Recall, Precision.

Our contribution is as follows:

- Proposed the deep learning neural network to identify the disease.
- A novel module, feature extraction, and image processing propose to enhance the different features from the datasets.
- Implementation of VGGNet-16,VGGNet-19 ,ResNet 50 and Custom net module on the ocular datasets. Also, utilized the Explainable AI to identify, whether this system is trustworthy or not. Fig 11 depicts the Explainable AI results, as you can see the results are not that accurate but in future we can improve

*Department of Computer and Communication Engineering, Manipal University Jaipur, India (Suman.bhakar@jaipur.manipal.edu)

†Department of Computer and Communication Engineering, Manipal University Jaipur, India (parthi.209303013@muj.manipal.edu

‡Department of Computer Science, SKNAU, Jobner, India (kundu.nidhi1990@gmail.com)

§Department of Computer and Communication Engineering, Manipal University Jaipur, India (Corresponding author: vijayshankar.sharma@jaipur.manipal.edu)

it.Substantially optimized the performance of the model.
- Adam optimizer is used to optimize the results of the model.
- Compared the performance matrices such as accuracy, F1 score of VGG 16, VGG 19, and ResNet-50 and custom net model also proved that Custom net model achieves better performance compared to other models.

2. Literature Review. Although, there are different existing imagining technology to detect ocular disease Optical Coherence Tomography (OCT) and Color Fundus Photography (CFP) are widely utilized nowadays [25]. Cross-section images based on retina created by the OCT, to measure the eye condition the retinal thickness is utilized. CFP maintains the interior parts of the eyes to examine the possible syndromes. CFP and OCT tools have been confirmed to identify the early stages of the ocular disease [8]. Although, CFP is costly but effective tool for the adults because periodically fundus assessment of the adults provides symptomless result. However, some ocular diseases viz. diabetic retinopathy, macular degeneration based on age-related and cataract with very few symptoms are very difficult to diagnose at an initial stage. Furthermore, manual examination of the created huge scale of CFP is a very time-consuming process. So, an automated system is developed to detect at the disease at an early stage and also to improve the accuracy of the detection [3].

Deep Neural networks (DNN), especially Convolutional Neural Networks (CNN) have proven to play a vital role in the field of medical imagining[4]. Moreover, CNN has proven to diagnose ocular disease into various levels of disease classification as well as detection of an object. The detection of the fovea centre in the Oct image has been performed by the pixel classification approach. The optic disk in the field of CFP [20] has been detected through the CNN model i.e. based on two-stages. The authors suggested adopting CNN to analyse the fluid in OCT images. Also,the encoder-Decoder based network has been developed to measure the various retinal layers and also computes the collected fluid in different OCT images. The portion of the retinal vessels based on CFP is identified by both combined and fully connected CNN. The improvement of the accuracy is achieved by the image level observation to identify the segmentation of the retinal image [19]. The classification of the ocular disease by CNN has been more attention rather than object detection and segmentation [7]. The authors classified the ocular disease according to the severity levels of the disease.

The transfer learning model [15] such as image net and inception network was observed to be very successful to ocular disease classification into a different stage. A better classification result could be attained by the Ensemble learning.

Furthermore, there are very few words that refer to the problem of ocular disease classification based on multi-label [29]. Also, one single patient can be affected by multiple types of ocular diseases. Additionally, there is a high probability of the patients who are affected by multiple ocular diseases. So, there is a need to find the optimal model to get a better result and more classification of the ocular disease [27]. The authors have identified the existence of myopia, runs the high rate of false-negative value in glaucoma patients. Subsequently, the existing technology is generated to give acceptable result for the specific activity. But this technology is not appropriate for real-time situations [14]. Moreover, for the mentioned issue there are very less works on the ocular disease classifications [16]. There are existing publications that are mainly focused on the analysis of the CFP images that are generated from the left as well as right eyes [12, 21, 10]. However, detection of the disease of patients through pre-screening is very risky. So, it is not recommended for the long-term procedure.

The authors [18] proposed a deep supervision-based network for Retinopathy of Prematurity (ROP) detection and classification into mild, moderate, and severe classes. The authors utilized two ROP detection datasets obtained from Guangzhou women and children's medical center. First dataset contained 7396 fundus images which were used for ROP detection. The second dataset contained 1337 fundus images that was used for classification into three classes. The data had been collected from year 2012 to 2015. The data was annotated by three experienced pediatric ophthalmologists that labeled the data for detection and classification purpose and only consistent resultant images were considered for the training and testing phase. The authors found consistent results in annotation for detection purpose but variable results in annotation done for classification purpose due to subjective evaluation. The authors utilized DenseNet121 CNN for feature extraction in the initial phase. In DenseNet121, every layer of the network was connected to the first layer that helped it to reuse the features. Later, they embedded multi semantic feature aggregation into CNN to handle the problems of local redundancy and complex global dependencies. Lastly, they used deep supervised learning strategy in which they added

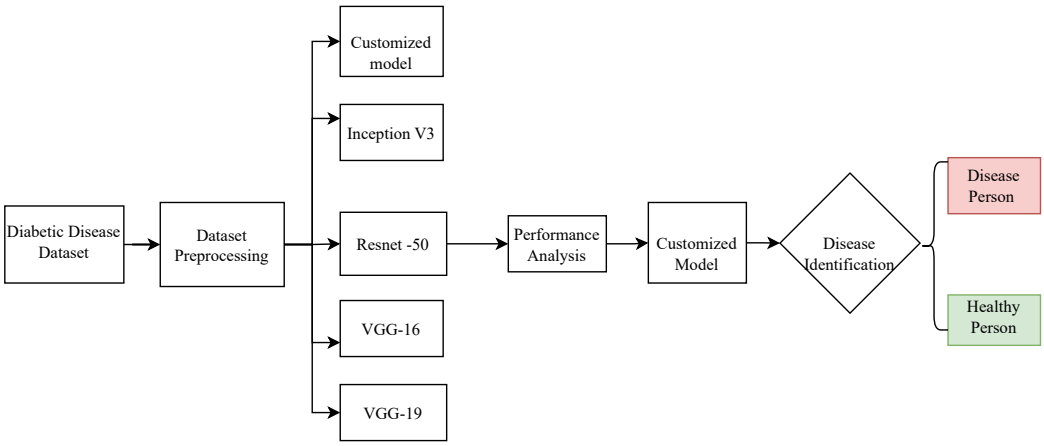


Fig. 3.1: Proposed Framework for Diabetic Disease Detection

two more auxiliary classifiers after the second and third stages in DenseNet121. This architecture benefited the system by providing full use of feature information in hidden layers and optimized the network. They used transfer learning to train the model due to small size of data. The system achieved an accuracy of 97.76% with recall and precision of 97.14% and 98.35% respectively. The proposed system outperformed many other CNN architectures such as DenseNet169, ResNet50, Resnext50 and InceptionV4. The classification process was tricky for experienced ophthalmologists and a challenge for system due to less amount of available data.

The authors [28] proposed an automated Aggressive Posterior-ROP (AP-ROP) diagnostic system that identified ROP from fundus images and classified them into normal ROP and AP-ROP. AP-ROP is a special type of ROP that evolved very rapidly in the fifth stage of ROP. The authors used dataset collected from Shenzhen Eye Hospital from 2009 to 2018 with 13,508 fundus images taken from RetCam3. The images were annotated by experienced ophthalmologists and shortlisted 12230 images based on good quality and clarity of ROP stages. Dataset consisted of 1698 AP-ROP, 4033 Regular ROP and 6499 Normal fundus images. it was divided into training, testing and validation set in the ratio of 54: 30:14. The authors used ResNet-18, 34, 50, 101 for feature extraction network and found all of them performing well in this case. The system used Hierarchical Bilinear Pooling (HBP) module that expanded the features of different layers to a high dimensional space by creating independent linear mappings in the network. Later, it integrated the features of different Convolutional layers through element-wise multiplication. Afterwards, the HBP module compressed high dimensional features using summation technique to obtain inter-layer interaction features. The system also used transfer learning to transfer the parameter learned by Network 1 to Network 2 to improve the classification performance of network 2. The system achieved accuracy of 98.88% with AUC value of 99.93% for task1. It had obtained an accuracy of 93.03% for task 2 using transfer learning mechanism.The authors proposed a solution for dealing with small files, the concept can be used in implementing database with Hadoop and can provide a simple way to store health data [22, 24, 23].

3. Methodology. In this manuscript, the authors propose a framework to detect the disease whether that person is healthy, disease diagnosed person. The proposed framework is categorized into three components. First component, Data pre-processing has been processed and then disease detection was done by different machine learning classifiers and custom net models. In this section, the authors explained about the proposed workflow, dataset, preprocessing, evaluation metrics and training details as shown in Fig. 3.1.

3.1. Dataset. Ocular Disease Intelligent Recognition(ODIR) is a structured ophthalmic of 5000 patients with Patient age, Patient sex, Left and Right Fundus, etc. data labels. The dataset represent real-life patient information collected by several doctors. The datasets classified as healthy,diseased (cataract, retina). At first, we downloaded the dataset from Kaggle, and required model weights were taken to indenfication of compatibility environment of VGG16 classification model. The image size should be 224*224 for the input of the model[17].

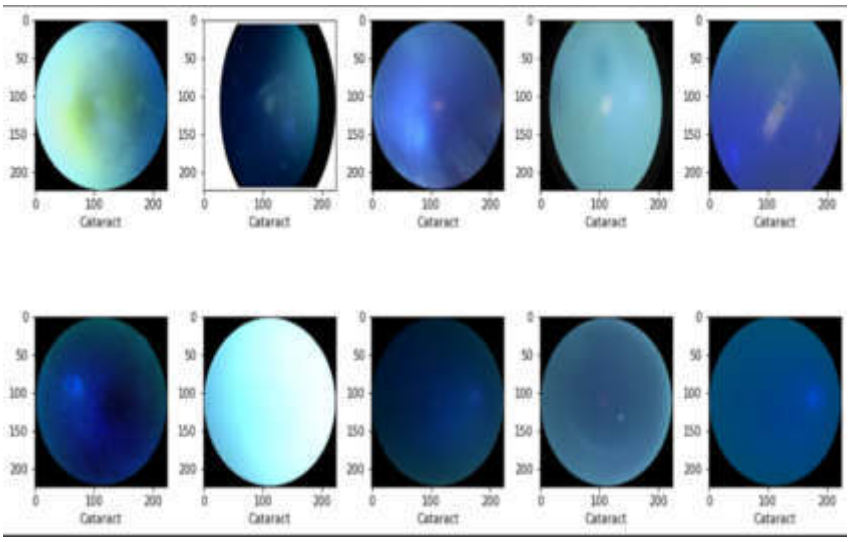


Fig. 3.2: Before augmentation

After that we performed some pre-processing on the images to make sure that the images are compatible with the model or not. It has been performed by the keras liberary with Data generator module.
To validate the system public ocular disease dataset is utilized. The dataset is divided into training set, testing set, and validation set as 80% training dataset, 10% as a testing set, and the remaining 10% for the validation set. The size of original datasets images is 3000*1700. Then by using image processing the size of the images is converted to 224* 224 dimensions.

3.2. Pre-processing. Cropped images are the input of data augmentation techniques. To get the left and right side of the image, augmentation is applied as shown in 3.2-3.3. Also, we applied the Gaussian filter to remove the noise and better visualization as shown in 3.4. The gaussian filter works on concept of Equation 3.1.

$$G(x) = \frac{1}{\sqrt{2\prod\sigma^2}} e^{\frac{x^2}{2\sigma^2}} \tag{3.1}$$

Here σ is the distribution of the standard deviation. Labeling performs a vital role to predict the classification of the image. It classifies the type of eye (either left or right) or type of disease as shown in Fig. 3.5.

3.3. Model architecture. The design of the "custom net" proposed to identify the diabetic disease dataset as shown in fig. consists of four convolution layers. Each convolution layer is supported by the max pooling layer. Additionally, the final max pooling layer is carried through the flattened and dense layer.ResNet model's different layers are used to extract the different features by utilizing CNN. The CNN [13] initialization is performed by the ImageNet. Ocular disease datasets are collections of 5000 images with different categories of the disease.

3.4. Evaluation Metrics. The classification of the ocular disease datasets is calculated with four evaluation matrices such as F1 score, TPR, FPR, accuracy,and Kappa score [9]-[6] as shown in Equation 3.2-3.8.

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{3.2}$$

$$P_0 = \frac{\sum_{c=1}^{c} TP_C}{\sum_{c=1}^{c}(TP_C + FN_c)} \tag{3.3}$$

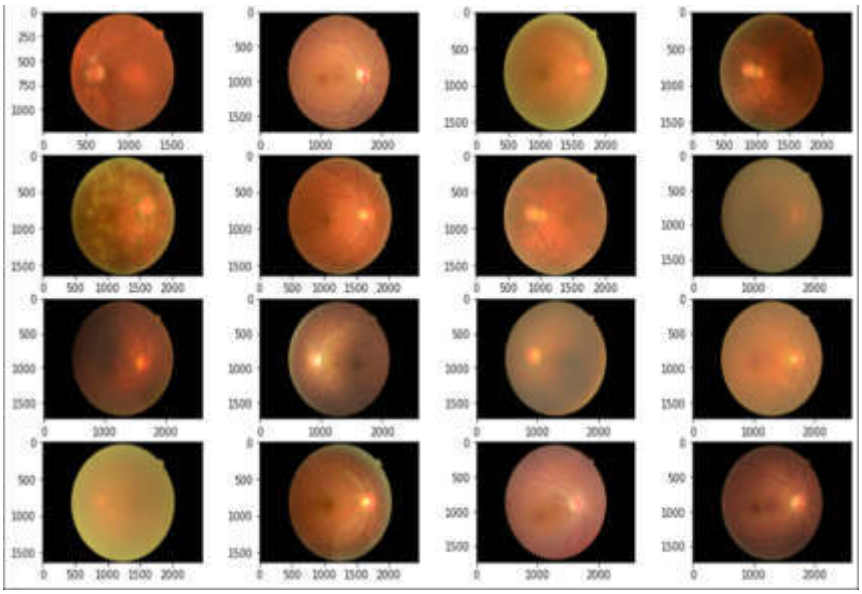


Fig. 3.3: After augmentation

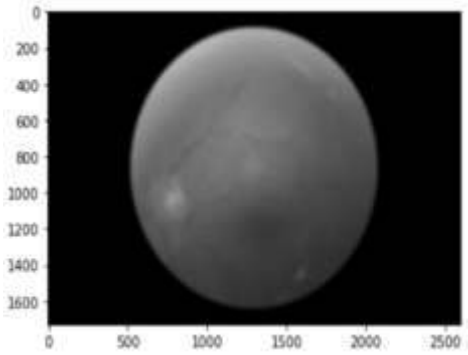


Fig. 3.4: Gaussian Filter for the thresholding

$$P_e = \frac{\sum_{c=1}^{c} TP_C * (Tp_c + FN_c)}{N^2} \tag{3.4}$$

$$TPR = \frac{TP}{TP + FN} \tag{3.5}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.6}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Preciison + Recall} \tag{3.7}$$

$$AVG = \frac{1}{k + F1 + AUC} \tag{3.8}$$

Fig. 3.5: Labeling prediction of left and right eye with the disease

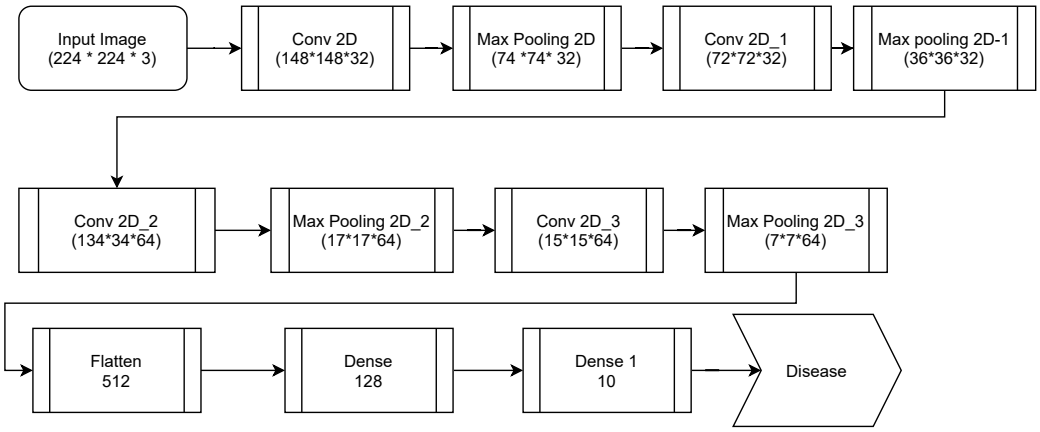


Fig. 3.6: Proposed Custom net Model

3.5. Training details. The Pytorch [17] is employed to implement the deep neural network. Models experiments are executed by NVIDIA 100Ti GPUs. For the optimization network, Adam optimizer is adopted to categorize into different classifications. The rate of learning is set as 0.0006, that is decay with decay policy lr = initially *(1-iter totaliter)power. The model power is set at 0.9. The complete experimental executed on 50 Epocs.

4. Experimental Results. In this section, the authors define the performance measurement of the custom net model based on the test data set, that is containing 5,000 images of dataset comprises of diabetic disease and healthy . They evaluate the pretrained and non-pre-trained versions of the deep learning model. The authors also evaluate the accuracy, F1 score, and precision and recall.

4.1. Average Accuracy. The Average accuracy calculates the pre-trained and non-pretrained versions of custom net mode and baseline models such as the Inception-V3, Resnet -50, and VGG -19,VGG-16.

The custom net model average accuracy based on non- pretrained version is 99.15%, and the pre-trained

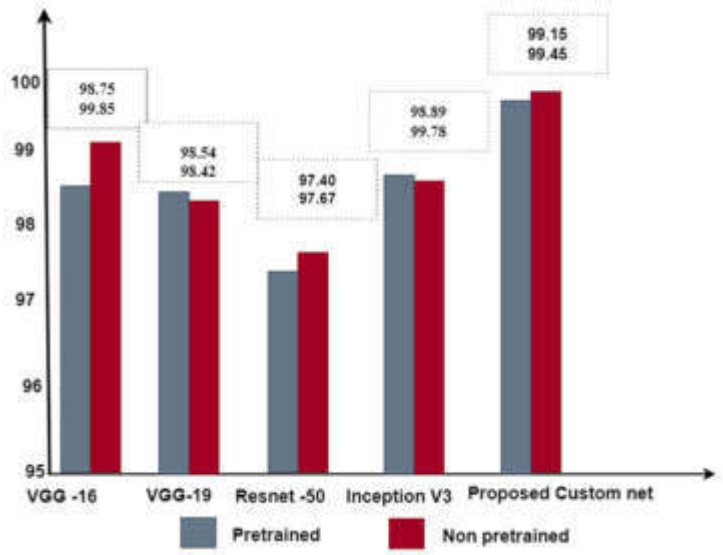


Fig. 4.1: Accuracy comparison with Custom net Model

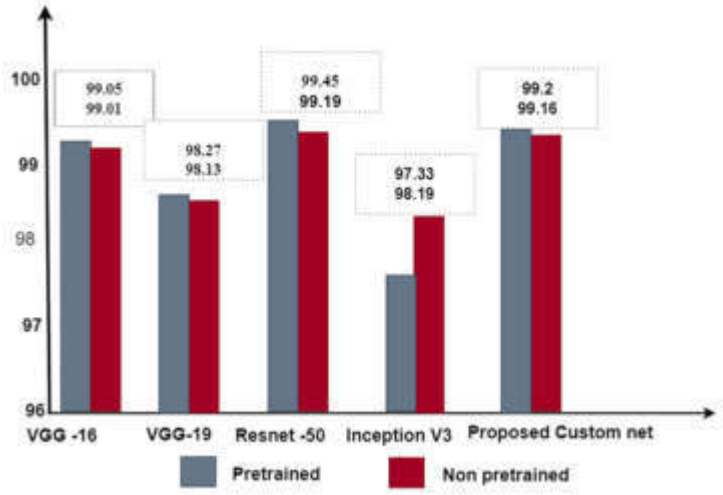


Fig. 4.2: Precision comparison with Custo net Model

model of custom net model accuracy is calculated at 99.45%. Similarly, the non-pre trained and pre-trained version of the Inception model V3 accuracy is reported as 98.89% and 98.79%. The Resnet -50 based on the non-pretrained model accuracy reported 88.30% in comparison to 97.40% of the pre-trained model. And the VGG -19 models report the average accuracy of pre-trained and non-pre-trained models are 98.54% and 98.42% as shown in Fig. 4.1.

4.2. Precision. The pre-trained model of the Resnet -50 provides the highest precision of 99.45% and there is a 0.26% percentage of the difference between the trained and non-pre-trained model of the Ressnet -50. Whereas the inception V3 model reports 97.33% accuracy of the pretrained model and 98.19% accuracy of the non-pre-trained model as shown in Fig. 4.2.

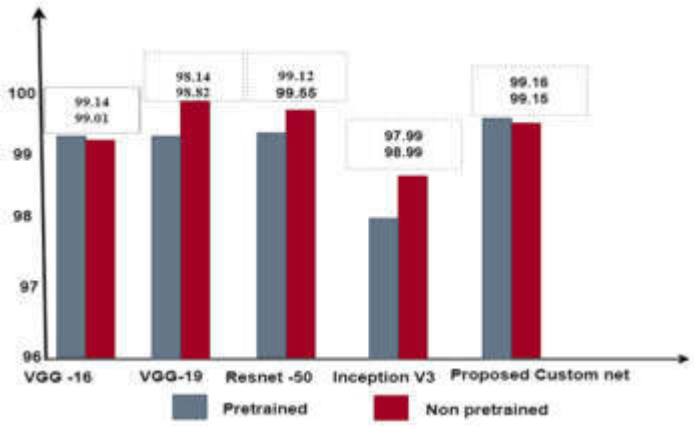


Fig. 4.3: Recall comparison with Custom net Model

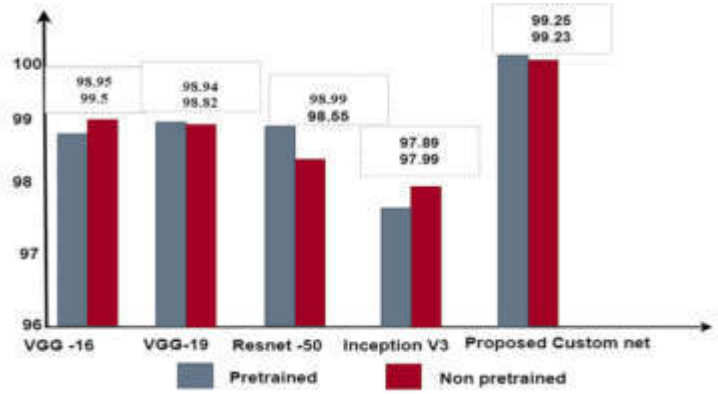


Fig. 4.4: F1-score comparison with Custom net Model

4.3. Recall. The custom net model average recall value of the non -pretrained version is 99.16%, and the pretrained model of custom net model accuracy is calculated as 99.15%. Similarly, the non- pretrained and pretrained version of the Inception V3 model accuracy is reported as 99.99% and 98.99%. The Resnet -50 based on non-pretrained model accuracy reported as 99.55% whereas 99.12% of pretrained model as depict in the Fig. 4.3.

4.4. F1-score. The custom-net model reports 99.25% of F1 score through pre trained model and 99.23% of accuracy through non pretrained model. Whereas inception V3, Resnet -50, VGG -16 and VGG -19 report the F1 score are 97.89%, 98.99%, 98.94%, 98.15% through pre-trained model. The proposed custom model achieves the highest F1-score among other models as shown in Fig. 4.4.

5. Conclusion. In this article, we proposed a method to detect the ocular disease and classify it into different phases according to the severity of the disease. The pre-processing phase include the feature extraction, augmentation, and labializing of the datasets. Additionally, we applied the gaussian filter to remove the unwanted noise. An ocular set of datasets is used for the predication of the disease. The classification of the disease is processed by the custom net model, VGG16, 19 and Resnet -50 and inception V3 and it is proved that proposed custom net model provides the better result compared to existing baseline model. still there are some limitations, that will be addressed in the future studies. Firstly, the proposed model can also be implemented on the other disease classification such as cancer, pneumonia disease. Secondly, this model was not compared

with other existing image processing methods for the better classification. Thirdly, it can be implemented on the rest of CNN module to identify the better accuracy.

REFERENCES

[1] R. Agrawal, S. Kulkarni, R. Walambe, and K. Kotecha, *Assistive framework for automatic detection of all the zones in retinopathy of prematurity using deep learning*, Journal of Digital Imaging, 34 (2021), pp. 932–947.

[2] S. Bhattacharya, P. K. Reddy Maddikunta, Q.-V. Pham, T. R. Gadekallu, S. R. Krishnan S, C. L. Chowdhary, M. Alazab, and M. Jalil Piran, *Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey*, Sustainable Cities and Society, 65 (2021), p. 102589.

[3] R. R. Bourne, G. A. Stevens, R. A. White, J. L. Smith, S. R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, K. Naidoo, et al., *Causes of vision loss worldwide, 1990–2010: a systematic analysis*, The lancet global health, 1 (2013), pp. e339–e349.

[4] J. P. Campbell, P. Singh, T. K. Redd, J. M. Brown, P. K. Shah, P. Subramanian, R. Rajan, N. Valikodath, E. Cole, S. Ostmo, et al., *Applications of artificial intelligence for retinopathy of prematurity screening*, Pediatrics, 147 (2021).

[5] J. Chen and L. E. Smith, *Retinopathy of prematurity*, Angiogenesis, 10 (2007), pp. 133–140.

[6] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, *Multi-label image recognition with graph convolutional networks*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5177–5186.

[7] C. L. Chowdhary and D. Acharjya, *Segmentation and feature extraction in medical imaging: A systematic review*, Procedia Computer Science, 167 (2020), pp. 26–36. International Conference on Computational Intelligence and Data Science.

[8] N. Congdon, B. O'Colmain, C. Klaver, R. Klein, B. Munoz, D. S. Friedman, J. Kempen, H. R. Taylor, P. Mitchell, et al., *Causes and prevalence of visual impairment among adults in the united states.*, Archives of Ophthalmology (Chicago, Ill.: 1960), 122 (2004), pp. 477–485.

[9] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] K. Hu, Z. Zhang, X. Niu, Y. Zhang, C. Cao, F. Xiao, and X. Gao, *Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function*, Neurocomputing, 309 (2018), pp. 179–191.

[11] S. J. Kim, A. D. Port, R. Swan, J. P. Campbell, R. P. Chan, and M. F. Chiang, *Retinopathy of prematurity: a review of risk factors and their clinical significance*, Survey of ophthalmology, 63 (2018), pp. 618–637.

[12] C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, *Deep-learning based, automated segmentation of macular edema in optical coherence tomography*, Biomedical optics express, 8 (2017), pp. 3440–3448.

[13] C. Li, J. Ye, J. He, S. Wang, Y. Qiao, and L. Gu, *Dense correlation network for automated multi-label ocular disease detection with paired color fundus photographs*, in 2020 IEEE 17th international symposium on biomedical imaging (ISBI), IEEE, 2020, pp. 1–4.

[14] B. Liefers, F. G. Venhuizen, V. Schreur, B. van Ginneken, C. Hoyng, S. Fauser, T. Theelen, and C. I. Sánchez, *Automatic detection of the foveal center in optical coherence tomography*, Biomedical Optics Express, 8 (2017), pp. 5160–5178.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, *A survey on deep learning in medical image analysis*, Medical image analysis, 42 (2017), pp. 60–88.

[16] X. Meng, X. Xi, L. Yang, G. Zhang, Y. Yin, and X. Chen, *Fast and effective optic disk localization based on convolutional neural network*, Neurocomputing, 312 (2018), pp. 285–295.

[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, in Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[18] Y. Peng, Z. Chen, W. Zhu, F. Shi, M. Wang, Y. Zhou, D. Xiang, X. Chen, and F. Chen, *Ads-net: attention-awareness and deep supervision based network for automatic detection of retinopathy of prematurity*, Biomedical Optics Express, 13 (2022), pp. 4087–4101.

[19] G. T. Reddy, S. Bhattacharya, S. Siva Ramakrishnan, C. L. Chowdhary, S. Hakak, R. Kaluri, and M. Praveen Kumar Reddy, *An ensemble based machine learning model for diabetic retinopathy classification*, in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–6.

[20] S. Rowe, C. H. MacLean, and P. G. Shekelle, *Preventing visual loss from chronic eye disease in primary care: scientific review*, Jama, 291 (2004), pp. 1487–1495.

[21] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, *Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks*, Biomedical optics express, 8 (2017), pp. 3627–3642.

[22] V. S. Sharma, A. Afthanorhan, N. C. Barwar, S. Singh, and H. Malik, *A dynamic repository approach for small file management with fast access time on hadoop cluster: Hash based extended hadoop archive*, IEEE Access, 10 (2022), pp. 36856–36867.

[23] V. S. Sharma and N. Barwar, *An efficient approach to enhance the scalability of the hdfs: Extended hadoop archive (ehar)*, in 2021 Emerging Trends in Industry 4.0 (ETI 4.0), 2021, pp. 1–6.

[24] V. S. Sharma and N. C. Barwar, *Performance evaluation of merging techniques for handling small size files in hdfs*, in Data Analytics and Management, A. Khanna, D. Gupta, Z. Pólkowski, S. Bhattacharyya, and O. Castillo, eds., Singapore,

2021, Springer Singapore, pp. 137–150.
[25] A. Sommer, J. M. Tielsch, J. Katz, H. A. Quigley, J. D. Gottsch, J. C. Javitt, J. F. Martone, R. M. Royall, K. A. Witt, and S. Ezrine, *Racial differences in the cause-specific prevalence of blindness in east baltimore*, New England journal of medicine, 325 (1991), pp. 1412–1417.
[26] S. P. Vijay Kumar Gurani, Chiranji Lal Chowdhary, *Exploring breast cancer classification of histopathology images from computer vision and image processing algorithms to deep learning*, International Journal of Advanced Science and Technology, 29 (2020), pp. 43 – 48.
[27] J. Yang, H. Deng, X. Huang, B. Ni, and Y. Xu, *Relational learning between multiple pulmonary nodules via deep set attention transformers*, in 2020 IEEE 17th international symposium on biomedical imaging (ISBI), IEEE, 2020, pp. 1875–1878.
[28] R. Zhang, J. Zhao, H. Xie, T. Wang, G. Chen, G. Zhang, and B. Lei, *Automatic diagnosis for aggressive posterior retinopathy of prematurity via deep attentive convolutional neural network*, Expert Systems with Applications, 187 (2022), p. 115843.
[29] W. Zhao, J. Yang, Y. Sun, C. Li, W. Wu, L. Jin, Z. Yang, B. Ni, P. Gao, P. Wang, et al., *3d deep learning from ct scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas*, Cancer research, 78 (2018), pp. 6881–6889.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: May 26, 2023
Accepted: Oct 10, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 187–203, DOI 10.12694/scpe.v25i1.2326

AN INSIGHT INTO VIABLE MACHINE LEARNING MODELS FOR EARLY DIAGNOSIS OF CARDIOVASCULAR DISEASE

MUKKOTI MARUTHI VENKATA CHALAPATHI*, DUDEKULA KHASIM VALI†, YELLAPRAGADA VENKATA PAVAN KUMAR‡, CHALLA PRADEEP REDDY§, AND PURNA PRAKASH KASARANENI¶

Abstract. Cardiovascular diseases (CVD) are a prominent source of death across the globe, and these deaths are taking place in low-to middle-income nations. Due to this, CVD prevention is a pressing issue that has already been the subject of extensive research. Innovative methodologies in machine learning (ML) can have a greater impact on the diagnosis of CVD, yet the research on CVD is more challenging and attracting more research indeed. In this paper, we investigate the differences between four distinct machine learning models, support vector machine (SVM), logistic regression, decision trees (DT), and artificial neural networks (ANN) in their classification accuracy and possible practicality in CVD classification. techniques such as ensemble learning and other model-specific optimizations are not part of this study, but more basic implementations of the various models were used. To implement abovementioned ML models, a subset of 14 features from the original heart disease dataset is considered and deemed relevant for classification where no individual feature data are missing. From the results, it is observed that there is no clear winner in the comparison of models. There is no significant difference in the average accuracy of models. The highest average hit rate is observed in SVM and ANN, however it is slightly lower in ANN. Even though the DT had lower accuracy, the fully trained model can be easily visualized and interpreted by humans. Hence, the DT is possibly the most practical model to use as a complement to doctors in their current methods of diagnosis.

Key words: Average Classification Accuracy, Cardiovascular Disease, Computer Aided Diagnostics, Machine Learning Algorithms, Test Data Split

1. Introduction. After covid-19 pandemic, the number of deaths due to cardiovascular disease (CVD) is increased worldwide. Over 26 million people have died of CVD in 2021 which corresponds to 39% of total deaths across the globe. Every year WHO suggests different methods to be followed to overcome the situation of CVD [1]. As per the WHO statistics, 75% of premature CVD, is avertable, and reducing risk factors can assist patients and healthcare providers to cope up with the observed high of CVD. The Interheart trial explicated the effects of CVD risk factors such as abdominal obesity, smoking, hypertension, diabetes, and dyslipidemia while demonstrating the beneficial advantages of eating fruits and vegetables and engaging in regular physical exercise. All demographics and socioeconomic levels evaluated had the same risk factors, demonstrating the validity of standardized methods for CVD primary prevention globally. Among them is the taxation of tobacco, better diet, and more physical exercise to name a few. However, these prevention methods need to be complemented together with early diagnostics and identification of high-risk patients which would have a substantial impact on public health [2].

Heart disease categories and learning machines have both been used to improve the heart failure analytical procedure. This research intends to investigate various machine learning methods and improve the utilization of healthcare data. The effectiveness of the classifier should increase. The circumstances of each person have an impact on their health problems, particularly the risk of heart failure (HF) rate. Machine learning techniques

*School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, Andhra Pradesh, India (mmv.chalapthi@vitap.ac.in)
†School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, Andhra Pradesh, India (d.khasimvali@vitap.ac.in)
‡School of Electronics Engineering, VIT-AP University, Amaravati 522237, Andhra Pradesh, India (pavankumar.yv@vitap.ac.in)
§School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, Andhra Pradesh, India (pradeep.ch@vitap.ac.in)
¶Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522302, Andhra Pradesh, India (kpurnaprakash@kluniversity.in)

are more effective in predicting the risk factors and predicting the possibility of high blood pressure [3]. The majority of heart failure situations can be linked to problems with the heart's physiology or anatomy. As a result, HF has been associated with reduced life satisfaction and less effort put into engaging in mental and physical activities. According to estimates, HF affects 10% of the elderly and 1% to 2% of the general population in wealthy nations. As our population ages, heart failure is anticipated to become more common. After leaving the hospital, a readmission rate was present in patients with heart failure of 56.6%. Ignoring high frequency now will result in significant problems later on. Currently, reducing re-admissions is one of the most urgent needs.

After the initial incident, patients with heart failure are frequently kept in the hospital for a long period. Patients regularly get blood drawn to collect various health statistics [4]. It is possible to collect non-hematological data like gender, age, and smoking history. Machine Learning models intend to gain knowledge of the surroundings and forecast upcoming occurrences using user-provided data. People who possess this quality are flexible and may make decisions today based on their perceptions of the past. These models are utilized in the diagnostic procedure for several disorders, including multiple sclerosis. The main objective of this research is to create a learning model for prognosticating the possibility of heart failure. In this research on early diagnosis of heart failure, the data analysis employs standalone machine learning models decision trees, support vector machine (SVM), logistic regression, and neural networks [5, 6, 7]. Socio-demographic factors, like gender, high blood pressure, smoking status, and the existence of chronic medical disorders, have an impact on survival rates as well. It is challenging to establish reliable prognostic forecasts for persons with heart disease because survival rates for them are so unpredictable [8].

With the rise of machine learning and deep learning, new techniques of medical diagnostic methods have been opened up to complement the expertise of physicians. Many studies such as Parkinson's disease [9, 10], covid-19 [11, 12] have already been made on a range of medical diagnostic topics with different levels of accuracy [3]. Ihe internet of things also playing a key role in health care systems [13]. Deep learning concepts also extended their implementation in smart televison for program recommendation based on user's choice of priority [14].

The objective of this present study is to compare machine learning models' accuracy and their viability in cardiovascular disease detection. This paper compared the classification accuracy of the presence of cardiovascular disease on the implementations of four different broad machine learning techniques and their viability on this disease. This was tested on the Cleveland heart disease dataset from the UC Irvine Machine Learning Repository (UCIMLR) [2]. A comparison of classification accuracy and viability in the detection of CVD between the machine learning models. This problem has previously been investigated using different approaches. Models like SVM [5], artificial neural networks (ANN) [6], logistic regression and decision trees among others have been used on the same heart disease dataset from UCIMLR. These studies mainly focus on optimizing the selected model to achieve maximum accuracy. This study instead focused on comparing different machine learning models in their basic form and analyzing their weaknesses and strengths for this particular problem.

The use of machine learning algorithms has improved the prognosis for patients' diagnoses using their medical information. To detect cardiac sickness in its earliest stages, machine learning techniques based on linear, ensemble, and boosting are used. In this paper, we have implemented the early detection of cardiovascular disease on linear machine-learning models. The implemented models are trained and tested on the heart disease dataset. The test data split has been considered into 30% and 20% and the performances are ascertained with the impact of the test data split on the model accuracy.

The study used a subset of 14 features from the dataset which were used in all previous studies. The features include information such as age, sex, blood sugar, cholesterol, blood pressure, and more. The different models were then applied to the dataset by using already existing tools and frameworks such as TensorFlow [8] and sklearn [16]. The paper did not use any image analysis or image-based data which also could be used in the diagnosis since it was out of scope for this paper. The results were measured with a percentage of prediction accuracy for each respective model. The data were divided into training and test subsets where the models tried to predict the presence of CVDs in the previously unseen data of the test set after training on the training set.

Basic preliminaries and related work are described in the remainder of the paper, which is organized similarly to section 2. Section 3 presents a thorough implementation of machine learning methods for early

cardiovascular disease diagnosis. The empirical analysis is deconstructed in Section 4, which is followed by a discussion in Section 5. Section 6 contains the conclusion.

2. Basic Preliminaries and Related work. This section discusses the basic preliminaries and the literature works.

2.1. Computer-Aided Diagnostics. With the rise of computational power, computer-aided diagnostics (CAD) has become a part of medicine. One example is in the detection of breast cancer [15] on mammograms where CADs are routinely used as a second opinion [17]. One commonly used technique for CAD is machine learning (ML). As data-gathering advances, the effectiveness of ML as a CAD has progressed and is predicted to have a significant effect on medicine [3]. In broad terms, CADs using ML perform analysis on a great amount of patient data with and without a disease. The data consists of attributes that are considered to have a contribution to this disease and are often called training data. Using this training data the ML model is modified slightly by each data sample resulting in a model that can, to some degree, correctly classify the samples in the training data. This model is then tested on test data which is the same kind of data as training data but previously unseen to the model. The caliber of the model is protracted by the classification accuracy of the test data.

2.2. Earlier techniques for diagnosing cardiovascular disease. Most applications rely on common clinical risk factors like diabetes and hypertension. Research has also been conducted in Brazil on locations that are at high risk for cardiovascular disease-related mortality. In addition, the study's analysis revealed that 50% of the participants had several, difficult-to-treat illnesses that raised their mortality risk. This subject hasn't been the subject of enough study [19]. It can be more difficult to set a baseline for normalcy when lab results are assessed more realistically. For calculating risk, using trustworthy survival models is essential [20].

Traditional proportional risk models can uncover more relevant predictors by automatically creating connections between their component values and large data response values using computational methods [1]. The medical industry has used a variety of machine-learning methods, including decision trees [22]. Nonparametric survival is an auxiliary for parametric and semi-parametric models since they are independent of time in characterizing relationships [23]. The ensemble approach and survival trees combine to produce more precise projections. The common CVD diagnostic methods are:

- Physical Examination - Patient medical history and lifestyle Blood pressure, patient weight, listening to heart and lungs, etc.
- Blood tests - Level of different electrolytes, proteins, and other biomarkers.
- Chest X-rays - Image of the chest, revealing enlarged heart, lung congestion, and other abnormalities [21].
- Echocardiography - Ultrasound images of a heart showing heart structure.
- Cardiac catheterization - X-ray images reveal if there are any blockages in the heart.
- Radionuclide Ventriculography - images that demonstrate how well the heart's blood supply and chambers are functioning.

There are multiple indicators or methods that physicians can use and combine in the diagnosis of CVD. Some of them require analysis of images after the test has been made and some do not. The CVD diagnostic methods contain different methods of diagnostics together with what the method result type is and if this paper uses any of the data the respective method generates in the training of the models. In our proposed research work, we have used physical examination, and blood tests to examine the dataset. This indicates what types of tests medical personnel must do to be able to use the models described in this paper. Note that Used does not mean every data point from the test is used. The detailed view of CVD data the models are trained in is presented in Table 3.1.

2.3. Machine Learning Models. Machine Learning models can be used as CAD tools. However, there exist many different ML models that can be used to perform CAD. The models vary in complexity and how well they perform on different kinds of data. There are various applications of machine learning techniques such as electric vehicle [24], smart home [25], and microgrid [26]. This section will describe three frequently used models that were also used in this study.

2.3.1. Support Vector Machine. The technique of using so-called support vectors in the data to find the widest margin between two classes [20]. With linearly separable data consisting of data points from two distinct classes, to find a hyperplane separating the classes with a considerable margin to the data points of each class, Support Vector Machine could be utilized [19]. Thus creating an optimal linear decision boundary given the data. The problem was that most data is not perfectly linearly separable which made the technique hard to use. The solution to this came in the 1990s when the so-called kernel trick was discovered. This made it possible to use Support Vector Machine for non-linearly separable data by applying a non-linear transformation on the data and finding a linear decision boundary on the transformed data [19]. Another development was the concept of soft margins. This made it possible to find solutions in data that is not linearly separable even after applying a kernel transformation by allowing a few points to lie on the other side of the edge. Soft margins also had the added benefit of allowing a few points to lie within the margin even though the data is linearly separable to gain a larger margin.

2.3.2. Decision Tree. Decision trees (DT) are used for predicting what target class or value an observation belongs to. When building a decision tree given some data, all possible values for the features are divided into several regions. Each region has a specific target class assigned to it. An observation, with specific values for all features, is then given its target class or value depending on what region it falls in, so it can be used as Regression and Classification [19]. Depending on the number of features the Decision Tree will have a different amount of region layers. The first feature splits the Decision Tree into regions J1 to Jn, then depending on the values of another feature each region Jk is split into regions Rk to Rm and so on. Because of this attribute, Decision Trees can easily be represented as trees. Here the observations have two features X1 and X2 and are split into five final regions. This entails that Decision Trees often can easily be interpreted by a human.

2.3.3. Artificial Neural Networks. The thought behind ANN is to create a model that resembles a nervous system that, in humans and animals, can solve complex problems with the help of a network of many interconnected cells [22]. ANN consist of an input layer where the values of the observations are fed forward to one or several hidden layers of neurons. Depending on the values from the input layers, weights, and thresholds of each neuron, different neurons will fire and result in a specific output value which ends up in the output layer. The process of generating an output value given observation is called forward propagation [22]. One way of training ANN is to modify the weights of the neurons to minimize the error of ANN that have given training data. The minimizing is done numerically by using gradient descent which aids in finding a local minimum of the error function. This process is called backpropagation [22].

2.3.4. Logistic Regression. Logistic regression can be employed for binary classification by shaping the probability of data points' affinity to a certain class, e.g. providing an anticipation of a person being healthy. This modeling is broadly done by regressing a sigmoid function (or an alternative function with a bound of 0 to 1) to a dataset [32]. This regression can be performed both with single variables and collective variables and the prompted function will need an input vector x with dimension n the count of features in the dataset. The regression treasure trove optimal values for the parameter vector β in the sigmoid function. Logistic regression is an elementary approach to classification that is speedy. The classifier deteriorates however when regressing composite relationships [3].

2.3.5. Cross-validation. A cross-validation is an approach where a model is trained and validated on the same data, proving a validation dataset unnecessary. This is achieved by splitting the data toward K different folds, the model is formally trained on K-1 of the folds and computed on the remaining fold, this has recurred K times for each fold and the median of the results is reported [36]. For implementing the proposed work, ten folds are considered. Cross-validation can be calculational expensive but erases the necessity of a validation dataset and more data can then be employed for training and testing.

2.4. Related work. The area of machine learning in health care is commonly split into physical and virtual implementations, where virtual experimentation mainly concentrates on analyzing patient data. Contemporary articles that reveal virtual implementations for analyzing data to predict heart failure will be illustrated in this section.

Machine learning has from an early stage been used for computer-aided medical diagnostics [3]. In this

field, CVD classification has been a prominent problem that has been thoroughly researched with a considerable amount of time being spent on optimizing different models.

K-nearest neighbor (KNN) and SVM, supervised learning algorithms are pragmatic to predict heart disease [33, 35]. The researchers contend that machine learning models can be cast-off to forecast cardiovascular diseases [35]. Numerous other algorithms have been investigated, such as random forest and logistic model tree (J48). The Cleveland Heartland Registry is used by UCIMLR researchers to validate heart disease in patients and screen them for it. The following categories of information are included in this dataset. Next, a recommended large-scale classification algorithm will be made. To provide patients with more precise diagnoses, machine learning can be utilized to discover connections and predict risk factors for occurring heart disease.

For patients with heart failure [34], machine learning appears to be able to predict survival duration with accuracy. Ejection fraction and blood creatinine readings are both excellent predictors of a patient's life expectancy, according to patient charts. With the aid of a novel technique, survival rates in heart failure patients might be predicted [35].

2.4.1. Results with Optimization of Classifiers on Heart Disease Data. A new measure of classification reliability guaranteed 100% accuracy on all patient data points it can classify, with the drawback that possibly only a subset of all the data points could be classified [5]. To test this proposed measure of classification reliability they used a Support Vector Machine on the UCIMLR heart disease dataset [19]. This resulted in an average classification accuracy of 73% with radial basis function (RBF) kernel and the polynomial of degree is 73.7%.

Optimizing accuracy is often the utmost sought-after end goal but there has been researched done trying to generate a greater accuracy while still keeping the model simple enough to be able to interpret it. Freund and Mason [17] propose a solution to this that they call an alternating decision tree (ADT) which achieves similar or better accuracy than C5 and is easy to interpret. Whereas the C5 decision tree implementation produces good accuracy at the cost of interpretability. The ADT achieves a classification accuracy of 83.0% which is better than the C5 classification accuracy of 79.8% on the same dataset.

The usage of learning models for heart disease prediction is discussed in this article [31]. Data analytics were employed in this study to look at heart disease. The researchers carried out a study to assess the reliability and accuracy of three different data analysis methods such as KNN, ANN, and SVM. The neural networks produce better results and enable faster model learning with 93% accuracy. The researchers demonstrated the use of machine learning to solve the problem of heart disease prediction [29]. Researchers from Stanford University have created a revolutionary algorithm that predicts a patient's likelihood of acquiring heart disease based on their medical history. Heart disease patients' outcomes were categorized and predicted using machine learning techniques like KNN and logistic regression. These actions led to the creation of a model that can predict cardiac events in a large population more accurately. The main advantage of this approach is the time and effort savings in assessing whether a classifier can accurately diagnose heart illness. You can save time and money by using the provided method for heart disease prediction. It is feasible to conclude the long-term health of people with heart disease. In this study, machine learning is utilized to identify key components for the detection of cardiovascular disease. A prediction model is built using several features and classification techniques. The authors assert that by uniting a random forest and a linear model, a 92% accuracy rate can be reached.

One of the highest accuracies on the heart disease dataset was achieved by [6] using an ensemble of independent multi-layered feed-forward neural networks. The accuracy achieved was 89.01% using three neural networks. Using an ensemble of more than three neural networks did not show any improvements. This paper uses models based on the same principles as the models mentioned above, however, instead of attempting to optimize a single model it investigates how they compare to each other in terms of classification accuracy and viability for cardiovascular disease.

3. Early diagnosis of cardiovascular disease using Machine Learning Methods. This section explains cardiovascular disease prediction and the process of constructing the feature dataset for the numerical experiments.

3.1. Dataset. The scope of the research primarily relies on a single dataset and it has been used extensively in research of computer-aided diagnostics [6, 7, 30]. Containing only 297 patients with 14 features makes it hard to draw any wide-stretching conclusions on which of the compared techniques is best suited for CVD diagnostics. The original dataset obtained from UCIMLR contains 76 features where the occurrence of heart disease is an integer-valued feature from 0, no presence of disease, to 4. All values greater than 0 indicate the presence of heart disease. All previous ML research on this particular dataset has not been done on all 76 features, but a subset of 14 features deemed relevant for classification and where no individual feature data was missing. The prediction values have then also been binary instead of integers valued up to 4 since it is out of scope to diagnose different kinds of CVD. The binary value will indicate angiographic disease status where value 0 corresponds to <50% diameter narrowing and value 1 corresponds to >50% diameter narrowing. This is called the "processed dataset", henceforth in this paper, it will be referred to as "dataset" and is the dataset that was used for this research work.

The different sources have varying degrees of missing features from some data samples. For example, almost all data samples from the Hungarian source are missing data from three features, among them the number of major vessels colored by fluoroscopy. The data samples from the other data source are missing even more features like blood sugar. Because of this, in this research, we have only used the data donated from the Cleveland source since that data contains very few data points with missing features.

The Cleveland source contains 303 samples in total with all features listed in Table 3.1. From these 303 samples, 6 samples have been removed because of missing values. The choice of using the same dataset and not cherry-picking data points from other datasets was made so the comparison between previous research would be fairer. The dataset was not altered or preprocessed in any other way and all models used the same data. This was done so the comparison is as fair as possible even though some models might benefit from further preprocessing. For example, artificial neural networks might benefit from bundling up the ages into ranges of ages (35-45, 45-60).

3.2. Implementation. This section will present the implementation of the proposed early diagnosis of cardiovascular disease. The different machine learning models used the same method for splitting the dataset into training data and test data. The data were randomly shuffled and then split into two categories. The training data was used to train/fit the different models to the problem and the test data was then used to assess the performance of the model to see how well it performs on new unseen data. The test size is a parameter that represents the proportion of the data that is used as test data compared to training. For example, a test size of 0.3 means training data is 70% and testing data is 30%. Because of the limited size of the dataset, only 30% and 20% of test data splits were tested, which represents the training sets of 70% and 80% respectively. A larger split would lead to even fewer training points and a lower one would lead to testing with less than 60 patients which increases the risk that a model that performs well was only "lucky".

The Python library is employed to implement each specific model. Logistic regression was executed using the class LogisticRegression from the linear_model package, neural network utilizing the class MLPClassifier from neural_network, decision tree applying the class DecisionTreeClassifier from the tree, and support vector machine harnessing the class support vector classification (SVC) from the SVM package. The linear regression, decision tree, and neural network models were all trained on the entire training dataset. The SVM model however was trained on a chunk of the training data due to poor performance, more on this in Discussion.

3.2.1. Decision Tree. The DT implementation was made using sci-kit learn, a machine learning library that has pre-made ML models [16]. Specifically, the DT classifier [23] from sci-kit learn was used for generating DT. The DT classifier uses an optimized version of the CART algorithm which is a high-performing general-purpose algorithm for building decision trees [27]. When building a Decision Tree many different parameters can be set which decide how the resulting tree will look and perform. The maximum depth of the tree, and the number of samples to consider when splitting a node, are used when split among others. In this study, only the max depth of the tree and the criterion for splitting were taken into consideration. These parameters were chosen as they have to most impact on the complexity and structure of the generated decision tree.

3.2.2. Support Vector Machines. The SVM was implemented using sci-kit learn, the same library used for decision trees. As described, SVMs use different kernels to transform the data to find optimal support

Table 3.1: CVD Dataset Features

S. No.	Feature	Description
1	Age: age in years	Integer value
2	Sex	1 = male, 0 = female
3	CP: Chest pain type	0: typical angina 1: atypical angina 2: non-anginal pain 3: asymptomatic
4	trestbps: Resting blood pressure	mm Hg
5	chol: Serum cholesterol	mg/dl
6	fbs: Fasting blood sugar >120 mg/dl	1 = true, 0 = false
7	restecg: Resting electrocardiographic results	0: Normal 1: having ST-T wave abnormality 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	thalach: maximum heart rate achieved	Integer value
9	exang: Exercise induced angina	1 = yes, 0 = no
10	oldpeak = ST depression induced by exercise relative to rest	Floating value
11	slope: the slope of the peak exercise ST segment	0: upsloping 1: flat 2: downsloping
12	ca: number of major vessels (0-3) colored by flourosopy	Integer Value 0-3
13	thal: Thalassemia	0: Normal 1: Fixed defect 2: Reversible defect
14	Condition	0 = no disease 1 = disease

vectors. The implementation of SVM include the usage of three different kernels, a linear, a polynomial, and a RBF [28]. Both the polynomial and radial basis function kernel can be modified by input parameters. When using the polynomial kernel, the degree of the polynomial, which determines the complexity of the model, needs to be specified. This study used polynomials of degree two and degree three. The RBF kernel requires a parameter gamma which specifies how much a single training point influences the kernel. Here, the default value of Nl. Lastly, there is C, which regulates how wide the margin between the different classes is allowed to be. A high value of C results in a strict margin that does not allow data points of different classes to be on the wrong side of the margin. As data from the dataset used most likely is not linearly separable a lower value of C is preferred for the classifier to be able to create a margin. Therefore C=1 was used in this study.

3.2.3. Artificial Neural Networks. All experimentation with ANN was conducted using the TensorFlow library [8]. TensorFlow is one of the most widely used libraries for building machine learning models in production. The library makes it easy to build and test different types of networks and network structures with different hyperparameters. The optimizer used for network training in this work was the optimizer with a batch size of 32 and a cross-entropy loss function.

Many different types of hyperparameters can be hard to set for artificial neural networks even when the type of network is known. These hyperparameters include the number of layers, the number of nodes in the hidden layers, the type of activation functions the number of training epochs in the training of the model.

This paper used a sequentially built feed-forward network. The input layer has the same size as the 13 features in the dataset and the rectified linear (ReLU) activation function used for the input layer. The ReLU activation function which has been proven to give better results than the sigmoid activation function [29]. The

output layer is a single node that is activated with the sigmoid function since it is a single binary output the network produces. There is no exact formula for selecting the number of hidden layers. In general, if the data are less complex 1 or 2 hidden layers are used, if the data are complex 3 or 5 hidden layers are used. When the number of hidden layers are increased it increases the capacity of the model, simultaneously it requires more samples. Every hidden layer within the network also uses the ReLU activation function. Each network was trained for 350 epochs. The structure of the network varied with different tests to examine which structure fits the problem best. The tested structures were selected in an ad-hoc way where rough testing first was done before selecting some of the best-performing structures. Note that the network depth includes the input and output layers in the results window.

3.3. Evaluation. The machine learning models were assessed based on the mean and standard deviation of average classification accuracy they have on the test data. After the model has been trained it is presented with previously unseen data samples from the test dataset and it then tries to forecast the presence of CVDs and then checks if the prediction was correct. The accuracy then becomes the ratio of correct predictions with the test dataset. As mentioned in previous sections the different methods depend on varying hyperparameters that can be tweaked and tuned and yield different results. Each model was implemented with multiple different hyperparameters to get a better overview of the model potential.

There are also elements of randomness in which data points the models get to train on depending on the train/test data split. The SVM and ANNs also depend on randomly initialized weights for their models which might affect the outcome of the training. To remedy this the training and evaluation of each implementation were done in 100 independent runs and the classification accuracy is assumed to follow a Gaussian distribution. This makes it possible to compare not only the mean (μ) but also the standard deviation (σ) of each implementation to see which implementation had the best consistency. Metrics employed in this study to correlate machine learning models are illustrated in this section. Multiple metrics are accustomed to evaluating the accomplishment of a network [28]. Exploring different metrics for a variety of tasks is meant to represent the network's ability to clarify a given problem. To define all metrics the listed terms will be explored: The evaluation metrics utilize true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The accuracy [28] ratio of cases that were accurately identified as a percentage of all instances can be calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Precision [28] ascertains with what precision the model fits in the positive instance category. Precision is calculated as follows.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall [28] stipulates how many positive occurrences the model recorded. The recall is calculated as follows.

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{3.3}$$

Precision expresses the number of positive guesses that were true. Sensitivity expresses the proportion of all positive illustrations that were apparently labeled. Specificity expresses the proportion of all negative examples that were labeled.

$$Specificity = \frac{TN}{TN + FP} \tag{3.4}$$

F1 score [28] is a metric that combines precision and sensitivity to a single metric. It is defined as follows.

$$F1Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{3.5}$$

Table 4.1: Average Classification Accuracy of Different Classifiers

Classifier	**Test Data Split in %**	**μ on Classification Accuracy %**	**σ on Classification Accuracy**	**Classification Accuracy (Max) in %**	**Classification Accuracy (Min) in %**
Decision Tree	20	76.53	0.053	88.33	61.66
Decision Tree	30	76.18	0.040	85.45	65.55
Logistic Regression	20	92.85	0.050	92.44	71.11
Logistic Regression	30	87.86	0.051	89.44	68.11
Support Vector Machine	20	83.18	0.044	91.66	70.00
Support Vector Machine	30	83.01	0.035	92.22	71.11
Artificial Neural Networks	20	82.16	0.042	91.66	66.66
Artificial Neural Networks	30	80.64	0.040	92.22	67.7

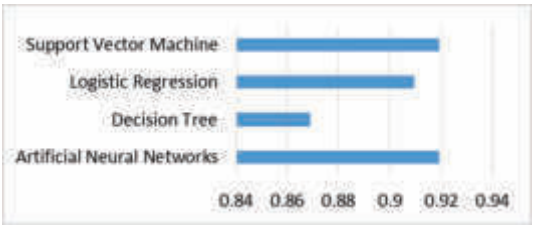


Fig. 4.1: Average of classification accuracy (maximum) by classifier

A receiver operating characteristic (ROC) curve is a graph that plots sensitivity and 1-specificity [18]. A good model has high sensitivity and specificity. A procedure to measure how well the model does this is to consider the area under the curve (AUC). A lofty AUC is better than a squat AUC. The top right of the graph used to gauge performance is where a model attempts to achieve high precision and high recall.

4. Experimental Results Analysis. The computer employed for the experimentation has an i5 processor at 1.8 GHz, with 8 GB of RAM, and a Windows 8.1 OS. The original data set [2] was exclusively used for classification experiments without any preceding feature selection technique. The investigation of these phases is described in full, along with a description of the results, in the immediate sections. In our research, two test data splits 30% and 20% are considered to test the accuracy of the machine learning models on test data.

Table 4.1 shows a general overview of the best average classification accuracy results of the different classifiers. The logistic regression has the highest mean (μ) classification accuracy 92.85 in both 30% and 20% of test data split together with the lowest $\sigma = 0.051$. The logistic regression and SVM also shared the highest accuracy 92.44% and 92.22% with ANN and the logistic regression and SVM had the highest lowest value 71.11%. The DT had the highest standard deviation of measured accuracy $\sigma = 0.053$. Figure 4.1 depicts the average of classification accuracy (Maximum) by the classifier.

4.1. Performance Evaluation on Decision Tree. Decision trees produce in all cases an average classification accuracy above 70% with the best mean result being an accuracy of 81.11%. As Table 4.2 indicates the depth of the tree has a greater impact on the classification accuracy than the splitting criterion. Setting a limit on the depth to 6 produces improvement in the classification accuracy compared to using the same test data split and criterion with no limit on the tree depth. Figure 4.2 is a graph that represents the average classification accuracy. Figure 4.3 shows what the decision tree looks like for the test with criterion entropy

Table 4.2: Average Classification Accuracy of Decision Tree Classifier

Max Depth	**Split Criterion**	**Test Data Split**	**μ on Classification Accuracy in %**	**σ standard Deviation on Classification Accuracy**	**Classification Accuracy (Max) in %**	**Classification Accuracy (Min) in %**
None	Gini index	30	74.56	0.043	83.33	61.11
None	Entropy	30	73.49	0.039	84.44	63.33
None	Gini index	20	72.23	0.049	86.66	60.00
None	Entropy	20	72.41	0.051	83.33	60.00
4	Gini index	30	80.73	0.045	85.55	64.44
4	Entropy	30	81.11	0.040	85.55	65.55
4	Gini index	20	74.62	0.054	86.66	56.55
4	Entropy	20	79.99	0.053	88.33	61.66

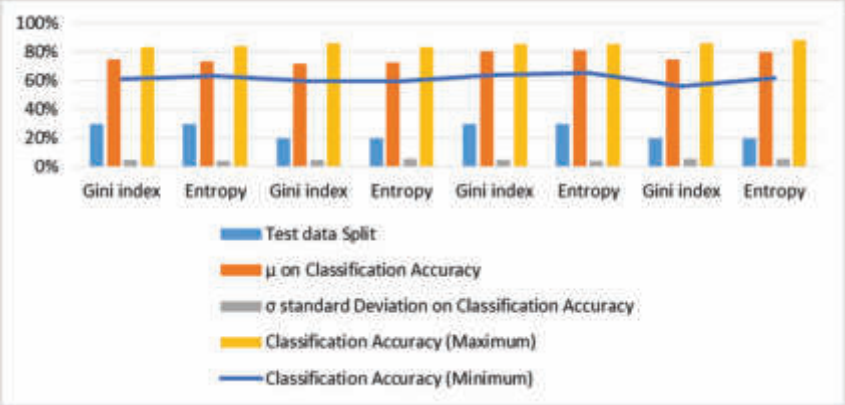


Fig. 4.2: Graph representation of average classification accuracy

and max depth 6 to classify a patient the Decision Tree first splits on the attribute thalassemia. This indicates that thalassemia might be one of the most influential factors of heart disease.

4.2. Performance Evaluation on Logistic Regression. Table 4.3 presents classification accuracy achieved with the different Regularization used for the Logistic Regression ranges from around 75% to around 91%. Table 4.3 shows that the regularization of Lasso performs well and has the highest μ accuracy of 93.18%, which is the highest overall mean classification accuracy achieved. At the same time the Ridge results in the lowest recorded accuracy, at 87.33%, across the different classifiers. Figure 4.4 shows the average classification accuracy of the logistic regression classifier.

4.3. Performance Evaluation on Artificial Neural Networks. Table 4.4, presents the shallower network had, in general, a higher μ classification accuracy than the deeper one for both 30% and 20% test data split and all hidden layer sizes. The networks had a higher μ together with a lower standard deviation σ for the test data split of 30% compared to the same type of network with a split of 20%. In general, the less complex networks with either smaller hidden layer sizes and/or shallower networks also had a higher μ and lower σ. The

Table 4.3: Average Classification Accuracy of the Logistic Regression Classifier

Regularization Type	**Test Data Split in %**	**μ on Classification Accuracy in %**	**σ standard Deviation on Classification Accuracy**	**Classification Accuracy (Max) in %**	**Classification Accuracy (Min) in %**
Lasso (L1)	20	93.18	0.054	91.66	70.00
Lasso (L1)	30	92.31	0.047	93.33	72.22
Ridge (L2)	20	88.96	0.054	87.77	67.77
Ridge(L2)	30	87.33	0.049	91.66	68.33

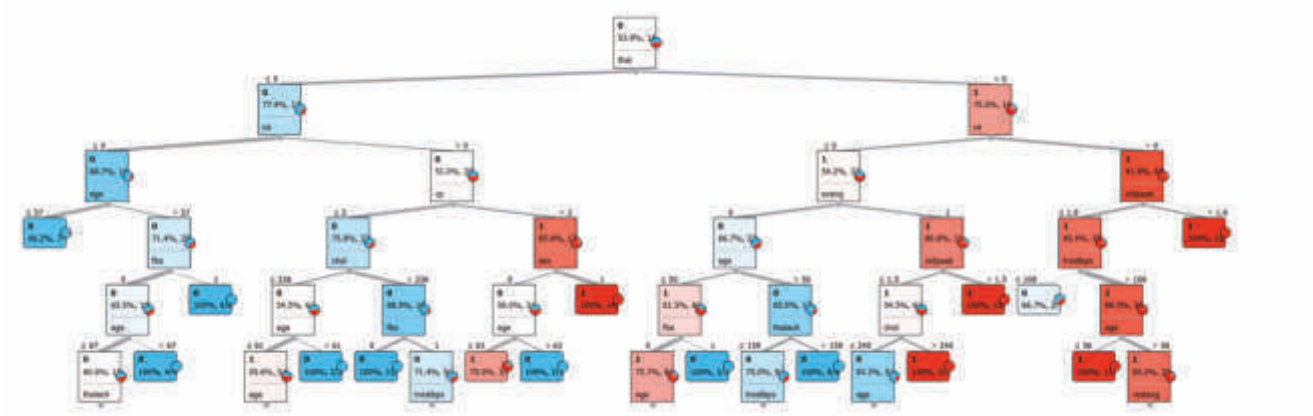


Fig. 4.3: The first six levels of the trained decision tree

network depth of 3 with 20% test data split and hidden layer size 14 had the highest μ. The corresponding network with a split of 30% had the most consistent results ($\sigma = 0.039$) and the highest accuracy (92.22%) came from the network with 4 layers with layer size 14 and 30% test data split.

4.4. Performance Evaluation on Support Vector Machine. The degree of classification accuracy attained using various kernels used for the SVM ranges from around 65% to around 83%. Table 4.5 shows that the polynomial degree of 2 kernels performs well and has the highest μ accuracy of 83.18%, which is the highest overall mean classification accuracy achieved. At the same time the radial basis function kernel results in the lowest recorded accuracy, at 65.61%, across the different classifiers. The kernels linear and RBF are having no degree information and is represented as not available (NA). This degree sometimes may not be available when there is no instance of the graph crossing the x-axis.

4.5. Cardiovascular Disease Detection using Machine Learning Models. The ML algorithms are trained and tested on two different test data split 30% and 20%. Cardiovascular disease detection on the four machine learning models is retrospect values measured by computing the sensitivity, precision, and the F1-score of each model. A ROC curve was depicted for all models. Figure 4.5 constitutes the Precision, specificity, sensitivity, and F1-score, achieved on the test data, for each model.

As shown in Table 4.6, the DT has achieved the highest precision and F1 score. Apparently, neither algorithm labeled a no disease instance incorrectly nor the logistic regression correctly predicted 85.8% of the heart disease cases while the artificial neural network truly predicted 84.5%.

Figure 4.6 represents the participation of each feature of the cardiovascular disease database on four different

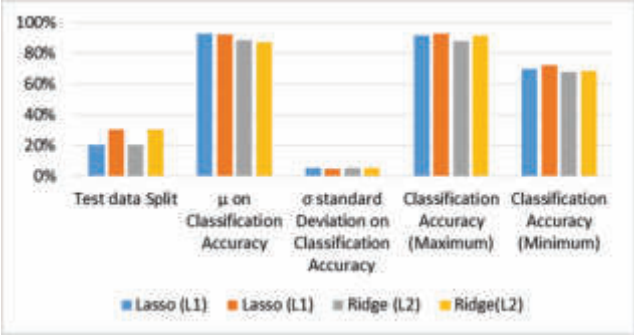


Fig. 4.4: Average Classification Accuracy of the Logistic Regression Classifier

Table 4.4: Average Classification Accuracy of the Artificial Neural Networks Classifier

Depth	**Hidden Layer Size**	**Test Data Split in %**	**μ on Classification Accuracy in %**	**σ standard Deviation on Classification Accuracy**	**Classification Accuracy (Max) in %**	**Classification Accuracy (Min) in %**
4	128	30	77.33	0.045	90.00	66.66
3	128	30	78.29	0.053	86.66	64.44
4	128	20	78.38	0.051	91.66	66.66
3	128	20	79.75	0.054	91.66	65.00
4	14	30	80.64	0.040	92.22	67.70
3	14	30	80.63	0.039	91.11	68.88
4	14	20	80.48	0.044	90.00	66.66
3	14	20	82.16	0.042	91.66	70.00

Table 4.5: Average Classification Accuracy of the Classifier Support Vector Machine

Kernel	**Degree**	**Test Data Split in %**	**μ on Classification Accuracy in %**	**σ standard Deviation on Classification Accuracy**	**Classification Accuracy (Max) in %**	**Classification Accuracy (Min) in %**
Linear	NA	20	82.56	0.041	93.33	75.00
Linear	NA	30	83.01	0.035	92.22	71.11
Polynomial	2	20	83.18	0.044	91.66	70.00
Polynomial	2	30	82.31	0.037	93.33	72.22
Polynomial	3	20	78.96	0.044	91.66	68.33
Polynomial	3	30	77.33	0.039	87.77	67.77
Radial Basis Function	NA	20	65.61	0.049	75.00	51.66
Radial Basis Function	NA	30	65.72	0.042	76.66	56.66

Table 4.6: Precision, Recall, and F1-score, achieved on the test data, for each model

Model	Precision	Recall	F1 Score
Logistic Regression	85.9	85.9	85.8
Artificial Neural Networks	84.5	84.5	84.5
Decision Tree	95.3	95.3	95.3
Support Vector Machine	69.2	67.0	66.7

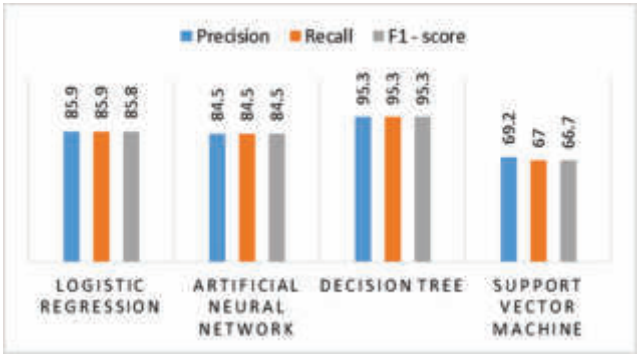


Fig. 4.5: Precision, Recall, and F1-score, achieved on the test data, for each model

ML algorithms. The values presented in the graph are mean, range, error rate, and hit rate on accuracy in the presence of disease across the attributes.

The DT earned a precision of 95.3% and a sensitivity of 95.3%. Truly predicting 95.3% of all cardiovascular occurrences but a false prediction of 5% of non-cardiovascular instances. The SVM realized a precision of 69.2% and 30% of predicted cardiovascular instances were not diagnosed.

The AUC is 0.923 for the logistic regression, 0.925 for the ANN, 0.989 for the DT, and 0.231 for the SVM. The erected value for AUC was achieved by the DT, followed by the ANN. Discover that the ROC curve for the DT classifier is a raw estimate due to the discrete essence of the classifier. The classifier only spawns a single false positive rate (FPR) and a single true positive rate (TPR) resulting in an atomic point on the graph, the ROC curve is estimated by combining this point with the points (0,0) and (1,1). The other algorithms rely on a continuous value and a threshold to classify data, this threshold can be altered to generate an endless relationship betwixt the FPR and TPR, which is depicted by the graph. The following PR curves were produced by the models on the test data.

4.6. Performance Differentiation of Accuracy on Machine Learning Algorithms. When applied to the whole dataset, DT can be observed to achieve 95.3% lofty accuracy than additional classification methods, while SVM achieves a modest accuracy of 67%. Figure 4.7 depicts the Accuracy of ML algorithms on cardiovascular disease data.

5. Discussion. The biggest benefit of the DT model is the interpretability of the resulting trained DT. The complete model can be visualized. This means that the results can be used by physicians as a complement to the current methods of diagnostics and patient examination. The tree in Figure 4.3 suggests that thalassemia has the biggest impact on the prediction of the presence of CVDs in the patient. This information can for example be used to make sure that all physical examinations of patients when checking for potential CVDs

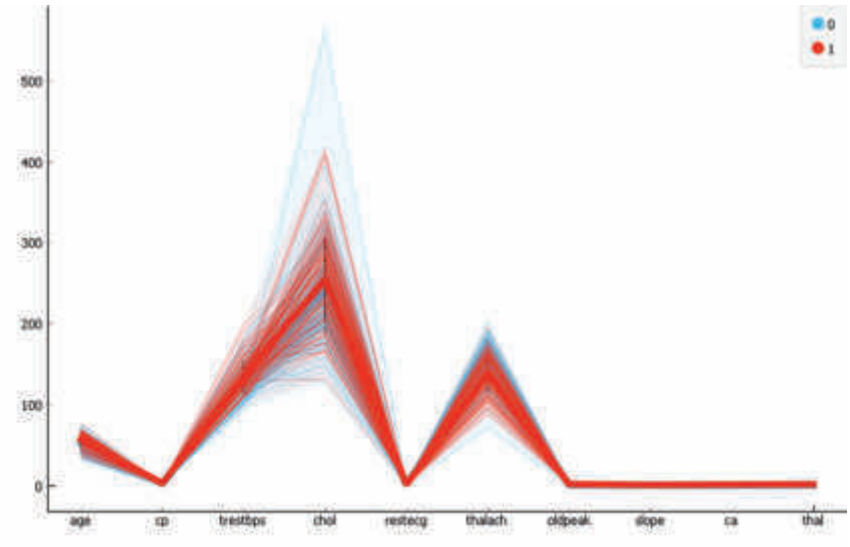


Fig. 4.6: Representing the mean, range, Error rate, and hit rate on features of the cardiovascular disease database

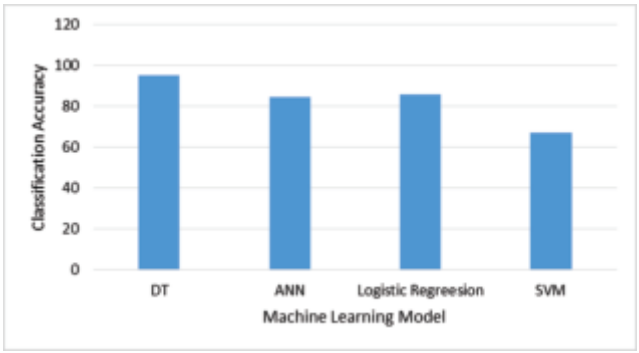


Fig. 4.7: Accuracy of machine learning algorithms on cardiovascular disease data

should include a test of patient thalassemia. This makes Decision Tree a very viable initial technique for CAD of CVDs since the risks are low and the final responsibility of diagnosis still firmly lies in the hands of the physicians and not in a yet unproven autonomous method of diagnosing CVDs. Not limiting the tree depth of a Decision Tree generates deep and complex trees that result in high variance and lower classification accuracy because of model overfitting. Allowing deeper or shallower trees than 4 levels did not improve the results as can be seen in table 3. As for the splitting criterion gini index performs slightly better than entropy however the difference is small and could depend on other factors.

ANN performed well where some individual runs gave a model with up to 92.22% accuracy. The best-averaged performance came from the 3-layered version with a hidden layer size of 14 and 20% test data split. This suggests that the higher complexity models might suffer from overfitting since the dataset only consists of 13 features with binary output. A single hidden layer that is not much larger than the input layer seems to suffice for this type of dataset. The higher test data split can also be an indicator of overfitting since a higher split gives the model fewer data points to train on which will keep the bias higher. Preliminary testing with the number of epochs also showed that more epochs than 350 almost always produced a worse result also indicating problems with overfitting. The standard deviation was generally in line with the SVM implementation indicating a more reliable result than the Decision Tree. The reliability is necessary for actual use on real patients but the dataset is too limited for it to translate to the general reliability of classification accuracy. More data points are needed with more features and with more diversity before the Artificial Neural Networks are truly ready as an autonomous diagnostics tool since the models themselves are not interpretable. How many data points and how diverse the data needs to be is hard to answer but probably in the size of millions of patients from around the globe and where medical experts can be a part of the feature selection.

Regarding SVMs, the two most complex kernels, with polynomial degree 3 and RBF, are the worst performing while the other two simpler kernels perform the best. This indicates that the worst-performing kernels are too complex for the data and that the kernels overfit the training data resulting in a lower classification accuracy just like the deeper ANN models. The highest mean accuracy by a small margin came from the 2nd degree polynomial kernel with a 20% test data split. However, it should be noted that the polynomial kernel had remarkably longer training time than the other kernels, especially compared to the linear kernels. This in combination that the 2nd degree polynomial kernel with 20% had a lower accuracy than the best linear kernel (with test data split at 30%) and a higher standard deviation suggesting that a linear kernel is the best choice for this particular dataset. Both in actual results previously mentioned and in practicality because of the much faster training time. The SVM suffers from the same problem of interpretability as ANN and a lack of data for autonomous use.

In general, the more complex models seem to have fared worse than the lower-complexity models. The deeper neural networks with larger hidden layers and higher complexity kernels in SVM all have lower mean accuracy than the shallower neural network and linear or low-degree polynomial kernels of the SVM. This does not necessarily mean that the problem of diagnosing CVDs can truly be described with a low-complexity model, a high-complexity model might very well outperform the lower-complexity ones with a large enough dataset. The SVM with a linear kernel and test data split of 30% is one of the best-performing models and fast to train however due to the limitations of the dataset it is very hard to know how well the model generalizes for other unseen real-world data. It is possible that the patients that have CVDs in the dataset used to share a hidden feature value, for example, smoking habits as previously mentioned, and that the SVM model instead finds the connecting relationship in another irrelevant feature. The range of classification accuracy for the various regularizations employed in the logistic regression is approximately 75% to approximately 91%. The regularization of Lasso works effectively and achieves the highest overall mean classification accuracy of 93.18%, as shown in Table 4.3. Ridge had the lowest documented accuracy of all the classifiers, at 87.33%, during the same period.

Because of the limitations in the dataset, the focus on mean accuracy lessens, especially since the different models also had very similar mean accuracy. The comparison then instead focuses on the interpretability of the model and viability as a tool for physicians. There are too many uncertainties for the resulting models in this paper to be used autonomously without a pathologist's supervision. A model that can be interpreted by a physician and used as a complement can be used even with its limitations previously described. This makes DT viable as a tool today. The SVM and ANN should be used with caution. They can of course also be used as complementary verification that the physicians can use but they should not be treated as equally accurate as the pathologists' judgment.

5.1. Comparison with earlier research. The previous studies mentioned in the literature have achieved a classification accuracy of around 70% - 90%. The classification accuracy attained in this paper is coherent with earlier reported algorithms as all the models are within this range. This may be a sign that consistently higher accuracy than this might be hard to get on a small dataset, even with specially adapted methods like

the ones used in the previous studies. One thing to note is that the SVM implemented in this paper has a 9.48 percentage point increase over the one implemented by [5]. This is most likely a consequence of [5] not focusing on optimizing the SVM but instead testing their new proposed measure of classification reliability.

The ANNs used in this paper reached a best mean accuracy of 82.16% which is worse than the previous study done by [6] which achieved an average accuracy of 89.01%. One of the main implementation differences is the use of ensemble methods which can potentially lower both bias and variance in the models and thus achieve higher overall accuracy. Their study used 14 nodes in the hidden layer which lowers the overall complexity of the model and is consistent with this paper's best ANN results. Overall slightly lower classification accuracy was achieved by the models in this paper which was expected as none of the models were optimized with a more advanced ensemble learning method which often results in higher accuracy and more reliable results.

5.2. Further Research. One of the biggest limiting factors in this paper, which has been mentioned before in previous sections, is the small size of the used dataset. With only around 300 samples it is hard to draw any general conclusions about the results as the models only get a small sample size to train on. Another factor, also related to the dataset, is the features the dataset contains. These 14 features are most likely not the only contributing factors to CVD and some of them may have very little or no contribution at all. With that said further research which uses a dataset containing a larger sample size and considers more features, such as whether the patient smokes or family history of CVD could achieve a more generalized result. All of the different image-based diagnostic methods could also benefit from machine learning which should be explored further.

Beyond a dataset with a larger sample size and more features, further research could also consider performing feature selection on the dataset that is used. With complex problems like diagnosing heart disease in patients taking into account many different features and then performing feature selection to only use, relevant features could yield better results.

6. Conclusion. The implementation of various machine learning models showed that there is no clear winner in the comparison of models. From the implementation, it is noticed that the support vector machines have achieved the highest average hit rate, while artificial neural networks achieved a similar highest hit rate. The support vector machine is best with regard to mean accuracy, highest accuracy, and lowest accuracy. Artificial neural networks along with the logistic regression gave the most reliable result with the lowest standard deviation. Although the decision tree achieved lower accuracy it can be visualized and interpreted easily by humans. These features led us to make the conclusion that the decision tree is the most practical model and also it is useful to doctors in their current methods of diagnosis. The scope of this work is primarily limited by the size of the dataset which contained a few patients, few features, and not enough diverse data. This must be taken into account while reviewing the relatively high mean accuracy of the models.

REFERENCES

[1] World Health Organization. Prevention of Cardiovascular Disease: Guidelines for Assessment and Management of Cardiovascular Risk. World Health Organization, 2022. https://www.cdc.gov/heartdisease/facts.htm
[2] Daniel Duprez. "Early detection of cardiovascular disease-the future of cardiology". In: EJ ESC Counc Cardiol Pract 4.19 (2006), p. 1.
[3] A. Ishaq, S. Sadiq, M. Umer et al., "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," IEEE Access, vol. 9, pp. 39707–39716, 2021.
[4] Bunyamin Ozaydin et al." Appropriate use of machine learning in healthcare Intelligence-Based Medicine", 2021.
[5] A. Jayakrishnan, R. Visakh, and K. T. Ratheesh, "Computational approach for heart disease prediction using machine learning," in 2021 International Conference on Communication, Control and Information Sciences (ICCISc), pp. 1–5, Idukki, India, 2021.
[6] S. Kathare and S. Gaikwad, "Practicability of heart attack prediction using machine learning," International Journal of Research Publication and Reviews, vol. 2, no. 7, pp. 1473– 1477, 2021.
[7] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," World Journal of Engineering and Technology, vol. 6, no. 4, pp. 854–873, 2018.
[8] TensorFlow. https://www.tensorflow.org/
[9] C. Lal Chowdhary, and R. Srivatsan, "Non-invasive detection of Parkinson's disease using deep learning," IJIGSP, vol. 14, no. 2, pp. 38–46, Apr. 2022.

[10] S. Padinjappurathu Gopalan, C. L. Chowdhary, C. Iwendi, M. A. Farid, and L. K. Ramasamy, "An efficient and privacy-preserving scheme for disease prediction in modern healthcare systems," Sensors, vol. 22, no. 15, pp. 5574, Jul. 2022.
[11] S. Bhattacharya et al., "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey," Sustainable Cities and Society, vol. 65, pp. 102589, Feb. 2021.
[12] B. Kumar Swain, M. Zubair Khan, C. Lal Chowdhary, and A. Alsaeedi, "SRC: Superior robustness of COVID-19 detection from noisy cough data using GFCC," Computer Systems Science and Engineering, vol. 46, no. 2, pp. 2337–2349, 2023.
[13] B. Singh, S. Bhattacharya, C. L. Chowdhary, and D. S. Jat, "A review on internet of things and its applications in healthcare," vol. 10, no. 1, 2017.
[14] K. V. Dudekula et al., "Convolutional neural network-based personalized program recommendation system for smart television users," Sustainability, vol. 15, no. 3, pp. 2206, Jan. 2023.
[15] C. L. Chowdhary, N. Khare, H. Patel, S. Koppu, R. Kaluri, and D. S. Rajput, "Past, present and future of gene feature selection for breast cancer classification – a survey," IJESMS, vol. 13, no. 2, p. 140, 2022.
[16] Scikit-learn. https://scikit-learn.org/stable/
[17] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conference Series: Materials Science and Engineering, Materials Science and Engineering, vol. 1022, no. 1, 2021.
[18] Junji Shiraishi et al. "Computer-aided diagnosis and artificial intelligence in clinical imaging". In: Seminars in nuclear medicine. Vol. 41, no. 6, pp. 449–462, 2011.
[19] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra, "A review on heart disease prediction using machine learning and data analytics approach," International Journal of Computers and Applications, vol. 181, no. 18, pp. 20–25, 2018.
[20] V. Kumar, G. S. Lalotra, P. Sasikala, et al., "Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques," Healthcare, vol. 10, no. 7, 2022.
[21] T. K. Das, C. L. Chowdhary, and X. Z. Gao, "Chest X-Ray investigation: A convolutional neural network approach," JBBBE, vol. 45, pp. 57–70, May 2020.
[22] A. Jamthikar et al, Cardiovascular/stroke risk predictive calculators: a comparison between statistical and machine learning models. Cardiovasc Diagn Therapy, vol. 10, no. 4, pp. 919–938, 2020.
[23] DecisionTreeClassifier. https://scikit- learn.org/stable/modules/generated /sklearn. tree.DecisionTreeClassifier.html
[24] B. Prasanth et al., "Maximizing regenerative braking energy harnessing in electric vehicles using machine learning techniques," Electronics, vol. 12, no. 5, pp. 1119, Feb. 2023.
[25] P. P. Kasaraneni, Y. Venkata Pavan Kumar, G. L. K. Moganti, and R. Kannan, "Machine learning-based ensemble classifiers for anomaly handling in smart home energy consumption data," Sensors, vol. 22, no. 23, pp. 9323, Nov. 2022.
[26] S. N. V. B. Rao et al., "Day-ahead load demand forecasting in urban community cluster microgrids using machine learning methods," Energies, vol. 15, no. 17, pp. 6124, Aug. 2022.
[27] Roger J Lewis. "An introduction to classification and regression tree (CART) analysis". In: Annual meeting of the society for academic emergency medicine in San Francisco, California. Vol. 14, 2000.
[28] Y. F. Khan, B. Kaushik, C. L. Chowdhary, and G. Srivastava, "Ensemble model for diagnostic classification of Alzheimer's disease based on brain anatomical magnetic resonance imaging," Diagnostics, vol. 12, no. 12, pp. 3193, Dec. 2022.
[29] Yoshua Bengio Xavier Glorot Antoine Bordes. "Rectifier and soft plus activation functions. The second one is a smooth version of the first." In: Deep Sparse Rectifier Neural Networks. 2011.
[30] Jesmin Nahar et al. "Association rule mining to detect factors which contribute to heart disease in males and females". In: Expert Systems with Applications, vol. 40, no. 4, pp. 1086–1093, 2013.
[31] M. W. Segar, M. Vaduganathan, K. V. Patel, et al., "Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score," Diabetes Care, vol. 42, no. 12, pp. 2298–2306, 2019.
[32] B. K. Turkmenoglu and O. Yildiz, "Predicting the survival of heart failure patients in unbalanced data sets," in 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1–4, Istanbul, Turkey, 2021.
[33] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, vol. 20, no. 1, 2020.
[34] P. A. Moreno-Sanchez, "Improvement of a prediction model for heart failure survival through explainable artificial intelligence," 2021.
[35] I. A. Marbaniang, N. A. Choudhury, and S. Moulik, "Cardiovascular disease (CVD) prediction using machine learning algorithms," in 2020 IEEE 17th India Council International Conference (INDICON), pp. 491–495, New Delhi, India, 2020.
[36] I. U. Haq, I. Haq, and B. Xu, "Artificial intelligence in personalized cardiovascular medicine and cardiovascular imaging," Cardiovascular Diagnosis and Therapy, vol. 11, no. 3, pp. 911–923, 2021.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Jun 2, 2023
Accepted: Sep 10, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 205–216, DOI 10.12694/scpe.v25i1.2236

RESEARCH ON MOOC CURRICULUM RECOMMENDATION MODEL OF HIGHER VOCATIONAL ENGLISH BASED ON IMPROVED INTENSIVE LEARNING NETWORK

YUXIA ZHENG* AND YANLEI MA†

Abstract. With the development and maturity of the Internet education industry, more and more vocational colleges have opened English teaching courses based on massive open online courses courses. However, there are many English-related courses on the massive open online courses course platform, and the use of scientific recommendation models can improve the teaching quality of such courses. Therefore, this research attempts to design two improved attention mechanisms and user-based embedded expression using meta-path technology. At the same time, these two are combined with reinforcement learning technology to design an improved massive open online courses English course recommendation model. The test results show that the hit rate of the model designed in this study is 89.84%, 74.28%, 70.81% and 71.35% respectively when the rank number is 20 and the parameter is 10. At this time, the cumulative income of normalized discount is 48.24%, 34.58%, 25.96% and 28.69% respectively. However, when the number of calculated samples reaches the maximum value of 1158609, the calculation time of the improved reinforcement learning recommendation model is 1867 seconds, which is also higher than the comparison model. The experimental results show that the curriculum recommendation accuracy of the massive open online courses recommendation model designed in this study is higher and the recommendation results are more reasonable. The results of this research have a certain application potential in the field of the construction of online education in colleges and universities.

Key words: Reinforcement learning; Metapath; Massive open online courses: Recommended model; Embedded expression; Attention mechanism

1. Introduction. After entering the 21st century, with the rapid development of computer technology, the online education industry has been recognized by more and more schools and teachers and students. In this context, a large number of vocational colleges have opened English teaching courses based on massive open online courses (MOOC). Moreover, MOOC, as one of the largest online education platforms in the world, has strong application value. It was also used by a large number of educational institutions at home and abroad during the COVID-19 epidemic [1, 2]. However, there are a variety of English courses on the MOOC platform. Even in some subdivisions of English education, such as oral English teaching, there are a large number of relevant courses to choose from. Too many choices have brought great difficulties to teachers who carry out teaching [3]. Teachers with less teaching experience cannot quickly select the appropriate curriculum design courseware and teaching process from a large number of courses [4]. Moreover, the self-study ability of students in higher vocational colleges is weaker than that of students in ordinary colleges [5]. Therefore, when they are faced with a large number of admirers, they may also have negative emotions such as confusion and helplessness. This may discourage students from learning. At the same time, in the process of self-study, if students choose inappropriate courses, it will greatly increase the learning difficulty and extend the learning hours, especially if they choose advanced courses with higher learning threshold [6]. It can be seen that the combination of higher vocational colleges and universities in English teaching has certain teaching value. However, it is necessary to use the MOOC recommendation model in the teaching system to recommend courses suitable for students' learning ability and learning stage for teachers and students. By this way, it can help teachers carry out teaching according to their aptitude and interest, and improve the learning effect of students. Under this background, this research attempts to combine attention mechanism, user embedded expression based on meta-path, and reinforcement learning (RL) technology. Based on this, a course intelligent recommendation model is designed. The model can more clearly extract key information such as learning interests and learning habits of users of

*Basic Teaching Department, Langfang Yanjing Vocational Technical College, Langfang, 065200, China
†Basic Teaching Department, Langfang Yanjing Vocational Technical College, Langfang, 065200, China (Yanlei_Ma2023@outlook.com)

English courses on the Moor platform.

2. Related Work. At present, online education is one of the key points of teaching reform in colleges and universities at home and abroad, and a large number of courses exist on the platform. This made experts realize the application value of the course recommendation model. Some experts have carried out relevant research on curriculum recommendation and general recommendation. Rabiu I and others believe that the recommendation system depends on the historical data purchased by users and their feedback to describe their preferences and make future recommendations [7]. Most of these systems usually use collaborative filtering models to analyze user ratings. At the same time, they infer the potential factors that show the characteristics of users and projects in the k-dimensional potential space. However, the historical rating data used for recommendation is usually sparse and unbalanced. Therefore, a new emotional scoring model based on long-term and short-term memory is proposed in the study. In order to alleviate the sparsity and imbalance of the data set, a combination function is designed in the experiment to capture the emotional bias between user ratings and comments. The test results of the design model using Amazon data show that the proposed model is superior to the existing static and dynamic models. Statistical tests show that all performance gains differ significantly. Roozbahani and others found that the current knowledge content on most mainstream knowledge sharing platforms is too complex. It is necessary to design a more intelligent recommendation model to help users quickly select knowledge content that is more in line with their needs. Therefore, the research team designed a recommendation model based on improved collaborative filtering algorithm. The test results show that the model can effectively improve the accuracy of content recommendation on the knowledge platform [8]. Zeng et al. found that the way of extracting information from the user's history is widely used to define the user's fine-grained preference to build an interpretable recommendation system. Because these aspects are extracted from the historical records, it is impossible to identify the aspects that represent the negative preferences of users. However, these potential aspects are also as important as the information representing the user's positive preference for building a recommendation system [9]. Choi et al. believed that the web-based courses used to teach the maintenance methods of mechanical components needed a more targeted recommendation system. The system can be recommended to engineers who need such services. For this reason, the research team has built a recommendation model for mechanical component maintenance course recommendation. The model combines K-means clustering algorithm and random forest classifier. The test results show that the recommendation results of this recommendation system are better than those of traditional methods [10]. Yang and others found that some commodity recommendation methods used by e-commerce platforms are designed based on traditional collaborative filtering algorithms. These recommendation methods have the shortcomings of data set reduction and coefficient matrix filling, and can't meet user needs well. Therefore, this study proposes an improved hybrid algorithm for online handicraft recommendation. The test results show that the model can effectively improve the effectiveness and exemption of online recommendation of handicrafts, and reduce the item score of candidates set users, which has certain application value [11]. Dat NV believes that the content-based recommendation algorithm has the problem of probability similarity calculation, so he proposes a recommendation algorithm based on Gaussian mixture model. The test results show that the recommended accuracy of the model is significantly higher than that of the comparison model. In practical application projects, the response time is shorter, the calculation speed is increased by 24.37% on average compared with the four comparison models, and the calculation results are the most stable and reliable [12].

To sum up, although many former scholars and scientists have designed a lot of improved recommendation systems to improve the efficiency and accuracy of online recommendation systems. However, at the same time, it is quite rare to consider more user interest characteristics and apply reinforcement learning to improve recommendation quality. Both of them have strong potential application value for more detailed mining of user information.

3. Design of MOOC Curriculum Recommendation Model Based on RL and Improved Attention Mechanism.

3.1. User Embedding Expression Mode and Node Layer Design Integrating Meta-path. This research is to design a recommendation model of MOOC that pays more attention to auxiliary information and user habits. On the one hand, it uses courses, knowledge points, and user data to build heterogeneous

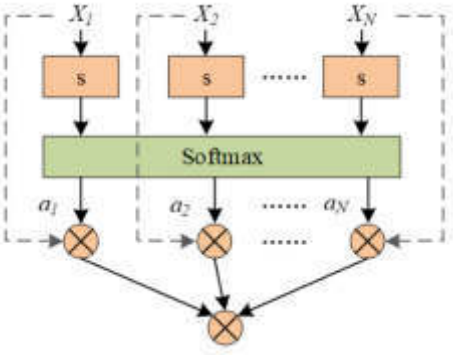


Fig. 3.1: General Calculation Mode of Attention Mechanism

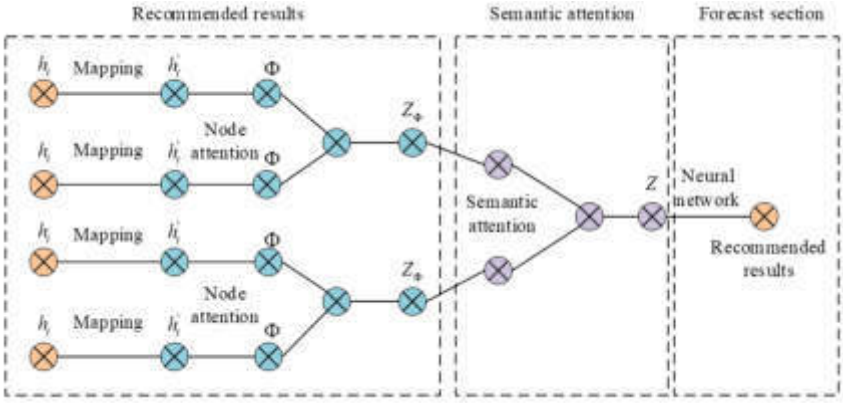


Fig. 3.2: Typical Structural Model of HAN

information networks, so as to use the two-level attention mechanism and meta-path sampling method to embed and express user data [9, 10]. On the other hand, RL is integrated into the recommendation model, which makes the model capture the user's interest features more accurately.

The improved MOOC recommendation model designed in this study is based on the attention mechanism and graph network. The general calculation model of the attention mechanism is shown in Figure 3.1. Figure 3.1, $\alpha_1, \alpha_2, \ldots, \alpha_N$ is the attention distribution, $X_1, X_2, \ldots, X_N$ is the input data, and is the data characteristics after query transformation [11]. As shown in Figure 3.1, the attention mechanism can make the neural network have the ability to focus computing resources on the specified feature subset, thus improving the feature extraction and expression ability of the neural network. The information with rich semantics and heterogeneity is the difficulty of graph representation in heterogeneous networks. The heterogeneous graph attention network (HAN) algorithm proposed by the predecessors has a higher precision in processing this data because it integrates semantic attention and node attention structure. The typical structure is shown in Figure 3.2 [12, 13]. As shown in Figure 3.2, the initial node feature h_i is first linearly mapped toh_i, and then fused through the node attention module composed of meta-path Φ , and then the node level attention embedding calculation is completed through the embedding expression Z_Φ of each meta-path [14]. The next step is to use the semantic layer attention to get the weight values under the embedded expression conditions of different meta paths, so as to calculate the final embedded expression for future calculation.

However, the spatial semantic information of the HAN model needs to be mined, and the user data in the MOOC recommendation model is complex and not applicable [15, 16, 17]. Moreover, HAN model can

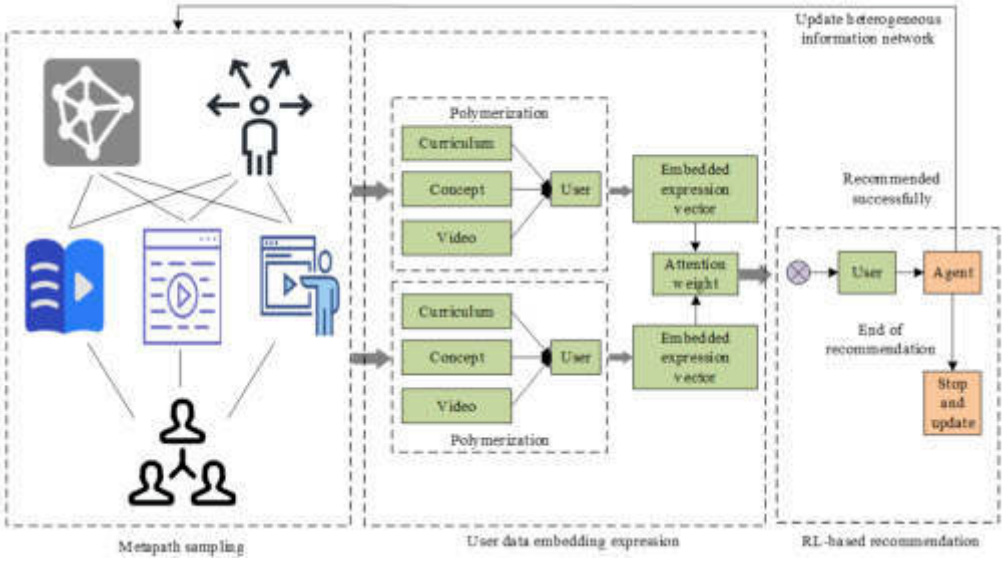


Fig. 3.3: RL-HAN Neural Network Model Calculation Framework

only learn data expressed by fixed heterogeneous information network at present. Its learning effect on non-fixed heterogeneous information network needs to be determined [18]. In view of the above shortcomings of HAN model, this research has designed a reinforcement learning combined with heterogeneous graph attention network (RL-HAN) model that is more suitable for the recommendation work of MOOC. The calculation framework of this model is shown in Figure 3.3. The RL-HAN model includes user embedded representation, meta-path sampling and enhanced knowledge point recommendation [19, 20]. In the meta-path sampling section, the algorithm will build a heterogeneous information network based on the concept of knowledge points, courses and user data. At the same time, the heterogeneous information network is sampled according to the random walk method, and the sampling is carried out according to the meta-path method [21, 22]. In the user embedded expression section, this research innovatively maps the obtained meta-path to the feature space through the hierarchical attention network. Subsequently, the self-attention mechanism is used to calculate the user's neighbor nodes and obtain the corresponding feature vector [23]. Then in the path layer, this research applies another attention level to fuse various semantic expressions and output the user's embedded expression features [24, 25]. In the strengthened knowledge recommendation module, this study referred to the enhanced learning technology to carry out user course recommendation. The following describes the meta-path sampling method first. After building a heterogeneous information network according to the MOOC data set, the same meta-path is sampled using the random walk method for the purpose of searching the network structure of the graph. Suppose that all users in user set U need to take N paths, so we can get $|U| \times N$ paths, put them into set M , and then complete the meta-path sampling.

As mentioned above, RL-HAN model calculates according to two attention mechanisms: node layer and path layer, and designs node layer attention mechanism. If a user has been assigned a meta-path type, because there are many corresponding nodes of user data in the heterogeneous information network. And the corresponding feature space of different types of nodes is also different. Therefore, a linear mapping is designed to map the corresponding nodes of user information to the same unified feature space. Suppose there is a type transfer matrix M_ϕ for each node type ϕ_i, and equation 3.1 shows the mapping method.

$$h_i' = M_{\phi_i} \cdot h_i \tag{3.1}$$

h_i' and h_i represent the characteristics of node i after and before mapping in equation 3.1. Since the user embedded expression of each node under the same meta path corresponds to different contribution weight

values, here we choose to use the self-attention mechanism to learn the weight values of different types of nodes, and calculate according to equation 3.2.

$$\alpha_{ij}^{\Phi} = \text{softmax}(e_{ij}^{\Phi}) = \frac{\exp(\sigma(a_{\Phi}^{T} \cdot [h_i' \& h_j']))}{\sum_{k \in N_i^{\Phi}} \exp(\sigma(a_{\Phi}^{T} \cdot [h_i' \& h_k']))} \tag{3.2}$$

σ represents the activation function, & is the calculation symbol defined in this study, which means the matrix splicing operation in equation 3.2. (a_{Φ}^{T}) , h_i' and h_j' represent the transposition of the attention vector of the node layer corresponding to the $\phi-$element path, the feature vector of the center node that completes the mapping, and the feature vector of the leader node after the mapping. In equation 3.3, the meta path embedding expression of node i can also be obtained through the fusion operation of adjacent nodes.

$$u_i^{\Phi} = \sigma\left(\sum_{j \in N_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot h_j'\right) \tag{3.3}$$

u_i^{Φ}represents the corresponding embedded expression of the node learned by the algorithm from the meta path in equation 3.3. And each embedded expression is related to the adjacent node representation. Considering the scale-free characteristics of heterogeneous graphs, the data variance may become large. Therefore, the attention mechanism of the node layer is adjusted to the multi-head attention mechanism to improve the robustness of the network training process. Therefore, it is necessary to copy the node layer attention K times to obtain different mapping features. And then splice these embedded expressions as the final user embedded expression vector, as shown in equation 3.4.

$$U_i^{\Phi} = \underset{k=1}{\overset{K}{\&}} \sigma\left(\sum_{j \in N_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot h_j'\right) \tag{3.4}$$

Assuming that the meta-path set is $\Phi_0, \Phi_1, \ldots, \Phi_P$, the embedded expression $U_{\Phi_0}, U_{\Phi_1}, \ldots, U_{\Phi_p}$ of node elements containing P specific meta-paths can be obtained by following the above operations.

3.2. Attention Mechanism of Fusion Path and Design of Recommendation Model Based on RL. After the input data set of the recommendation model is processed by the node-level attention mechanism, the embedded expression of the meta-path node is output. But the expression scale of this information is still limited, because the user's preferences on different meta-path are different. Therefore, the attention mechanism of the path layer is also designed here. Assuming that there is an embedded expression matrix of P specific meta-path nodes, the corresponding weight value $\beta_{\beta_0}, U_{\beta_1}, \ldots, U_{\beta_p}$ can be calculated according to equation 3.5.

$$(\beta_{\Phi_0}, \beta_{\Phi_1}, \ldots, \beta_{\Phi_P}) = \text{att}_{path}(U_{\Phi_0}, U_{\Phi_1}, \ldots, U_{\Phi_P}) \tag{3.5}$$

att_{path} is the path-level attention in the neural network structure in equation 3.5, which is used to learn different types of deep semantic information in heterogeneous information networks. Assuming that is the attention matrix of the path layer, the embedded expression vector needs to go through nonlinear mapping and internal product calculation with q , and equation 3.6 shows the calculation method of normalized output weight w_{Φ_i}.

$$w_{\Phi_i} = \frac{1}{|\nu|} \sum_{i \in \nu} q^{T} \cdot \tanh(W \cdot u_i^{\Phi} + b) \tag{3.6}$$

W and b respectively represent the parameters that need to be trained and optimized in equation 3.6. They are shared in all meta-paths and path-level attention. So far, the corresponding weight coefficients of each meta-path are obtained. The next step is to normalize all weights according to the softmax function. See equation 3.7 for the calculation method.

$$\beta_{\Phi_i} = \frac{\exp(w_{\Phi_i})}{\sum_{i=1}^{P} \exp(w_{\Phi_i})} \tag{3.7}$$

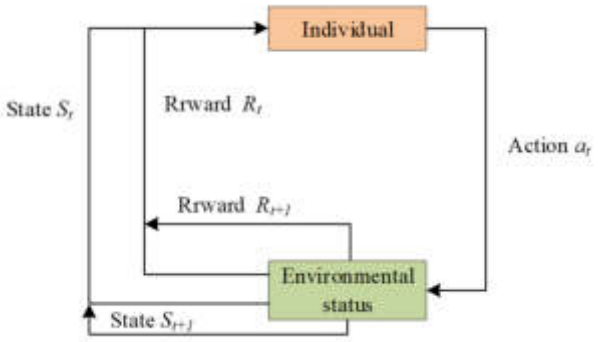


Fig. 3.4: Schematic Diagram of Traditional RL Calculation Process

β_{Φ_i} represents the weight vector obtained after normalization. This indicator can be used to show the importance of meta-path in formulation. Specifically, the larger the value, the more important it is to represent this meta path. The final user embedded expression can be obtained by support, see equation 3.8.

$$U = \sum_{i=1}^{P} \beta_{\Phi_i} \cdot U_{\Phi_i} \tag{3.8}$$

Traditional recommendation systems often build a model to minimize the distance between the quantitative data of users' real behavior and the prediction results [26]. The optimization process is realized by loss function. However, this recommendation model does not take into account the long-term interest of the recommended person. At the same time, the user's interest will also change with time and the observed behavior. Even in some cases, displaying or hiding specific items can be used to guide users' interests. However, in this case, the recommendation results are often unsatisfactory. Therefore, this study designed a recommendation model based on RL. Because the model better considers the long-term interests of users and the dynamic embedded expression characteristics of users, it has the potential to improve the recommendation performance. The typical overall task flow of RL is shown in Figure 3.4. Since this structure has been more commonly used, details will not be described here.

The optimization purpose of reinforcement learning is to find a strategy that can maximize the expectation of cumulative rewards, as shown in equation 3.9.

$$L_{\mathrm{RL}}(\theta) = \mathbb{E}_{\pi_\theta(C_t \mid u)} \sum_{t=1}^{T} r_t(c_t \mid u) \tag{3.9}$$

In equation 3.9, π_0 and r respectively represent optimization strategies and immediate rewards. Considering that the recommended task in the study has no RL environment, the self-designed method is selected to obtain the environment. Take 1 and -1 as the reward points, which respectively represent the output of the environment when the model prediction result is user behavior or prediction recognition. And when the recommended course is correct, the model will modify the heterogeneous information network to connect the recommended course c_t with user u . After the modification of the heterogeneous network, the new embedded information expression u_{t+1} can be obtained. If the recommended course is reasonable, the RL-HAN model can continue to recommend until the guidance recommendation fails. The model uses the embedded expression information of user as the input state of the reinforcement learning module. If the recommendation result is wrong, the predicted Q_{t+1} network will become consistent with Q_t , which is not suitable for the MOOC recommendation task. Therefore, the strategy gradient method is selected as the optimization strategy of reinforcement learning. According to

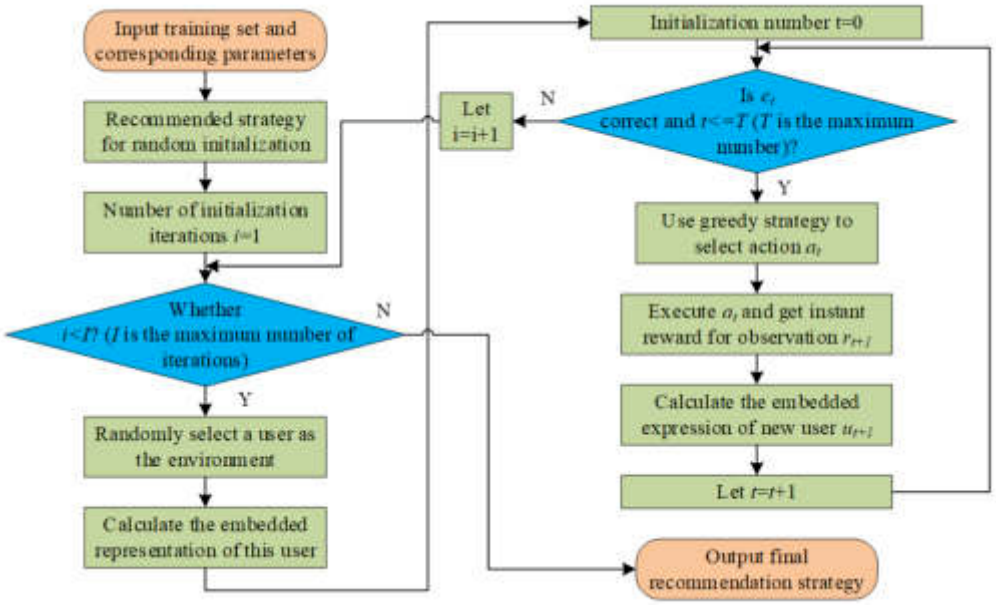


Fig. 3.5: Calculation Flow of English MOOC Recommendation Model Based on RL-HAN Algorithm

equation 3.10, its cumulative reward expectation gradient ($\nabla_\theta L_{\mathrm{RL}}(\theta)$) is calculated.

$$\nabla_\theta L_{\mathrm{RL}}(\theta) = -\sum_{t=1}^{T}[\nabla_\theta \log \pi_\theta(c_t \,|\, u_t)] \cdot r_t \tag{3.10}$$

To improve the learning speed and quality of recommendation model and improve its global search ability, entropy regularization is used as the regularization term ($H[\pi_\theta(c_t, |, u_t)]$ of the optimization index. See equation 3.11 for the calculation method.

$$H[\pi_\theta(c_t, |, u_t)] = -\sum_{t=1}^{T}\sum_{c_i \in C} \log(\pi_\theta(c_t \,|\, u_t))\pi_\theta(c_t \,|\, u_t) \tag{3.11}$$

Therefore, according to equation 3.12, the objective function of the recommended model can be calculated.

$$E_{c\sim\pi_\theta(c\,|\,u)} L_{\mathrm{RL}}(\theta) + \lambda H[\pi_\theta(c \,|\, u)] \tag{3.12}$$

λ represents the regularization coefficient in equation 3.12. The design of the English MOOC recommendation model based on RL-HAN algorithm is completed, and its overall calculation process is shown in Figure 3.5. The contents in Figure 3.5 have been shown completely and have been repeatedly mentioned, and it will not be repeated here.

Subsequent tests will be carried out to verify the performance of the design model. In the test, the hit rate $HR\#K$ with rank k , the normalized discount cumulative income $NDCG\#K$, and the mean reciprocal ranking MRR are used as evaluation indicators. Their calculation methods are shown in equation 3.13 to equation 3.15.

$$HR\#K = \frac{\mathrm{Num\_Hits}K}{|GT|} \tag{3.13}$$

Table 4.1: Detailed information of the test experiment dataset.

Node No	Entity node	Number of nodes	Node edge number	Node edge type	Number of node edges
#01	Video	98552	*01	Concept+Video	12846
			*02	User+video	54185142
#02	Curriculum	7424	*03	Video+course	852654
			*04	User+course	17564281
			*05	Concept+course	70154
#03	Concept	2588	*06	Course+concept	22512
			*07	Video+concept	12404
#04	User	3862031	*08	Course+user	16538410
			*09	Video+user	16538410

$$NDCG\#K = \frac{1}{Q}\sum_{q=1}^{|Q|} Z_{kq} \sum_{j=1}^{k} \frac{2^{r(j)}-1}{\log(1+j)} \tag{3.14}$$

$$MRR = \frac{1}{Q}\sum_{q=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{3.15}$$

Num_{HitsK} and $|GT|$ respectively represent the sum of elements belonging to the test set and the size of the test set in the TOP-K recommendation list of each user in equation 3.13. In equation 3.14, Z_{kq} is the regularization factor. In equation 3.15, $|Q|$ and $rank_{\ i}$ represent the size of the candidate set and the corresponding ranking.

4. Performance Test of Improved MOOC Course Recommendation Model.

4.1. Test Experiment Scheme Design and Model Parameter Setting. To verify the performance of the recommended model designed in this study, a test experiment is designed here. The data required for the experiment is from the MOOC platform of Tsinghua University in China, which includes 7424 courses, 98552 teaching videos, 2588 concepts, 3862031 users and 154266250 edges connecting information entities. For details, see Table 4.1. The data set is divided into training sets and test sets according to the 7:3 ratio.

The improved Neural Architecture Search with Reinforcement Learning (NASR) algorithm based on the Gated Recurrent Unit (GRU) neural network, the Multilayer Perceptron (MLP) algorithm based on the shallow neural network, and the Batch-material Requirement Planning (BRP) algorithm based on Bayesian estimation were selected as the comparative recommendation model. The hit rate $HR\#K$ of the evaluation index specifically selects *HR#3* , *HR#5* , *HR#10* , *HR#15* and *HR#20*. The evaluation index of normalized discount cumulative income *NDCG#K* is specifically selected as *NDCG#3* , *NDCG#5* , *NDCG#10* , *NDCG#15* and *NDCG#20* . In addition to the hit rate $HR\#K$ with rank k , the normalized discount cumulative income *NDCG#K* , and the mean reciprocal ranking MRR, the study also selected Area Under Curve (AUC), loss function, and calculation time as evaluation indicators.

5. Analysis of test results. First, the change rule of the loss function of each recommended model in the training process is compared. See Figure 6 for the statistical results. The horizontal axis represents the number of iterations, and the vertical axis represents the loss function value. In Figure 5.1, with the increase of the number of iterations, the loss function of each recommended model decreases rapidly and gradually converges. However, their convergence rate and the value after convergence are different. Specifically, the convergence speed of BRP recommendation model and MLP recommendation model is relatively fast, but the loss function value after convergence is large. RL-HAN and NASR recommended models have a slow convergence rate, but the loss function value after convergence is low. When the number of iterations exceeds 300, all model's complete convergence. The loss functions of RL-HAN, BRP, MLP and NASR models are 1.28, 5.74, 3.42 and 1.35 respectively.

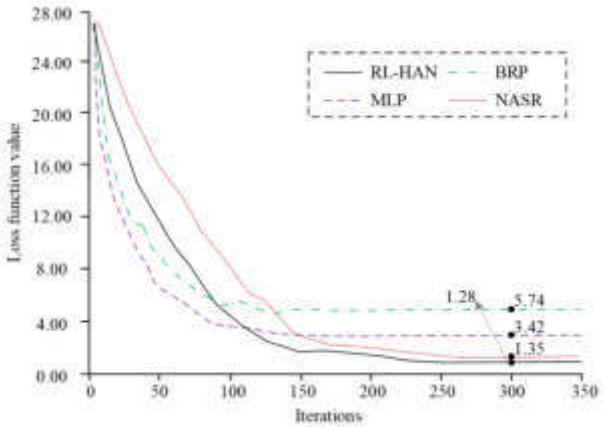


Fig. 5.1: Change law of loss function during training

Table 5.1: Calculation Results of Various Indicators for Each Recommended Model

Evaluating Indicator	RL-HAN	BRP	MLP	NASR
NDCG#3	47.08%	27.55%	15.77%	20.36%
NDCG#5	46.93%	29.51%	16.82%	22.40%
NDCG#10	48.24%	34.58%	25.96%	28.69%
NDCG#15	49.85%	36.72%	29.07%	31.75%
NDCG#20	52.91%	38.49%	31.48%	33.48%

In the following experiment, the quality of the recommended results of each model after the training is analyzed. The calculation results of various indicators are shown in Table 2. The larger the selected rank, the higher the hit rate, which is also consistent with the calculation logic of hit rate. The horizontal comparison shows that when the rank number is the same, the RL-HAN recommendation model has the highest hit rate, followed by the BRP model. For example, when the rank number is determined as 20, the hit rates of RL-HAN, BRP, MLP and NASR recommended models are 89.84%, 74.28%, 70.81% and 71.35% respectively.

Then in the experiment, the normalized discount cumulative income *NDCG#K* index is used to evaluate each model, and the statistical results are shown in Table 5.1. The larger the parameter k, the higher the corresponding normalized discount cumulative income. Through horizontal comparison, with the same parameter , the of RL-HAN recommended model is still the highest, and the corresponding value of MLP model is the lowest. For example, when parameter is 10, the normalized discount cumulative income of RL-HAN, BRP, MLP and NASR recommendation models is 48.24%, 34.58%, 25.96% and 28.69 respectively.

In the following experiment, the index *MRR* of mean reciprocal ranking is used to evaluate each model, and the statistical results are shown in Figure 5.2. The horizontal axis in Figure 5.2 is used to show different recommended models, the vertical axis represents the *MRR* value, and the different icons represent different scale test data sets. On the whole, the stability of the model recommended by NASR is the best. Because this model has the smallest numerical difference in different scale test data sets, and the stability of MLP and BRP models is poor. From the perspective of MRR value, the RL-HAN recommended model has the largest MRR value as a whole. For example, when using 100% test data set, the values of RL-HAN, BRP, MLP and NASR recommended models are 0.3756, 0.3086, 0.2069 and 0.2024 respectively.

The statistical results of AUC values of each recommended model are shown in Figure 5.3. The horizontal axis represents the false positive rate and the vertical axis represents the true positive rate in Fig. 8. The curves of different styles in the figure represent different recommended models. All curves in the figure are receiver

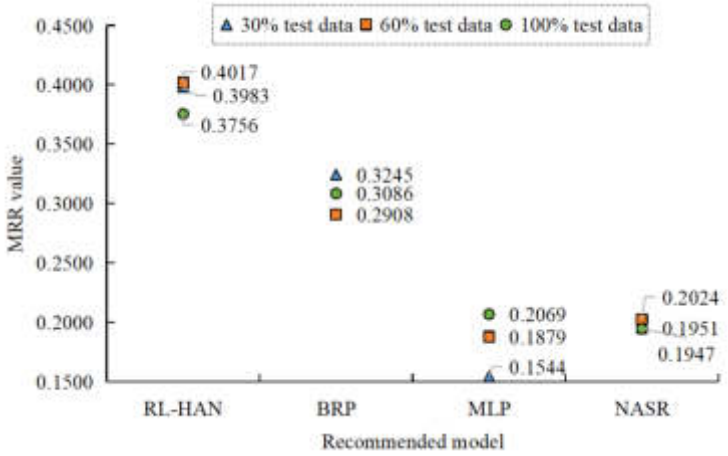


Fig. 5.2: Statistical Results of Index of Each Model

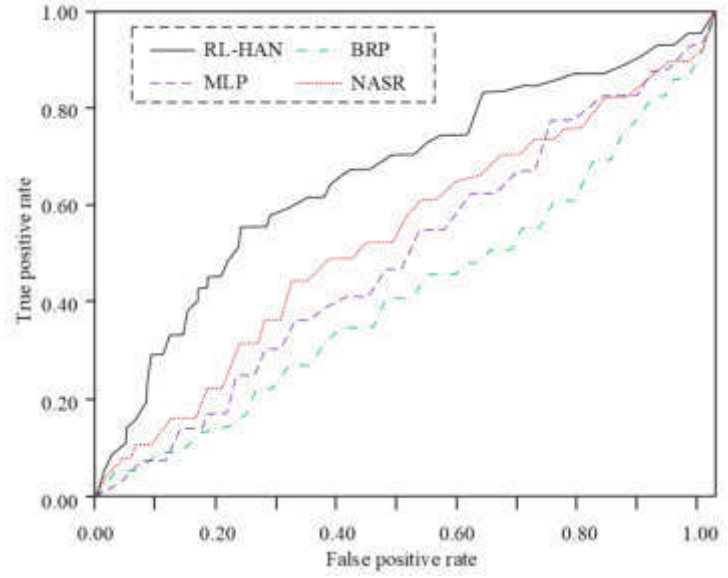


Fig. 5.3: Statistical Results of AUC Indicators of Each Model

operating characteristic curve (ROC) curves. The area under the ROC curve and enclosed by the coordinate axis is AUC. According to Figure 5.3, the ROC curve of RL-HAN recommended model is always above the ROC curve of other models, and this model's AUC area is the largest. The AUC values of RL-HAN, BRP, MLP and NASR recommended models are 0.681, 0.463, 0.517 and 0.586 respectively.

Finally, the recommendation efficiency of each recommendation model is analyzed in the experiment, and the calculation time is used to evaluate the performance in Figure 5.4. The horizontal axis represents the number of samples participating in the test, and the vertical axis represents the calculation time, in seconds. Different icons represent different recommended algorithms, and different linetypes represent different fitting curves. With the increase of calculation samples, the calculation time of each model increases. However, the calculation time of NASR model shows an exponential growth trend, and the calculation time of other models

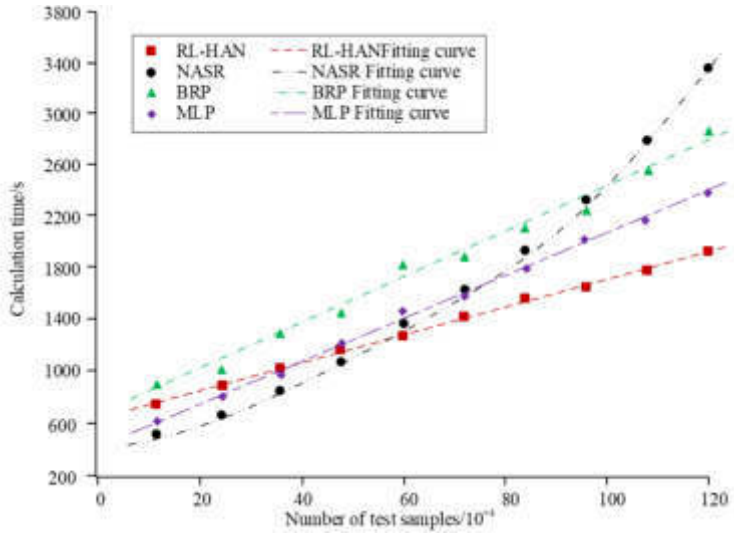


Fig. 5.4: Comparison of Calculation Time of Each Model

increases according to a linear rule. When the calculation sample is small, the efficiency of the model designed in this study is low. When the calculation sample is large, the calculation efficiency of the design model in this study is the highest. When the number of calculated samples reaches the maximum of 1158609, the calculation time of RL-HAN, BRP, MLP and NASR recommended models is 1867 s, 2866 s, 2349 s and 3980 s respectively.

6. Conclusion. Aiming at the inaccurate problem of curriculum recommendation in higher vocational education, this study designed an improved recommendation model integrating meta-path and reinforcement learning technology. When the number of iterations exceeds 300 in the training process, all model's complete convergence. The loss functions of RL-HAN, BRP, MLP and NASR models are 1.28, 5.74, 3.42 and 1.35 respectively. When the rank number is the same, the RL-HAN recommendation model has the highest hit rate. When the rank number is determined as 20, the hit rates of RL-HAN, BRP, MLP and NASR recommended models are 89.84%, 74.28%, 70.81% and 71.35% respectively. With the same parameter , the of RL-HAN recommended model is still the highest, and the corresponding value of MLP model is the lowest. When parameter is 10, the normalized discount cumulative income of RL-HAN, BRP, MLP and NASR recommendation models is 48.24%, 34.58%, 25.96% and 28.69% respectively. From the perspective of MRR value, the RL-HAN recommended model has the largest MRR value as a whole. When using 100% test data set, the AUC values of RL-HAN, BRP, MLP, and NASR recommended models are 0.3756, 0.3086, 0.2069, 0.2024. The AUC values of RL-HAN, BRP, MLP, and NASR recommended models are 0.681, 0.463, 0.517, and 0.586, respectively. Moreover, when the number of calculated samples reaches the maximum of 1158609, the calculation time of RL-HAN, BRP, MLP and NASR recommended models is 1867 s, 2866 s, 2349 s and 3980 s respectively. This shows that the recommendation accuracy of the improved MOOC recommendation model designed in this study is higher than that of the common models. However, its disadvantage is that its computational efficiency is not excellent, which is also an aspect that can be further improved in subsequent research.

REFERENCES

[1] Jiang, N., Gao, L., Duan, F., Jie, W., Tao, W. & Honglong, C. SAN: Attention-based social aggregation neural networks for recommendation system. *International Journal Of Intelligent Systems*. **37**, 3373-3393 (2021)

[2] Chiu, M., Huang, J., Gupta, S. & Akman, G. Developing a personalized recommendation system in a smart product service system based on unsupervised learning model. *Computers In Industry*. **128**, 1-10342 (2021)

[3] Muuzuri, A., Otero-Cacho, A. & Mira, J. Ventilation time recommendation system incorporating local meteorological data. *Indoor And Built Environment*. **31**, 1418-1437 (2022)

[4] Sharma, B., Hashmi, A., Gupta, C., Osamah, I., Abdulsahib, G. & MM., I. Hybrid Sparrow Clustered (HSC) Algorithm for Top-N Recommendation System. *Symmetry*. **14**, 793-798 (2022)
[5] Daneshvar, H. & Ravanmehr, R. social hybrid recommendation system using LSTM and CNN. *Concurrency And Computation: Practice And Experience*. **34** pp. 18 (2022)
[6] Gwadabe, T. & Liu, Y. Improving graph neural network for session-based recommendation system via non-sequential interactions. *Neurocomputing*. **468**, 111-122 (2022)
[7] Rabiu, I., Salim, N., Da'U, A. & Nasser, M. Modeling sentimental bias and temporal dynamics for adaptive deep recommendation system. *Expert Systems With Applications*. **191**, 1-11626 (2022)
[8] Roozbahani, Z., Rezaeenour, J., Katanforoush, A. & Bidgoly, A. Personalization of the collaborator recommendation system in multi-layer scientific social networks: A case study of ResearchGate. *Expert Systems*. **39**, 1-12932 (2021)
[9] Zeng, Z., Shi, Y., Pieptea, L. & Ding, J. Using latent features for building an interpretable recommendation system. *The Electronic Library*. **39**, 281-295 (2021)
[10] Choi, Y., Lee, J. & Yang, J. Development of a service parts recommendation system using clustering and classification of machine learning. *Expert Systems With Applications*. **188**, 1-11608 (2022)
[11] Yang, Y. & And, K. and Application of Handicraft Recommendation System Based on Improved Hybrid Algorithm. *International Journal Of Pattern Recognition And Artificial Intelligence*. **36**, 1-22500 (2022)
[12] Dat, N. & Toan P. V. Thanh, T. Solving distribution problems in content-based recommendation system with gaussian mixture model. *Applied Intelligence: The International Journal Of Artificial Intelligence, Neural Networks, And Complex Problem-Solving Technologies*. **52**, 1602-1614 (2022)
[13] Ephzibah, E., Sujatha, R., Chatterjee, J. & An, M. adaptive neuro-fuzzy inference for blockchain-based smart job recommendation system. *International Journal Of Information And Decision Sciences*. **14**, 5-14 (2022)
[14] Madhavi, A., Nagesh, A. & Govardhan, A. Study on E-Learning and Recommendation System. *Recent Advances In Computer Science And Communications*. **15**, 748-764 (2022)
[15] Ziarani, R. & Ravanmehr, R. Deep neural network approach for a serendipity-oriented recommendation system. *Expert Systems With Application*. **185**, 1-11566 (2021)
[16] Zeeshan, Z., Ain, Q., Bhatti, U., Memon, W., Ali, S., SA., N., Nizamani, M., Mehmood, A., Bhatti, M. & Shoukat, M. Feature-based multi-criteria recommendation system using a weighted approach with ranking correlation. *Intelligent Data Analysis*. **25**, 1013-1029 (2021)
[17] Salina, A., Ilavarasan, E. & Rao, K. Enabled Machine Learning Framework for Social Media Content Based Recommendation System. *International Journal Of Vehicle Information And Communication Systems*. **7**, 161-175 (2021)
[18] Bhuvaneshwari, P. & Rao, A. Product recommendation system using optimal switching hybrid algorithm. *International Journal Of Intelligent Enterprise*. **8**, 185-204 (2021)
[19] Sundari, P. & Subaji, M. comparative study to recognize fake ratings in recommendation system using classification techniques. *Intelligent Decision Technologies: An International Journal*. **15**, 443-450 (2021)
[20] Cui, Y. Intelligent Recommendation System Based on Mathematical Modeling in Personalized Data Mining. *Mathematical Problems In Engineering*. **2021**, 1-66720 (2021)
[21] Gupta, S. & Dave, M. Product Recommendation System Using Tunicate Swarm Magnetic Optimization Algorithm-Based Black Hole Renyi Entropy Fuzzy Clustering and K-Nearest Neighbour. (Journal of Information & Knowledge Management,2021)
[22] Saraswathi, K., Mohanraj, V., Suresh, Y. & Senthilkumar, J. Hybrid Multi-Feature Semantic Similarity Based Online Social Recommendation System Using CNN. *International Journal Of Uncertainty, Fuzziness And Knowledge-based Systems: IJUFKS*. **29** pp. 333-352 (2021)
[23] Fu, Y., Yang, M. & Han, D. Interactive Marketing E-Commerce Recommendation System Driven by Big Data Technology. *Hindawi Limited*. **3059**, 1-38730 (2021)
[24] Alagarsamy, R., Arunpraksh, R., Ganapathy, S., Rajagopal, A. & Kavitha, R. fuzzy content recommendation system using similarity analysis, content ranking and clustering. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **41**, 6429-6441 (2021)
[25] Zhu, W. Topic recommendation system using personalized fuzzy logic interest set. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **40**, 2891-2901 (2021)
[26] Kotciuba, I., Shikov, A. & Voitekhovsky, Y. Recommendation system for finding improved coworking area based on intelligent information technologies. *World Journal Of Engineering*. **18**, 621-629 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Apr 12, 2023
Accepted: Nov 17, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 217–228, DOI 10.12694/scpe.v25i1.2286

RESEARCH ON STUDENT BEHAVIOR ANALYSIS AND GRADE PREDICTION SYSTEM BASED ON STUDENT BEHAVIOR CHARACTERISTICS

QIANG FU*

Abstract. In the era of big data, the traditional governance model of student behavior management gradually shows the disadvantage of "post positioning". Therefore, the research uses relevant data mining algorithms to extract and analyze students' behavior characteristics, and constructs a SPC system model for students' behavior analysis and performance prediction. At the same time, the experiment verifies its effectiveness. The experimental results show that in the factor analysis, the overall variance of the first seven indicators is 69.942%, which shows that there is a significant correlation between students' learning behavior and academic performance. In the model performance analysis of SPC system, the accuracy of the five algorithms is kept at 40% - 60%, while SVM has higher stability than other methods. In addition, the prediction accuracy and response rate of SPC model reached 86.90% and 81.57% respectively. When the sequence length was increased from 1 to 20, the accuracy of SPC model exceeded 70%; However, when the feature dimension exceeds 50, the model representation ability will decline. Therefore, appropriate feature dimensions are needed to predict student performance. To sum up, the SPC model built by the research is effective in analyzing student behavior and predicting students' performance, and practical in actual student management.

Key words: Internet of Things; path Behavior characteristics; Performance prediction; Data mining; SPC model

1. Introduction. The progress of science and technology has made people's lives increasingly networked and information-based, and the information management system has also replaced the traditional text recording method with the development of technology, so as to promote the exponential growth of the recorded data on the trajectory of human daily behavior [1]. In education management, the characteristics of students' behavior data are effective indicators to promote the improvement of school teaching. As an effective tool for extracting data features, data mining technology has been widely used in education management [2]. Trakunphutthirak et al. used a progressive temporal data mining method of educational information on the basis of data mining technology to effectively improve the prediction accuracy of students' academic performance [3]. To effectively evaluate the learning achievements of students in online distance education, Bütüner et al. conducted in-depth research on artificial neural networks and deep learning algorithms in data mining methods, so as to effectively improve the prediction accuracy of students' learning achievements [4]. Shreem et al. have improved on the basis of genetic algorithm to achieve effective mining of educational data and effective prediction of student performance, thus effectively realizing the prediction classification of student learning performance [5]. Under this background, based on the recurrent neural network (RNN), a sequence-based performance classifier (SPC) was built by using support vector machines (SVM) and a hybrid encoder decoder network (HRNN) built on the basis of attention. The purpose is to effectively analyze students' behavior, so as to effectively predict their achievements, so as to provide help for teachers to improve teaching and help students to improve their achievements.

2. Related Work. The promotion of digital application system enables students' learning dynamics and life trajectory to be recorded in an all-round way in the form of digital information. These data are of great help to education managers in analyzing students' activity trajectory [6]. In the field of education management, the ultimate purpose of analyzing students' behavior characteristics is to improve students' academic performance and promote their all-round development, and to teach based on the characteristics of each student [7]. Therefore, the majority of scholars at home and abroad have conducted in-depth research on students' performance prediction. To realize the early prediction of students' performance, Xiong et al. organically combined convolutional neural network and recurrent neural network to propose a mixed depth learning model,

*Principal's Office, Zhengzhou Finance School, Zhengzhou, 450007, Henan China (hq181151397130126.com)

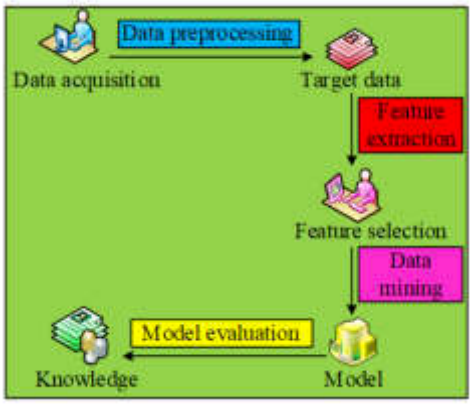


Fig. 3.1: The process of data mining

which accurately predicted students' behavior and improved their learning performance [8]. To promote the development of education, Keser et al. proposed a new hybrid integrated learning algorithm using data mining technology, so as to achieve effective prediction of student performance and help the development of education [9]. To explore the impact of social media on students' performance, Nti et al. built a prediction framework for learning students' performance by using the learning algorithms of decision tree (DT) and random forest (RF), thus realizing effective prediction of students' performance [10]. Teng M F et al. conducted a confirmatory analysis of students' English writing scores by using the data collection method based on the metacognitive academic writing strategy questionnaire, thus achieving an effective prediction of students' English writing performance [11]. In addition, to improve the quality of students' academic achievements in the School of Education, Britum and others conducted statistical tests to solve the problem of students' self-esteem, so as to effectively improve students' achievements on the basis of improving their sense of self-efficacy [12]. Serrano et al. analyzed in detail the role of the five-factor model on the personality level to improve the accuracy of the prediction of students' academic performance [13]. Su has effectively improved the data processing algorithm on the basis of machine learning and neural network algorithm to improve the accuracy of learning achievement prediction [14]. To improve students' academic performance, Deeba et al. analyzed the self-concept of English majors in detail, thus providing help for improving students' self-concept and achieving effective prediction of performance [15].

From the research of scholars at home and abroad, the current research method is mainly to manually extract the statistical characteristics of data, which cannot make deep use of effective information. Therefore, on the basis of correlation analysis of campus behavior data, the proposed performance prediction modeling method based on student behavior sequence makes full use of the attributes of education data. On the basis of using data mining technology, it fully makes up for the shortcomings of traditional methods, and is innovative to a certain extent.

3. Research on Student Behavior Analysis and Grade Prediction System Based on Student Behavior Characteristics.

3.1. Analysis of relevant theoretical algorithms of student behavior characteristics. To help teachers improve students' academic performance in a personalized way, the research constructs a classification system model of performance prediction based on students' behavior by extracting students' behavior characteristics. Data mining technology is widely used in student behavior feature extraction. Data mining refers to the use of algorithms to find information hidden in the massive data. It is usually deeply related to computer science, and can effectively mine data through statistical analysis, machine learning, etc [16]. The current data mining process has formed a relatively mature process system, as shown in Figure 3.1.

From Figure 3.1, the process of data mining starts with data collection; Secondly, the target data is

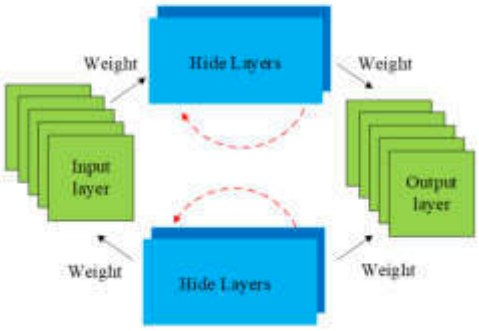


Fig. 3.2: Structure diagram of recurrent neural network

obtained after data preprocessing; Then the feature of target data is extracted to complete the feature selection; Then the corresponding model is obtained through data mining; Finally, the model is evaluated to obtain the corresponding knowledge. In the actual data mining, SVM is selected to analyze and predict the student's behavior sequence and performance. SVM is a generalized binary classification method based on supervised learning. Its decision boundary is the maximum margin hyperplane. On the basis of structural risk minimization, it is formalized to solve a convex quadratic programming problem, and its local optimal solution must be the global optimal solution [17]. In addition, when SVM is linearly nonseparable, a kernel function is introduced to solve the problem of sample classification. On the sequence of students' learning behavior, the research chooses the sequence model of RNN to model the characteristics of students' short-term campus behavior sequence. RNN can capture the characteristics of the input sequence and adapt to various length changes by iteratively updating its internal hidden state. Its structure is shown in Figure 3.2.

From Figure 3.2, a basic RNN is composed of three layers, namely the input layer, hidden layer and output layer. In RNN, the connecting line between the input layer, hidden layer and output layer not only appears among the three, but also exists between multiple hidden layers in the upper time dimension. Among them, the expression of the update value formula of the hidden layer is shown in equation 3.1.

$$h_1 = f(h_{t-1}, x_t) \tag{3.1}$$

In equation 3.1, h represents the hidden state; $f(\dot{})$ represents the nonlinear function; represents time; x represents the input sequence. Specifically, the expression of the state calculation process of the hidden layer is shown in equation 3.2.

$$h_t = f(Ph_{t-1} + Qx_t + b) \tag{3.2}$$

In equation 3.2, P represents the weight matrix of the state of the hidden layer; Q represents the weight matrix of the state from the hidden layer to the input layer; b represents the bias item. Among them, the parameters of RNN model have the same weight in the time dimension, that is, under different time scales, each parameter of the model has the same consistency, thus reducing the training parameters of the model [18]. The learning of network parameters is usually updated by back-propagation algorithm. Therefore, RNN has the ability of short-term memory when processing time series samples of any length, and its optimal variable is the optimal parameter of the recurrent neural network.

3.2. Research on short-term behavior feature extraction of students based on attention. The research chooses to mine and analyze the data from the perspective of classification and statistical characteristics, so as to explore the inner relationship between students' behavioral characteristics and academic performance and realize the prediction of their performance. On this basis, the study uses relevant statistical analysis

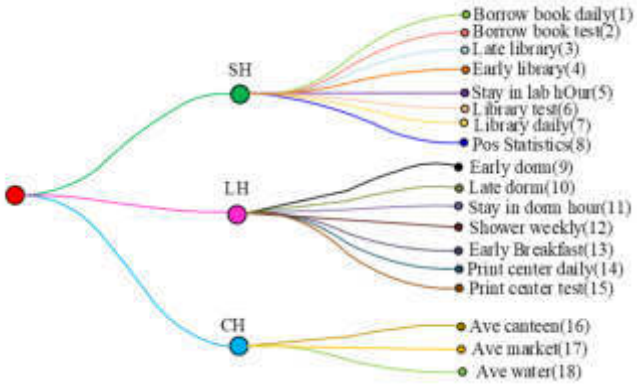


Fig. 3.3: Types of student characteristics proposed by traditional methods

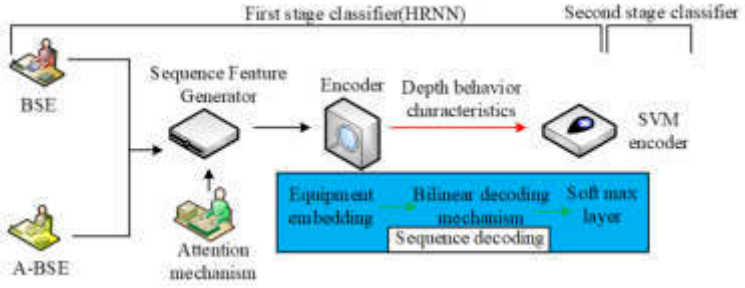


Fig. 3.4: Schematic Diagram of SPC Model Structure

methods and selects student behavior attributes related to grade rankings to extract 18 characteristics, the contents of which are shown in Figure 3.3.

From Figure 3.3 that students' behavior characteristics are mainly divided into three categories, namely, study habits (SH), living habits (LH) and consumption habits (CH). Among them, SH includes the times of borrowing books in the library during non exam and exam time, the times of entering the library before 8 a.m., the times of leaving the library at 22 p.m., the average daily time of staying in the library, the times of going to the library during non exam and exam time, and the times of using point of sales (POS) machines during class. The LH includes the times of leaving the dormitory between 8 a.m., returning to the dormitory after 22 p.m., daily average times of staying in the dormitory, weekly times of getting wet, times of eating breakfast between 8 a.m., and times of going to the printing center during non examination and examination time. CH includes the daily average times of canteen consumption, supermarket consumption and water fetching. In addition, on the basis of sequences, the study introduced a Hybrid Recurrent Neural Network (HRNN) model based on attention in the first stage of student performance prediction, which is used as the first learning classifier to learn the sequence characteristics of students' behavior. The Sequence based Performance Classifier (SPC) is shown in Figure 3.4.

From Figure 3.4 that the SPC model consists of two-stage classifiers. Among them, the base sequence encoding is combined with the output of the attention sequence encoder, which is input into the sequence feature generator. Both are basic networks built on the basis of Gate Recurrent Unit (GRU). The specific process expressions are shown in equation 3.3 and equation 3.4.

$$\begin{cases} z_t &= \rho(W^{(z)}x_t + P^{(z)}h_{t-1}) \\ r_t &= \rho(W^{(r)}x_t + P^{(r)}h_{t-1}) \end{cases} \tag{3.3}$$

In equation 3.3, Z_t represents the update gate; ρ represents the S -type growth curve function (Sigmoid); W represents the weight matrix of the state between hidden layers; r represents the reset gate.

$$\begin{cases} h'_t = \tan(Wx_t + r_t \odot Ph_{t-1}) \\ h_t z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \end{cases} \tag{3.4}$$

According to the analysis of equation 3.3 and equation 3.4, this process linearly interpolates the existing hidden state with the existing hidden state, and the final hidden state contains all the information of all the original sequences. In fact, the last implicit state is used to express the behavior order of students, which is called Basic Sequence Encoder (BSE). The expression is shown in equation 3.5.

$$c_t^g = h_t = h_t^g \tag{3.5}$$

In equation 3.5, c_t^g denotes the base sequence encoder; h_t^g denotes the final hidden state. According to equation 3.5, it can be found that not all student behaviors are linked to their learning performance, so in the actual prediction, the SPC model can pay more attention to the behavioral interaction related to student performance. Based on this, the attention-based sequence encoder (Attention-based Sequence Encoder, A-BSE) is studied, and its expression is shown in equation 3.6.

$$c_t^l = \sum^t \alpha_t i h_i \tag{3.6}$$

In equation 3.6, c_t^l represents the context vector; a_{ti} represents the weighting factor; i represents a specific moment of time. Wherein, c_t^l enables the decoder to dynamically generate and linearly combine different parts of the input sequence. The calculation expression of the attention mechanism function is shown in equation 3.7.

$$\alpha_{ti} = \rho(W_\alpha[h_t; h_i]) \tag{3.7}$$

In equation 3.7, ρ the expressed Sigmoid function does not add up all hidden states learned from the RNN network to represent the student's behavior sequence. Instead, it is important for coders to use weight coefficients to represent hidden states / interactions. Finally, the student's behavior sequence is expressed by weighting and summing these hidden states. In order to better understand the context vector, its modified expression is shown in equation 3.8.

$$c_t^l = \sum^t \alpha_{ti} h_i = \sum^t \alpha_{ti} h_t^l \tag{3.8}$$

In equation 3.8, the last hidden state of BSE is used to encode the entire sequence behavior, and A-BSE is used to calculate the attention weight value of the previous hidden state. This hybrid approach can be built into the same generator, the sequential feature generator in the SPC model. Its expression is shown in equation 3.9.

$$c_t = \left[c_t^g; c_t^l\right] = \left[h_t^g; \sum_{i=1}^t \alpha_{ti} h_t^l\right] \tag{3.9}$$

On the basis of equation 3.9, the study applies an optional bilinear decoding mechanism between the current implicit representation and each campus card device to calculate the similarity score. The expression of this score is shown in equation 3.10.

$$S_i = \mathrm{emb}_i^m T c_t \tag{3.10}$$

In equation 3.10, S_i represents the similarity score; T represents the combination matrix composed of the campus card embedding dimension and the sequence representation dimension. In sequence prediction, BSE is used to summarize students' weekly activities, while A-BSE can automatically select corresponding behaviors according to students' learning habits, so as to understand students' learning motivation.

3.3. Analysis of grade prediction classification model based on student behavior. In the SPC model constructed by the study, HRNN was selected for the first stage classifier study, while the SVM classifier was selected for the second stage study. It can find an optimal compromise between model complexity and machine learning ability by using limited sample data, so that it is more suitable for linear inseparable problems in multi-category situations. Among them, in the actual situation of linear separability, SVM will try to find the optimal classification hyperplane to maximize the interval separation. The expression of the hyperplane is shown in equation 3.11.

$$w^T \cdot x' + b = 0 \tag{3.11}$$

In equation 3.11, w denotes normal vector; T denotes transposition; x' denotes feature. Before finding the hyperplane, a quadratic programming problem needs to be solved first, and the expression of the problem is shown in equation 3.12.

$$\begin{cases} \min\Phi(w) & \frac{1}{2}\|w\|^2 \\ \text{subject to} & y_j\left[(w^T \cdot x'_j + b) - 1\right] \geq 0 \end{cases} \tag{3.12}$$

In equation 3.12, y_j represents the result label; j represents the dimension. At this time, the quadratic programming problem can be solved according to the Lagrangian dual solution method, and its expression is shown in equation 3.13.

$$\min \quad L(w, b, \alpha') = \frac{1}{2}\|w\|^2 - \sum_{j=1}^{l} \alpha'_j[(w^T \cdot x'_j + b) - 1] \tag{3.13}$$

In equation 3.13, α' represents the Lagrangian multiplier. According to equation 3.13, the relationship between normal vector, offset term and Lagrangian multiplier can be obtained by using differential formula and reduction method, and its expression is shown in equation 3.14.

$$\begin{cases} \max & W(\alpha') = \sum_{j=1}^{l} \alpha'_j - \frac{1}{2}\sum_{k,j=1}^{l} \alpha'_j\alpha'_k y_j y_k x_j^T x_k \\ \text{subject to} & \sum_{j=1}^{l} \alpha'_j, \alpha'_j \geq 0 \end{cases} \tag{3.14}$$

Equation 3.14 is a quadratic function optimization problem with inequality constraints. Its objective function and linear constraint are convex functions, and there is a unique solution. The final expression of the optimal classification hyperplane is shown in equation 3.15.

$$f(x) = \sum^{l} \left(\alpha_j^* y_j x_j^T x + b^*\right) \tag{3.15}$$

In equation 3.16, $f(x)$ represents the optimal classification hyperplane function; $\alpha_j'^*$ represents the support vector point. In the case of linear inseparability, you can choose to use kernel functions to map features into higher dimensions to achieve the goal of linear separability. Therefore, using a nonlinear mapping method to map the training samples to a high-dimensional feature space can make the nonlinear classification become a linear classification in the input space. The expression of this category is shown in equation 3.16.

$$\begin{cases} \min\Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{j=1}^{l} \xi_j C > 0 \\ \text{subject to}\xi_j > 0, \quad y_j(w^T \cdot x'_j + b) \geq 1 \end{cases} \tag{3.16}$$

In equation 3.16, Φ represents the nonlinear mapping; $C\sum_{j=1}^{l} \xi_j$ represents the penalty term. The solution process of equation 3.16 is similar to the case of linear separability. At this time, in the high-dimensional space, according to the corresponding kernel function, a linear classifier is used to implicitly construct a classification surface in the high-dimensional space. Based on the theory of structural risk minimization, SVM constructs the optimal segmentation hyperplane in the feature space, so that learners can obtain the global optimal solution,

Table 4.1: Data field information content of experiment part

-	Student Library Borrowing Data						
-	1	2	3	4	5	6	7
Field Description	Student ID	Hours of loan service	Title	Call number	-	-	-
-	Data of student all-in-one card system						
Field Description	Student ID	Consumption category	Consumption place	Consumption mode	Dissipate	Consumption amount	Remaining amount
-	Student dormitory access control data						
Field Description	Student ID	Time of entry and exit	In and out direction (0 in and 1 out)	-	-	-	-
-	Access control data of student library						
Field Description	Student ID	Time of entry and exit	Access control No	-	-	-	-
-	Student performance ranking data						
Field Description	Student ID	College No	Score ranking	-	-	-	-

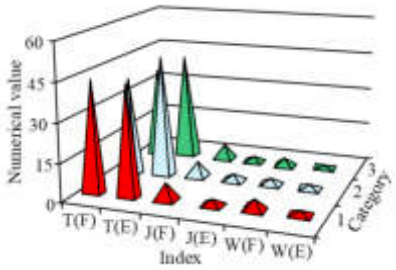
and the required overall risk meets a certain upper bound. In solving multi-classification problems, the research chooses to classify every two categories. According to Figure 4, the actual input of SCM is the complete sequence behavior feature constructed in the HENN classifier network model, and the output is the real grade ranking of students. Therefore, students' learning performance can be regarded as a short-term sequential model. By analyzing the weekly behavior of a specific student and using the SPC two-level classification system to predict it, students with learning crisis can be detected in time.

4. Student behavior data and grade prediction analysis based on student behavior characteristics. To better explain the behavior data extracted from the study, the study used Chint distribution method to divide students' scores into three grades, namely, good, medium and poor, which are represented by A, B and C. Before the experiment, the relevant data were collected and cleaned. It is worth noting that the content of the data set used in the study is the behavior data of students using the card in and out of the campus within two school years and the student score ranking data in the school's teaching management system. Some data field information is shown in Table 4.1.

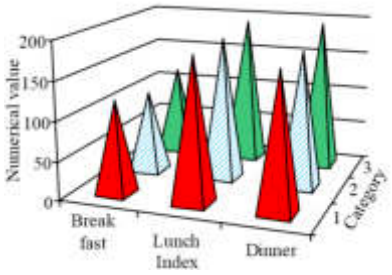
In Table 4.1, the numbers 1-7 are the field serial numbers. From Table 4.1, part of the data field information is mainly divided into four categories, namely, student library borrowing data, student card system data, student dormitory access control data, student library access control data, and student grades ranking data. Among them, the "consumption pattern" field in the all-in-one card data is the focus of research. When there is a "default" in the original data, the row where the "user" is located should be "deleted". If the borrowing time, consumption time and entry and exit time fields are the same twice, the row where it is located will be deleted. Because repeated records will analyze the data, resulting in errors. On this basis, the study combines the feature classification results in Figure 3 to give three levels of students who go to the library (T), borrow books (J) and go to the printing center (W) during the examination and non-examination periods. The number of times was compared, and the three meals of the students of the three levels were compared at the same time, and the results are shown in Figure 4.2.

In Figure 4.2, E and F represent examination and non-examination respectively. Figure 4.1(b) In order to reduce the actual error, the students of No. 1 college of this school were selected for analysis. From Figure 4.1(a) that the students of grade A borrow 52 books on average, the students of grade B borrow 48 books, and the students of grade C borrow 42 books on average. Likewise, as the grade decreases, the number of activities such as studying in the library decreases gradually. It shows that the better the students' academic performance, the more practical and feasible they are in their daily life on campus. From Figure 4.1(b) that the lunch of the three types of students is much larger than the breakfast, indicating that there is an irregular diet in the student group. In addition, in the comparison of the three types of students, students with grade A have the most breakfasts, which shows that students with academic performance have relatively good eating habits and are more self-disciplined. To verify the correlation between the behavior characteristics of students in Figure 5 and their academic performance, the study further used the statistical methods of factor analysis and principal component analysis to interpret the correlation between the two. Among them, the study used the measure (Kaiser-Meyer-Olkin, KMO) and Bartlett's sphericity (Bartlett's) test to select features when conducting factor analysis. The result is shown in Figure 4.4.

From Figure 4.3(a) that the statistic of KMO is 0.720, which is greater than 0.6, indicating that the research experiment can be used for factor analysis. At the same time, the approximate chi square value obtained by

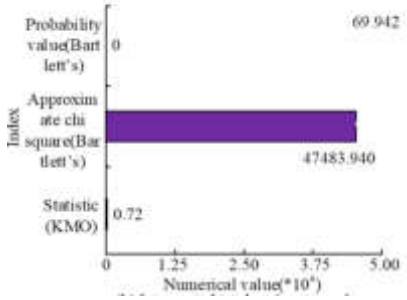


((a)) Results of behavior comparison among three types of students

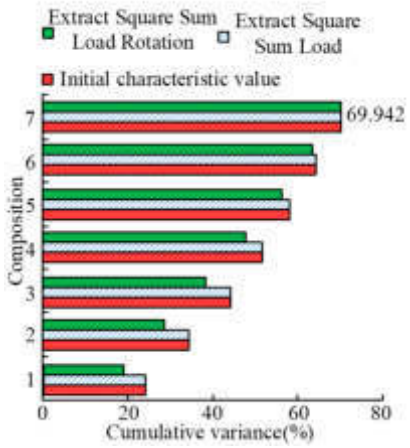


((b)) Three Meals for Three Kinds of Students

Fig. 4.2: The number of times that three types of students go to three places and the diet comparison results of three meals



((a)) Interpreted total variance results

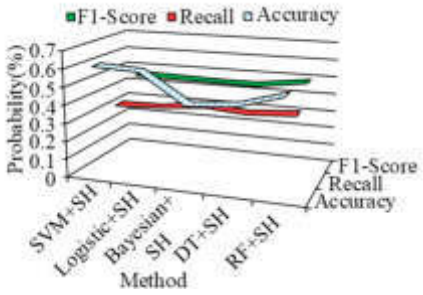


((b)) Interpreted total variance results

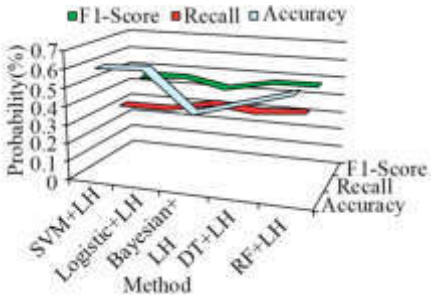
Fig. 4.4: KMO and Bartlett's test results and explained total variance results

Bartlett's test is 47483.940, and the probability value at this time is 0.000, less than 0.05, reaching the level of significance. The result shows that there is correlation between variables. Therefore, according to Figure 4.3(b) obtained by factor analysis in Figure 4.3(a), the overall variance of the first seven factors is 69.942%, which meets the actual needs, effectively reflects the overall information and shows that there is a significant relationship between students' learning behavior and academic performance. On this basis, based on the results obtained in Figure 4.4 and the traditional behavior characteristics, the research built a multi category prediction model of Bayesian, logistic, DT, RF and SVM, and used three evaluation methods, namely, accuracy, recall and F1 score, to evaluate the prediction models of these students' scores. The result is shown in Figure 4.6.

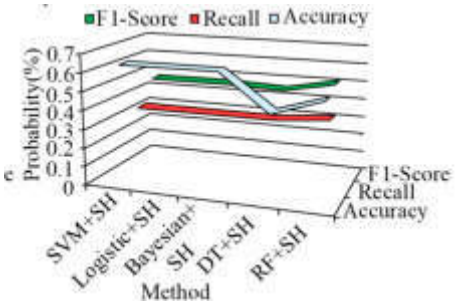
From Figure 4.6 that the accuracy of the five methods is maintained between 40% and 60%, the recall rate is maintained at about 35%, and F1 Core is maintained at about 45%. In general, Bayesian model has the worst classification effect, the overall accuracy of logical regression and SVM is relatively more stable, and has a stronger mathematical interpretation of the research data. At the same time, it also proves that it is feasible to predict students' performance by using their behavior data, and verifies the effectiveness of SVM. After verifying the feasibility of the research direction, the research continues to analyze the actual performance of



((a)) Classification of learning habits by different methods



((b)) Classification of living habits by different methods



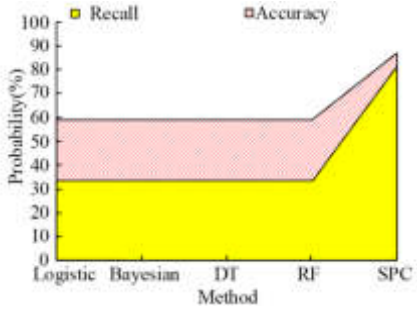
((c)) Classification of consumption habits by different methods

Fig. 4.6: Results of the five traditional classification methods

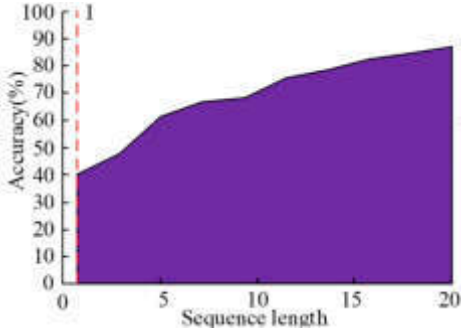
the SPC model system and its feature modeling effect on the student behavior sequence. To reduce the impact of uncertain data on the actual results, the study extracted the behavior records of attacking more than 9000 students in 29 weeks. The performance verification test results are shown in Figure 4.8.

From Figure 4.8 that the SVM based on RNN can handle the sequences in multiple sequences well and complete the classification task efficiently in the process of processing multiple sequences. After considering the students' sequential behavior and the main goals of learning, the SPC proposed in the study can exceed all the benchmarks. The relative performance accuracy is 86.90%, and the response rate is 81.57%. In addition, when the length of behavior sequence increases from 1 to 20, the performance of SPC model exceeds 70% accuracy, indicating that it can achieve better prediction. On this basis, the research experiment analyzes the accuracy rate and recall rate of SPC model system in different cycles and dimensions, and the results are shown in Figure 4.10.

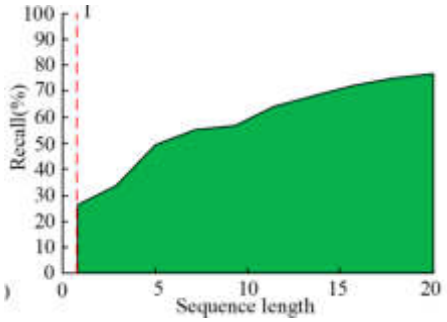
From Figure 4.10 that when the sequence length of the SPC model is divided by month, the excessive sequence length makes it impossible for managers to intervene in students' learning in time, resulting in a decline in accuracy, which is only 71.40%. It also shows that short-term continuity modeling can better predict students' learning situation and performance. In addition, as the feature dimension increases from 5 to 100, the performance of the model has been greatly improved, with the maximum accuracy rate exceeding 75% and the recall rate exceeding 60%. At the same time, the increase in the number of hidden units also prompted the curve of the two indicators to gradually show a gentle trend. Therefore, although the increase of hidden layer feature dimension can more comprehensively grasp the situation of students, the representation ability of the model will decline when the dimension exceeds 50. Therefore, it is necessary to set appropriate feature dimensions to better predict students' academic performance. And let educators better help students.



((a)) Comparison of accuracy and recall of different methods



((b)) Accuracy results of SPC model under different sequence lengths

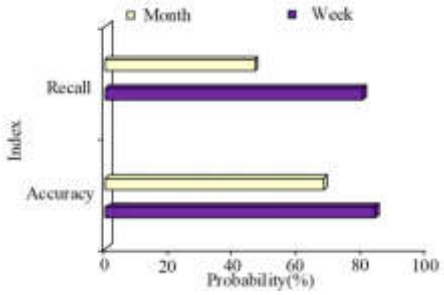


((c)) Recall rate results of SPC model under different sequence lengths

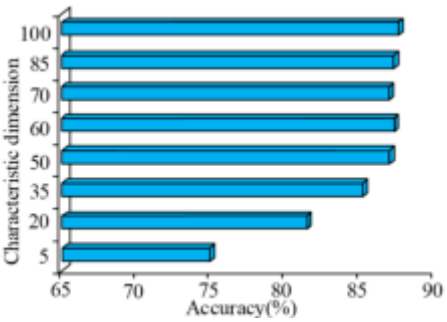
Fig. 4.8: SPC model performance test results

5. Research limitations and prospects. The selection of data mining in research still needs to be optimized. With the rise and development of artificial intelligence technology, applying artificial intelligence algorithms to SPC models may further improve their performance. At the same time, due to many restrictions, the research only applies it to education data. In fact, in the era of Big data, it can be popularized in other fields, such as short video user data, e-commerce, but it is undeniable that the research data can only be applied in education, and is affected by different educational environments. Therefore, in the future, the application of AI algorithms, Big data technology, cloud computing and other cutting-edge technologies in the SPC model will further enhance its applicability in the educational environment and expand.

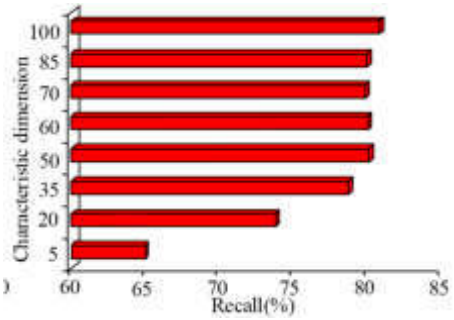
6. Conclusion. To realize the effective prediction of students' grades and help teachers to improve students, the study extracts the behavior characteristics of students, and constructs the SPC system model on the basis of HRNN and SVM, and analyzes the behavior of students. The features and performance of the SPC model were analyzed experimentally. The results of the experiment show that the average number of books borrowed by grade A students is 52, which is higher than that of the other two grades. At the same time, the number of breakfasts is also the most, which shows that students with good academic performance have good study habits. In the factor analysis, the overall variance of the first seven factors reached 69.942%, indicating that there is a significant correlation between students' learning behavior and academic performance. Therefore, in the performance analysis of the SPC system model, the accuracy rate of the five methods is maintained between 40% and 60%, the recall rate is maintained at about 35%, and the F1-Score is maintained at about 45%. At the same time, the stability of SVM is the highest, indicating that it has high effectiveness. In addition, the accuracy rate of the SPC model reached 86.90%, and the response rate reached 81.57%, both



((a)) SPC Model Accuracy and Recall Results in Different Periods



((b)) SPC model accuracy results under different feature dimensions



((c)) SPC model recall results under different feature dimensions

Fig. 4.10: Accuracy and recall results of SPC model in different periods and dimensions

higher than other methods. When the sequence length and feature dimension are increased to 20 and 100, respectively, the accuracy of the model exceeds 70%, and the highest exceeds 75%. On the whole, the SPC system model constructed by the research is effective in extracting student behavior characteristics for performance prediction. However, the selection and application of data mining technology in the research has yet to be optimized and needs to be improved in the future. of SVM is the highest, indicating that it has high effectiveness. In addition, the accuracy rate of the SPC model reached 86.90%, and the response rate reached 81.57%, both higher than other methods. When the sequence length and feature dimension are increased to 20 and 100, respectively, the accuracy of the model exceeds 70%, and the highest exceeds 75%. On the whole, the SPC system model constructed by the research is effective in extracting student behavior characteristics for performance prediction. However, the selection and application of data mining technology in the research has yet to be optimized and needs to be improved in the future.

REFERENCES

[1] Umer, R., Khan, S., Ren, J., Umer, S. & Shaykat, A. Prediction of students' failure using VLE and demographic data: case study on Open University data. *International Journal Of Business Intelligence And Data Mining*. **20**, 235-249 (2022)

[2] AdrianChin, Y., JosephNg, P., Eaw, H., Loh, Y. & And, S. and learning behaviour dubious relationship. *International Journal Of Business Information Systems*. **41**, 548-568 (2022)

[3] Trakunphutthirak, R. Lee V C S. Application of educational data mining approach for student academic performance prediction using progressive temporal data. *Journal Of Educational Computing Research*. **60**, 742-776 (2022)

[4] B"ut"uner, R. & Calp, M. Estimation of the Academic Performance of Students in Distance Education Using Data Mining Methods. *International Journal Of Assessment Tools In Education*. **9**, 410-429 (2022)

[5] Shreem, S., Turabieh, H., Al Azwari, S. & Baothmn, F. Enhanced binary genetic algorithm as a feature selection to predict

student performance. *Soft Computing*. **26**, 1811-1823 (2022)
[6] Mingyu, Z., Sutong, W., Yanzhang, W. & Dujuan, W. An interpretable prediction method for university student academic crisis warning. *Complex & Intelligent Systems*. **8**, 323-336 (2022)
[7] Kittur, J., Bekki, J. & Brunhaver, S. Development of a student engagement score for online undergraduate engineering courses using learning management system interaction data. *Computer Applications In Engineering Education*. **30**, 661-677 (2022)
[8] Xiong, S. Gasim E F M, Ying C X, Wah K K, Ha L M. *A Proposed Hybrid CNN-RNN Architecture For Student Performance Prediction*. **10**, 347-355 (2022)
[9] Keser, S. & Aghalarova S. Hela:, A. novel hybrid ensemble learning algorithm for predicting academic performance of students. *Education And Information Technologies*. **27**, 4521-4552 (2022)
[10] Nti, I., Akyeramfo-Sam, S., Bediako-Kyeremeh, B. & Agyemang, S. Prediction of social media effects on students' academic performance using Machine Learning Algorithms (MLAs). *Journal Of Computers In Education*. **9**, 195-223 (2022)
[11] Teng, M., Qin, C. & Wang, C. Validation of metacognitive academic writing strategies and the predictive effects on academic writing performance in a foreign language context. *Metacognition And Learning*. **17**, 167-190 (2022)
[12] Britwum, F., Amoah, S., Acheampong, H. & Adiei, E. Self-esteem as a predictor of students' academic achievement in the colleges of education. *International Journal Of Learning And Teaching*. **14**, 29-40 (2022)
[13] Serrano, C., Murgui, S. & Andreu, Y. Improving the prediction and understanding of academic success: The role of personality facets and academic engagement. *Revista De Psicodidáctica (English Ed.* pp. 21-28 (2022)
[14] Su, Y., Wang, S. & Li, Y. Research on the improvement effect of machine learning and neural network algorithms on the prediction of learning achievement. *Neural Computing And Applications*. **34**, 9369-9383 (2022)
[15] Deeba, F., Ullah, S. & Saleem, A. An Analysis of Students' Academic Self-Concept in English as Predictor of their Academic Performance. *Pakistan Journal Of Humanities And Social Sciences*. **10**, 403-415 (2022)
[16] Hidalgo, Á., Ger, P. & Valentín, L. Using Meta-Learning to Predict Student Performance in Virtual Learning Environments. *Applied Intelligence*. **52**, 3352-3365 (2022)
[17] Ajibade, S., Dayupay, J., Ngo-Hoang, D., Oyebode, O. & Sasan, J. Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students. *Journal Of Optoelectronics Laser*. **41**, 48-54 (2022)
[18] Shin, J., Chen, F., Lu, C. & Bulut, O. Analyzing students' performance in computerized formative assessments to optimize teachers' test administration decisions using deep learning frameworks. *Journal Of Computers In Education*. **9**, 71-91 (2022)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 15, 2023
Accepted: Nov 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 229–240, DOI 10.12694/scpe.v25i1.2287

RESEARCH ON THE APPLICATION OF PROJECT TEACHING METHOD IN THE NEW MODEL OF SOFTWARE ENGINEERING COURSE

LI MA* AND LEI HUANG†

Abstract. Software engineering course is one of the core courses of computer science. The students trained in the current teaching mode can no longer meet the market demand for high-technology talents. Based on this, the research attempts to optimize the traditional software engineering teaching mode by using the project teaching method (PTM). According to the basic concept of PTM and course characteristics, the reform path of PTM in software engineering course is explored in the experiment. Then in the experiment, the indicators that affect teaching reform effect is selected and a evaluation model is built. And GA-BP algorithm is used to evaluate the effect of the evaluation model. To verify the performance of the built model and the final evaluation effect, the research results are tested from the fitness of the algorithm, error performance, prediction in the data set and other aspects. GA-BP algorithm converged when it iterated to the 18th generation, and the final fitness value was 0.61. The average error square value of GA-BP was 0.35 and the minimum error square sum of GA-BP was 0.48. Its prediction accuracy in test set and training set was 93.4% and 94.1% respectively. The maximum prediction error in the training sample was only 0.015, and the performance of the above data was better than the other three algorithms. To sum up, applying PTM to software engineering curriculum reform can achieve better teaching results.

Key words: Project teaching method; Software engineering; GA-BP; Evaluation model

1. Introduction. Software engineering is one of the core courses of computer colleges in colleges and universities. This requires students to master the relevant theoretical knowledge of software technology in daily learning. And students' engineering awareness and practical hands-on operation design ability need to be cultivated [1]. PTM advocates that under the guidance of the teacher, a relatively independent curriculum project should be handed over to the students themselves. In this process, students' ability to process and collect information, design schemes and successfully implement planned projects can be cultivated [2]. In the process of implementation, there are three prominent features in PTM. The training cycle of the whole teaching process is short, effective and controllable. Students can participate in the project and complete the project design according to the established teaching objectives [3]. BP neural network has many advantages in solving nonlinear problems. In view of the problem that BP neural network is prone to fall into local minimum at the later stage of iteration, many scholars have adopted various algorithms to optimize it [4, 5]. The software engineering teaching in major universities still are concerned about teaching mode and the explanation of basic theoretical knowledge. The whole teaching content is boring. Therefore, students are prone to problems such as boredom, boredom and lack of real participation in learning. About the teaching problems in traditional software engineering courses, the research attempts to reform the traditional software engineering courses with a new teaching method - PTM. GA-BP can evaluate the reformed teaching model, which provides a new reference for innovating the teaching mode of software engineering courses in colleges and universities.

2. Related Works. PTM is to build a real teaching environment around various problems or social issues in life. Under the constructed teaching environment, it can help students turn these issues into driving questions and use them in teaching activities. Given that the current teaching environment is changing, the traditional one-to-one knowledge transfer learning mode no longer meets the requirements of high-quality and high-level talent cultivation [6]. Zak et al. applied the PTM to the teaching and cultivation of engineering students, trying to improve students' learning ability, hands-on ability, cooperation ability, etc. To better show the reform results of PTM applied to engineering teaching, the course completion quality of chemical engineering students was

*School of Artificial Intelligence, Chongqing Youth Vocational & Technical College, Chongqing, 400712, China (mie8787@163.com)
†General Education College, Chongqing Youth Vocational & Technical College, Chongqing, 400712, China (huangsweet@163.com)

evaluated by virtual demonstration and face-to-face demonstration. Students had better performance in the chemical engineering curriculum optimized by the PTM [7]. Vascocelos et al. applied the PTM to the chemical engineering curriculum design project. Students needed to complete the three processes of pre-fermentation, fermentation and distillation in a group cooperation way, and finally report the results before the auditors. In this design project, students and teachers were highly satisfied with the proposed teaching plan [8]. To improve the operation ability of civil engineering students in the actual project design and construction, AMR Pasanda and others carried out optimization experiments on the traditional civil engineering teaching mode. The teaching program combining the PTM and the traditional teaching program were used to teach the students of civil engineering. The teaching scheme combined with the PTM could cultivate more students who have strong movement ability, and their ability to solve problems was also stronger in the actual project operation [9].

Gao et al. used GA-BP to predict the fetal weight. This method aimed to determine whether the fetus could develop healthily and whether the pregnant woman could give birth smoothly. The accurate prediction of fetal weight through the establishment of fetal weight prediction model could provide a new guarantee method for the safety of fetus and pregnant women [10]. Zou M et al. used GA-BP to identify lunar soil's shear parameters, thus providing data support for the path planning, traction control and risk avoidance of the lunar rover. GA-BP algorithm could accurately and effectively identify the shear parameters of lunar soil [11]. Wang Y et al. proposed a model predictive control method combining recurrent back-propagation neural network and genetic algorithm for nonlinear systems with time-delay and uncertainty. In the offline modeling stage, GA-BP network was introduced as the prediction model and used to train parameters. The method proposed in the experiment could effectively reduce the computational load of the nonlinear control system [12].

According to previous research, many colleges and universities had applied PTM to engineering teaching [13]. To solve problems such as path planning and risk avoidance through prediction values, many scholars also used GA-BP neural network to build prediction models [14]. Based on this, the research attempts to apply the PTM to the reform of software engineering curriculum, and constructs the corresponding index evaluation model. The results are used to improve the deficiencies in current software engineering teaching.

3. Research on the application of PTM in software engineering courses and the construction of evaluation index system.

3.1. Path exploration of PTM in software engineering curriculum reform. Software engineering course not only needs to cultivate students' engineering awareness, but also needs to cultivate students' practical ability in teaching [15, 16]. Each new teaching model must refer to some previous teaching ideas and theoretical research when it is proposed and applied. The PTM is to build a real teaching environment around various problems or social issues in life, so that students can turn these issues into driving problems and use them in teaching activities under the constructed teaching environment. In the specific project teaching process, students are not only required to study independently, but also need to cooperate. This research carries out path reform and exploration from the following five aspects.

In Figure 3.1, it mainly explores the application mode of PTM in software engineering courses from the five major theories. They are constructivism learning theory, human comprehensive development theory, learning transfer theory, pragmatism learning theory, and the theory of recent development zone. According to the constructivist school, teachers should care about the dynamic characteristics of knowledge in daily teaching. The individual differences of students' experience should be cared about. And students' interactivity, scene and self-construction in classroom learning should be enriched. The PTM also dominates students' autonomous learning in the whole teaching process, so it conforms to the idea of constructivism school. The core idea of pragmatism advocates "learning by doing", with experience as the core, children as the main body and activities as the center. The PTM can not only combine the theory of pragmatism and emphasize the connection between knowledge, but also emphasize the self-development of students. The recent development zone theory believes that when teaching, students' actual and potential development difference must be paid attention. The difference between the two is also called the nearest development zone. The PTM relies on this theory to develop the modern teaching concept in an all-round way. So that every student who is required to take the software engineering course can learn to build their own understanding of the course. Then, under the leadership of the teachers, the level of their own development has been improved. The theory of all-round development of human beings points out that people with all-round development are those who have achieved

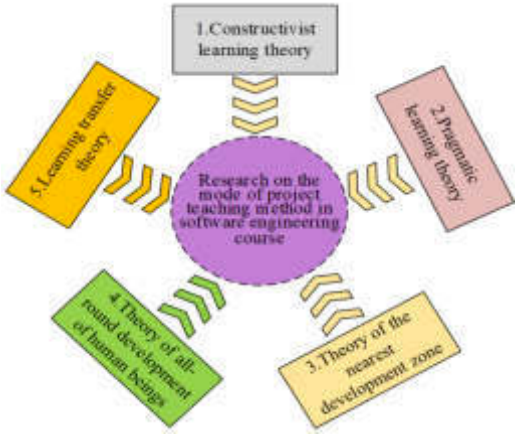


Fig. 3.1: Theoretical basis of PTM in software engineering curriculum reform

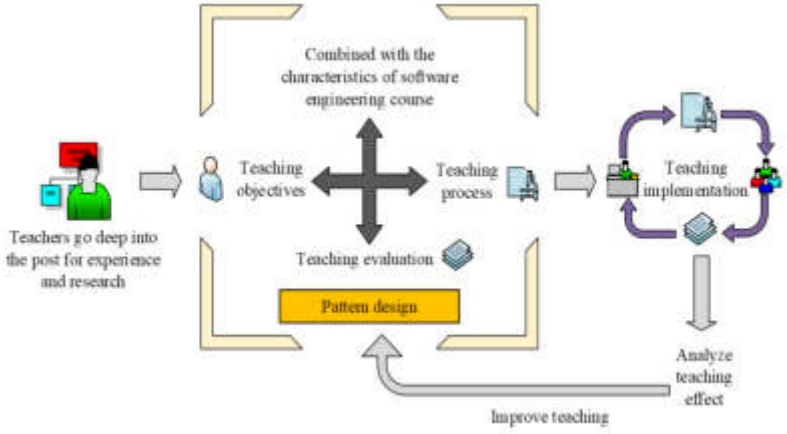


Fig. 3.2: The reform path planning of PTM in software engineering course

coordinated development in physical strength and intelligence. According to this idea, the PTM emphasizes that students should make full use of their potential in all aspects during the teaching process. It should promote students to become people with all-round development of morality, intelligence, physique, beauty and labor in the learning process. The theory of learning transfer believes that, based on the premise of mastering old knowledge, new knowledge's acquisition can be carried out. Because new and old knowledge has a certain transfer function, teachers should consciously give correct guidance to students, and lead them to transferring the knowledge that they have learned in the teaching process. So that a complete knowledge system can be built and students can cultivate their creativity and hands-on ability. Through the above theoretical knowledge, the PTM implementation plan suitable for the software engineering course is formulated. Figure 3.2 shows the reform path planning of PTM in software engineering course.

Before applying PTM to software engineering teaching, it is necessary to investigate the teachers. This is to ensure that teachers have excellent teaching ability, can go deep into the project posts and have sufficient teaching experience. The mode design principle of PTM in software engineering course includes three modules: teaching objectives, teaching process and teaching evaluation. When designing patterns, we should fully combine the characteristics of the software engineering course, so that students can really simulate the development of software under the guidance of teachers. After the model design is completed, it will be applied to the actual

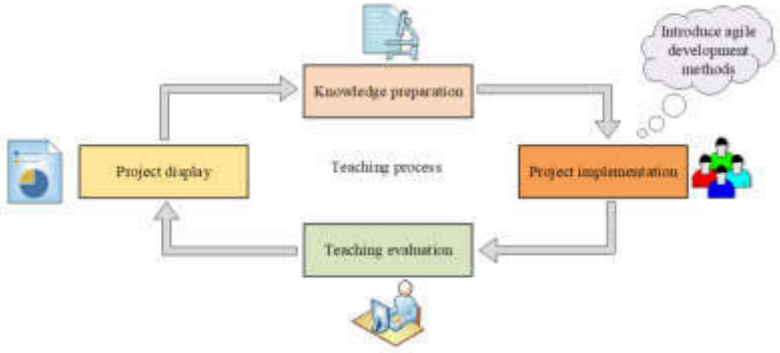


Fig. 3.3: Teaching flow chart of software engineering course combined with PTM

education and teaching. Finally, it will reflect on the teaching results, propose improvement methods for the shortcomings of this teaching, and change the original model design. According to the above reform planning path, the traditional software engineering teaching mode is repeatedly modified until the final teaching model can achieve good evaluation results.

In view of the most important link in the reform path, the construction of teaching process and teaching evaluation index model in the model design is crucial. The construction of teaching evaluation index model will be described in detail in the next section. The design of teaching process is mainly divided into four modules: project presentation, knowledge preparation, project implementation and teaching evaluation. In the project implementation stage, it needs introduce agile development methods to optimize the traditional teaching project implementation methods. The agile development method advocates to focus on research and development, and teaching is only auxiliary. Returning the final R & D results to students can stimulate students' potential for independent learning. At the same time, they can incubate high-quality technical talents for enterprises in advance.

3.2. Construction of PTM used in the software engineering curriculum reform. The last section explored the reform path of PTM in software engineering courses. Next, we will evaluate the implementation effect of PTM by building an evaluation index model. By analyzing the results of the questionnaire and the participation of students, the index factors that affect the PTM in the software engineering curriculum reform are determined.

The evaluation indicators of the PTM used in the software engineering curriculum reform are shown in Table 3.1. Next, the research will optimize the traditional BP neural network, use genetic algorithm to improve the shortcomings of BP neural network, and propose GA-BP neural network model to evaluate the constructed index system. In Figure 3.4, as the basic component of neural network, the structure of artificial neuron is shown [17].

The main components of artificial neural network are the selection of connection weights, summation steps and activation functions in Figure 3.4. The connection weight can combine neurons of different strength. The value of each neuron is weighted and summed by the adder. Finally, the activation functions of different thresholds are set to limit the output amplitude of neurons so that the output signals are consistent.

$$Y_j = f\left\{\left[\sum_{i=1^n} W_{ij}X_i - \theta_j\right]\right\} \tag{3.1}$$

Equation 3.1is the mathematical expression of the basic flow of the above neural network. j and i represent neurons. X_i is neuron input. W_{ij} represents the connection weight value. θ_j is the threshold value. Y_j represents neuron output. $f(\bullet)$ is the activation function. The selection of activation function is generally

Table 3.1: Evaluation index of PTM used in software engineering curriculum reform

Evaluation system	Main indicators	Secondary indicators	Indicator code
PBL used in software engineering curriculum reform	Teacher indicators	Teaching experience	Q1
		Teaching style	Q2
		Knowledge mastery and familiarity	Q3
	Student indicators	Learning attitude	Q4
		learning interest	Q5
		Daily attendance	Q6
		Proficiency in using relevant software	Q7
		Ability to use the theoretical knowledge of software engineering course for hands-on design	Q8
		Ability to apply theoretical knowledge of software engineering course	Q9
		Thinking ability	Q10
		Communication ability	Q11
		Software project design effect	Q12
	Content of courses	Theoretical knowledge learning	Q13
		Structured design	Q14
		Software project management	Q15
		After-class Q&A	Q16
	Teaching equipment	Multi-functional media device	Q17

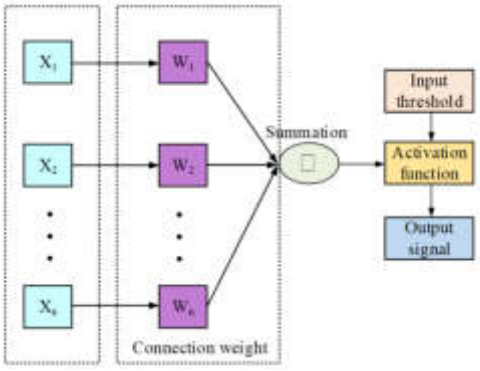


Fig. 3.4: Structure of artificial neuron

divided into threshold function, piecewise linear function and sigmoid function.

$$f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{3.2}$$

Equation 3.2 is the expression of threshold function. Its function is to map the input value to two output values. When $x \geq 0$, its mapping output value is 1, which means that the corresponding neuron is in an excited

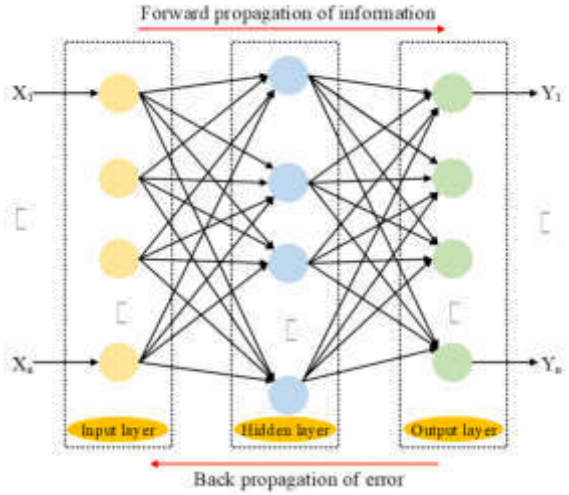


Fig. 3.5: BP neural network structure

state. When $x < 0$, its mapping output value is 0, and the neuron is in the inhibition state.

$$f(x) = \begin{cases} 1, x \geq 1 \\ x, -1 < x < 1 \\ -1, x \leq 1 \end{cases} \tag{3.3}$$

Equation 3.3 is the expression of piecewise linear function. The linear function can amplify the input signal according to the neuron model. In the interval (- 1,1), the linear function can be used as a linear combiner. When the coefficient in the linear interval is infinitely amplified, the linear coefficient becomes a threshold function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

Equation 3.4 is the expression of the sigmoid function. Because of its smooth and easy derivation, it can reduce the amount of model calculation and is often used as the activation function [18, 19]. BP has strong learning ability and can be used to solve nonlinear problems with complex internal mechanism. It can still work normally in case of local damage and can correctly classify the objectives. So, it is widely used to solve problems in different fields.

Figure 3.5shows the general structure of BP neural network, which is mainly composed of three layers. BP neural network's training process includes: forward propagation of information and back propagation of error. When the output result is not within the error range, the model will be automatically transferred back. The output error is reduced by modifying weight and threshold. The model will not stop training until the error finally meets the predetermined range. A three-layer BP is taken, and it needs be initialized first. The number of neurons in the input layer is set as . The number of neurons in the hidden layer is set as n . The number of neurons in the output layer is set as q . After setting the error function, it needs calculate the precision value and the maximum learning times. And the random number within the weight (- 1,1) range is given.

$$\begin{cases} X(k) = (X_1(k), X_2(k), \ldots, X_n(k)) \\ D_o(k) = (D_1(k), D_2(k), \ldots, D_q(k)) \end{cases} \tag{3.5}$$

Equation 3.5 shows the input sample X_k of input layer k and the expected output $D_o(k)$ corresponding to each input sample.

$$\begin{cases} H_h(k) = \sum_{i=1}^{n} W_{ih} X_i(k) - b_h \\ H'_h(k) = f(H_h(k)) \end{cases} \tag{3.6}$$

Equation 3.6 shows the expression formula of hidden layer input vector and output vector. $H_h(k)$ represents hidden layer's input sample, and $H'_h(k)$ represents hidden layer's output.W_{ih} represents input and hidden layer's connection weight. b_h represents the threshold value of the hidden layer.

$$\begin{cases} I_o(k) = \sum_{H=1}^{P} W_{ho} H_i(k) - b_o \\ I'_o(k) = f(I_o(k)) \end{cases} \tag{3.7}$$

$I_o(k)$ represents output layer's input sample, and $I'_o(k)$ represents output layer's output sample. W_{ho} represents hidden and output layer's connection weight. b_o represents the threshold value of the output layer.

$$e = \frac{1}{2} \sum_{o=1}^{q} (D_o(k) - I'_o(k))^2 \tag{3.8}$$

Equation 3.8 is the expression of error function e. The partial derivatives of the hidden layer and the output layer can be calculated through the error function.

$$\delta_o(k) = (D_o(k) - I'_o(k) f'(I_o) k) \tag{3.9}$$

Equation 3.9 is the calculation formula of partial derivative of output layer weight $\delta_o(k)$.

$$\delta_h(k) = -(\sum_{o=1}^{q} \delta_o(k)) W_{ho} f'(H_h(k)) \tag{3.10}$$

Equation 3.10 is the calculation formula of partial derivative of hidden layer weight $\delta_h(k)$. Combining formula 3.9 and 3.10 can correct W_{ho} and b_o .

$$W_{ho}^{(N+1)}(k) = W_{ho}^{N}(k) + \eta \delta_o(k) H'_o(k) \tag{3.11}$$

Equation 3.11 is the calculation formula of the corrected connection weight $W_{ho}^{(N+1)}$. η represents the learning step. $W_{ho}^{N}(k)$ is the value before correction.

$$b_o^{(N+1)}(k) = b_o^{N}(k) + \eta \delta_o(k) \tag{3.12}$$

Equation 3.12 represents the calculation formula of corrected threshold $b_o^{(N+1)}(k)$. $b_o^{N}(k)$ is the threshold value before correction. Wih and b_h are modified in the same way, and the global error is calculated at last.

$$E = \frac{1}{2m} \sum_{k=1}^{m} \sum_{o=1}^{q} (D_o(k) - I'_o(k))^2 \tag{3.13}$$

Equation 3.13 is the calculation formula of global error, and m represents the learning times. If the final global error result reaches the set accuracy value or its learning times exceed the maximum learning times, the algorithm will stop running. Otherwise, new samples will be selected for the next round of learning. Although BP neural network has many advantages in solving nonlinear problems, it is easy to have local minimum problems due to the use of gradient descent algorithm for training. The method to optimize BP neural network is a algorithm which has strong global optimization probability, namely genetic algorithm (GA). It can optimize BP neural network's weight and threshold. And it can accelerate the convergence speed of the whole model by optimizing the individuals in BP neural network.

Genetic algorithm is an optimization tool that simulates biological evolution, the algorithm simulates the collective evolutionary behavior of a population where each individual represents an approximate solution to the search space of the problem. Starting from an arbitrary initial population, genetic algorithm effectively implements a stable and optimized breeding and selection process through individual inheritance and mutation, so that the population evolves to a better range of search space. Genetic algorithm optimization of BP neural network is mainly divided into the following three parts. The first is to use the algorithm to determine the

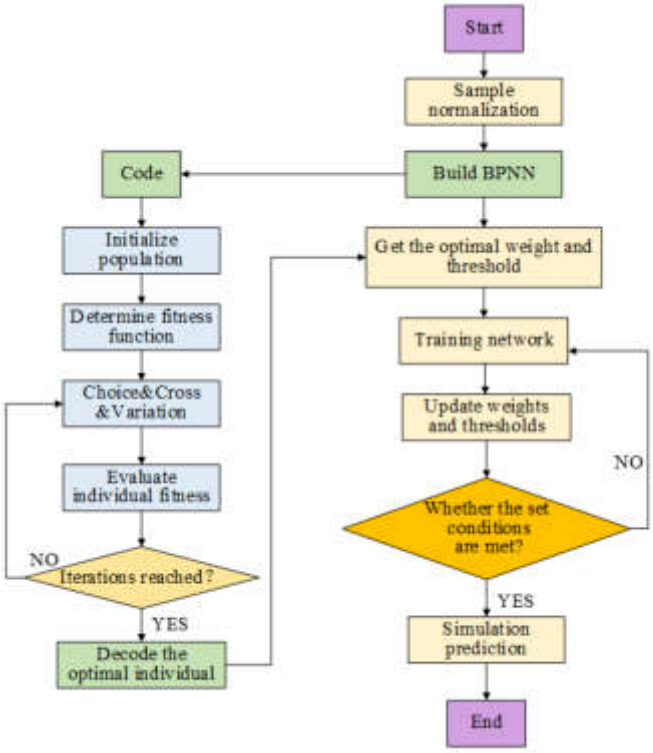


Fig. 3.6: Flow chart of GA-BP optimization algorithm

specific structure of the BP neural network, including the number of nodes in the final input layer as well as the number of neurons in the output layer ultimately determined according to the evaluation results. The second is to use the algorithm to optimize the weights and thresholds of the BP neural network. The algorithm is capable of randomly generating a population whose individuals represent the network weights and thresholds, and then the fitness function is utilized to calculate the fitness value, and finally the optimal individuals are found through selection, crossover and mutation operations. Finally, BP neural network (Genetic Algorithm- Back propagation, GA-BP) under GA optimization is used for prediction. After initialization with the optimal individuals, the weights and thresholds of the BP neural network can be locally optimized again during the training process, thus making the GA-BP neural network have better prediction accuracy and prediction efficiency. The indicators in the evaluation index system are used as the input of GA-BP, and after a series of operations, the predicted value of the model can be obtained. Comparing the predicted values of each index, we can get the degree of influence of each factor on the software engineering teaching mode, so as to assist colleges and universities to take corresponding measures to carry out educational reform.The operation flow chart of the GA-BP algorithm is shown in Figure 3.6.

The flowchart of the optimized GA-BP algorithm operation is shown in Figure 3.6. The positive and negative propagation mode of BP neural network can adjust the neuron weight and threshold. Therefore, the model will evaluate the effect of PTM applied to software engineering curriculum reform according to the algorithm process [20]. The evaluation results of the model can determine the factors that have a greater impact on the software engineering curriculum reform by using PTM. The teaching model can be further improved by modifying the relevant indicators.

4. Evaluation effect of PTM in software engineering curriculum reform. The result analysis part evaluated the effect of the PTM used in the software engineering curriculum reform. The performance of the index evaluation model was tested first. Through Matlab, the PTM using GA-BP optimization algorithm was

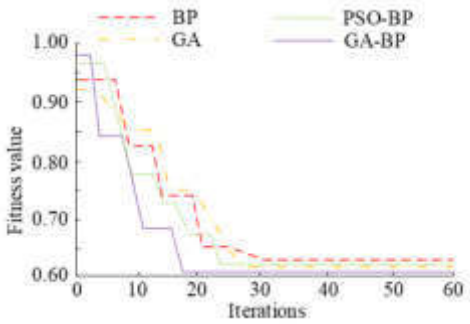
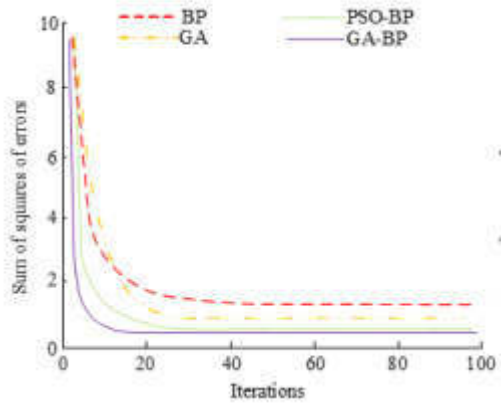
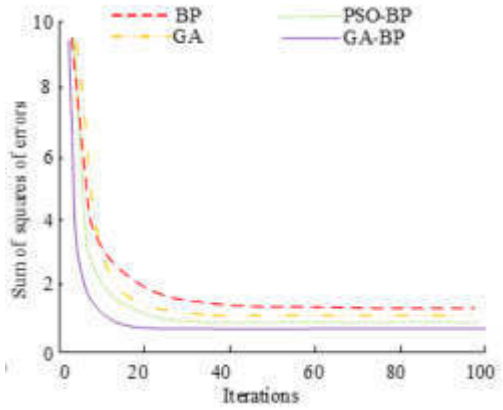


Fig. 4.1: Adaptability changes of different algorithms under different iterations



((a)) Sum of squares of average errors of different algorithms



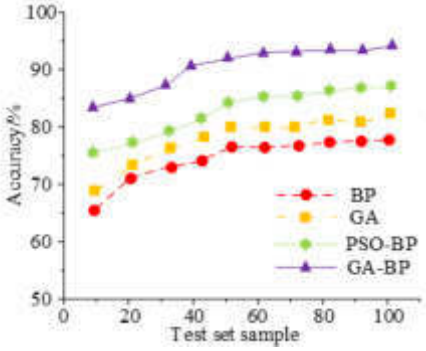
((b)) Sum of squares of minimum errors of different algorithms

Fig. 4.3: Error performance of different algorithms

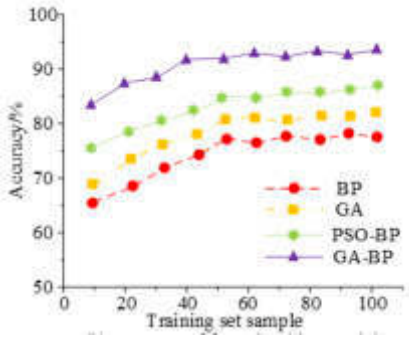
used to simulate the effect of software engineering curriculum reform.The performance of the algorithms was tested using a home-made dataset, which was divided into a training dataset and a test dataset in a ratio of 9:1, to train and test each algorithm.

Figure 4.1 showed the fitness changes of the four algorithms under different iterations. The GA-BP algorithm used in this study was compared with the traditional BP neural network, the GA algorithm, and the improved BP network based on PSO algorithm (PSO-BP). With the increase of the number of iterations, the four algorithms could finally converge to a stable state. Among them, the GA-BP algorithm began to converge at the 18th generation, and the final fitness value was 0.61. The BP neural network began to converge in about 32 generations, and its final stability fitness value was 0.64. The iteration of GA algorithm was better than that of BP neural network. It started to converge after 28 iterations, and the final fitness value remained at 0.62. PSO-BP algorithm began to converge around 24 generations, and its fitness value was 0.63 when it converged to a stable state. Comparing the fitness values of four different algorithms, we could find that GA-BP algorithm could converge to the optimal fitness value as soon as possible. This showed that the algorithm has better optimization ability, and could better avoid the model falling into the local optimal solution when other three algorithms are compared with it.

Figure 4.3 showed four algorithms' error performance in the iteration process. Figure 4.2(a) showed the square sum of the average errors of the four algorithms. Figure 4.2(b) showed the sum of squares of the



((a)) Accuracy of four algorithm models in test set samples



((b)) Accuracy of four algorithm models in training set samples

Fig. 4.5: Prediction of different algorithms in test set and training set

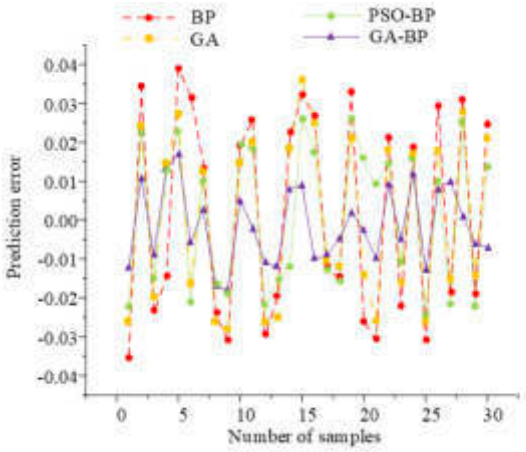


Fig. 4.6: Evaluation error results of different algorithms in training samples

minimum errors of the four algorithms. According to the iteration curves of the four algorithms, we could see that GA-BP can iterate to convergence state faster. The final average error square value and the minimum error square sum of the convergence state were 0.35 and 0.48, respectively. The error performance of the GA-BP algorithm was far better than the other algorithms, so the GA-BP algorithm had better global optimization ability at runtime.

Figure 4.4(a) showed the prediction of the four algorithms in the test set. Figure 4.4(b) showed the prediction of the four algorithms in the training set. When the samples number increased, the prediction accuracy of the four algorithms in the test set and the training set had increased and could eventually become stable. When the stable prediction accuracy was reached, the optimal prediction accuracy of BP, GA, PSO-BP and GA-BP in the test sample set was 74.6%, 80.5%, 85.1% and 93.4% respectively. The optimal prediction accuracy of BP, GA, PSO-BP and GA-BP in the training sample set was 75.2%, 80.8%, 86.2% and 94.1% respectively. Whether in the test set or in the training set, the prediction accuracy of GA-BP algorithm was higher.

Figure 4.6 showed the evaluation error performance of the four algorithms in training samples. The four algorithms could evaluate the given training samples, but the evaluation effect of GA-BP algorithm was far

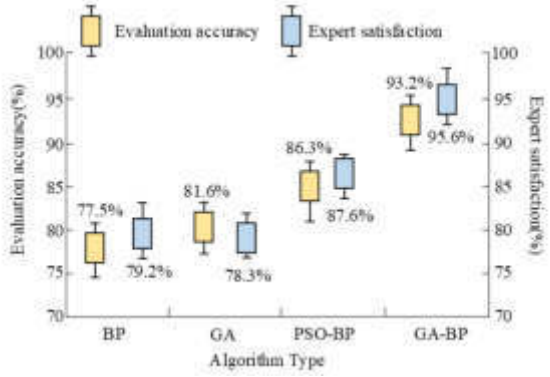


Fig. 4.7: Accuracy and student satisfaction of different algorithms in the evaluation model of PTM reform

better. The prediction error of GA-BP algorithm could be controlled within, so it had good prediction accuracy. The maximum prediction error of GA-BP algorithm was only 0.015. The prediction effect of BP neural network was the worst, its prediction error is within, and the maximum prediction error was 0.37. The prediction effect of GA algorithm was second only to BP neural network, its prediction error could be controlled within, and the maximum prediction error value was 0.035. Although the prediction effect of PSO-BP algorithm was better, its control of prediction error range is still insufficient. The prediction error of PSO-BP algorithm is within , and PSO-BP algorithm's highest prediction error is 0.024. According to the results in Figure 4.5and Figure 4.6, GA-BP algorithm had a good prediction effect in the sample data set. Next, it is used to evaluation model's construction of PTM reform, and the actual evaluation effect of different algorithms used in the PTM reform software engineering teaching is observed.

Figure 4.7 showed the accuracy rate and student satisfaction of the four algorithms in the evaluation model of PTM reform. Four algorithms were used in the evaluation model of PTM reform. According to the evaluation results, GA-BP algorithm had the best evaluation accuracy and could accurately evaluate the parameters and indicators that affected the teaching reform, with the evaluation accuracy of 93.2%. The evaluation accuracy of BP, GA and PSO-BP were 77.5%, 81.6% and 86.3% respectively. In addition, we also collected the satisfaction of software engineering students with the evaluation results of the four algorithm models. Student satisfaction could provide more effective ideas for the follow-up curriculum reform. The degree of satisfaction of software engineering students with BP, GA, PSO-BP and GA-BP was 79.2%, 78.3%, 87.6% and 95.6% respectively. According to the satisfaction results, students were most satisfied with the evaluation effect of using GA-BP algorithm in the evaluation model.

5. Conclusion. In view of a series of problems existing in the current teaching mode of software engineering course, the PTM is proposed to reform the software engineering course. And an effect evaluation model is built to verify reformed course's teaching effect. About BP neural network's defects, genetic algorithm is used to optimize it. And the optimized GA-BP algorithm is used in the index evaluation model to test the teaching method of the project and the reform effect of the software engineering course. To prove the performance of GA-BP algorithm and the effect of its application in the evaluation index model, a series of algorithm comparison experiments are set up and the comparison results are analyzed. Compared with BP, GA and PSO-BP, GA-BP has better convergence effect. When the algorithm is iterated to 18 generations, the fitness value of GA-BP algorithm reaches a stable value of 0.61. In addition, GA-BP algorithm has smaller average error square value and minimum error square sum value, which are 0.35 and 0.48 respectively. In the test set and training set, the prediction accuracy of the four algorithms is tested. The prediction accuracy of GA-BP algorithm is as high as 93.4% and 94.1%, and its prediction error range is also smaller, and the final control is within. Finally, four algorithms are used to evaluate the evaluation accuracy and student satisfaction of the test model in the evaluation model. The evaluation accuracy of BP, GA, PSO-BP and GA-BP are 77.5%, 81.6%, 86.3% and 93.2% respectively, and the student satisfaction is 79.2%, 78.3%, 87.6% and 95.6% respectively. The

above experimental results prove the role of project-based learning in software engineering curriculum reform. However, since the determination of evaluation indicators is often affected by many factors, there is still room for modification in the selection of indicators of the evaluation model.

REFERENCES

[1] Zhang, H. & Tian, Z. Failure analysis of corroded high-strength pipeline subject to hydrogen damage based on FEM and GA-BP neural network. *International Journal Of Hydrogen Energy*. **47**, 4741-4758 (2022)
[2] Rungsirisakun, R., Intarapong, P., Lekprasert, B. & Tantiwisawaruji, S. Using Project Based Learning in a Fundamental Chemistry Course: An Experience Report. *Psychology Research*. **9**, 6-10 (2019)
[3] Wang, S., Zhang, F., Gong, Q., Bolati, D. & Ding, J. Research on PBL teaching of immunology based on network teaching platform. *Procedia Computer Science*. **183**, 750-753 (2021)
[4] Wang, H., Li, J. & And, L. and weld forming control based on GA-BP algorithm for riveting-welding hybrid bonding between magnesium and CFRP. *Journal Of Manufacturing Processes*. **70**, 97-107 (2021)
[5] Wang, H., Zhang, Z. & And, L. and fitting of weld morphology of Al alloy-CFRP welding-rivet hybrid bonding joint based on GA-BP neural network. *Journal Of Manufacturing Processes*. **63**, 109-120 (2021)
[6] Wang, T., Liang, L. & Zheng, M. Application of formative evaluation and teaching feedback in PBL teaching of Medical Genetics. *Hereditas (Beijing)*. **42**, 810-816 (2020)
[7] Zak, A., Bugada, L., Ma, X. & Wen, F. Virtual versus In-Person Presentation as a Project Deliverable Differentially Impacts Student Engaged-Learning Outcomes in a Chemical Engineering Core Course. *Journal Of Chemical Education*. **98**, 1174-1181 (2021)
[8] Vasconcelos, M., Guedes, M., Melo, P., Amore, C. & Linares, J. An example of project-based learning with the support of a process simulator applied to the chemical engineering final course project. *Computer Applications In Engineering Education*. **30**, 490-504 (2022)
[9] Pasandín, A. & Pérez, I. Developing theory from practice: A case study in civil engineering airport design problem-based learning. *Computer Applications In Engineering Education*. **29**, 1112-1131 (2021)
[10] Gao, H., Wu, C., Huang, D., Zha, D. & Zhou, C. Prediction of fetal weight based on back propagation neural network optimized by genetic algorithm. *Mathematical Biosciences And Engineering: MBE*. **18**, 4402-4410 (2021)
[11] Zou, M., Xue, L., Gai, H., Dang, Z. & Xu, P. Identification of the shear parameters for lunar regolith based on a GA-BP neural network. *Journal Of Terramechanics*. **89**, 21-29 (2020)
[12] Wang, Y. & Qing, D. Model Predictive Control of Nonlinear System Based on GA-RBP Neural Network and Improved Gradient Descent Method. *Complexity*. **2021**, 491-495 (2021)
[13] Yan, C., Li, M. & Liu, W. Transformer Fault Diagnosis Based on BP-Adaboost and PNN Series Connection. *Mathematical Problems In Engineering*. **2019**, 341-351 (2019)
[14] Saleh, A., Chen, Y., Hmelo-Silver, C., Glazewski, K., Mott, B. & Lester, J. Coordinating scaffolds for collaborative inquiry in a game-based learning environment. *Journal Of Research In Science Teaching*. **57**, 1490-1518 (2020)
[15] Ueda, K. & Oguni, M. Reliability of the Phonon Density of States Determined by Real-Coded Genetic Algorithm from Heat Capacities of Benzoic Acid Crystals. *The Journal Of Physical Chemistry B*. **125**, 6322-6329 (2021)
[16] Fernandes, H. From student to tutor: A journey in problem-based learning. *Currents In Pharmacy Teaching And Learning*. **13**, 1706-1709 (2021)
[17] Chmelárová, Z. & Onková, A. Project Based Learning from the Point of View of Economics Students. *TEM Journal*. **10**, 832-838 (2021)
[18] Nasrabadi, A. & And, M. and optimization of a biosensor-based microfluidic microbial fuel cell using both genetic algorithm and neural network PSO. *International Journal Of Hydrogen Energy*. **47**, 4854-4867 (2022)
[19] Sevilgen, G., Bulut, E., Albak, E. & And, Z. and optimization of the design decisions of liquid cooling systems of battery modules using artificial neural networks. *International Journal Of Energy Research*. **46**, 7293-7308 (2022)
[20] Ham, Y., Lee, J. & Lee, S. Study on Evaluation in College Mathematics Education in the New Normal Era. *Communications Of Mathematical Education*. **34**, 421-437 (2020)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 15, 2023
Accepted: Sep 5, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 241–253, DOI 10.12694/scpe.v25i1.2290

CLUSTERING ALGORITHM IN DIGITAL MANAGEMENT AND SUSTAINABLE SYSTEM CONSTRUCTION FOR URBAN RAIL TRANSPORTATION STUDENT EDUCATION

YIJIA LI*

Abstract. With the rapid growth of the national economy, people's demand for transportation is becoming increasingly strong. The rail transit business is booming in large and medium-sized cities, and the education management of urban rail transit students needs further reform. At the same time, digital information technology is widely used in various fields, and digital management of education has become one of the major development directions of education reform. The study proposes a specific construction path based on the analysis of the necessity of digital management of education for urban rail transportation majors, and then optimizes the K-medoids algorithm in the clustering algorithm and validates its education digital management effect. The outcomes show that the clustering precision of the upgraded K-medoids algorithm in the selected dataset is up to 92.68%, and the running time is all below 5s, with the lowest value being 3.9s; In the digital management of urban rail transit majors in universities, the precision obtained by the algorithm is all maintained at around 95%, and the satisfaction rate is all higher than 90%. The effectiveness of the proposed method has been verified, providing a new method for the management of digital education systems for urban rail transit students. It can better understand the needs and characteristics of students, help improve their learning effectiveness and educational quality, and achieve more targeted allocation of educational resources.

Key words: Internet of Things; path optimization; time windows; dynamic demand; Clustering; K-medoids algorithm; Rail transportation; Education; Digital management cost

1. Introduction. With the acceleration of urbanization and the rapid growth of population, the construction and management of urban rail transit system is facing more and more challenges [1]. To better meet the travel needs of urban residents, improve transportation efficiency, and enhance the quality of students' education, digital management and sustainable system construction have become one of the important directions of urban rail transit development. As an important data mining technology, clustering algorithm has increasingly attracted attention in its application in digital management and sustainable system construction of urban rail transit student education. As an efficient, fast and environmentally friendly transportation mode, urban rail transit is of great significance to the travel needs of urban residents [2]. With the rapid development of information technology, digital management has become a trend in modern management. Cluster algorithm, as a commonly used data mining technique, can classify and analyze data by dividing it into groups with similar characteristics. In the application of digital management and sustainable system construction in urban rail transit student education, efficient processing technologies and methods are urgently needed due to the complexity of teaching content. At the same time, with the continuous expansion of urban rail transit system, its management is facing more and more challenges [3]. Digital management of urban rail transit education resources is an important aspect of urban rail transit management. In addition, the K-medoids algorithm in clustering algorithms has a simple principle, but tends to fall into local optima. Further improvement is needed to meet the requirements of digital education management [4]. Therefore, this study optimizes the K-medoids algorithm by introducing the Artificial Bee Colony Algorithm (ABC) algorithm and applies it to the digital management of urban students' education. Through the method proposed in this article, it is expected to use clustering algorithms to rationally allocate educational resources, improve resource utilization efficiency, promote sustainable system construction, and promote the sustainable development of urban rail student education.

*College of Railway Transportation Management and Economics, Hunan Vocational College of Railway Technology, Zhuzhou 412000, China (yj543127986@126.com)

2. Related works. Digital management of education is an inevitable trend in the development of information technology, and in recent years, breakthroughs have been made in numerous studies. Raimundo's team summarized the application of blockchain technology in higher education management. Blockchain has become an important concept at the intersection of ICT and higher education. The results show that blockchain technology is widely used in education to optimize the efficiency of educational data management, and improve the effectiveness and security of the system. It also poses certain challenges for future research directions [5]. Scholars like Mohamed have developed a qualitative model for the digital transformation of higher education and management. A decision support system is used to effectively generate and manage an index of the importance of student experience and learning expectations. The study achieves effective student management by developing a transformation roadmap for the strategic management of universities and the changes in influencing factors [6]. Williamson B suggests that the management of higher education programs is beginning to move towards online platforms and towards digital and data. The process found that the Pearson method has begun to become the definitive approach to higher education platforms. Digital management is gradually linking higher education with the benefits of digitalization [3].

Shaturaev J. has developed a causal model to mitigate the potential impact on economic and educational management as a result of the Fourth Industrial Revolution. Operating under this model, the sustainable development of economic and educational management becomes more directional. It is also recommended that economic and educational organizations adopt the model in a rational way in order to remain competitive. A reference for the subsequent educational management and development of students is provided [8]. Barakina E Y and other scholars conducted research to address the use of intelligent robots in educational management. The study found a significant relationship between artificial intelligence and the sustainability of educational management. The results show that educational management can be effective in training students with the help of intelligent methods and using the talents developed for the research of new intelligent technologies.

The factors that hinder the implementation of technology are analyzed and appropriate recommendations for the education sector are developed [10]. Syahputra Y H et al. applied the K-means clustering algorithm to classify top students to further improve their performance, in response to the difficulty of searching and processing student data. The results indicate that this method can provide schools with the information and solutions needed to classify and determine advantageous classes, thereby improving the academic performance of students in school [11]. In order to conduct Big data analysis on innovative talent education mode, Luo Y's team designed an innovative talent education mode based on segmented information fusion regression statistical analysis, and integrated educational resources through data mining and information processing. The results show that the quantitative evaluation accuracy of this method is high and the convergence is better [11].

The application of clustering algorithms in digital education management has gradually become a research hotspot in this field. The improvement of the algorithm provides necessary methods and technical references for the information management and sustainable system construction of urban rail transit students. Hu's team proposed a fuzzy-based clustering algorithm (FCAN) in order to better apply the clustering algorithm. The process was followed by combining fast fuzzy and clustering algorithms with each other as Fast Fuzzy Clustering Algorithm (F2-CAN) to solve the shortcomings of slow convergence of FCAN. Using five data sets, F2-CAN was found to be more efficient in terms of both convergence speed and clustering accuracy. It provides a reference for future solutions for large-scale complex industrial networks [12]. Researchers such as Bindhu V propose to incorporate artificial intelligence techniques to reduce the response time and cost of the system. In the course of the experiments, the use of subspace clustering is proposed to handle the connectivity and sparsity between factors. A comparison is made between IoT and traditional image-based techniques, as well as between the proposed methods, to verify the effectiveness and widespread use. The outcomes demonstrate that in the IoT environment, the research proposed method is not only effective in dealing with noise, but also that subspace clustering helps to find the desired optimal strategy [13].

Hassan B A et al. propose an architecture that utilizes evolutionary clustering algorithms in order to reduce the ambiguity of frame form contexts in order to address the problem of time-consuming generation of formal contexts in educational data corpora. The outcomes demonstrate that the quality of the semantic concept hierarchy under the proposed method of the study can be maintained at a stable 89% compared to the traditional concept lattice, and some simplification of the data is achieved. And the clustering algorithm

performs the concept lattice faster than other algorithms at different filling rates [14]. Chen et al. proposed a CNN-clustering algorithm-based model for image segmentation for license plate photo recognition to monitor the reasonable use of cars. The process also uses algorithms for localization and monitoring to optimize the detection accuracy. The process used the collected dataset for simulation experiments and the results show that the model proposed by the researcher outperformed all the traditional methods [15].

Shin D et al. experimentally applied the clustering algorithm to a mathematics education course to fully understand students' behavior and thought processes during the learning process and to automatically generate reports on students' performance on classroom assignments. The results show that the clustering algorithm performed well. And it is not limited to the field of mathematics education, largely provides a reference for future science education and indicates potential research directions [16].

In summary, the clustering algorithm has demonstrated good application in the processing of massive data, while most studies have upgraded it for its convergence speed problem, but there is less analysis on the stability and classification effect of this algorithm. The proposed K-medoids algorithm introduces the Gaussian similarity function, which increases the stability of the cluster center and makes it more robust, but it has the characteristic of strong randomness in selecting the centroid of the initial cluster. Therefore, the improved ABC algorithm is further introduced to achieve global point search, and the optimal solution is the initial centroid of the K-medoids algorithm. Meanwhile, the study uses the ABC algorithm to optimize the K-medoids algorithm in the clustering algorithm and validates it in the digital management of education for urban rail transportation majors, with a view to providing technical support for improving the learning effect of students in this major and improving the digital management system of education.

3. Digital management and sustainable system construction of urban rail transportation student education based on clustering algorithm. In this chapter, the digital management and sustainable system construction of urban rail transit student education are firstly carried out, and then the digital management of education is completed by improving the K-medoids algorithm.

3.1. Digital management and sustainable system construction for urban rail transportation student education. In mass transit education, the student is the main target and the focus is on learning. The digital management of education and the construction of sustainable systems must give priority to students [17]. The digital management of existing urban railway education is a necessity and a way to build a sustainable system. Firstly, digital technology has penetrated into many fields, including education. The educational management model of universities must comply with this progressive trend. In the process of improving overall level, providing higher quality management services can continuously meet one's own development needs [18]. Second, the need for high-quality development in universities has become increasingly important. As a transactional project, the education management of urban railway professionals, although centered on students, also affects all university staff. Taking advantage of digital technology, exploring and building advanced education management models and sustainable systems has also become one of the driving forces for sustainable school progress and students' professional empowerment. Finally, as urban railway students, they must have the ability to adapt to new technologies and create a convenient and effective learning environment with the help of digital media platforms for teaching and learning to achieve better learning outcomes. The necessity and impact relationship between digital management and sustainable system construction for urban railway students' education is shown in Figure 3.1.

In Figure 3.1, the necessity of digital management of urban rail transit student education and sustainable system construction is student-centered. Under the two dotted lines centered on students, there are mainly two objective requirements, namely the penetration of digital technology and the need for high-quality development of universities. Under the urgent needs of these two aspects, a relationship diagram within the dashed box in the figure has been formed. That is to say, the development of urban rail transit vocational education and the informatization of education management are two parts, which jointly respond to the development trend of digital education management. Sharing, openness and consistency are the three principles that must be followed in the student-centred digital management of education in the urban rail transport profession. Sharing requires that the digital education management system of this profession must take the whole school education management as the goal to achieve unified management and supervision. All functional modules in the system can provide and share information in real time to optimize the effectiveness and scientific nature of education

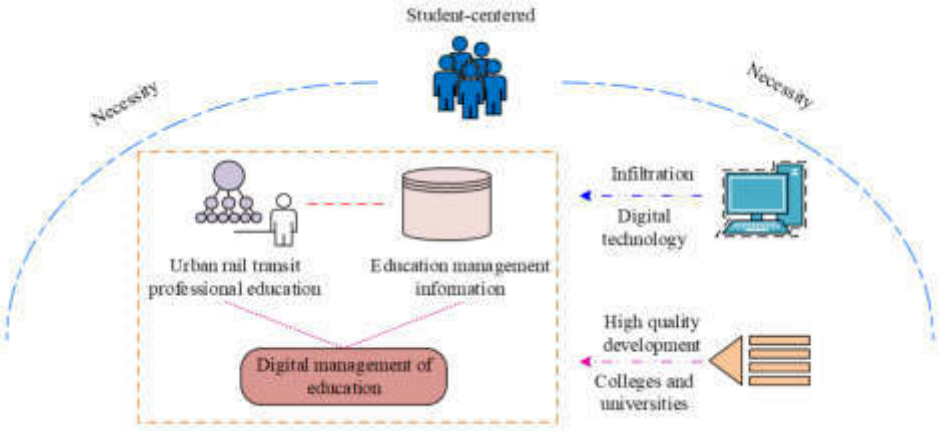


Fig. 3.1: The necessity and influence of digital management and sustainable system construction of urban rail transit students' education

management, and data security is guaranteed by identity verification and database encryption. To facilitate the construction of an education management platform and sustainable system, multiple open information modules are set up to achieve openness in a variety of integrated functions such as student outcomes enquiry, handling of various affairs and course scheduling, i.e. the principle of openness [19]. The principle of consistency requires standardized standards for the construction of educational digital platforms. By standardizing and ensuring the uniformity of educational management processes, and naming the information fields such as courses, teachers, and students in urban rail transit in a unique way, the relevance and accuracy of educational management are optimized. According to these three principles, to build a digital management and sustainable development system for urban rail transit professional education, it is first necessary to innovate digital technology in management methods. It includes scheduling professional courses through intelligent scheduling software, and integrated publishing of major events through WeChat official account. Secondly, to build a sharing platform for urban rail transportation teaching management, through digital teaching management, to achieve the sharing of quality education resources among teachers and students. The platform will also be maintained and upgraded to fully grasp the laws of teaching and coordinate the management of various departments, thus improving the decision-making level of education managers. In addition, the evaluation system of urban rail transit education management is further upgraded through multimedia technology. With the support of new media technology, education management evaluation can include multiple aspects of the subject, and the evaluation content is more social and diversified, providing a platform for urban rail transit students to showcase [20] . The path to build a digital management and sustainable system for urban rail transit student education is demonstrated in Figure 3.4.

3.2. Digital management of education based on the K-medoids algorithm. The K-medoids algorithm is based on the K-means algorithm, which uses random initialization to select the reference points for clustering [21]. Unlike the K-means algorithm, which uses the mean of the current cluster samples as the representative object, the K-medoids algorithm chooses the actual object as the representative of the cluster, thus reducing the degree to which the clustering effect is influenced by edge outliers. At the same time, the K-medoids algorithm optimizes the clustering effect by moving the samples between the division blocks, with objects contained within the same class cluster being highly similar and objects contained between different class clusters being significantly different. The K-medoids algorithm first selects a class cluster that represents the class of an actual object, and for the remaining objects, it divides them into the corresponding class clusters according to the Euclidean distance between sample points. Finally, clustering is achieved by iterating over other sample points instead of centroids, and the clustering results are compared for every two iterations,

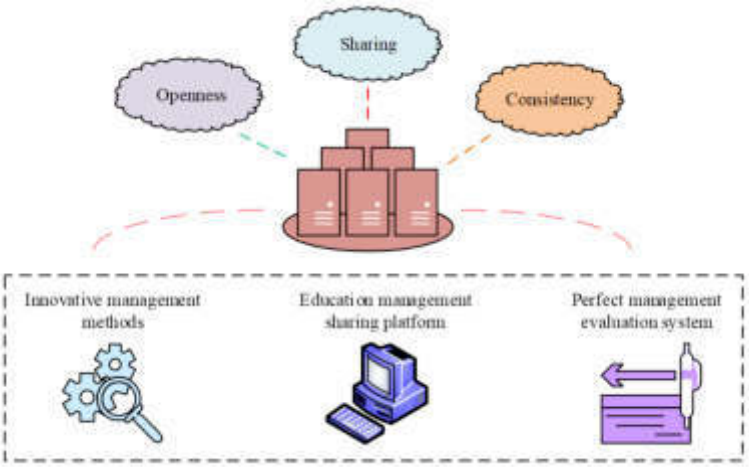


Fig. 3.2: Digital management and sustainable system construction path of urban rail transit students' education

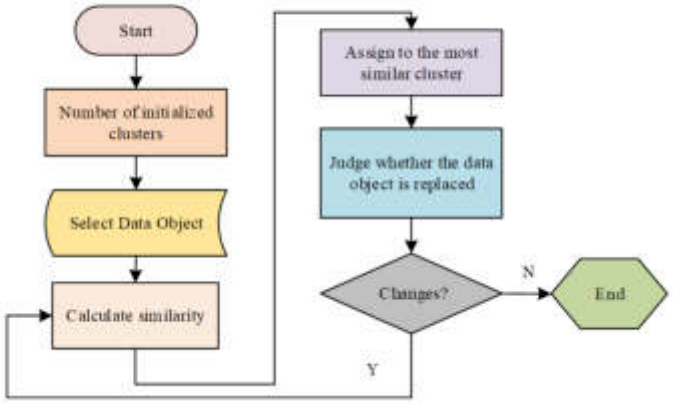


Fig. 3.3: Basic flow of K-medoids algorithm

keeping the best result as output [22]. The basic flow of the K-medoids algorithm is shown in Figure 3.3.

The spatially true distance of two sample points in the K-medoids algorithm is calculated as demonstrated in equation 3.1.

$$d(x_i, c_j) = \sqrt{\sum_{a=1}^{m} (x_{ai} - c_{aj})^2}, \quad j = 1, 2, 3...n; \quad i = 1, 2, 3, ...n \tag{3.1}$$

In equation 3.1, x_i and C_J are data objects, m represents the feature dimension, and ^{n} represents the first n data object. The mass of the cluster centroids is calculated by summing the squares of the errors, as demonstrated in equation 3.1.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon c_j} d(p, c_j) \tag{3.2}$$

In equation 3.2, E denotes the data set,C_J denotes the class clusters.P is the object,and k denotes the number

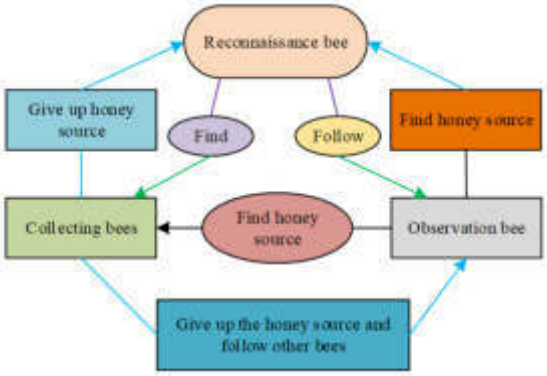


Fig. 3.4: Relationship diagram of bee species transformation of ABC algorithm

of class clusters into which the data set is divided. The K-medoids algorithm evaluates whether the quality of the clusters is optimized by a cost function as shown in equation 3.3.

$$E = E_2 - E_1 \tag{3.3}$$

In equation 3.3, E_2 is the sum of the absolute error values of all representative points in the replacement centroid data set and the new centroids of the class clusters. S is the total difference before and after replacement, and E_1 is the sum of the absolute error values of all representative points and the centroids of the class clusters before replacement is performed. If S is less than 0, the old centroid is replaced with the new centroid and the rest of the data samples are reclassified into the class cluster represented by the nearest centroid. If S is greater than or equal to 0, the current centroids are maintained. Although the K-medoids algorithm reduces the effect of noisy points and outliers on the clustering results, it still needs to mitigate the damage to the results caused by edge outliers [23]. The Gaussian similarity function can find the similarity of the data samples and increase the stability of the class cluster centers. Therefore, the study introduces Gaussian similarity to optimize the objective function of this algorithm, as shown in equation 3.4.

$$d(x_i, c_j) = e\frac{-\|x_i - x_j\|^2}{\sigma^2} \tag{3.4}$$

In equation 3.4, σ denotes the Gaussian kernel function parameter. After improving the objective function of the K-medoids algorithm by using Gaussian similarity, which is more robust in the face of isolated points, and considering its feature of random initial selection of initial clustering centroids, the study further optimizes it by means of the ABC algorithm. The ABC algorithm uses honey sources to represent the possible solutions in the sample solution space, and reflects the superiority of solutions by the good or bad honey sources, and in its description of the honey harvesting process, the bee species. The conversion is demonstrated in Figure 3.4.

In the starting phase of the ABC algorithm, the bees do not have any information about the sample, and a number of honey sources are generated by the formula demonstrated in equation 3.5.

$$X_i j = (X^j_{max}) - (X^j_{min}) rand(0,1) + (X^j_{min}) \tag{3.5}$$

In equation 3.5, j represents a dimension and belongs to the D dimension.X^j_{max} is the upper limit of the searchable solution space.X^j_{min} is the lower limit, and $X_i j$ represents the randomly obtained feasible solution space. The nectar volume of the nectar source is actually the relative size of the fitness value, which is calculated according to equation 3.6

$$fitness_i = \left\{ \begin{matrix} |fi|, f_1 < 0 \\ \frac{1}{f_i+1}, f_i \geq 0 \end{matrix} \right\} \tag{3.6}$$

Once the bee species transformation is complete, a neighbourhood search is performed as shown in equation 3.7.

$$V_ij = rand()(X_{ij} - Xkj), i \neq k \tag{3.7}$$

In equation 3.7, V_ij represents the new nectar source found around the original source X_ij k=1,2.....N_e, $j = \{1, 2, \ldots, D\}$, i and k are unequal and are obtained in a random way.rand() is a randomly selected value in [0,-1] and X_{kj} represents the nectar source location with the index value k . It is then evaluated whether the new nectar location is better than the old one, as shown in equation 3.8.

$$V_{ij}^{'} = \begin{Bmatrix} X_{ij}, f_v > f_x \\ V_{ij}, f_v \leq f_x \end{Bmatrix} \tag{3.8}$$

The observer bee decides whether to follow or not based on the fitness value passed by the nectar collecting bee, calculated by the probability obtained from equation 3.9.

$$p_i = \frac{fitness}{\sum_i fitness_i} \tag{3.9}$$

In equation 3.9, p_i represents the fitness value. In case the total number of neighborhood searches performed by any of the observation and honey collecting bees is greater than a limited number and the current nectar source location is satisfied, the honey collection at the current location is stopped and the original bee species is transformed into a scout bee, at which point the new nectar source location is found by equation 3.10.

$$X_i(n) = rand(0, 1)(X_{max} - X_{min})Basi \geq Limit + X_{min} \tag{3.10}$$

In equation 3.10,$Basi$ represents the total number of near-region searches, and $Limit$ is the finite number. The ABC algorithm runs to completion and outputs the global optimal solution when the swarm search combines other stopping criteria or reaches the maximum number of iterations. Although the ABC algorithm is simple and easy to implement with few parameters, it has obvious shortcomings such as easy premature termination and slow convergence in the later stages. Tent chaos mapping results in a flat and uniform distribution of mapping values, which can enhance swarm diversity. The IABC algorithm is based on the tent chaos mapping to obtain the initial nectar source, as shown in equation 3.11.

$$y_{i,j+1} = \begin{Bmatrix} \frac{1-x_{ij}}{1-u}, u \leq y_{ij} \leq 1 \\ \frac{y_{ij}}{u}, 0 \leq y_{i,j+1} < u \end{Bmatrix} \tag{3.11}$$

In equation 3.11, i represents the population size number.j represents the chaos number. u is the chaos parameter in the range of [0,2], and is the random number in [0,1]. The outcomeing population initialization formula for the IABC algorithm is demonstrated in equation (12).

$$X_{ij} = y_{i,j+1} (X_{max}^{j} - X_{min}^{j}) + X_{min}^{j} \tag{3.12}$$

In Eq. (12), X_{ij} denotes the randomly obtained feasible solution space. The IABC algorithm performs a binomial crossover of the global optimum with the new solution obtained from the honey bee neighbourhood search, as shown in Eq. 3.13

$$V_{ij} = \begin{Bmatrix} v_{ij,rand<cr} \\ x_j^{Global}, other \end{Bmatrix} \tag{3.13}$$

In equation 3.13, X_J^{Global} is the global optimum factor. The swarm is also constructed by adding a term to the equation, and the resulting position update equation is given by equation 3.14.

$$V_{ij} = \begin{Bmatrix} v_{ij,rand<cr} \\ x_j^{Global} + (x_j^{Global} - v_{ij}), other \end{Bmatrix} \tag{3.14}$$

Table 3.1: Software and Parameter Settings for test of IABC algorithm and ABC algorithm

Set Item	Specific Situation
Population Size	20
Maximum Number of Searches	100
Maximum Number of Iterations of Program	2000
Collecting Bees	10
Observation Bees	10
Experimental Software	MATLAB 2016

In equation 3.14, cr is the coefficient, whose main role is to coordinate the exploration and development capability of the algorithm. Finally, the K-medoids algorithm is fused with the IABC algorithm to obtain the final clustering algorithm as the IABCK-medoids algorithm. The IABCK-medoids algorithm first performs a global merit search on the dataset by utilizing the merit search advantage of the IABC algorithm, and the resulting optimal solution is the initial centroid of the K-medoids algorithm. The clustering results of the K-medoids algorithm are then passed to the IABC algorithm, which updates the swarm and iteratively updates it to achieve optimal clustering. The objective function of the algorithm is the standard absolute error formula, which is also used as a method to compute the fitness value, as shown in equation 3.15

$$fitness_i = E \tag{3.15}$$

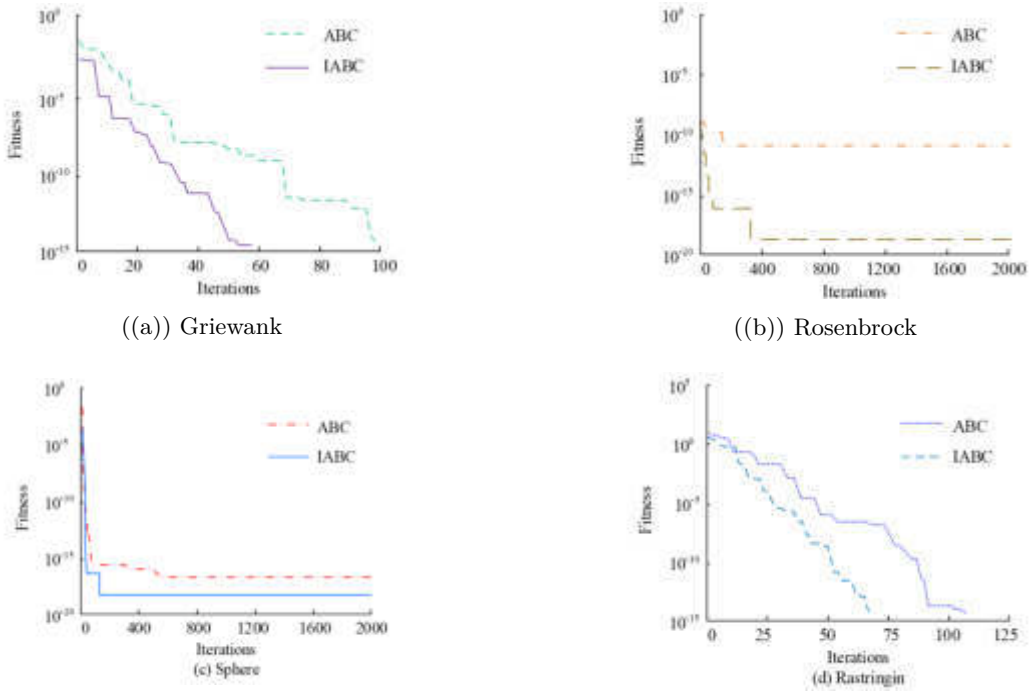
Finally, the IABCK-medoids algorithm is applied to the digital management of urban rail transportation student education, clustering student information and teaching evaluation, so as to achieve efficient management and sustainable system construction.

3.3. Improving the effectiveness of the K-medoids algorithm in the digital management of education. This chapter mainly tests the performance of the improved K-medoids algorithm, compares it with the traditional K-medoids algorithm and other algorithms, and finally verifies its practical application effect in the digital education management of urban rail transit students. The study uses the IABC algorithm to optimize the K-medoids algorithm to obtain the IABCK-medoids algorithm and apply it to the digital management and sustainable system construction of urban rail transit students' education. To analyze the application effect of this algorithm, the performance of the IABC algorithm is first verified. The IABC algorithm was compared with the ABC algorithm. To avoid overfitting and underfitting problems, the parameters were set with reference to relevant data and previous experience. The experimental software and associated parameters were set as shown in Table 3.1.

Four test functions, Sphere, Rastringin, Griewank and Rosenbrock, were selected to experiment with the IABC algorithm and the ABC algorithm, and the outcomes obtained are demonstrated in Figure 3.6. In Figure 3.6, subplots (a), (b), (c) and (d) correspond to the processing outcomes of the two algorithms in the Griewank, Rosenbrock, Sphere and Rastringin test functions, respectively. From Figure 3.6 (a) and (d), the IABC algorithm has fewer iterations and a better starting point for finding the optimum in both the Griewank and Rastringin test functions, and both achieve the optimum value. In Figure 3.6(b) and (c), for the Rosenbrock test function, which has a smaller fitness value, the convergence values of the ABC and IABC algorithms are 10-11 and 10-19, respectively; For the sphere test function, the IABC algorithm converges in about 130 iterations with a fitness value of about 10-18, while the ABC algorithm converges in about 540 iterations. Overall, the IABC algorithm is faster, can jump out of local extremes, and has a stronger search capability.

The performance of the IABCK-medoids algorithm is then verified by comparing it with the ABCK-medoids algorithm and the K-medoids algorithm. The experimental configuration is a Windows 10 64-bit operating system with a central processor of Intel Core i5-10600KF @ 4.10GHz and 8G RAM. The selected datasets were Iris, Wine, Glass, and Segmentation, and the details of the four datasets are shown in Table 3.2.

The clustering precision statistics for the three algorithms on the selected datasets are demonstrated in Figure 3.8. Figure 3.8 (a), (b) and (c) represent the clustering precision outcomes of the K-medoids, ABCK-medoids and IABCK-medoids algorithms in the four datasets, respectively. From Figure 3.8(a), the precision



((a)) Griewank

((b)) Rosenbrock

((c)) Sphere

((d)) Rastringin

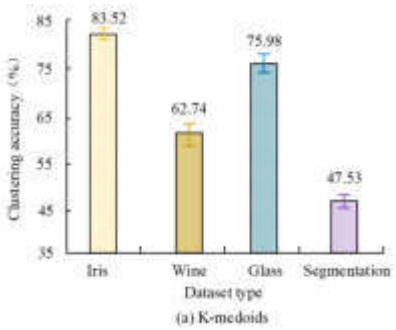
Fig. 3.6: Performance test outcomes of IABC algorithm and ABC algorithm in Sphere, Rastring, Griewink and Rosenbrook test functions

Table 3.2: Details of Iris, Wine, Glass, and Segmentation Datasets

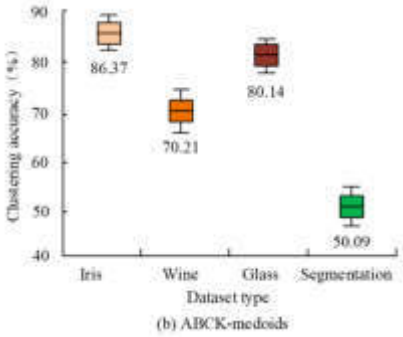
Dataset name	Number of clusters	Attribute dimension quantity	Number of objects
Iris	3	4	150
Segmentation	7	19	2310
Wine	3	13	178
Glass	6	9	214

of the K-medoids algorithm is 83.52%, 62.74%, 75.98% and 47.53% in the Iris, Wine, Glass and Segmentation datasets respectively, and the clustering precision in the dataset Segmentation is significantly lower. From Figure 3.8(b), the ABCK-medoids algorithm has the highest precision of 86.37% in the Iris dataset and the next highest precision of 80.14% in the Glass dataset. From Figure 3.8(c), the proposed IABCK-medoids algorithm achieves a clustering precision of 92.68% in the Iris dataset and still achieves a precision of over 70% in the Segmentation dataset. The comparison shows that the clustering precision of the IABCK-medoids algorithm in the Segmentation dataset is higher than the other two algorithms, with a maximum of 27.32%, and two datasets reach more than 90%, which is a clear advantage.

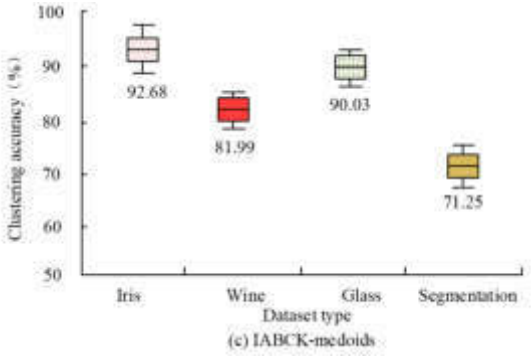
A comparison of the running times of the three algorithms on the selected datasets is shown in Figure 3.10. From Figure 3.10, the running time of the K-medoids algorithm is shorter in all four datasets, with the shortest in Iris at 1.4s and the highest in the Segmentation dataset at 2.4s. The running time of the ABCK-medoids



((a)) K-medoids



((b)) ABCK-medoids

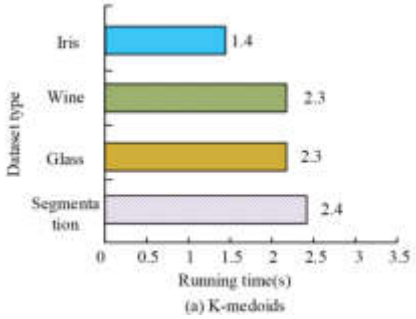


((c)) IABCK-medoids

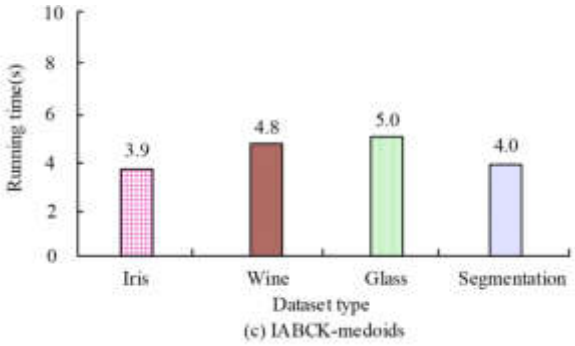
Fig. 3.8: Statistical outcomes of clustering precision of the three algorithms in the selected data set

algorithm is longer in all cases, with the highest at 9.1s and all over 7s. IABCK-medoids the IABCK-medoids algorithm has introduced the tent chaos mapping to improve the swarm diversity and the global optimality factor to achieve faster convergence of the algorithm, which inevitably prolongs the running time to some extent. However, it is still lower than the ABCK-medoids algorithm, and the accuracy is greatly improved compared to the K-medoids algorithm, with better clustering results and performance.

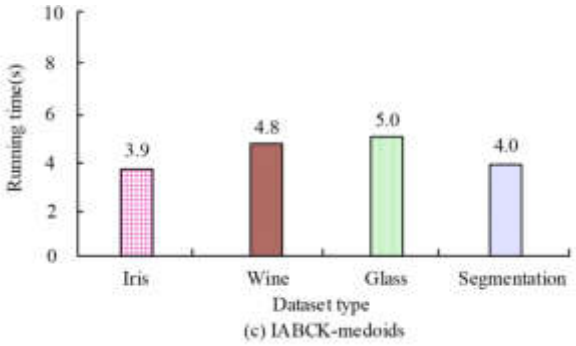
Finally, the method was applied to the digital management of urban rail transit students in a university, and the clustering effect of the three algorithms was evaluated in four aspects: student information, teaching data, course arrangement and grade management, and the usage evaluation of the four subjects: teachers, students, administrators and experts was statistically evaluated, and the obtained outcomes are demonstrated in Figure 3.12. Figure 3.12(a) demonstrates the comparison of the clustering precision of the three methods, and Figure 3.12(b) indicates the usage evaluation of the four subjects. From Figure 3.12(a), the precision of the K-medoids algorithm was generally in the range of 75% to 85%, with a minimum of 78% and a maximum of approximately 83%. The ABCK-medoids algorithm was generally in the range of 80% to 90%, with a maximum and minimum of 90% and 84%, respectively. The IABCK-medoids algorithm, on the other hand, fluctuated mostly around 95%, with a maximum of 97% and a minimum of over 90%. From Figure 3.12(b), the K-medoids algorithm is only slightly more satisfied than the AABCK-medoids algorithm in one of the four subject evaluations, while the other three are lower than the other two methods. The AABCK-medoids algorithm has a higher satisfaction rate, up to 90% in the teacher's evaluation, with an average satisfaction rate of around 85%. The IABCK-medoids algorithm, on the other hand, remained above 90%, with the highest rating of up to 96% for managers, having a better application.



((a)) K-medoids

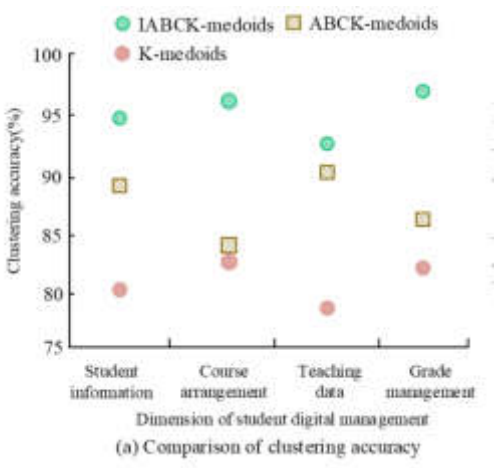


((b)) ABCK-medoids

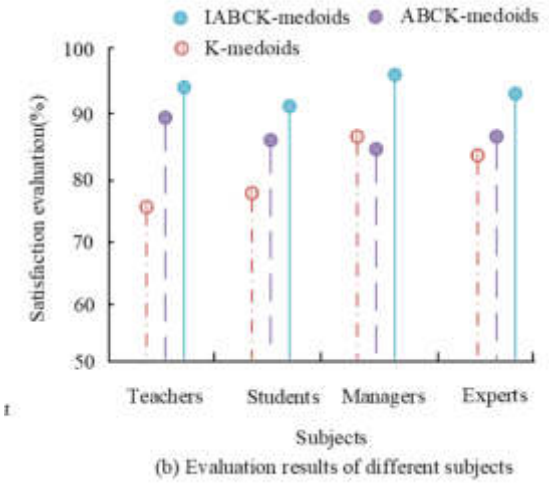


((c)) IABCK-medoids

Fig. 3.10: Comparison of the running time of the three algorithms in the selected dataset



((a)) Comparison of clustering accuracy



((b)) Evaluation results of different subjects

Fig. 3.12: Three methods of educational digital clustering outcomes and satisfaction evaluation

4. Conclusion. The rapid development of urban rail transit scale has put forward higher requirements for the urban rail transit profession. The digital management of student education in this major has become one of the important means of talent cultivation in the rail transit industry. The study takes the urban railway students as an educational subject and uses information technology to achieve digital management; Then, the K-medoids algorithm is further optimized by improving the ABC algorithm to form the IABCK-medoids algorithm, and the algorithm is used to digitally manage education. The results show that the improved ABC algorithm converges to the Sphere test function at the 130th iteration, with an adaptive value of approximately 10-18, and converges to the Rosenbrook test function, demonstrating strong optimization ability; The IABCK medoids algorithm has a minimum accuracy of over 70% on Iris, Wine, Glass, and Segmentation datasets, which is 27.32% and 96% higher than the K-medoids and ABCK medoids algorithms. It can optimize the digital management of urban rail transit students. In conclusion, the method proposed in this study can help educational administrators to better understand the needs and characteristics of students in order to formulate targeted educational strategies. At the same time, it can help educational administrators to allocate resources rationally and improve the efficiency of resource utilization. In terms of theoretical significance, the proposed method can enrich the application scenarios of data mining in the field of student education and promote the development of data mining in the field of education. At the same time, it can help optimize the distribution and utilization of educational resources for urban rail transit students, thereby improving the sustainable development of the education system and providing theoretical support for the future development of urban rail transit student education. However, clustering algorithms need to rely on a large amount of student data and relevant feature information. If the amount of data is small or the feature information is not comprehensive, the effect of the clustering algorithm may be affected. At the same time, the clustering algorithm is sensitive to the selection of initial parameters, and different parameter settings may lead to different clustering results. Therefore, from the perspective of digital management of urban railway student education and building a sustainable system, further research will be conducted on how to combine other machine learning algorithms with clustering algorithms to improve the accuracy of student classification and resource allocation. At the same time, we can study how to use clustering algorithms to predict students' learning needs and behaviors to provide more personalized and accurate learning services. In addition, in the future, it is necessary to study how to apply clustering algorithms to optimize the operational efficiency of the urban railway student education digital management system, for example, by optimizing student grouping and resource allocation to improve the overall educational quality of schools.

5. Fundings. The research is supported by: Hunan University Student Innovation and Entrepreneurship Training Program Project; "Poetic Herbal" Humanities and Nature Notes, China (S202110541003X).

REFERENCES

[1] Decuypere, M., Grimaldi, E. & Introduction, L. Critical studies of digital education platforms. *Critical Studies In Education*. **62**, 1-16 (2021)
[2] Hakimi, L., Eynon, R. & Murphy, V. The ethics of using digital trace data in education: A thematic review of the research landscape. *Review Of Educational Research*. **91**, 671-717 (2021)
[3] Williamson, B. Education technology seizes a pandemic opening. *Current History*. **120**, 15-20 (2021)
[4] Ytre-Arne, B. & Moe, H. Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society*. **43**, 807-824 (2021)
[5] Raimundo, R. & Rosário, A. Blockchain system in the higher education. *European Journal Of Investigation In Health, Psychology And Education*. **11**, 276-293 (2021)
[6] Mohamed Hashim, M., Tlemsani, I. & Matthews, R. Higher education strategy in digital transformation. *Education And Information Technologies*. **27**, 3171-3195 (2022)
[7] Williamson, B. Making markets through digital platforms: Pearson, edu-business, and the (e) valuation of higher education. *Critical Studies In Education*. **62**, 50-66 (2021)
[8] And, S. and Management as A Result of The Fourth Industrial Revolution: An Education Perspective. *Indonesian Journal Of Educational Research And Technology*. **3**, 51-58 (2022)
[9] Barakina, E., Popova, A., Gorokhova, S., Technologies, V. & Education, A. European Journal of Contemporary Education. (2021)
[10] Syahputra, Y. & Hutagalung, J. Superior class to improve student achievement using the K-means algorithm. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*. **7**, 891-899 (2022)

[11] Luo, Y. & An, Z. Research on self-learning system with "Internet+ Education" innovative talents education mode under big data background. *Computer Applications In Engineering Education*. **31**, 662-675 (2023)
[12] Hu, L., Pan, X., Tang, Z. & Luo, X. fast fuzzy clustering algorithm for complex networks via a generalized momentum method IEEE Transactions on Fuzzy Systems. (2021)
[13] Bindhu, V. & Ranganathan, G. Hyperspectral image processing in internet of things model using clustering algorithm. *Journal Of ISMAC*. **3** pp. 02 (2021)
[14] Hassan, B., Rashid, T. & Mirjalili, S. Formal context reduction in deriving concept hierarchies from corpora using adaptive evolutionary clustering algorithm star. *Complex & Intelligent Systems*. **7**, 2383-2398 (2021)
[15] Chen, J. & Zong, J. Automatic vehicle license plate detection using K-means clustering algorithm and CNN. *Journal Of Electrical Engineering And Automation*. **3**, 15-23 (2021)
[16] Shin, D. & Shim, J. systematic review on data mining for mathematics and science education. *International Journal Of Science And Mathematics Education*. **19**, 639-659 (2021)
[17] Gao, P., Li, J. & Liu, S. An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. *Mobile Networks And Applications*. **26**, 2123-2126 (2021)
[18] Zaring, O., Gifford, E. & McKelvey, M. Strategic choices in the design of entrepreneurship education: an explorative study of Swedish higher education institutions. *Studies In Higher Education*. **46**, 343-358 (2021)
[19] Zhou, M., Dong, H., Zhao, Y., Ioannou, P. & Wang, F. Optimization of crowd evacuation with leaders in urban rail transit stations. *IEEE Transactions On Intelligent Transportation Systems*. **20**, 4476-4487 (2019)
[20] Yang, J. Research on Group Cooperative Learning Method Teaching Based on Urban Rail Transit Comprehensive Training Course. *International Journal Of Social Science And Education Research*. **4**, 202-206 (2021)
[21] Mashrabovich, M. The role of digital technologies in improving the quality of higher education. *ACADEMICIA: An International Multidisciplinary Research Journal*. **12**, 23-26 (2022)
[22] Dafir, Z., Lamari, Y. & Slaoui, S. survey on parallel clustering algorithms for big data. *Artificial Intelligence Review*. **54**, 2411-2443 (2021)
[23] Majhi, S. Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. *Evolutionary Intelligence*. **14**, 35-46 (2021)
[24] Bhattacharjee, P. & Mitra, P. survey of density based clustering algorithms. *Frontiers Of Computer Science*. **15**, 1-27 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Sep 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 255–270, DOI 10.12694/scpe.v25i1.2292

ENGLISH DISTANCE TEACHING BASED ON SPOC CLASSROOM AND ONLINE MIXED TEACHING MODE

MEIJUAN ZHANG* AND XIAOLI ZHU†

Abstract. At present, the English Small Private Online Course (SPOC) online mixed teaching model has problems in evaluating students' learning and organizing teaching papers. For example, the evaluation is chaotic and unable to meet the key points of organizing the paper. Starting from the thinking chain that accepting learning outcomes can promote learning behavior, a score prediction method and test paper generation algorithm (TPGA) based on a learning evaluation diagnostic model are designed. Among them, the performance prediction algorithm is designed by combining multiple linear regression (MLR) and random forest (RF). The TPGA is based on students' learning status. The research results show that most of the predicted values output by the performance prediction model are not significantly different from the actual values. They are within a reasonable range. Meanwhile, under the influence of TPGA, the number of students in the experimental group is higher in the 70-80 and 80-90 segments, with 27 and 6, respectively. The experimental group has a higher average score rate on each type of question and knowledge point. Both models have high student satisfaction, indicating that the results oriented online mixed learning strategy designed in the study can effectively improve students' learning outcomes.

Key words: SPOC classroom; English online teaching; Teaching system design; Mixed teaching

1. Introduction. With the development of computer and network technology, the limitations of traditional classroom English education are gradually becoming apparent. These limitations are mainly reflected in the time and space, as well as significant limitations on students and classroom forms [1, 2, 3, 4]. However, English distance teaching based on online teaching technology and computer technology can solve the limitations of traditional classroom English education. This educational approach not only solves time and physical limitations, making teaching more flexible, but also makes the teaching process more intelligent, breaking away from the manual handover part of traditional classroom teaching [5, 6]. However, online English teaching also has certain drawbacks. For example, due to the main use of online video connections in education, it is more difficulty to grasp students' classroom learning status. The learning effect is unstable. The hybrid mode of online teaching solves this problem. This method redeploys online learning activities through activity and resource oriented, forming a more stable and flexible teaching system [7]. SPOC (Small Private Online Course) classroom refers to a small-scale restricted online classroom, which means that students participating in classroom education need to use online learning and must be restricted by classroom access conditions [8]. Only students who meet the requirements can be allowed to enter the classroom. The main limitations of SPOC teaching methods include group work, classroom cohesion, and acceptance of learning achievements. Therefore, by combining SPOC teaching with online mixed teaching mode, this study mainly focuses on evaluating the learning outcomes of students in teaching [9]. Then a teaching achievement evaluation system is established to improve the effectiveness of SPOC online mixed teaching system. The main motivation and novelty of this study lies in the focus not only on the advantages of online teaching, but also on the existing problems. It is attempted to solve these problems through the combination of hybrid mode and SPOC teaching. The contribution lies in proposing a new teaching achievement evaluation system to improve the effectiveness of SPOC online blended teaching system. By deeply understanding students' learning processes and outcomes, teaching effectiveness can be truly improved [10]. Therefore, this special method is chosen. The first part of the study introduces research purposes. The second part designs a learning evaluation model and paper formation strategy based on SPOC classroom theory. The third part conducts experimental verification on the learning

*School of Foreign Languages, Jiangsu Open University, Nanjing, 210036, China (Corresponding author: Meijuan_Zhang2023@outlook.com)
†Department of Student Affairs Management, Nanjing Institute of Technology, Nanjing, 211167, China

evaluation model and paper formation strategy research based on SPOC classroom theory. The fourth part draws research conclusions.

2. Related Work. In recent years, the research on SPOC classroom has been gradually enriched. Aiming at the drawbacks of cramming teaching in traditional college English translation courses, Sun H proposed a SPOC teaching method. It is named "Rain Classroom", which is beneficial for improving classroom learning efficiency. This teaching mode integrates the WeChat platform with other multimedia intelligent teaching tools. The learning process is divided into three steps, namely before, during and after class. The study findings revealed that the method has a positive effect on the reform of English translation teaching [10]. Zhang's team designed a teaching method based on the concept of conclusion-based interpretation and SPOC flipped classroom. This model follows the flipped classroom journey of teaching objectives-teaching activities-learning evaluation. A two-year experiment is carried out in the teaching of engineering cost specialty. According to the findings, the teaching model can improve learning efficiency [8]. Law et al. designed a SPOC flipped classroom teaching mode for private teaching. This mode fully utilizes online teaching mode, video teaching, layered teaching, interactive discussion and other learning forms. They are integrated into a complete teaching framework. Under the influence of this teaching mode, students' total grades are positively correlated with their classroom participation [12]. The team established a hierarchical learning method based on the SPOC flipped classroom. The model is divided into three main parts, namely the cognitive layer, the design layer and the application layer. For the teaching, the teaching form centered on activities and resources is adopted. A classification and grading evaluation method is used in evaluating teaching effectiveness. Compared with the traditional teaching mode, students are more satisfied with the improved teaching mode [13]. The Kastrati team proposed an exploratory model for the value of customized machine learning SPOC education models. Students' cognition and attitudes towards the SPOC teaching model are analyzed. Study findings demonstrate that the SPOC teaching mode more directly affects the students' knowledge and skills [14].

Computer technology is being used more wisely in classroom planning. The Gitinabard team designed a corresponding student performance prediction model for the hybrid online-offline teaching approach. The method combined the performance of students in offline classrooms with the evaluation results of online classrooms to form early predictions. The research results show that the method achieves stable cross-category data prediction [15] . Liu Q's team proposed an intelligent education evaluation model for the performance prediction of students in the physical education. The model is mainly based on the LSTM model. It tracks the students' exercise activities and extracts sufficient information from the exercise activity to predict performance. The results show that the model is effective [16]. Chen's team proposed an automatic TPGA based on genetic algorithm. It adjusts the difficulty factor according to the online learning data. The research results show that the model can effectively control the difficulty of the test paper [17]. Computers are more inclined towards intelligent improvement of traditional teaching procedures in teaching. Therefore, the research also conducts a model for teaching evaluation and test-taking strategies. A one-stop automated online teaching system is designed, providing a new realization path for SPOC online blended teaching.

The Soufiane O team mainly explored the three basic concepts of SPOC, blended learning, and flipped classroom. A working method for developing online training in the form of SPOC is proposed. This study mainly focused on issues such as student evaluation and teaching paper organization in the online blended teaching mode. A score prediction method and test paper generation algorithm based on a learning evaluation diagnostic model are designed. Therefore, this study has innovation in practical application and design of practical methods [18]. Zhou Y focused on the implementation of SPOC blended learning mode in the choir command teaching system. Although all research is focused on the SPOC blended learning model, this study is to address student learning evaluation and organization of teaching papers. Reference 2 focuses on how to improve the effectiveness of practical teaching in choir conducting. The characteristic of this study is to focus on the student learning evaluation and generate solutions through machine learning prediction [19]. The Ram í rez Monoso team studied a gamified mobile collaboration tool to improve the utilization of online learning resources. It is based on a learning evaluation diagnostic model, which uses complex machine learning algorithms to predict students' learning outcomes and organize teaching papers. Both are efforts to improve students' learning outcomes. But the exploration of evaluation systems and machine learning applications in this study is more in-depth and innovative [20].

Table 2.1: Literature Content

Author Name	Research Contents	Findings
Sun H [10]	Proposed a SPOC teaching method called "Rain Classroom" to improve classroom learning efficiency	The designed methods have positive impacts on the reform of English translation teaching
Zhang et al [8]	Designed a teaching method based on conclusion interpretation concepts and SPOC flipped classroom	The designed teaching model can improve students' learning efficiency
Law et al. [12]	Designed a SPOC flipped classroom teaching model for private teaching	Under the influence of this teaching model, students' total grades are positively correlated with their classroom participation
An X [13]	Established a layered learning method based on SPOC flipped classroom	Compared to traditional teaching models, students are more satisfied with the improved teaching model
Kastrati et al [14]	Proposed an exploratory model for SPOC education model for applying machine learning	The SPOC teaching model directly affects students' mastery of knowledge and skills
Gitinabard et al [15]	Designed a corresponding student performance prediction model for hybrid online offline teaching methods	This method achieves stable cross-category data prediction
Liu Q et al [16]	Proposed an intelligent education evaluation model for predicting the performance of sports students	The designed model is effective
Chen et al [7]	Proposed an automatic TPGA based on genetic algorithm to adjust difficulty factors based on students' online learning data	This model can effectively control the difficulty of the test paper
Soufiane O et al [18]	Studied the basic concepts of SPOC, blended learning, and flipped classroom. Proposed a working method for conducting online training in the form of SPOC. Mainly focused on the student evaluation and organization of teaching papers in online blended teaching mode, and designed a score prediction method and paper generation algorithm based on a learning evaluation diagnostic mode	Innovative in practical applications and design practices
Zhou Y [19]	Studied the implementation of SPOC blended learning mode in the choir command teaching system.	Focusing on solving the student learning evaluation and teaching paper organization, and proposed solutions to student learning evaluation problems through machine learning prediction and algorithm generation
Ramirez Monoso et al [20]	Researched a gamified mobile collaboration tool to enhance utilization of online learning resources.	Based on a learning evaluation diagnostic model, complex machine learning algorithms are used to predict learning outcomes and organize teaching papers

This study conducts in-depth research on the student learning evaluation and teaching paper organization. A new scoring prediction method and test paper generation algorithm are designed, which have obvious innovation in specific problem-solving methods, machine learning applications, and practical applications. Literature contents are shown in Table 2.1.

In recent years, research on SPOC classrooms has gradually enriched. Many research teams have designed and implemented various SPOC based teaching models. The results demonstrate their positive impact on improving students' learning efficiency and engagement. However, although these studies have made some important discoveries, there are still many unexplored areas in how to apply computer technology more intelligently to classroom planning. For example, how to design models that can predict student performance. How to adjust the difficulty of exam papers based on students' online learning data. They are currently important

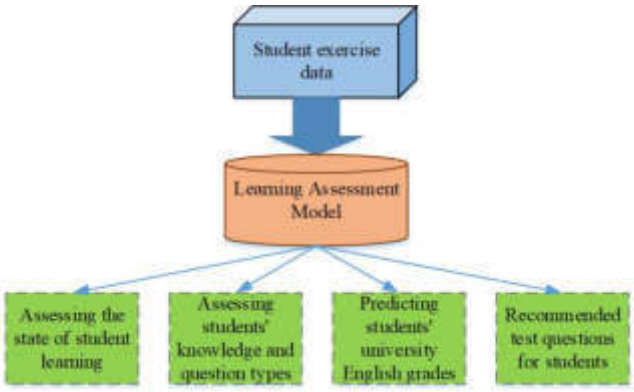


Fig. 3.1: Framework Diagram of Learning Evaluation Model Based on SPOC Classroom and Online Mixed teaching Mode

research directions. Therefore, this study will design a model for teaching evaluation and exam strategies. A one-stop automated online teaching system of teaching evaluation product prediction automatic exam is implemented, providing a new implementation path for SPOC online blended teaching. This will be the novelty and contribution of this study.

When designing scoring prediction methods and test paper generation algorithms (TPGA), relying on learning evaluation diagnostic models, some new and not yet widely applied technologies such as deep learning and neural networks are also introduced. This will increase the innovation of the research. The research results not only demonstrate the accuracy of the prediction model, but also reveal the positive impact of TPGA on students' academic performance. This discovery has not been fully explored in previous studies. Therefore, the research provides a new perspective for the online blended learning strategies. The research results indicate that results-based online blended learning strategies can effectively improve students' learning outcomes. This conclusion not only validates the research hypothesis, but also provides a new practical strategy for the online blended teaching model, which has a certain degree of innovation.

3. Diagnostic model and verification of test paper formation strategy.

3.1. Construction of Learning Evaluation Model Based on SPOC Classroom Theory. The goal of this research is to create a teaching system that can analyze students' learning status, knowledge point relevance and predict college English performance based on SPOC classroom theory, providing personalized learning suggestions for students in English distance teaching. Figure 3.1 is the essential framework diagram of the learning evaluation model based on the SPOC classroom and online mixed teaching mode proposed in this study.

In Figure 3.1, firstly, a learning evaluation model is built. Then the student learning data is preprocessed to build a sub-model from the three dimensions, namely learning status, knowledge point correlation, and college English grade test scores. Finally, the three sub-models are combined. A learning evaluation model according to the SPOC theory is obtained. Student problem chart (S-P table) analysis method is used to analyze the learning situation of students. Among them, the student's attention coefficient is an important parameter in the analysis. The question type and knowledge point (QTKP) score are extracted from the database to construct the SP table. For values higher than or equal to the average, they are recorded as 1. Those that are lower than the average value are recorded as 0. The S-P table analysis method is used to analyze students' familiarity with QTKP, and calculate students' attention coefficient. An example of an S-P table is shown in Figure 3.2

The calculation method of student's attention coefficient is shown in equation 3.1.

$$AF_a = \frac{SumSL_0 - SumSR_1}{SumSL_R - TS_a \cdot SAS} \tag{3.1}$$

	Title 7	Title 1	Title 10	Title 6	Title 3	Title 2	Title 5	Title 9	Title 8	Title 4	Total score	%
Student 2	1	1	1	1	1	1	1	1	1	1	10	100
Student 5	1	1	1	1	1	1	1	0	1	1	9	90
Student 9	1	0	1	1	1	0	1	1	1	0	7	70
Student 4	1	1	1	1	0	0	0	1	1	1	7	70
Student 8	1	1	1	1	1	1	0	0	0	0	6	60
Student 3	1	0	1	0	0	1	1	1	1	0	6	60
Student 1	1	1	1	0	1	0	1	0	0	1	6	60
Student 6	1	1	1	1	0	1	1	0	0	0	6	60
Student 10	1	1	0	1	1	0	0	0	0	1	5	50
Student 7	0	1	0	0	0	1	0	1	0	0	3	30
Total score	9	8	8	7	6	6	6	5	5	5		

Fig. 3.2: Schematic Diagram of S-P table

In equation 3.1, $SUMSL_0$ represents the student's quantity who answered wrongly on the left side of the S curve. $SumSR_1$ is the student's quantity who answered correctly on the right side of the S curve. TS is the students quantity who answered correctly on the left side of the S curve. is the total score of the students a $.SAS$ is the average score of all students. The larger the value of the attention coefficient is, the more unstable the learning state of the students. Teachers need to pay more attention. In normal conditions,AF_a the value is less than 0.5. If the value of AF_a is higher than 0.5, teachers should pay attention to the state of students. If it is higher than 0.7, teachers need to pay more attention. According to the AF_a value and the score of each knowledge point, the learning status is evaluated and analyzed. The specific formula for students' familiarity with each QTKP is shown in equation 3.2.

$$MS = \frac{KPQ_{GE}\text{ No.}}{KPQ\text{No.}} \times 100\% \tag{3.2}$$

The method used in the study is based on students' score rates in various knowledge points and different question types. By calculating familiarity and attention coefficient, students' college English grades are predicted. The main variables include students' score rates on various knowledge points and different question types, test scores, and the total number of correct answers in the knowledge point question types. The effective parameters for the main variable include the average score of knowledge points and question types, as well as the percentage of students' average score. Firstly, the student's score rates for each knowledge point and different question types are obtained. They are stored in a matrix format. Then, familiarity with each question type and knowledge point, and attention coefficients are calculated. Finally, the students are determined based on their familiarity and attention coefficient. The percentage of students' scores in knowledge point question types and attention coefficients are output.

The learning state evaluation algorithm (LSEA) is used to calculate students' familiarity with each QTKP and the student's attention coefficient, so as to provide personalized teaching guidance to students according to the learning style. The following is the specific calculation procedure. The initial step is to obtain the scores of students in each knowledge point and different question types (QTs). They are stored in a matrix form. There are A students in the database. Each student's scoring rate is B. The specific calculation method of the matrix is shown in equation 3.3.

$$y_{ab} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ y_{o1} & y_{o1} & \cdots & y_{op} \end{bmatrix} \tag{3.3}$$

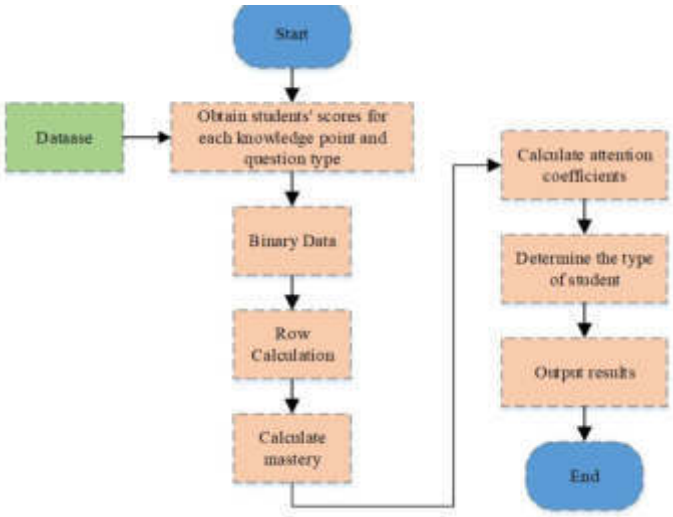


Fig. 3.3: Algorithm Steps for Assessing Students' Learning Status

The LSEA can only deal with binary data, but the scoring rates of each knowledge point and different QTs are numbers distributed within the range [0,1]. Therefore, it must be processed uniformly. The average of the QTKP score is assumed to be the threshold. It is also used as a criterion for determining familiarity. Those greater than or equal to the threshold are recorded as familiar and marked as 1, while unfamiliar ones are marked as 0. The specific formula is shown in equation 3.4.

$$Line(y_{ab}) = \begin{cases} 0, & \text{if } y < \frac{\sum_{a=1}^{o} y_{ab}}{o} \\ 1, & \text{if } x \geq \frac{\sum_{a=1}^{o} y_{ab}}{o} \end{cases} \tag{3.4}$$

In equation 3.4, y_ab is the knowledge score percentage of the a student on the QTb . The score of student a is y_a. The number of students with correct answers of the QTb is y_b. The specific expression is shown in equation 3.5.

$$\begin{cases} y_a = \sum_{b=1}^{p} y_{ab}, \\ y_b = \sum_{a=1}^{o} y_{ab}, \end{cases} \tag{3.5}$$

Then the familiarity with each QTKP, and attention coefficient are calculated. y_ab is the student's score percentage in the QTKP. y_a is the student a's test score.y_b is the total number of students who answered correctly in QTb. η is the student's average score percentage. The expression of the student's attention coefficient $AF_a$is shown in equation 3.6.

$$AF_a = 1 - \frac{\sum_{b=1}^{o}(y_{ab})(y_b) - (y_a)(\eta)}{\sum_{b=1}^{y_a} y_b - (y_b)(\eta)} \tag{3.6}$$

Then the student type is determined. The students' familiarity and attention coefficient obtained in the above calculation are judged according to the student classification standard. Finally, the student's score percentage in the QTKP, the student's attention coefficient and the student type are output. The working method of LSEA is described above. Figure 3.3 displays the specific flowchart.

In Figure 3.3, firstly, it is necessary to extract all data related to the user's usage process from the database, including the number of questions done, exam score rate, knowledge point score rate, question type score rate, and CET-4 score. Next, the collected data is processed, which includes steps such as cleaning the data and processing invalid data. In the feature selection stage, the information related to the predicted target is used

Table 3.1: Examples of Student Scores for Each Question Type

Student number	Essay writing	Short news	Long conversation	Listening chapters	Vocabulary comprehension	Long read	Read carefully	Chinese to English
00100	F	F	F	F	P	F	F	F
00101	P	P	F	P	F	P	T	P
00102	F	F	P	P	P	P	F	P
00103	F	P	F	P	P	P	F	P
00104	P	F	P	F	F	P	P	F
00105	P	P	F	P	P	F	P	P
00106	P	F	P	F	P	P	F	F

as a feature, which should be related to English practice. At the same time, by observing the data, hidden features are discovered to increase feature dimensions. In the construction step of the prediction model, the dataset is divided into the training set and the validation set. Then, a random forest model and a multiple linear regression model are used to construct prediction models for the training set, respectively. The model fusion is carried out through voting method. Finally, the prediction model is validated through a validation set. Once verified, the model can be launched for users to use. This is the entire process of constructing a prediction model for college English Test Band 4 scores.

At present, the student's question score is a continuous value distributed between [0,1]. The Apriori algorithm is a Bool-type attribute association analysis algorithm. Therefore, the score should be replaced by a binary Boolean data P and F . The difficulty of each QT varies. Based on the average score of each QT, the data is performed binary-boolean operations. If the student's score on a QT is higher than or equivalent to the average, then it is defined as P . If the student's score on a QT is less than the average, then it is defined as F .CR_ab is the percentage of student 's score on QT b. The specific expression is shown in equation 3.7.

$$CR_{ab} = \begin{cases} F, & \text{if } CR_{ab} < \overline{CR_{ab}} \\ P, & \text{if } CR_{ab} \geq \overline{CR_{ab}} \end{cases} \tag{3.7}$$

The above cleaned data in Table 3.1 are used for hierarchical and symmetric analysis, i.e. analyzing the association between P and F , high and low scoring questions. Then they are combined to produce an association rule table.

Data are analyzed using the Apriori algorithm in SPSS. SPSS converts data into executable data for Apriori algorithm through data flow. The specific operation process is shown in Figure 3.4.

According to Figure 3.4, the data is first filtered, followed by binary conversion. Finally, the Apeior algorithm is introduced to obtain the analysis results.

3.2. Construction of the Prediction Model for College English Scores Based on SPOC and Online Mixed Teaching Mode. Predicting CET-4 scores is not a feature of the current college English practice diagnosis system. Many students who have just entered college are unfamiliar with the QTKP of CET-4. They can't accurately estimate their ability. Therefore, they are unsure if they can pass level 4. In this study, when using the original system for field research, 58% of students believe that even though they complete some CET-4 exercises, they cannot fully confirm their English skills. They are not sure whether they can pass the college English proficiency test. The flowchart of the algorithm is shown in Figure 3.5.

In Fig. 3.5, this study collects the data during the initial use of the system and the scores left by students for college English college exams. Following data processing and feature screening, the data are trained with MLR and RF respectively. The vote method is used for model fusion. After verification, a more accurate predictive model of college English grades is finally obtained. The above data alone cannot accurately predict CET scores, and hidden features must also be extracted. To improve the prediction accuracy, the authors uses the following hidden features. The student's test score is the basis of evaluation, so this score should be added

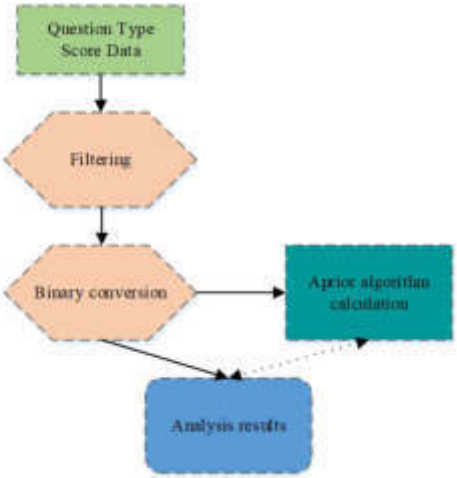


Fig. 3.4: Flow Chart of SPSS Data Flow Operation

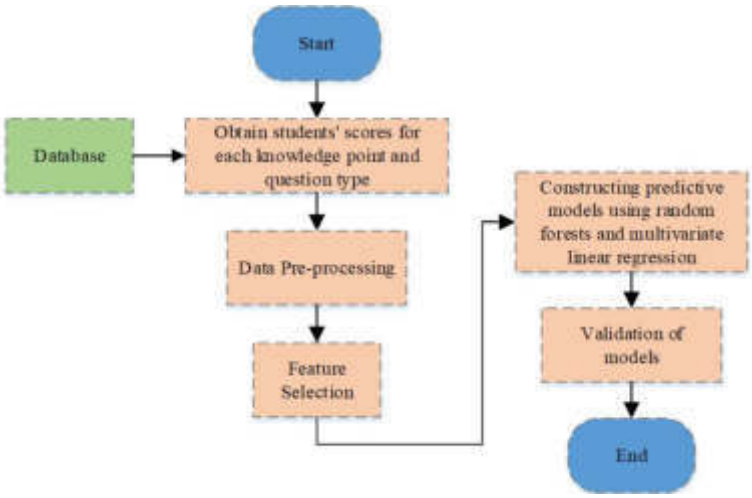


Fig. 3.5: Flow chart of college English grade score prediction

to the feature set. The specific calculation formula is shown in equation 3.8.

$$SR = \frac{AC}{T} \tag{3.8}$$

In equation 3.8,AC represents the number of questions answered correctly. T represents the question quantity in the test paper. After adding hidden features to the feature set, interactive variables need to be built at the same time. The purpose of adding interaction variables is to explore the internal relationship between features. The prediction model can be created once the feature selection and data processing are finished. RF and MLR are used in this study. In addition, the root mean square error (RMSE) and root mean square error are used to verify the model.

$$\begin{cases} RMSE = \sqrt{\frac{\sum_{a=1}^{o}(Y_{PV,a}-Y_{RV,a})^2}{o}}, \\ RMSLE = \sqrt{\frac{\sum_{a=1}^{o}(\log(Y_{PV,a}+1)-\log(Y_{RV,a}+1))^2}{o}} \end{cases} \tag{3.9}$$

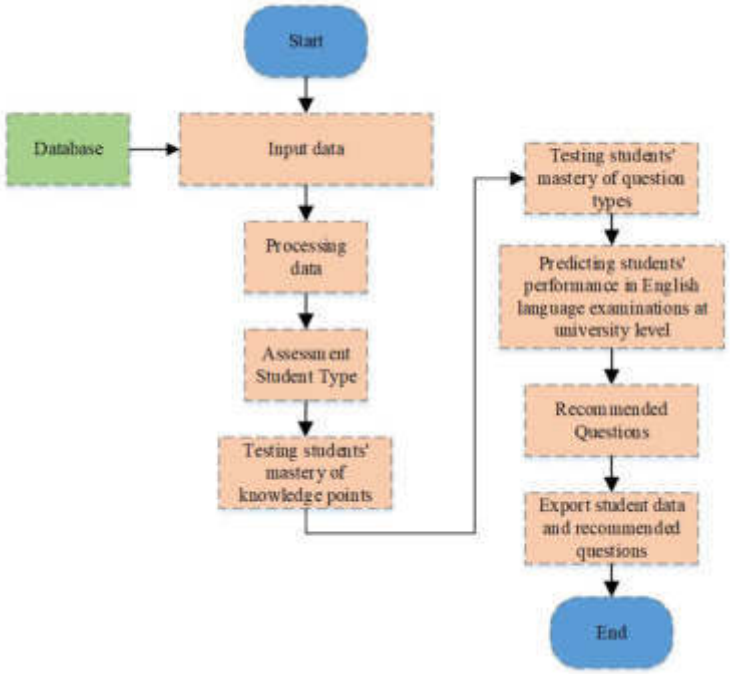


Fig. 3.6: Flow chart of student learning diagnostic evaluation model

In equation 3.9, $Y_{PV,a}$ represents the predicted value and $Y_{RV,a}$ represents the real value. In summary, the above is an overview of the operational steps involved in the comprehensive diagnostic evaluation model. The flow diagram of the overall model is shown in Figure 3.6.

In Figure 3.6, when extracting data, the user data to be evaluated is obtained from the database. When processing data, to meet the requirements of the model, user data needs to be cleaned and transformed. When evaluating learner types, a learning state evaluation model should be used to calculate the user's mastery of knowledge point types and attention coefficients to determine the learner type. When diagnosing the mastery of user knowledge points, the knowledge point association rule table is adopted. Based on the user's knowledge point scoring rate, the user's knowledge strengths and weaknesses are determined. When diagnosing the mastery of user question types, the question type association rule table is used to determine the strong or weak item question types based on the user's question type scoring rate. When predicting user scores, a score prediction model is used to estimate user scores. When recommending test questions, targeted and personalized test question recommendations should be provided to users.

3.3. The Design Method of the TPGA for the Learning Evaluation Model Based on the SPOC Classroom Theory. At present, there are six algorithms widely used in the college English diagnosis system, namely practice question set, random question set, QT set, knowledge point set, knowledge point maximization set and knowledge point association rule set. Except for knowledge point maximization and knowledge point association, all these algorithms are relatively simple. Most of them require students to set parameters by themselves. According to the results of the statistical study presented in the prior section, the research has established a more accurate model for evaluating learning status. Therefore, this part proposes a test recommendation algorithm based on students' learning status. The algorithm considers the student's abilities and learning stability, reasonably allocates the number of test questions with different difficulties. Then it is implemented in the system, so that students can use it when updating the system. According to the goal of the document clustering algorithm based on students' learning status, the design idea of the algorithm is as follows.

1. Correspond the student type calculated by the student status assessment model with the difficulty of the work.
2. Read the attention factors calculated through the assessment model. The total number of test questions is shown in equation 3.10.

$$AT = AT_{four} \cdot (1 + AF) \tag{3.10}$$

   To prevent decimals from appearing in the calculation of the questions, this study adjusts them to integers in the algorithm.
3. Students' knowledge scores are distributed in 5 ranges from the highest score to the lowest score to match the 5 ranges in the distribution coefficient of the test difficulty. That is, knowledge points with higher scores are assigned to more difficult topics. Knowledge points with lower scores are assigned to easier topics.
4. The number of sub-tests assigned to each knowledge item is shown in equation 3.11.

$$AKP_{eachD} = \frac{AT}{AKP} \tag{3.11}$$

In the previous study, this study collects data from multiple colleges and universities in China. It is compared with the college English proficiency test questions in this study. It is concluded that the English level of some students is at the middle level. There is a little gap with the ability level of college students. Therefore, the student data of the school is used for research. The traditional educational measurement method adopts the scoring rate to judge the difficulty of test questions. The difficulty coefficient is applied to explain the difficulty. The specific calculation formula of the difficulty coefficient is shown in equation 3.12.

$$L_a = \frac{R_a}{E_a} \tag{3.12}$$

In equation 3.12,R_a represents the student quantity who answered the question correctly. E_a represents student quantity answering the question. Therefore, the difficulty coefficient P for the entire test paper is defined as shown in equation 3.13.

$$P = \frac{\overline{A}}{H} \tag{3.13}$$

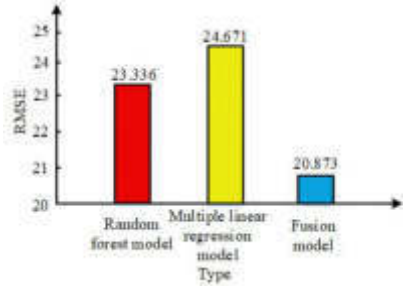
In equation 3.13, A is the average score of the test papers,$\overline{A}$ and is the full score of test papers.

From the score model of the college English proficiency test, the college English proficiency test adopts a normal distribution model to allocate the number of test questions with different difficulties. The normal probability density function formula is shown in equation 3.14.

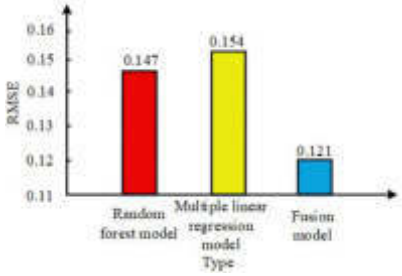
$$f(x) = \frac{1}{\sqrt{2\phi\gamma}} e^{-\frac{(x-\phi)^2}{2\gamma^2}} \tag{3.14}$$

ϕ is the expected value and γ is the variance of the normal distribution. The four-level score interval [0, 710] is divided into 5 main intervals, corresponding to the difficulty of the 5-level test questions. Specifically, they are [0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], and [0.8, 1.0]. The critical values are 0.2, 0.4, 0.6, 0.8, and 1.0 respectively. The corresponding four-level score values are 332, 364, 391, and 421 respectively. According to the obtained four-level score values, the score interval [0, 710] is divided into five score intervals. Each interval corresponds to a different probability interval of test difficulty, as shown in equation 3.15.

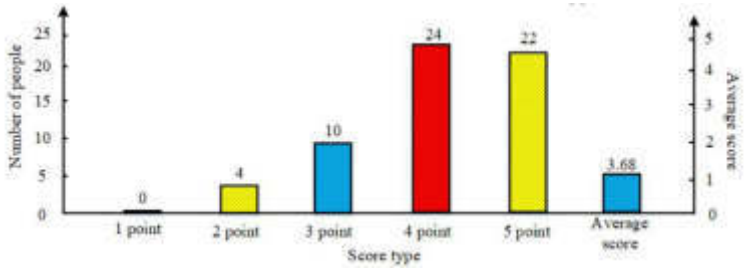
$$\begin{cases} L_1 = \int_0^{332} \sigma_{\phi,\gamma}\, dx \\ L_2 = \int_{332}^{364} \sigma_{\phi,\gamma}\, dx \\ L_3 = \int_{364}^{391} \sigma_{\phi,\gamma}\, dx \\ L_4 = \int_{391}^{421} \sigma_{\phi,\gamma}\, dx \\ L_5 = \int_{421}^{710} \sigma_{\phi,\gamma}\, dx \\ L = \sum_{a=1}^{5} L_a = L_1 + L_2 + L_3 + L_4 + L_5 \end{cases} \tag{3.15}$$



((a)) RMSE



((b)) RMSLE



((c)) Evaluation Score

Fig. 4.2: Comparative Analysis Results and Satisfaction Results

4. The effect analysis of the diagnostic model and test-taking strategy.

4.1. Effect Analysis of Diagnostic Model. In terms of parameter configuration for random forest models, grid search and cross validation methods are used to determine the optimal parameter combination. The main parameters to be adjusted include the number of trees (n_estimators), selected from [100, 200, 300, 400, 500], max depth, selected from [10, 20, 30, None], and maximum number of features (max_features), selected from ['auto ',' sqrt ']. For the multiple linear regression model, the regularization parameter (alpha) is mainly adjusted and selected from [0.1, 0.01, 0.001, 0.0001]. In the diagnostic effect analysis, the RMSE and root mean square logarithmic error (RMSLE) are used for comparative analysis of different models. A functional evaluation questionnaire is designed to investigate the satisfaction of students who use the evaluation model with the model. The comparative analysis results and satisfaction results are shown in Figure 4.2.

From Figure 4.2, in the comparison of RMSE indicators, the RMSE value of the RF model is 23.336, and the RMSE value of the MLR model is 245.671. The RMSE value of the model incorporating voting methods is 20.873. The RMSE value of the fusion model designed by the research is the lowest. In the RMLSE index, the RMSE value of the RF model is 0.147. The RMLSE value of the MLR model is 0.154. After fusing the two models through voting, the RMLSE is 0.121. The designed fusion model has the lowest RMSE value, which is more accurate. In terms of satisfaction rating, the satisfaction score of 1 is 0 people. The satisfaction score of 2 is 4 people. The satisfaction score of 3 is 10 people. The satisfaction score of 4 is 24 people. The satisfaction score of 5 is 22 people. Most students are satisfied with the model effect.

In the comparison of RMSE indicators, the RMSE value of the RF model is 23.336, and the RMSE value of the MLR is 245.671. However, the RMSE value of the model designed in the study that fused these two models through voting method is 20.873. This means that the fusion model performs better in prediction error. As the RMSE value decreases, the prediction error decreases. The prediction accuracy of the model is higher. This is very important for practical applications. The model predicts the results as accurately as possible to make correct decisions based on the predicted results. After comparing the prediction models, the comparative results of the predicted value and the actual value obtained from the verification set data are shown in Figure 4.3.

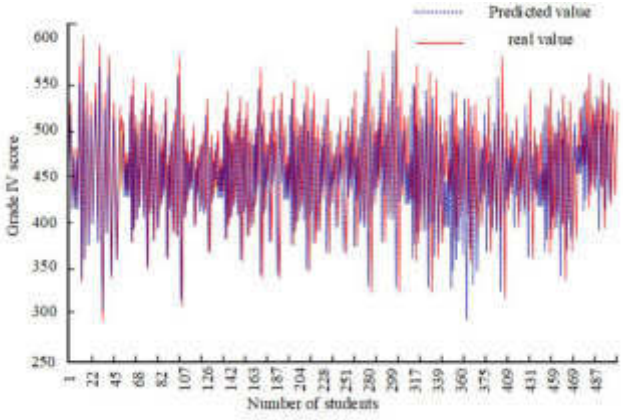


Fig. 4.3: Comparison chart of predicted value and actual value

In Figure 4.3, the abscissa indicates the total number of students. The ordinate represents the student's score. The majority of predicted values are relatively close to the actual values, which are all within a reasonable range. This shows that the model designed in the study is accurate and stable. It can still guarantee accurate performance prediction under large-scale data.

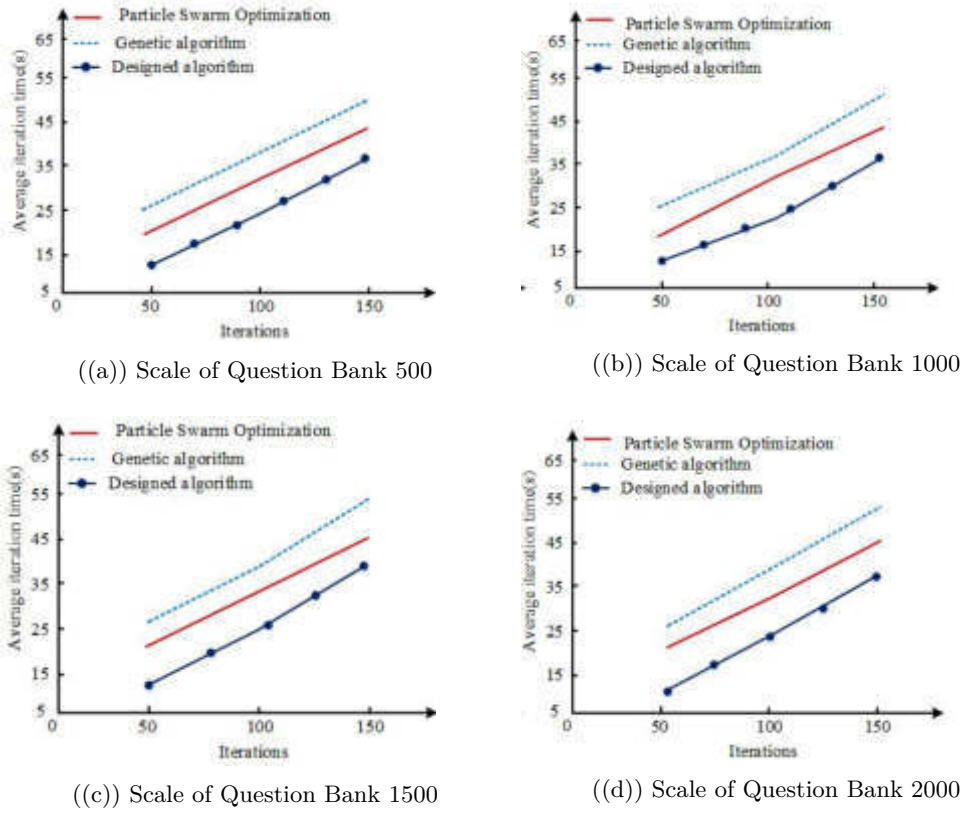
4.2. Experimental Verification of the Algorithm for TPGA. In terms of parameter configuration for the test paper generation algorithm, the following parameters are mainly adjusted, including population size, selected from [50, 100, 150, 200], and iterations, selected from [100, 200, 300, 400, 500]. The fitness function is set based on the difficulty, differentiation, and covered knowledge points of the test paper. Through multiple experiments and comparisons, the optimal parameter combination is ultimately determined. In verifying the TPGA, the running time and application effect of the algorithm in different question bank sizes are compared. The scale of the question bank mainly chooses four scales of 500, 1000, 15000 and 2000. The TPGA designed in the research is compared with the particle swarm TPGA and the genetic TPGA under different question bank scales. The running time comparison is shown in Figure 4.5.

From Figure 4.5, under the four scales of question bank scales of 500, 1000, 15000 and 2000, with an increase in iteration steps, the running time of the TPGA has a positive correlation with the number of iteration steps. Compared with the particle swarm and the genetic algorithm, the designed algorithm has the lowest dashed line position. It demonstrates that the duration of the TPGA is the shortest.

Under four scales of question bank size 500, 1000, 15000, and 2000, as the iteration steps increase, the variation in the running time formed by the designed test paper generation algorithm shows a positive correlation with the iteration steps. This indicates that the test paper generation algorithm designed in the study has higher operational efficiency when processing large-scale data. This is very important for practical applications, as it usually need to process a large amount of data. If the algorithm runs for too long, the efficiency in practical applications will be greatly reduced. The application effect is shown in Table 4.1.

From Table 4.1, the difficulty coefficient of the designed TPGA is 0.593, which is similar to the other two comparison algorithms. The error rate formed is only 0.837%, which is much smaller than other algorithms. The knowledge points cover rate is higher, reaching 98.31%. In terms of the TPGA, the scores of the algorithm designed in the research are 7.98, 27.65, 29.57, and 30.73 in memorization ability, comprehension ability, simple application ability and comprehensive application ability respectively, all of which are higher than other algorithms. The score of innovation ability is 14.28, which is close to 14.29 of genetic test paper. However, the TPGA designed by the research has higher scores in memorization ability, comprehension ability, simple application ability, and comprehensive application ability while maintaining the same innovation ability as other models. The effect is better. The model needs to run in a server environment. The running efficiency in the server is also analyzed, as shown in Figure 4.7.

From Figure 4.7, the time for the test model in the identical machine testing circumstances is 1.49s, while



((a)) Scale of Question Bank 500

((b)) Scale of Question Bank 1000

((c)) Scale of Question Bank 1500

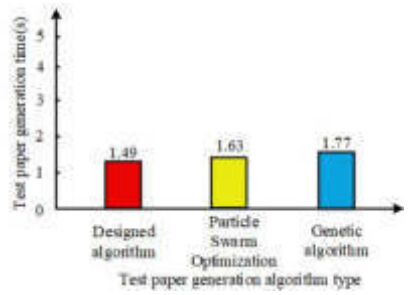
((d)) Scale of Question Bank 2000

Fig. 4.5: Runtime Comparison

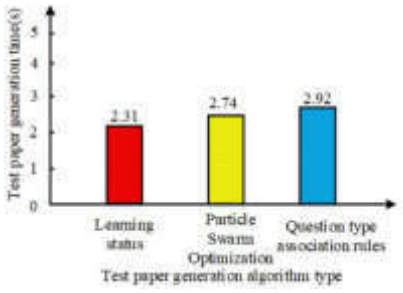
Table 4.1: Algorithm Application Effect Comparison

Algorithm		Genetic algorithm	Particle swarm algorithm	Research design algorithm
Degree of difficulty		0.586	0.626	0.593
Error rate (%)		2.168	4.002	0.837
Coverage of knowledge points		96.48	87.52	98.31
Ability Level Rating	Memorization ability	6.37	5.47	7.98
	Comprehension	24.93	26.75	27.65
	Simple Application Capability	27.63	23.48	29.57
	Comprehensive ability	26.85	27.57	30.73
	Creativity	14.29	11.82	14.28

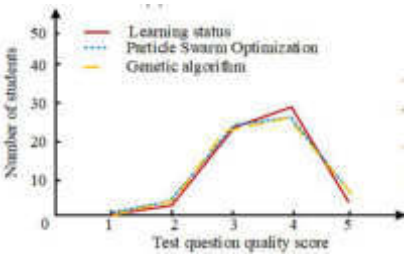
the particle swarm test model and the genetic test model in the local test environment are 1.63s and 1.77s. The designed model takes less time. It has higher calculation efficiency under the identical machine testing circumstances. In the server testing circumstances, the time for the designed model is 2.31s, while the particle swarm test model and the genetic test model are 2.74s and 2.92s respectively. The learning state of the TPGA designed by the research also takes less time to operate, which has higher operation efficiency under the same server test environment. In addition, in terms of the quality and quantity scores of the test papers, the number of people who scored between the quality scores of 1 and 5 in the TPGA designed by the study are 0, 2, 24,



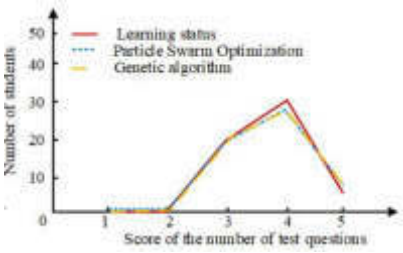
((a)) Native Test Environment



((b)) Server Test Environment



((c)) Change of Test Question Quality Score



((d)) Change of The Number of Test Question

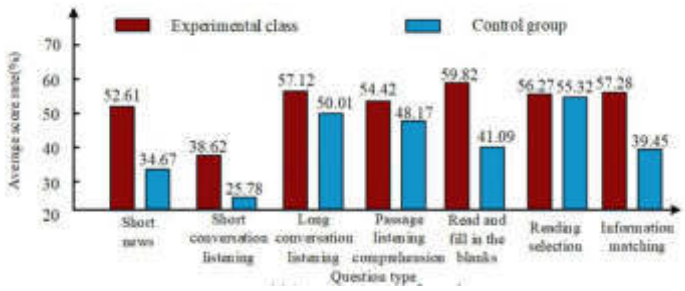
Fig. 4.7: Efficiency and quality analysis results of test papers

30 and 4 respectively. The number of people scoring between 1 point and 5 points are 0, 0, 24, 32 and 4 respectively. Compared with other models, more 4-point ratings and 5-point ratings are obtained. Student satisfaction is relatively higher. Finally, how the TPGA improves students' academic performance is examined. The knowledge points and question scoring rates are used to measure it. The specific results are shown in Figure 4.9.

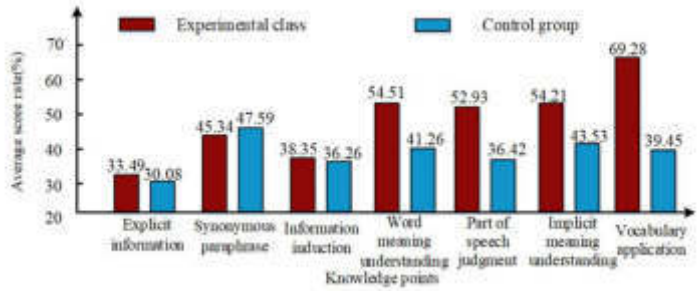
From Figure 4.9, with the help of the designed learning status TPGA, the students in the experimental group performs seven tasks in short news, brief conversation listening, long conversation listening, short listening comprehension, reading comprehension, reading selection and information matching. The average scoring rates in the QTs are 52.61%, 38.62%, 57.12%, 54.42%, 59.82%, 56.27% and 57.28%. In the control group, the average score who adopts the traditional learning system in the seven QTs are 34.67%, 25.78%, 50.01%, 48.17%, 41.09%, 55.32% and 39.45%. In the experimental group, the average score rates in the information disclosure, synonymous reporting, information induction, word meaning understanding, part of speech judgment, implicit meaning understanding and vocabulary application are 33.49%, 45.34%, 38.35%, 54.51%, 52.93%, 54.21% and 69.28% respectively. In the control group, the average score rates in the seven knowledge points are 30.08%, 47.59%, 36.26%, 41.26%, 36.42%, 43.53%, and 39.45%. In each QTKP, the average score rate of the experimental group students is relatively high. The learning status of the TPGA designed by the research can help students improve their grades.

With the help of the learning state test paper model designed in the research, the average score rate of the experimental group students in each question type and knowledge point is higher. This indicates that the learning state test paper generation model designed in the research can effectively improve their academic performance. This is very important for practical teaching, because the goal is to improve their academic performance. If a model can effectively improve their academic performance, then it is a successful model.

5. Conclusion. The research mainly starts from the achievement promotion behavior in the English SPOC online mixed teaching mode. A joint teaching strategy of learning diagnostic assessment + study performance



((a)) Average Score rate of Question Type



((b)) Average Score of Question Points

Fig. 4.9: Knowledge points and score rate of question types

prediction + test paper teaching model is designed. The MLR technique and the RF algorithm are combined to design a performance prediction model. The learning status is taken as the main consideration when designing the TPGA. According to the performance prediction model, the research results show that 24 people have a satisfaction rating of 4 points for the prediction model. 22 people have a satisfaction rating of 5 points. These results indicate that most people express high satisfaction with the accuracy and effectiveness of the prediction model. In terms of TPGA, the designed TPGA receives more 4 and 5 points, and student satisfaction is relatively high. This indicates that the TPGA has been widely recognized by students. It plays a positive role in improving their academic performance. In the experimental group, there are more students in sections 70-80 and 80-90, with 27 and 6 students respectively. This result further proves the effectiveness of the teaching strategy and predictive model. The designed performance prediction model and TPGA are effective, which can encourage students to improve their grades. This conclusion not only proves the effectiveness of the research method, but also provides new ideas and directions for future teaching models. While discussing the research results, the methods and models of this study are mainly based on the SPOC online blended teaching mode. Therefore, the applicability may be influenced by the teaching mode and student learning status. Regarding the limitations of case studies, the research mainly focuses on the student population with scores ranging from 70 to 90, which may overlook the learning situation of students in other grades. Therefore, future research should consider a wider student population to improve the universality of the model and teaching strategies. For future work, it is recommended to further optimize and adjust the learning diagnosis assessment, learning performance prediction, and exam teaching models to adapt to more teaching modes and student types. More research is expected to validate and improve the models and strategies of this study.

Funding. The research is supported by: "The Design and Practice of Teaching English Courses at Open Universities ", 2023 Teaching Reform and Research Project of Jiangsu Open University(23-YB-08); "Exploration and Practice on Improving the Effectiveness of Continuing Education Courses from the Perspective

of Educational Psychology", 2022 Higher Education Scientific Research and Planning Project, the Chinese Association of Higher Education (22JX0413).

REFERENCES

[1] Chakraborty, P., Mittal, P., Gupta, M., Yadav, S. & Arora, A. Opinion of students on online education during the COVID-19 pandemic. *Human Behavior And Emerging Technologies.* **3**, 357-365 (2021)
[2] Dhawan S. Online learning:, A. panacea in the time of COVID-19 crisis. *Journal Of Educational Technology Systems.* **49**, 5-22 (2020)
[3] Sekeroglu, B., Dimililer, K. & Tuncal, K. . *Student Performance Prediction And Classification Using Machine Learning Algorithms[C]//Proceedings Of The 2019 8th International Conference On Educational And Information Technology.* pp. 7-11 (2019)
[4] Mahmood, S. Instructional strategies for online teaching in COVID-19 pandemic. *Human Behavior And Emerging Technologies.* **3**, 199-203 (2021)
[5] Nambiar, D. The impact of online learning during COVID-19: students' and teachers' perspective. *The International Journal Of Indian Psychology.* **8**, 783-793 (2020)
[6] Fauzi, I. Khusuma I H S. *Teachers' Elementary School In Online Learning Of COVID-.* **5**, 58-70 (2020)
[7] Chen, J. SPOC-based flipped learning model applied in interpreting teaching. *International Journal Of Emerging Technologies In Learning (iJET).* **15**, 4-13 (2020)
[8] Zhang, L., Xuan, Y. & And, Z. and application of spoc-based flipped classroom teaching mode in installation engineering cost curriculum based on obe concept. *Computer Applications In Engineering Education.* **28**, 1503-1519 (2020)
[9] Chen, J. SPOC-based flipped learning model applied in interpreting teaching. *International Journal Of Emerging Technologies In Learning (iJET).* **15**, 4-13 (2020)
[10] Sun, H. A SPOC teaching mode of college english translation based on" Rain Classroom". *International Journal Of Emerging Technologies In Learning (iJET).* **14**, 182-193 (2019)
[11] Zhang, L., Xuan, Y. & And, Z. and application of spoc-based flipped classroom teaching mode in installation engineering cost curriculum based on obe concept. *Computer Applications In Engineering Education.* **28**, 1503-1519 (2020)
[12] Law, L., Hafiz, M., Kwong, T. & Eva Wong, .. Enhancing SPOC-flipped classroom learning by using student-centred mobile learning tools. *Emerging Technologies And Pedagogies In The Curriculum.* pp. 315-333 (2020)
[13] An, X. & Qu, C. hierarchical learning model based on deep learning and its application in a SPOC and flipped classroom. *International Journal Of Emerging Technologies In Learning (iJET).* **16**, 76-93 (2021)
[14] Kastrati, Z., Kurti, A. & Hagelb"ack, J. . *The Effect Of A Flipped Classroom In A SPOC: Students' Perceptions And Attitudes[C]//Proceedings Of The 2019 11th International Conference On Education Technology And Computers.* pp. 246-249 (2019)
[15] Gitinabard, N., Xu, Y., Heckman, S. & Lynch, C. How widely can prediction models be generalized? Performance prediction in blended courses. *IEEE Transactions On Learning Technologies.* **12**, 184-197 (2019)
[16] Liu, Q. & Huang, Z. Yin Y,Hui Xiong,Yu Su, Guoping Hu. *Ekt: Exercise-aware Knowledge Tracing For Student Performance Prediction.* **33**, 100-115 (2019)
[17] Chen, X., Zhong, D., Liu, Y., Li, Y., Liu, S., Deng, N. & Others Auto-generating examination paper based on genetic algorithms[C]//Advances on P2P. *Parallel.* pp. 887-893 (2020)
[18] Soufiane, O., Class, M. & Learning[j], B. Higher Education. (2021)
[19] Zhou, Y. Exploration of the Implementation of SPOC Blended Learning Model in the Teaching System of Choral Conducting[J]. *Curriculum And Teaching Methodology.* **6**, 51-57 (2023)
[20] Ramírez-Donoso, L., Pérez-Sanagustín, M., Neyem, A. & Others Fostering the use of online learning resources: results of using a mobile collaboration tool based on gamification in a blended course[J]. *Interactive Learning Environments.* **31**, 1564-1578 (2023)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Oct 7, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 271–284, DOI 10.12694/scpe.v25i1.2294

CREATION OF DEEP LEARNING SCENARIOS IN THE NETWORK TEACHING OF PHYSICAL EDUCATION TECHNICAL COURSES

FANGYU LI*

Abstract. The network teaching evaluation of sports professional technical courses has positive significance for the sustainable development of education. And how to establish an effective evaluation model is the key part. The research introduces the creation of in-depth learning scenarios (LS) into the network teaching of sports professional technical courses, and then constructs a new network teaching mode of sports professional technical courses. The Particle Swarm Optimization algorithm - Attention- Long Short-Term Memory network (PSO-Attention-LSTM) Chinese Emotion Classification Model (ECM) is constructed to classify the online evaluation text to realize the evaluation of online teaching. This model combines the improved PSO, Attention and LSTM classification models. The optimal number of hidden layer nodes for LSTM model is about 100, and the optimal data size for batch processing is 25. The overall error rate of online teaching teachers of male and female sports professional skill courses is 10.1%, and the overall error rate of online teaching teachers of general and advanced sports professional skill courses is 12.8%. The application effect of the creation of in-depth learning scene in the network teaching of physical education technical courses is shown. When the classification threshold is 0.6 and 0.8 respectively, the AUC of PSO-Attention-LSTM Chinese ECM is 0.821 and 0.809 respectively. The research institute has put forward that the online teaching platform of sports professional technical courses has extremely high practical application value and has been unanimously praised by network users.

Key words: DL; LS; Sports major; Network teaching; PSO; Attention mechanism; LSTM

1. Introduction. With the development and popularization of network technology, more and more physical education courses have begun to adopt the network teaching mode. This teaching mode can provide more flexible learning time and place, which is convenient for students to conduct independent learning and course selection. And due to the uneven distribution of teaching resources, online teaching has become a common teaching mode for students to obtain better teaching resources. The cultivation of students' core literacy needs to take Deep Learning (DL) as the goal, and teachers need to create learning situations to realize students' DL. The process of DL emphasizes students' deep understanding of core curriculum knowledge and their ability to apply this knowledge to real problems and situations. It also emphasizes whether learners can transfer and apply in similar situations or in new situations. In recent years, curriculum teaching in various disciplines has been actively exploring how to implement the core quality of students' development, and research results have emerged in endlessly [1, 2, 12]. Different countries have different expressions of the core literacy content of sports discipline. But the constituent elements are basically the same, including health education, personality development, sports ability and social adaptation. In the existing literature research, most of them are about the strategy research of the creation of sports in-depth learning situation. And in practice, there are many materials for the creation of in-depth sports learning situations, such as the problem situation, real learning situation, classic historical event situation, and successful experience situation. However, it is relatively scattered and has not formed a theory, which requires comprehensive, detailed and in-depth research and refinement. At the same time, Long Short-Term Memory Network (LSTM) and Particle Swarm Optimization (PSO), as common classification algorithms, have been widely used in education, finance and other industries, and have achieved outstanding results [10, 14]. Based on this, a Chinese ECM combined with multiple classification algorithms is proposed to classify the content of online teaching evaluation of sports technical courses. The purpose is to complete the physical education classroom teaching with a teaching scheme suitable for their own in-depth learning situation.

*School of Culture and Tourism, Luzhou Vocational and Technical College, Luzhou 646000, China (fangyu05li@outlook.com)

2. Related works. Scenarios emphasize more on various objective scenes, things and events that can stimulate students to actively learn. It is divided into two types: passive and mechanical. They have no substantive connection with the learning theme and lack certain depth and breadth. Haerens L and other researchers believe that situational teaching is a combination of vivid intuition and language description. They create a typical scene to stimulate children's enthusiasm for learning and thus promote their active participation in the teaching process. This model has been widely used in the actual teaching process [15]. Lentillon-Kaestner V et al. revealed the connotation of the learning situation of the core quality of the discipline. It includes four aspects: context-based learning, emphasizing students' interaction and participation, paying attention to students' experience and exploration spirit, and paying attention to the leading and guiding role of physical education teachers [16]. Roure C and other researchers created "combination exercises", "game games", "problem tasks" and "group cooperation" situations to promote students to achieve DL and develop students' core quality of sports discipline. Situational teaching is not limited to the middle and lower grades of primary school. As long as the design is reasonable, it is also applicable to senior high school students and college students. Situational teaching includes game competition, imitation and music accompaniment [17]. When physical education network teaching evaluating, the methods based on emotion dictionary and machine learning have some shortcomings. In this regard, Sun Z and other researchers used the Laplacian smoothing algorithm of mutual information of emotional inclination points to expand the dictionary. At the same time, they analyzed the positive and negative effects of different sentence patterns on sentence emotion [18].

Jingchao H et al. found that judging the students' listening state in the classroom is a more difficult problem, and the current intelligent models to recognize the students' state in the classroom are not accurate, to address this problem, the research team proposes a two-stage state detection framework based on deep learning and HMM feature recognition algorithms, which is capable of recognizing the facial expressions of students in the classroom, and judging their listening the research results show that the proposed model has a relatively good performance in classroom student state feature recognition [19]. At present, there is a strong subjectivity problem in online courses teaching evaluation. In this regard, scholars He H have built the most basic Chinese emotion dictionary by using English seed dictionary and machine translation technology. However, the coverage of emotional words is low, which cannot be transferred in combination with the context, which is easy to lead to ambiguity [20]. To realize the improvement of teaching methods in the network teaching process of sports technical courses, Reckhow S scholars and other scholars have expanded Hownet in Chinese. On this basis, they proposed a method based on semantic similarity and a method based on semantic correlation field. The experiment can achieve more than 80% accuracy in the common word set [21]. To evaluate the current common network teaching evaluation methods, Lau E T team selected support vector machine as the basis and compared the bagging method, lifting method and random subspace method. The experiment shows that the integrated learning model has better effect [22]. Peng W et al. designed a model based on three classifiers: the first two are naive Bayesian and maximum entropy models based on statistics. And the last is a knowledge-based tool that can conduct in-depth analysis of natural language sentences. This model has higher practicability and feasibility, and can realize online teaching evaluation [23]. To obtain an objective analysis method for online teaching evaluation, Yang C scholars used two-way LSTM combined with attention model to encode and express the micro-blog text and its emoticons. The model performs better than the known model on multiple tasks [24]. Wang X and other researchers have established a keyword thesaurus based on LSTM, which can further explore the potential information in the depth of the text and improve the judgment ability of emotional orientation. The constructed keyword thesaurus is helpful to objectively and fairly evaluate the text content of online evaluation [25].

DL situation has become an important way and learning method to implement students' core literacy development. And it involves the creation of different situations in primary school, junior high school and senior high school, as well as in the network teaching of sports technical courses. In summary, many scholars have already had research in this field, through similar semantics and LSTM and other methods of learning scenarios construction and evaluation to detect the quality of learning, the experimental results show that the proposed methods can basically achieve the expected results, but there is also a large space for improvement. This study in the sports deep learning scenarios created after the introduction of physical education professional arts course network teaching to build a In this study, we constructed a new online teaching model for sports

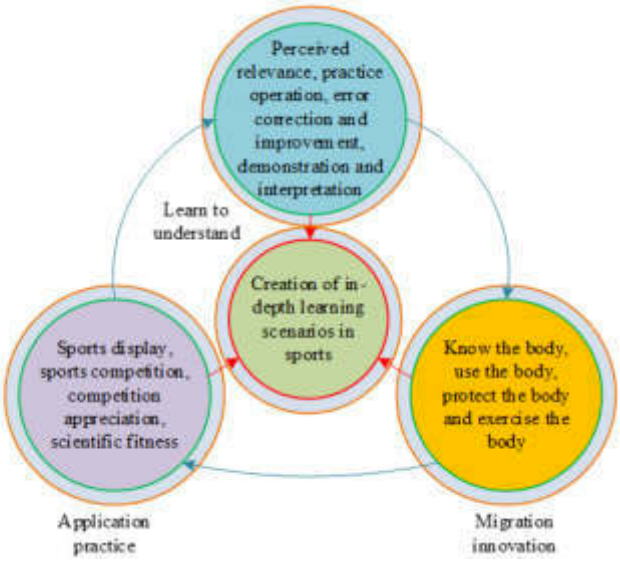


Fig. 3.1: Creation of physical education DL scene in the network teaching

professional arts courses after introducing sports deep learning scenario creation into the online teaching of sports professional arts courses, and constructed PSO-Attention-LSTM Chinese emotion classification model to classify the online evaluation text in order to realize the evaluation of online teaching. This experiment is a cross-section of the basic theoretical research on deep learning contexts and the creation of deep learning contexts in different disciplines at home and abroad, aiming to provide ideas and references for subsequent research.

3. The creation of sports depth LS in the network teaching of sports technical courses.

3.1. Sports depth LS creation. International research focuses on the use of different methods to study the connotation, structure, formation, development and evaluation of the core literacy of sports discipline. It puts forward the understanding and requirements of sports core literacy. The creation of sports DL situation is to make students' sports ability, healthy behavior and sports morality developed. It is a meaningful learning activity created by linking the teaching content of physical education and health with students' learning, life reality and social practice. It has distinct characteristics of subjectivity, inquiry, guidance and openness. Figure 1 refers to the process of creating sports depth LS in the network teaching of sports technical courses [26]. The creation of sports depth LS is based on three dimensions: the cognitive style, activity experience and cognitive level of sports and health discipline. It includes three learning stages: learning understanding, application practice and transfer innovation. Subject knowledge needs to go through learning and understanding, application practice, migration and innovation and other key ability activities to complete the external orientation, independent operation and conscious internalization from specific knowledge to cognitive methods. Learning and understanding is the ability of students to input, store, process, relate and systematize subject knowledge. It is embodied in the ability to complete DL tasks such as recall and extraction, identification and confirmation, generalization and correlation, explanation and demonstration in the cognitive process. Only through DL can students form core literacy, and creating sports DL context is the necessary path to develop students' core literacy. Learning and understanding includes perception and relevance, practice and operation, error correction and improvement, and demonstration and interpretation. Application practice mainly includes sports display, sports competition, competition appreciation and scientific fitness. Migration innovation mainly includes understanding the body, using the body, protecting the body and exercising the body.

Lewin's action model is adopted for the specific action research of the network teaching of sports profes-

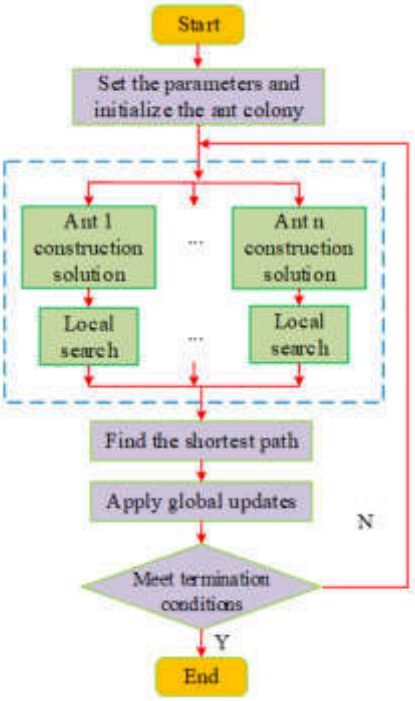


Fig. 3.2: The principle of ant colony optimization algorithm.

sional technical courses, which is the most representative and operational research model and starts from the problem [3]. The experiment plans to carry out two rounds of action research, and carry out the first round of action research according to the relevant theoretical content of the creation of sports DL situation. Observation and reflection are made based on students' classroom performance and feedback, and the quality of students' homework after class. On this basis, the students' homework materials are modified and improved, and the typical cases of the creation of sports DL situations with relatively high quality are selected. The ideas and methods of the creation of sports DL situations are initially condensed. The second round of action research reflects and summarizes the first round of action research, improved the deficiencies, and focuses on how students create the sports DL situation. The first round of action research includes three classes. In the first class, the teacher tells the theoretical knowledge about the creation of sports DL situation, and arranges the homework after class according to the knowledge content in the class. And the experiment allows students to create a sports DL situation based on learning understanding, application practice, and transfer innovation, correct their homework, and screen out excellent homework. In the second class, students rating as excellent homework are selected to share their homework after class and how they created the sports DL situation. Other students and teachers make comments to revise and improve the homework materials. For the third class, students go to the stage to exchange and share the revised and improved homework materials, and other students and teachers comment again to enrich the case materials. The typical cases of creating high quality sports DL situations are screened, and the ideas and methods of creating sports DL situations are initially condensed.

3.2. Improvement of PSO and LSTM algorithm. As an intelligent bionic algorithm, PSO has many advantages, such as heuristic search, strong robustness, positive information feedback, self-organization, distributed computing, and so on. It is often used to find the best path [4]. Figure 3.2shows the principle of PSO. PSO actually imitates the ability of ants to find the shortest foraging path through information exchange. Ants secrete pheromones during foraging, and pheromones complete information exchange between ant groups. The shorter the path length, the higher the ranking order of the ants and the larger the weight value. The

pheromone update of the top ants is required, and formula 3.1 is the calculation [5].

$$\tau_i j(t+1) = (1-\rho)\tau_i j + \sum_{k=2}^{w} \Delta\tau_{ij}^{k}(t) + \Delta\tau_{ij}^{*}(t)) \tag{3.1}$$

In formula 3.1, the initial pheromone volatilization factor is , and its value range is (0,1). The pheromone update of the second to w ants is $\sum_{k=2}^{w} \Delta\tau_{ij}^{k}$, and the pheromone update of the best ants is $\Delta\tau_{ij}^{*}(t)$. PSO can solve many linear and nonlinear problems with good convergence speed. But basic PSO is prone to local optimum when particle searching, which reduces the diversity of particles. To solve this problem, an improved algorithm based on PSO is proposed to make particle population's diversity improved and avoid local optimization [6, 7, 8]. In the M-dimensional target search space, a population including n_2 particles can be obtained randomly. In search space, V refers to the speed of the particles, and U represents particles position. The corresponding fitness value can be calculated through the objective function. $P_I = P_{i1}, P_{i2}, ...P_{iM}, P_g = P_{g1}, P_{g2}.....P_{gM}$ represent individual extremum and group extremum respectively. During each iteration, the particle updates its speed and position by comparing the fitness value of the new particle with the fitness value of the current individual extreme value and the population extreme value [9, 10]. The particle speed depends on the position information of the current particle and the particle in the last iteration. Formula 3.2is the update equation.

$$V_{im}^{k+1} = wV_{im}^{k} + c_1 r_1 (P_{im}^{k} - (1+\beta_1 U_{im}^{k-1)} + \beta_1 U_{im}^{k}) + c_2 r_2 (P_{gm}^{k} - (1+\beta_2) U_{im}^{k} + \beta_2 U_{im}^{k-i}) \tag{3.2}$$

In equation 3.2, $m = 1, 2.....M, i = 1, 2.....n_2$. k is the current number of iterations. $c_1, c_2 > 0$ are the acceleration factors. r_1, r_2 are random number between 0-1, and w is the inertia weight coefficient. When iterations number k increasing, the non-linearity decreases [11]. Formula 3.3 is its expression.

$$w = w_{ini} - (w_{ini} - w_{end}) \left(\frac{k}{k_{max}}\right)^2 \tag{3.3}$$

w_{ini} and w_{end} are the initial and ending values of w in formula 3.3. $\beta_i < \frac{\sqrt[2]{c_i - 1}}{ci}, i = 1, 2, ...n_2$ in the early stage $\beta_i \geq \frac{\sqrt[2]{ci-1}}{c_i}, i = 1, 2.....n_2$ in the later stage is to make global search ability enhanced in the early stage. Optimization ability's improving purpose in the later stage is to achieve detailed search. The principle of particle position update is formula 3.4.

$$U_{im}^{k+1} = \begin{cases} U_{im}^{k} + V_{im}^{k+1}, 2d^k, \sum_{i \neq j, j=1}^{n} \left\| u_{im}^{k} - u_{jm}^{k} \right\| < d^k \\ U_{im}^{k} + V_{im}^{k+1}, \sum_{i \neq j, j \neq 1}^{n} \left\| u_{im}^{k} - u_{jm}^{k} \right\| \geq d^k \end{cases} \tag{3.4}$$

$d^k = d_{ini} - (d_{ini} - d_{end}) \left(\frac{k}{k_{max}}\right)^2$ is the minimum distance allowed between particles in formula 3.4. d_ini ,d_{end} and d^k are the initial value and the end value respectively. Figure 3.3 is LSTM structure. The specific steps includes following : first, determine the information forgotten by neurons. At time , it is assumed that samples number is n,batch data X_t of x is the vector, h and H_t are hidden layer's length and state. H_t at previous time is represented by H_{t-1} . Equation 3.5 is the expression of forgetting gate at time t.

$$f_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \tag{3.5}$$

σ is sigmoid function. W_f and b_f are learnable weight and offset vector in formula 3.5. Secondly, it determines the neural unit information. And it uses sigmoid function to obtain updated value in network layer in formula 3.6

$$i_t = \sigma(X_t W_{xi} + H_{t_1} W_{hi} + b_t) \tag{3.6}$$

In formula 3.6, W_i is update door weight. b_t is update door offset. It uses hyperbolic tangent function to obtain candidate value in tach layer in formula 3.7.

$$\overline{C_t} = tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \tag{3.7}$$

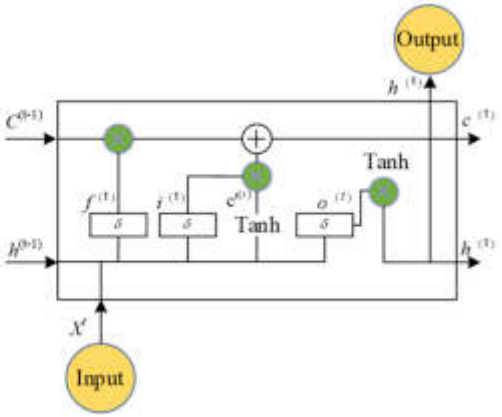


Fig. 3.3: Schematic diagram of LSTM neural network structure

Then, the memory state was updated. The state is updated by point multiplication. Information flow is controlled by output and forgetting gate. Finally it can get updated state in formula 3.8.

$$C_t = f_t \Theta tanh(C_t) \tag{3.8}$$

When the forgetting door is close to 1 and the input door is close to 0, old state memory unit can be recorded to current time in formula 3.9. LSTM can alleviate circulatory nerve's gradient disappearance. Finally, output state memory unit can be got.

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \tag{3.9}$$

In formula 3.9, W_o and b_0 are output gate weight and offset. Equation 3.10 is hidden layer state H_t 's calculation at time t .

$$H_t = O_t \Theta tanh(C_t) \tag{3.10}$$

3.3. Chinese ECM of PSO-Attention-LSTM. To realize the evaluation of the network teaching of sports professional technical courses under the creation of sports depth LS, the selection method PSO-Attention-LSTM Chinese ECM is studied. The model has short construction time and memory function, and has good practical value in the actual process. In view of the long distance dependence problem of attention mechanism and the shortcomings of recurrent neural network, we study the use of self-attention mechanism. This mechanism can effectively extract features and get the correlation of words in sentences. It can make long-distance dependence issue solved, carry out parallel operations, reduce the computational difficulty of each layer, and optimize model performance. Self-attention mechanism is mainly divided into Multi-Head Attention and Scaled Dot-Product Attention, as shown in Figure 3.4.

As can be seen in Figure 3.4, the structure consists of multiple attention heads, each with the same structure. In each attention header, the input is divided into three parts: query (Q), key (K) and value (V). These three parts transform the input into different representations by means of linear transformations. Attention weights are then assigned to the different values by computing the dot-product attention scores of the query with respect to the keys. The attention weights indicate the correlation between the query and the key, where higher weights indicate more important information. The attention weights computed by each attention head are multiplied with the corresponding values and these products are summed to get the final attention output. Finally, the outputs of all the attention heads are concatenated together for final transformation by another linear transformation. The computation process of Scaled Dot-Product Attention starts with the inputs consisting of query vectors (Q), key vectors (K) and value vectors (V). The dot product between the query vector and the key vector is computed to get the attention score. The result of dot product indicates

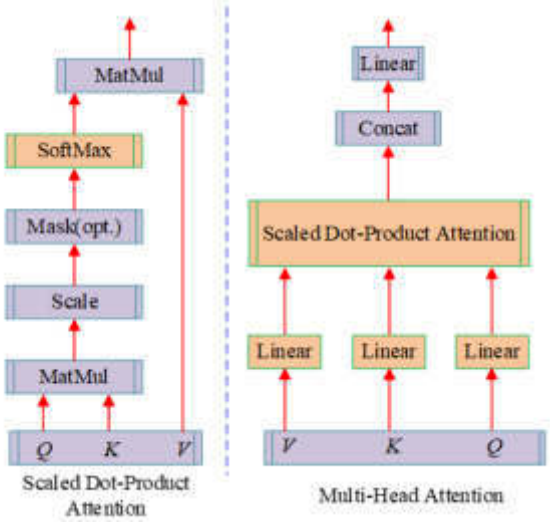


Fig. 3.4: Structure of self-attention mechanism

the similarity between the query vector and the key vector. Then, a scaling operation is performed on the attention score to avoid the value of the dot product being too large. The scaling factor is generally taken as the square root of the inverse of the dimension of the query vector. Next, the attention weights are obtained by normalizing the scaled attention scores by the Softmax function. Finally, the attention weights are applied to the value vectors to obtain the weighted value vectors. In the left part of Figure 3.4, v represents value, K represents key, and Q represents query. The three vectors represent the sentence itself in the algorithm. In the coding, they are obtained by multiplying the input vector X and the weight matrix, as shown in formula 3.11.

$$\begin{cases} Q = W^Q X \\ K = W^K X \\ V = W^V X \end{cases} \tag{3.11}$$

After obtaining the values of the three vectors, the Scaled Dot-Product Attention is calculated by formula 3.12.

$$Attention(Q, K, V) = softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)V \tag{3.12}$$

d_k represents the word vector dimension of k and Q in formula 3.12. $\frac{1}{\sqrt{d_k}}$ plays a regulating role to avoid excessive internal product of K and Q . The probability distribution is normalized by softmax, and the weight relative to V is obtained. The weighted sum is the result of multiplying by V . In the structure shown in the right part of Figure 3.4, it is necessary to project V , K and Q linearly for h times, and then calculate the h times through equation 3.13 to obtain the Multi-Head Attention mechanism, as shown in equation 3.13

$$\begin{cases} MultiHead = Concat(head_1, head_2, ...head_n)W^o \\ \quad headi = Attention(QW_i^Q, KW_i^k, VW_i^v) \end{cases} \tag{3.13}$$

The vector dimensions of model words are $W_i^Q \epsilon R^{dm \times dk}, W_i^k \epsilon R^{dm \times d_k}, W_i^v \epsilon R^{dm \times dv}, W^o \epsilon R^h d_v \times d_m$ and $d_m in$ Equation 3.13. By using softmax function, it can get attention distribution vector, which is ‘s weight, as shown in Equation 3.14.

$$e_i j = \alpha(s_i - 1, 0_j) = v^T tanh(w_s s_i - 1 + w_0 O_j) \tag{3.14}$$

Figure 3.5 is the schematic diagram of the Chinese ECM of PSO-Attention-LSTM. The model consists of four parts. First, it is the preprocessing of text, including sorting and loading text, text cleaning and word

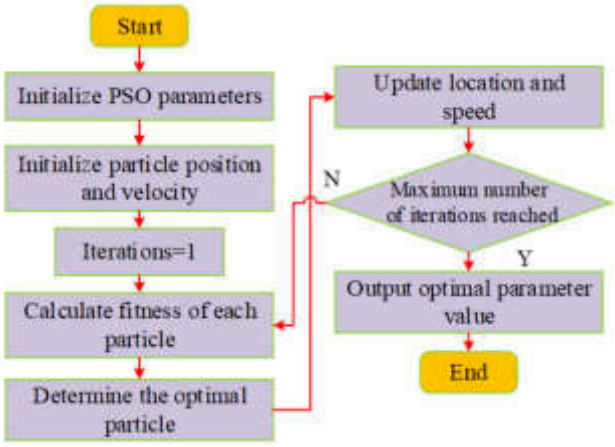


Fig. 3.5: Schematic diagram of Chinese ECM of PSO-Attention-LSTM

segmentation, text standardization and converting text into word vector representation. Then there is the hierarchical structure of the model, which consists of four parts, namely input layer, LSTM layer and attention layer, namely the fully connected softmax layer. The improved PSO is applied to the whole neural network model for parameter optimization. First, the original Chinese comment anticipation in the online teaching process of sports professional technical courses is obtained, and the text preprocessing operations such as word segmentation are carried out. Secondly, the sentence is mapped to the corresponding word embedding vector according to the word vector model, and it is referred to by the word vector of the input layer. Thirdly, the improved PSO is initialized. The parameters include random number, learning factor, inertia weight, fitness function, etc. Thirdly, the optimized object of the improved PSO is set to initialize the speed and position of particles. Fourth, the fitness of particles is calculated according to iterations number, and particles' speed and position at the current time are continuously updated. Fifthly, the improved PSO's iteration is ended, and optimal parameter value is determined. The value is used for model training, and finally the emotional classification of the evaluation text is obtained. Students of a school were selected for this study and the survey was conducted by randomly distributing questionnaires. The data collected from the survey was subjected to data preprocessing, a process designed to screen out data from the raw data that do not fit into the model and to correct the data or eliminate useless data. Such as gender, height, weight, etc.In the questionnaire, there is a large amount of personal information, and if not handled properly, it is highly likely to lead to the leakage of personal privacy. Based on the content of this study, eliminate significantly unrelated information such as blood type, height, weight, etc.

4. The application effect of the creation of sports depth LS in the network teaching of sports technical courses. The system required for the experiment is Windows 10, the processor is Inter (R) Core (TM) i7-6700, the memory is 4.00G, the application software version is MATLAB R2022, the selected experimental data is 10000 groups, and the ratio of training set and test set is 9:1. Table 4.1 refers to the parameter settings of Chinese ECM.

The research first analyzes the model test results of PSO. Figure 4.2 (a) and (b) refer to the sum of error squares before and after the improvement. From Figure 4.2 (a), the sum of squares of the minimum error squares of the PSO indicates a high convergence rate before the number of iterations is 10. However, the convergence rate gradually slows down when iterations number is 10-20, and finally tends to converge when iterations number is 20. Its stable value of the sum of squares of errors is about 0.9. The convergence speed of the improved PSO is faster. When iterations number is about 6, convergence occurs. The convergence value of the sum of squares of errors is 0.21, the convergence speed is increased by 75%, and the average sum of squares

Table 4.1: Parameter setting of Chinese ECM

Parameter	Value
LSTM Batch	30 pieces
LSTM Maximum Iterations	200 pieces
LSTM learning rate	0.001
optimizer	Adam

Parameter	Value
PSO inertia weight	0.68
PSO random number	0.6, 0.3
PSO learning factor	1.7, 1.7
Dropout	0.5

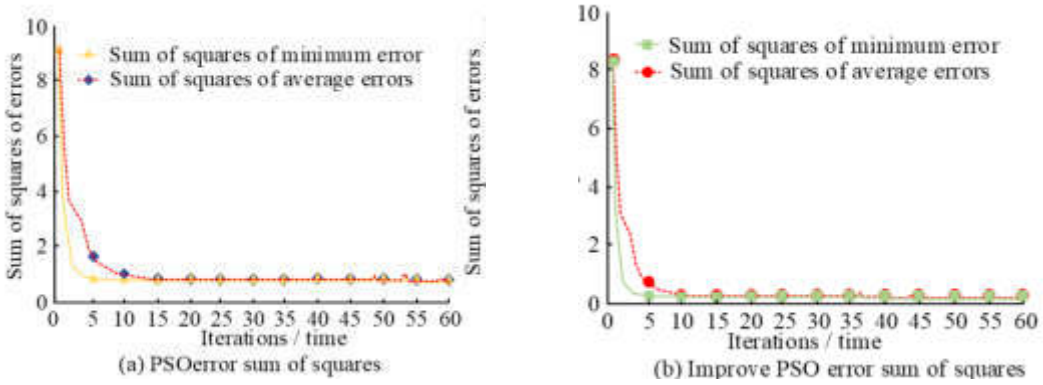
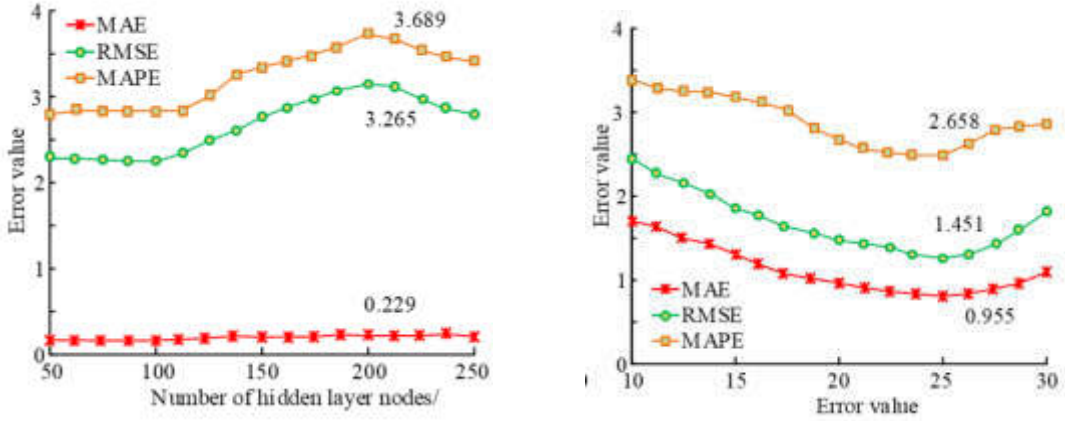


Fig. 4.2: Sum of square errors of GA-BP and AGA-BP neural network algorithms

of errors is reduced by 80%.

The research first determines the optimal model parameters of the proposed LSTM model. The loss of the model is calculated using three indicators: mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). The loss value of the model under different number of hidden layer nodes is shown in Figure 4.4 (a). When hidden layer nodes number is less than 100, the values of MAE, RMSE and MAPE are larger. When the number of nodes exceeds 100, the values of the three indicators gradually increase. When the value exceeds 200, the values of the three indicators gradually decrease. Therefore, the optimal number of hidden layer nodes is about 100, and the error values of MAE, RMSE and MAPE are 0.229, 3.265 and 3.689 respectively. Figure 4.4 (b) refers to the error values under different batch processing data scales., The optimal data size for batch processing is 25. First the three error indicators decrease. Then they increase gradually with the

The research applies the proposed model to the evaluation of the network teaching content of the technical courses of physical education specialty. It sets 956 teaching evaluation contents, and the teaching score is 0-100 points. The teaching effect level is low, medium and high, and the corresponding score is less than 30 points, [30, 70] points and more than 70 points. Figure 4.5 is the result of the evaluation of the network teaching content of the technical courses of physical education. Of the 956 teaching content evaluations, 265, 348 and 343 are rated as low, medium and high, with an average teaching score of (69.68 $\pm$ 7.56) points. Therefore, the vast majority of teaching evaluations believe that the network teaching of sports professional technical courses has achieved the ideal teaching effect. Figure 4.6(a) refers to the level of teaching evaluation content of teachers of different genders in sports professional technical courses. From Figure 9 (a), the teaching evaluation effect of male teachers of network teaching in the technical courses of physical education is better than that of female teachers. The average teaching scores of male and female teachers of network teaching in the technical courses of mathematics and physical education are (72.36 $\pm$ 7.89) points and (64.26 $\pm$ 5.86) points. Figure 4.6(b) refers to the teaching evaluation level of network teaching teachers of different levels of sports professional skills courses. The teaching evaluation effect of senior teachers is better than that of ordinary teachers. The average teaching score of network teaching teachers of ordinary and senior sports professional technical courses is (64.26



((a)) No of Hidden Layer Nodes ((b)) Error Value

Fig. 4.4: Performance of LSTM model

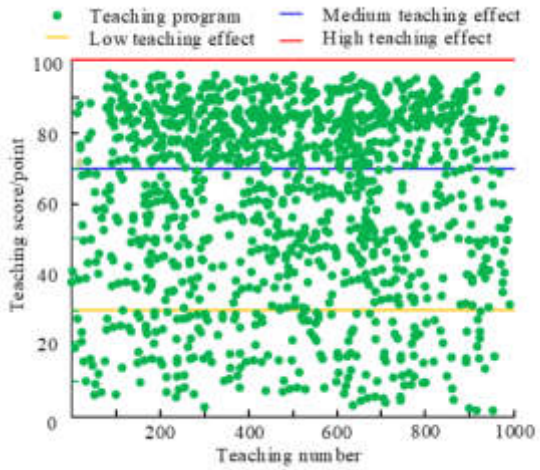


Fig. 4.5: Results of the evaluation of the network teaching content of the technical courses of physical education

± 5.36) points and (76.25 ± 6.35) points.

The research compares the results obtained with the evaluation results of teaching experts. Figure 4.7shows the error rate of different types of teaching grade evaluation. In terms of gender, the overall error rate of male teachers of network teaching in the three teaching evaluations is higher than that of female teachers, with the error rate of 5.3% and 4.8% respectively, and the overall error rate of 10.1%. For the position, the overall error rate of the network teaching teachers of the general physical education professional technical courses in the three teaching evaluations is higher than that of the senior teachers. Its error rate is 7.0% and 5.8% respectively, and the overall error rate is 12.8%. The experiment further verified the application effect of the creation of sports depth LS in the network teaching of sports technical courses. The research sets up a comparison algorithm for verification. The results of regional convolutional neural network (R-CNN), YOLO (You Only Look Once) are shown in Figure 4.9. Figure 4.9 (a) and (b) refer to the receiver operating characteristic curve (ROC)

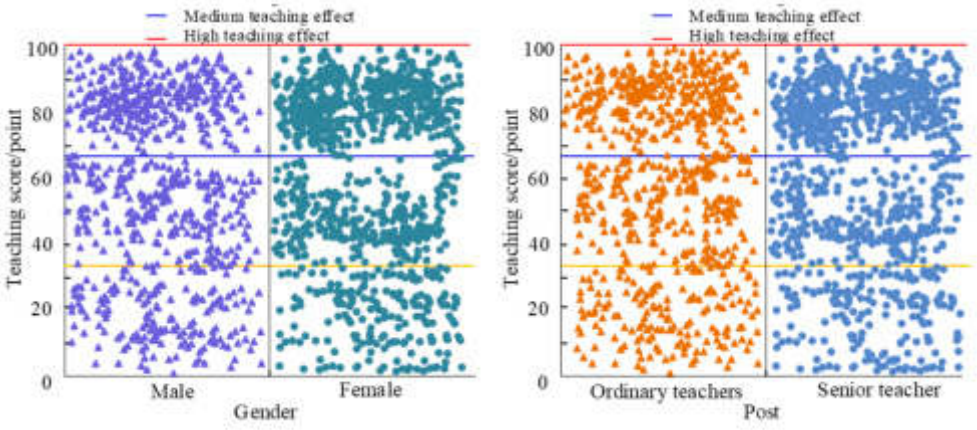


Fig. 4.6: The grade of teaching evaluation content of teachers of different genders in physical education

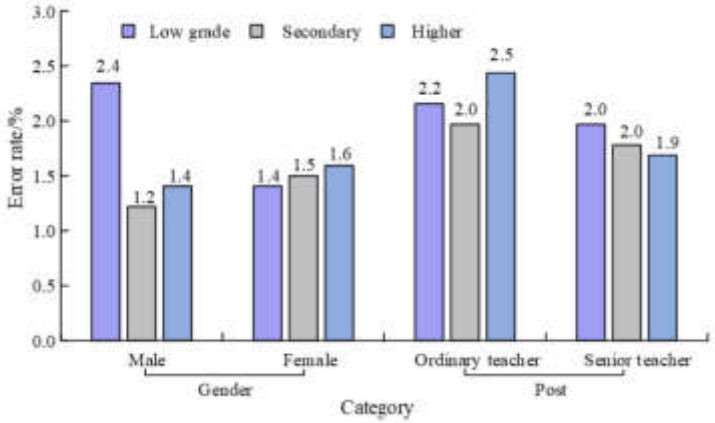


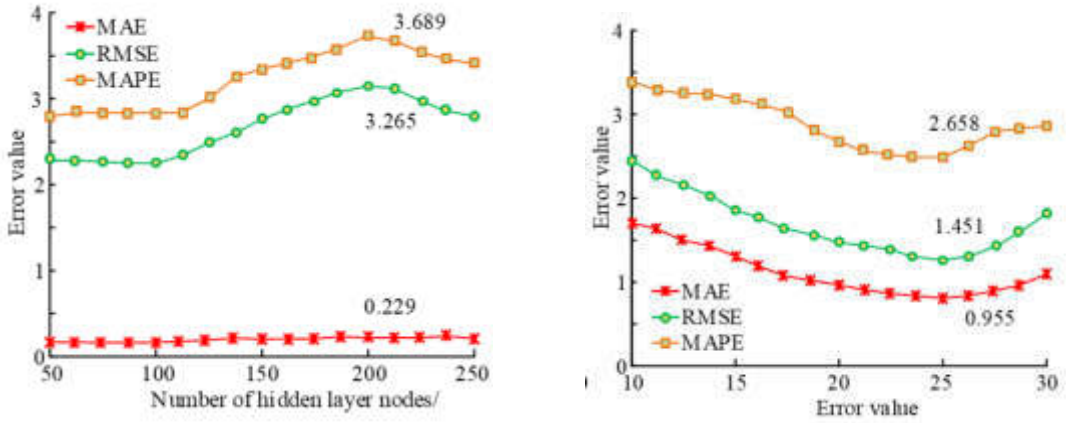
Fig. 4.7: Error rate of evaluation of different teaching levels

with classification threshold of 0.6 and 0.8 respectively. The area under curve (AUC) of the ROC curve of the PSO-Attention-LSTM Chinese ECM is the largest. When the classification threshold is 0.6 and 0.8, the AUC is 0.821 and 0.809 respectively. PSO-Attention-LSTM Chinese ECM performs better than other classification models of the same type.

In order to further verify the performance of the hybrid model, three models were selected to compare their F1 values and recall values, as shown in Figure 4.11. From Figure 4.11, it can be seen that the F1 values and recall of the PSO-Attention LSTM model are relatively high compared to other models, and tend to stabilize at a dataset size of around 6000. The experimental results indicate that the PSO Attention LSTM model performs better than other models.

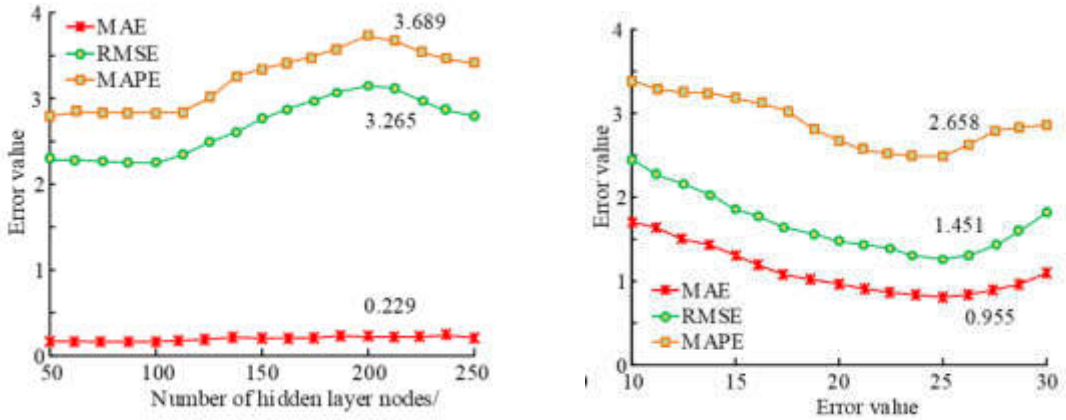
The common deep learning models used for sentiment analysis and text categorization tasks are LSTM and Attention-LSTM, and the two models are analyzed and compared with PSO-Attention-LSTM to compare their performance. The results are shown in Fig. 4.12. From Fig. 4.12, it can be seen that the accuracy of the three deep learning models increases as the dataset increases, among which, the accuracy of this proposed model is the highest and stabilizes at a dataset size of 700.

Randomly select 10 students from 3 groups and 10 teachers from 3 groups, and collect their evaluation results on the model as shown in Table 2. From Table 2, it can be seen that teacher 1, teacher 2, teacher 3,



((a)) Classification threshold 0.6 ((b)) Classification threshold 0.8

Fig. 4.9: Performance comparison of PSO-Attention-LSTM Chinese ECM



((a)) F1 values of different algorithms model ((b)) Recall values of different algorithms model

Fig. 4.11: F1 values and recall rates under four models

and student 1, student 2, and student 3 have PSO-Attention LSTM scores of 8.7, 9.2, 8.9, 9.7, 9.5, 9.3, and the average evaluation scores for the four different models are 9.22, 8.40, 8.22, and 7.73, respectively. The experimental results indicate that the proposed PSO Attention LSTM is more highly praised by users.

5. Conclusions. To explore the application effect of sports depth LS creation in online teaching, a PSO-Attention-LSTM Chinese ECM is proposed to classify the platform comment text. The PSO finally converges when iterations number is 20, and the sum of squares of errors is stable of 0.9. The improved PSO converges faster, and converges when iterations number is about 6. The sum of error squares' convergence value is 0.21, and the convergence speed is increased by 75%. The optimal number of hidden layer nodes is about 100, and the error values of MAE, RMSE and MAPE are 0.229, 3.265 and 3.689 respectively. The optimal batch processing data size is 25, and the corresponding three error values are 0.955, 1.451 and 2.658 respectively. Of the 956

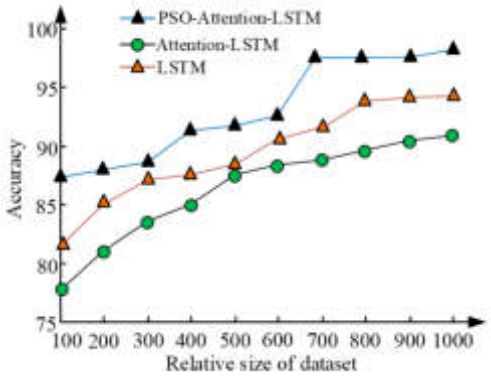


Fig. 4.12: Performance comparison of the three models

teaching content evaluations, 265, 348 and 343 are rated as low, medium and high, with an average teaching score of (69.68 ± 7.56) points. In the three teaching evaluations of professional sports courses, the overall error rate of male teachers of network teaching is higher than that of female teachers, with the error rate of 5.3% and 4.8% respectively. In the three teaching evaluations of the technical courses of general physical education, the overall error rate of the network teaching teachers is higher than that of senior teachers, with the error rate of 7.0% and 5.8% respectively. PSO-Attention-LSTM Chinese ECM performs better than other classification models of the same type. PSO-Attention-LSTM model can effectively classify the evaluation text, and has practical significance in providing teaching feedback for users of online education platform.Haerens L and other researchers believe that situational teaching is a combination of vivid intuition and language description, and have proposed a model. Compared to the research of other scholars, the method model proposed in this study has higher performance. However, there are also shortcomings in the research. Later, more scientific and accurate evaluation can be carried out from the perspective of sentence level of online evaluation text. Moreover, the PSO-Attention-LSTM model has a long training time, which is not advantageous in terms of time, and reducing the training time is a subsequent problem that needs to be investigated, and at the same time, this study was conducted in a laboratory environment, and if its performance can be tested in a real environment, it will be able to show the superiority of the proposed method. The proposed method can be applied not only in physical education, but can be generalized to all online teaching after improving the model, by evaluating the level of learning of students in other teaching for the subject and evaluating the teaching.

REFERENCES

[1] Farias, C., Harvey, S., Hastie, P. & Mesquita, I. Effects of situational constraints on students' game-play development over three consecutive Sport Education seasons of invasion games. *Physical Education And Sport Pedagogy.* **24**, 267-286 (2019)

[2] Casey, A. & Quennerstedt, M. Cooperative learning in physical education encountering Dewey's educational theory. *European Physical Education Review.* **26**, 1023-1037 (2020)

[3] Maia, M. Lapo L V . *Articipatory Action Research For Global Antibiotic Stewardship Network In CPLP: Mixed-method Study.* **29**, 187-192 (2019)

[4] Shou, Z., Lu, X., Wu, Z., Lai, J. & Chen, P. Learning path planning algorithm based on kl divergence and d-value matrix similarity. *ICIC Express Letters.* **15**, 49-56 (2021)

[5] Li, D., Yin, W., Wong, W., Jian, M. & Chau, M. Quality-oriented hybrid path planning based on a* and q-learning for unmanned aerial vehicle. *IEEE Access.* **10** pp. 7664-7674 (2021)

[6] Xiong, S., Zhang, Y., Wu, C., Chen, Z., Peng, J. & Zhang, M. Energy management strategy of intelligent plug-in split hybrid electric vehicle based on deep reinforcement learning with optimized path planning algorithm. *Proceedings Of The Institution Of Mechanical Engineers.* pp. 3287-3298 (2021)

[7] Low, E., Ong, P. & Cheah, K. Solving the optimal path planning of a mobile robot using improved Q-learning. *Robotics And Autonomous Systems.* **115** pp. 143-161 (2019)

[8] Wang, J., Hirota, K., Wu, X., Dai, Y. & Jia, Z. Hybrid bidirectional rapidly exploring random tree path planning algorithm with reinforcement learning. *Journal Of Advanced Computational Intelligence And Intelligent Informatics.* **25**, 121-129

(2021)
[9] Pan, Y., Yang, Y. & Li, W. A deep learning trained by genetic algorithm to improve the efficiency of path planning for data collection with multi-UAV. *Ieee Access*. **9** pp. 7994-8005 (2021)
[10] Xu, X., Cai, P., Ahmed, Z., Yellapu, V. & Zhang, W. Path planning and dynamic collision avoidance algorithm under COLREGs via deep reinforcement learning. *Neurocomputing*. **468** pp. 181-197 (2022)
[11] Chen, P., Pei, J., Lu, W. & Li, M. A deep reinforcement learning based method for real-time path planning and dynamic obstacle avoidance. *Neurocomputing*. **497** pp. 64-75 (2022)
[12] Roure, C., Méard, J., Lentillon-Kaestner, V., Flamme, X., Devillers, Y. & Dupont, J. The effects of video feedback on students' situational interest in gymnastics. *Technology, Pedagogy And Education*. **28**, 563-574 (2019)
[13] Xu, X., Li, D., Sun, M., Yang, S., Yu, S. & Manogaran, G. ... & Mavromoustakis, C. *X. Research On Key Technologies Of Smart Campus Teaching Platform Based On*. **5** pp. 20664-20675 (2019)
[14] Jiang, J., Chen, M. & Fan, J. Deep neural networks for the evaluation and design of photonic devices. *Nature Reviews Materials*. **6**, 679-700 (2021)
[15] Haerens, L., Krijgsman, C., Mouratidis, A., Borghouts, L., Cardon, G. & Aelterman, N. How does knowledge about the criteria for an upcoming test relate to adolescents' situational motivation in physical education?. *A Self-determination Theory Approach*. **25**, 983-1001 (2019)
[16] Lentillon-Kaestner, V. & Roure, C. Coeducational and single-sex physical education: students' situational interest in learning tasks centred on technical skills. *Physical Education And Sport Pedagogy*. **24**, 287-300 (2019)
[17] Roure, C., Méard, J., Lentillon-Kaestner, V., Flamme, X., Devillers, Y. & Dupont, J. The effects of video feedback on students' situational interest in gymnastics. *Technology, Pedagogy And Education*. **28**, 563-574 (2019)
[18] Sun, Z., Anbarasan, M. & Praveen Kumar, D. Design of online intelligent English teaching platform based on artificial intelligence techniques. *Computational Intelligence*. **37**, 1166-1180 (2021)
[19] Jingchao, H. & Zhang, H. Recognition of classroom student state features based on deep learning algorithms and machine learning. *Journal Of Intelligent And Fuzzy Systems*. **40**, 2361-2372 (2021)
[20] He, H., Yan, H. & Liu, W. Intelligent teaching ability of contemporary college talents based on BP neural network and fuzzy mathematical model. *Journal of Intelligent & Fuzzy Systems*. **39**, 4913-4923 (2020)
[21] Reckhow, S., Tompkins-Stange, M. & Galey-Horn, S. How the political economy of knowledge production shapes education policy: The case of teacher evaluation in federal policy discourse. *Educational Evaluation And Policy Analysis*. **43**, 472-494 (2021)
[22] Lau, E., Sun, L. & Yang, Q. Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*. **1** pp. 1-10 (2019)
[23] Peng, W. Construction and application of accounting computerization skills teaching resource database under the background of. *Curriculum And Teaching Methodology*. **2**, 1-4 (2019)
[24] Yang, C., Xie, L., Qiao, S. & Yuille, A. July). *Training Deep Neural Networks In Generations: A More Tolerant Teacher Educates Better Students*. pp. 5628-5635 (0)
[25] Wang, X., Lin, X. & Dang, X. Supervised learning in spiking neural networks: A review of algorithms and evaluations. *Neural Networks*. **125** pp. 258-280 (2020)
[26] Konar, A., Chakraborty, I., Singh, S., Jain, L. & Nagar, A. A deterministic improved Q-learning for path planning of a mobile robot. *IEEE Transactions On Systems, Man, And Cybernetics: Systems*. **43**, 1141-1153 (2013)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Sep 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 285–298, DOI 10.12694/scpe.v25i1.2295

RESEARCH ON DETECTION TECHNOLOGY OF ABNORMAL DATA IN COLLEGE PHYSICAL EDUCATION NETWORK TEACHING TEST RESULTS

FENG SHAN*AND DONGQI LI†

Abstract. In recent years, with the rapid development of big data technology, more and more data are continuously generated with the summary of university systems. Therefore, how to use these educational data to provide more scientific decision-making information for university information builders is very important. This research collected various educational data through the network teaching system and campus information system. After that, the collected data was used to analyze the students' online behavior and to excavate valuable behavioral characteristics. In addition, the experiment also proposed indicators to describe students' behavior, such as network course behavior, network viscosity and life regularity, to provide the basis for the subsequent abnormal performance prediction model. Finally, the experiment used DBSCAN algorithm based on distance optimization for clustering analysis, and constructed a NA model based on multiple classifiers. The research results showed that when, the SC value was 0.711, which was the optimal solution of D-DBSCAN algorithm. At this time, the corresponding number of clusters was 4. When N=2, that was, the base classifier of NA model is composed of C4.5 model and SVM model, the prediction accuracy and time consumption are the most appropriate. The accuracy, recall and F1 values of NA model were 98.16%, 97.26% and 0.958 respectively, which was better than that of single model. To sum up, the NA model based on classifiers proposed in the study had higher accuracy and better model performance, can effectively reflect students' academic level, and could provide accurate abnormal performance data for college sports online teaching tests.

Key words: Abnormal data; Achievements; Network education; Distance optimization; Cluster analysis

1. Introduction. With the rapid development of modern information technology, the construction of network education has gradually improved. Therefore, a large number of education data are gathered [1]. How to excavate valuable information from these educational data has become a problem faced by university informatization builders. Education Data Mining (EDM) can provide more scientific and accurate basis for university administrators' management decisions [2]. Social psychology theory shows that human behavior is determined by subjective norms, attitudes and perceived behavior control [3]. Therefore, through the analysis of students' behavior, we can reflect on individual behavior attitudes and tendencies. In the construction of university informatization, there will be a large number of student behavior data such as achievement data, library related data and network log data. These data reflect students' learning attitude and learning status, and can provide data basis for the analysis of students' behavior logs [4].

At present, most colleges and universities have realized the construction of online teaching platform, so that students can obtain tutorial resources. For students with strong self-discipline ability, this can help them improve their academic level [5, 6]. However, for students with poor self-control, disordered lifestyles and improper internet use have caused significant negative impacts that cannot be ignored.

Students spend a lot of time on socializing, entertainment, and games, exacerbating the risk of entrepreneurial failure, and in severe cases, even psychological problems may occur [7]. In addition, most of the current research on the analysis of academic performance in various universities is based on shallow level analysis of simple data and models, with a single field oriented approach, such as campus card consumption data and online education platform data.

There has been no research on the impact of combining online behavior with other data on academic performance. Therefore, it is very important to accurately predict students' academic performance and pay attention to students' abnormal performance data. Through the above means, students' online behavior can

* School of Physical Education, Shanghai University of Sport, Shanghai 200438, China
†College of Physical Education and Health, Hunan University of Technology and Business, Changsha 410000, China; College of Education, Emilio Aguinaldo College, Manila 1000, Philippines (lidongqi123456789@163.com)

also be predicted and potential risks can be effectively prevented.

In order to carry out personalized teaching for different student groups, discover the abnormal behavior of students in time, and evaluate students comprehensively and objectively, this study conducted data mining on the network use behavior and campus behavior log of students in a university's physical education network teaching. The experiment digitally analyzes the focus of students' academic level analysis, and finally proposes a N-Adaptive Boosting (NA) model based on multiple classifiers to predict students' grades.

2. Related Work. In recent years, the rapid development of big data technology has promoted changes again and again, laying a technical foundation for data mining in the field of education. EDM focuses on students and changes the traditional collective education mode to personalized learning mode. This method prepared individual learning reports by recording students' learning behaviors and data in online education. EDM mainly analyzed the potential laws in students' learning process, so as to achieve the goal of promoting students' effective learning [8]. To make effective project decisions, Yahya A A first thoroughly understood the internal relationship and mutual relationship between the project education objectives (PEO) and student outcomes (SO), and proposed a method based on data mining to mine relevant knowledge. In the experiment, Apriori algorithm was applied to the dataset to generate management rules. The experiment finally confirmed the effectiveness of the mined knowledge for engineering education decision-making [9].

To explore the relationship between students' online courses and final exam scores, Kerzic D and others selected first-year undergraduate students from the School of Administration of the University of Ljubljana to carry out the experiment. Orange data mining software was used for two prediction modeling tasks. The research results showed that there was a strong relationship between students' performance in online education tests and their final results [10].

To analyze the utility and applicability of deep learning in EDM and learning analysis, Doleck T and others compared the prediction accuracy of current mainstream deep learning algorithms. The research results showed that the deep learning method showed the same performance as other machine learning methods [11]. Fan J et al. applied data mining technology to the development of university information management system based on the role of modern management in cultivating talents and serving the society. The research results showed that data mining could greatly improve the data analysis ability and management level of managers in the application of university informatization [12].

With the improvement of university information system construction, it has accumulated a huge amount of student learning data. This provides a data basis for the analysis and modeling of students' learning behavior under the condition of big data [13]. However, how to use a large number of student behavior data for modeling to further achieve the analysis and evaluation of academic level is still concerned by many researchers. Joshi A et al. proposed a new integrated machine learning model (CatBoost) to predict students' academic performance. The experimental results showed that the accuracy of the model is 92.27%, which verified its reliability. The proposed model helped educators identify students at early risk [14].

Ade R proposed a classifier that combines fuzzy ARTMAP and Bayesian ARTMAP classifiers, and predicted students' learning achievements. The experimental results verified the good accuracy of this method in predicting students' performance [15]. Deepika K proposed a hybrid feature selection method of random forest (RFBF-FE) based on unused education data, which combined Relief-F and budget tree. Compared with the existing logistic regression model, the SAP accuracy of this method had increased by 6.85% [16]. To improve students' academic performance, Yusuf A proposed a performance prediction model using stack classifier and composite minority oversampling technology. The research results showed that this technology improved the performance of data mining models [17].

To sum up, data mining in the field of education has important practical significance for teaching management and the prediction of students' academic level. With the continuous improvement of online education and university information construction, online education behavior has also become an important influencing factor. In order to explore the impact of college students' online behavior data on the prediction of students' abnormal academic performance, this experiment depicts students' behavior portraits from a new perspective. The experiment uses DBSCAN clustering method to classify different student behavior portraits, and finally constructs a prediction model of student performance anomalies based on multiple classifiers.

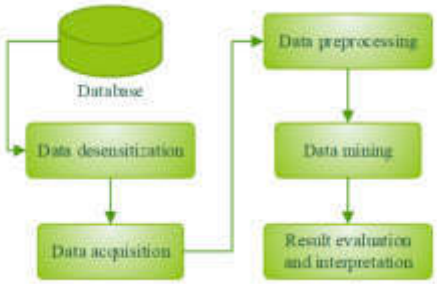


Fig. 3.1: EDM Process

3. The Construction of Student Achievement Anomaly Prediction Model Based on Multiple Classifiers.

3.1. Analysis of College Students' Behavior Data. With the continuous promotion of digital construction in colleges and universities, online education has rapidly entered the plan of college administrators[18]. Student behavior data analysis mainly refers to the use of data mining technology to mine the hidden information and patterns in student behavior log data, and then extract effective features to predict academic level. EDM is a unique branch of the education field. It is a mining technology that uses computer science and data mining technology to obtain special types of data in the education system. This technology extracts and analyzes valuable information under the guidance of psychology and planned behavior, and then discovers students' learning patterns.

EDM inherits the complete process of data mining technology, including data collection, desensitization, preprocessing, mining, and result evaluation and interpretation. See Figure 3.1 for the detailed process. According to data types and mining purposes, EDM methods are mainly divided into four types: clustering algorithm, association algorithm, classification algorithm and regression algorithm. The classification algorithm belongs to supervised learning. It mainly mines the relevance between daily data records and their labels through the known target categories in the data, and classifies them into corresponding categories [19]. After comprehensive consideration, the current mainstream decision tree model (DT), support vector machine (SVM) and integrated learning algorithm (IL) are used in this study [20, 21, 22]. The most classical algorithm in DT is ID3 algorithm. However, in the case of obvious differences and small sample size, the characteristics will be ignored. To solve the above problems, the experiment uses C4.5 algorithm to optimize ID3 algorithm, and uses information gain rate as the standard of feature selection. C4.5 algorithm not only improves the prediction ability of feature points in missing value processing, but also can process and predict discrete and continuous eigenvalues. C4.5 algorithm uses information gain rate to select the optimal partition attribute, see 3.1.

$$\begin{cases} GR(B|A) = \frac{IG(B|A)}{IV(A)} \\ IV(A) = \sum_{m=1}^{M} \frac{|B_m|}{|B|} \log_2 \frac{|B_m|}{|B|} \end{cases} \tag{3.1}$$

In equation 3.1, B is the sample set; $A = a_1, a_2, \ldots, a_m$ has a total of values. If B is divided by A , M branch nodes will be generated. Among them, the m^{th} node contains samples with the value of a_m on all attributes A in B , which is recorded as B_m. $IGB|A)$ is defined as the information gain of attribute A to B. $IV(A)$ is the intrinsic value of property a. The larger the value of M , the greater the value is. C4.5 algorithm can provide effective decision-making for students' behavior analysis, but small changes in data will cause changes in feature selection, and ultimately lead to sudden changes in decision-making logic. SVM is a kind of supervised learning model, which can be used not only for classification problems, but also for nonlinear regression problems. The classification principle of SVM is shown in Figure 3.2.

For linearly separable samples, the calculation of the optimal hyperplane HP is actually a convex quadratic programming problem. In Figure 3.2, there is a sample data set D. And hyperplane HP_1 and HP_2 are expressed

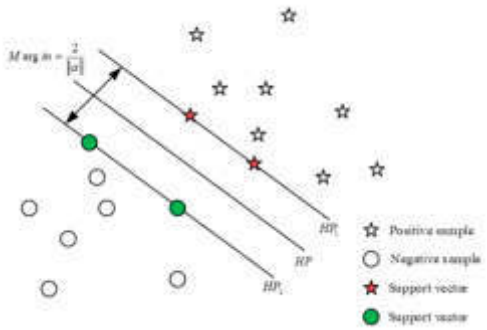


Fig. 3.2: Principle of SVM Model

as equation 3.2.

$$\begin{cases} \text{HP:} & \alpha^T u + g = 0 \\ \text{HP}_1: & \alpha^T u + g = 1 \\ \text{HP}_2: & \alpha^T u + g = -1 \end{cases} \tag{3.2}$$

In equation 3.2, α is the normal vector of the hyperplane; g is the distance between the hyperplane and the coordinate origin. The support vector is consistent with HP_1 and HP_2. The classification interval $Margin$ is the projection of the difference of heterogeneous support vectors at α, as shown in equation 3.3.

$$M = \frac{2}{\|\alpha\|} \tag{3.3}$$

To maximize $Margin$, it is need to solve the convex quadratic programming problem. For optimization problems with constraints, Lagrange function optimization is usually used, that is, adding Lagrange multiplier $\delta_j \geq 0$ to each constraint. The original problem can be transformed into equation 3.4.

$$\begin{cases} \max \sum_{j=1}^{n} \delta_j - \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \delta_j \delta_k v_j v_k u_j u_k \\ \text{s.t.} \quad \sum_{j=1}^{n} \delta_j v_j = 0, \qquad\qquad j = 1, 2, \ldots, n. \end{cases} \tag{3.4}$$

According to equation 3.4, the expression of SVM model can be obtained, see equation 3.5.

$$f(u) = \sum_{j=1}^{n} \delta_j v_j u_j^T u + g \tag{3.5}$$

According to equation 3.2, the result of SVM model is only related to support vector. However, in practical applications, there are many factors that can cause nonlinear classification of sample data. Therefore, nonlinear SVM model came into being. The sample data of low latitude is transformed into high dimensional space through kernel function, so that the sample data becomes linearly separable in high dimensional space. The principle of nonlinear SVM model is as follows. Assuming that there is a mapping function $\zeta(u)$ for vector u , the hyperplane expression can be obtained, as shown in equation 3.6.

$$f(u) = \alpha^T \zeta(u) + g \tag{3.6}$$

Then the operation is similar to linear SVM model, add δ_j . The minimum value problem can be transformed into the maximum value problem under limited conditions. It is difficult to calculate $\zeta(u)$ in the feature space, so the mapping relationship can be transformed. That is, the inner product of u_j and u_k in the mapping space

is equal to the function value calculated by function $\psi(u)$ in the original space. Finally, the nonlinear SVM model can be obtained, see equation 3.7.

$$f(u) = \sum_{j=1}^{n} \delta_j v_j \psi(u_j, u_k) + g \tag{3.7}$$

In equation 3.7, $\psi(u)$ is the kernel function. Different kernel functions can be used for micro-mapping in experiments. At present, the commonly used kernel functions include Gaussian kernel function and sigmoid kernel function. The principle of IL algorithm is to generate multiple weak classifiers through training data. Then, according to a certain rule or integration strategy, multiple weak classifiers are combined into a strong classifier, and then the final decision is made. At present, the most common IL algorithms are Bagging algorithm and Boosting algorithm. For the classification problem of class prediction, the IL algorithm integrates the results of the base classifier into a voting strategy, including simple voting and weighted voting. See equation 3.8 for corresponding discrimination results.

$$\begin{cases} L_1(u)L_1(u) = \text{argmax}(N_j) \\ L_2(u) = \textbf{sign}(\sum_{j=1}^{n} \vartheta_j l_j(u)) \end{cases} \tag{3.8}$$

In equation 3.8, $L_1(u)$ and $L_2(u)$ are the final results of simple voting and weighted voting respectively; N_j is the number of base classifiers whose output result is category j ; ϑ_j is the weight assigned to the j^{th} base classifier; l_j is the judgment result of the j base classifier.

4. Design of Student Achievement Anomaly Prediction Model Based on Distance Optimization and Multiple Classifiers. The data of the study comes from the campus all-in-one card data of a university student and the test result data of online physical education teaching. The study collected relevant data from 53 universities across the province from October 2022 to June 2023 for four grades, making the behavioral analysis results more effective. After data acquisition, desensitization of data records is required. The desensitization process includes five types of fields: name, electronic account, student number, IP address and URL. After data desensitization, data pretreatment is also required. The preprocessing process is as follows: First, delete the vacant data record in the dataset directly; Then filter the duplicate values and keep the first record. The test result data of students' physical education online teaching includes the test results and credits of each course.To carry out a comprehensive assessment and evaluation index of students' academic level, the research selects the test results of compulsory and optional courses in physical education network teaching for weighted average processing. According to the final weighted average score, students are divided into four performance groups, namely, abnormal group, passing group, excellent group and non-excellent group. The abnormal group is the student with less than 60 scores, and the passing group is the student with more than 60 scores; The excellent group is the students with more than 90 points, and vice versa. The abnormal group can detect the students who have the risk of abnormal test results; The excellent group can test the behavioral characteristics of students with excellent academic level. Finally, according to the above standards, students' grades are given corresponding labels. Through the above analysis, we can get three behavioral construction indicators of network behavior, network viscosity and life regularity, and then build a feature library of student behavior portraits. See Table 4.1 for details.

According to the above student portrait description indicators, the research uses the Density-Based Spatial Clustering of Applications with Noise based on Distance Optimization (D-DBSCAN) algorithm to cluster college student groups. This method divides students into groups with different performance differences to explore the differences between students' network behaviors under the condition of different academic achievements. The principle of DBSCAN algorithm is shown in Figure 4.1.

In Figure 4.1, u_1 and u_2 are the core points; u_3 and u_4 are boundary points; u_2 direct from u_1 density; u_3 direct from u_2 density; u_4 direct from u_1 density; u_4 connected to (A^{u_3}) density. DBSCAN algorithm has the following advantages: the clustering process is not affected by the noise in the sample set; The number of clusters does not need to be given in advance; The clustering results are not biased. However, DBSCAN algorithm has the following shortcomings: it is difficult to select the initial parameter neighborhood radius χ

Table 4.1: Three Aspects of Behavioral Evaluation Indicators

Behavior	**Field name**	**Field type**	**Field Description**
-	Network behavior	Enum	Evaluate students' preference for online physical education courses
-	Network stickiness	Enum	Evaluate students' dependence on the network
-	Regularity of life	Enum	Evaluate whether students' self-study is regular
Network behavior	Frequency of sports video courses	Numerical type	Frequency of visits to online physical education courses every month
	Knowledge frequency	Numerical type	Frequency of visiting knowledge websites every month
	Game frequency	Numerical type	Frequency of visiting game websites every month
	Social frequency	Numerical type	Monthly visits to social networking sites
Network stickiness	Online time	Numerical type	Distribution range of online time of students every month
	Online duration	Numerical type	Average online time of students per month
	Online days	Numerical type	Number of days students are online per month
Regularity of life	Self-study duration	Numerical type	Average time spent on self-study in the online library every month
	Number of days to enter the library website	Numerical type	Number of days to access the library website every month

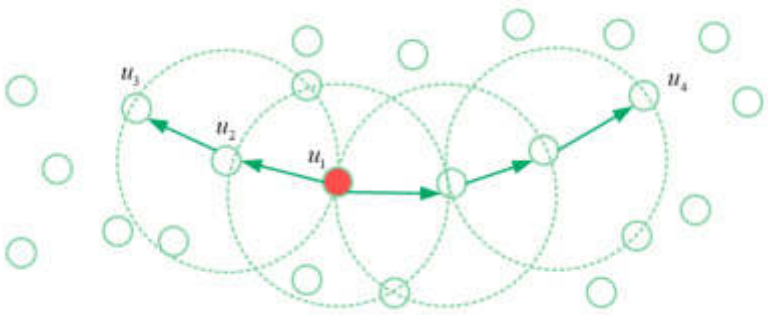


Fig. 4.1: Principle of Dbscan Algorithm

and density threshold $minPts$; Not suitable for sample sets with uneven density and large distance space; In the case of high dimension of sample set, accurate clustering cannot be achieved. In view of the above problems, the D-DBSCAN algorithm is proposed. The algorithm automatically selects the value of χ according to the characteristics of $minPts$ and the data distribution density in the data set. Assuming the existence of sample set $Q = u_1, u_2, \ldots, u_n$, the density can be obtained. For $u_j \in Q$, the density of u_j is in the neighborhood of u_j, and the number of data points is shown in equation 4.1.

$$N(u_j) = \left\{ u_k \in Q | 0 < distance(u_j, u_k) < \chi \right. \tag{4.1}$$

In equation 4.1, $distance(u_j, u_k)$ is the distance between u_j and u_k. For the sample point u_j in the neighborhood of the core point u_k, the distance coefficient between u_k and $u_j(A)$ can be obtained as equation 4.2.

$$\gamma = \frac{N(u_k)}{N(u_j)} \tag{4.2}$$

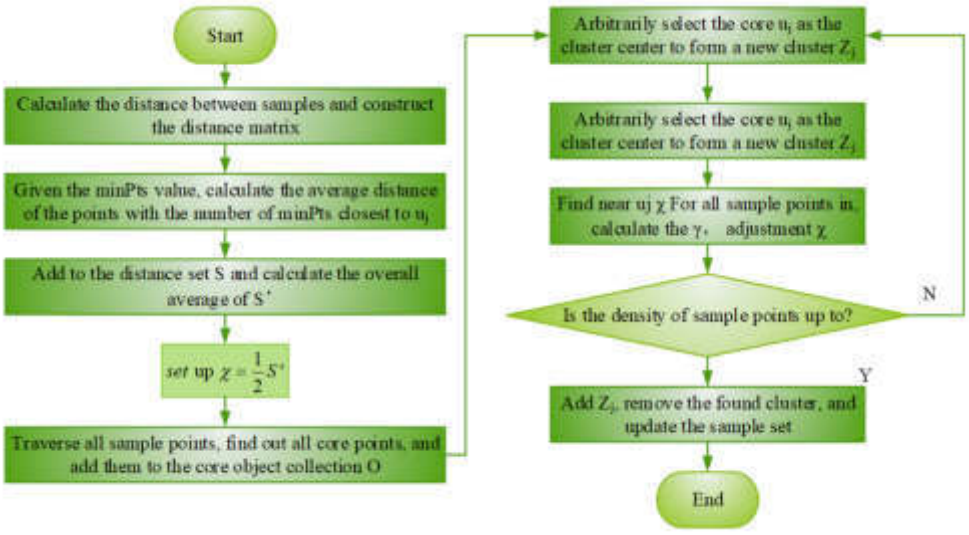


Fig. 4.2: D-DBSCAN Algorithm Flow

According to the above definition, the D-DBSCAN algorithm flow can be obtained, as shown in Figure 4.2.

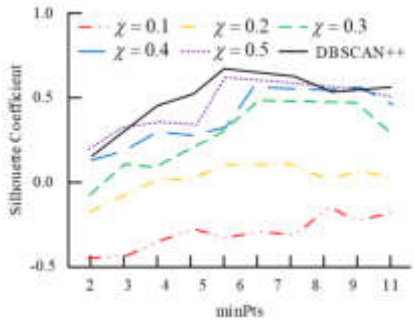
According to the extracted student behavior characteristic data set, the meaning, dimension and order of magnitude of each characteristic index in the multidimensional characteristic data set are different. It is very inappropriate to directly conduct data mining without considering the dimension of feature vectors before knowing the influence of feature vectors on the calculation results. Therefore, in order to ensure the reliability of the results and the validity of the model, before entering the model, it is necessary to standardize the data set of the original features, so that each feature has an equal amount of influence factors in the initial state. The commonly used standardization methods include deviation standardization (DS) and Z-csore standardization. In DS, for sequence $U = u_1, u_2, \ldots, u_j$, $j \in 1, 2, \ldots, n$, if the characteristic index is the greater the better type index or the smaller the better type index, ewuation 4.3 can be obtained.

$$\begin{cases} v_j = \frac{u_j - \min(U)}{\max(U) - \min(U)} \\ v_j = \frac{\max(U) - u_j}{\max(U) - \min(U)} \end{cases} \tag{4.3}$$

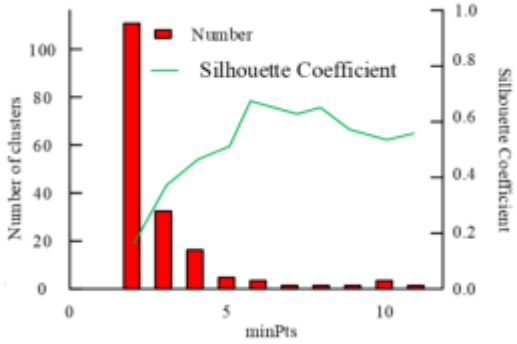
DS method is simple and easy to implement, but it is easy to be affected by outliers or outliers, and easy to increase the calculation amount. The standardized expression of Z-csore is shown in equation 4.4.

$$v_j = \frac{u_j - \bar{u}}{\sigma} \tag{4.4}$$

In equation 4.4, $\bar{u}$ and σ are the average and standard deviation of sequence U respectively. This method is applicable to the case of outliers in the feature data set. AB (Adaptive Boosting, AB) model is one of the most classical algorithms in Boosting model. It adopts the idea of joint decision to improve the classification accuracy. However, due to the same type of base classifier, the model still has the limitations of a single classifier in the learning process. The principle of AB model is studied, and a NA model based on multiple classifiers is established. NA models no longer use a single classifier as a base learner, but integrate multiple classifier models to avoid the problem that similar classifiers perform well in a certain aspect, where N represents the number of classifier models. In the training process, the base learner is composed of multiple classifier models. Each classifier model will learn and classify the training samples, and the obtained training results are decided by several classifier models using simple voting. During each iteration, the training sample data set is used to pass several classifier models and the model is fitted using the same weights. By integrating different classifier models,



((a)) Comparison of contour coefficients of two algorithms under different parameters



((b)) Values of D-DBSCAN algorithm under different minPts

Fig. 5.2: The Results of the Contour Coefficients of the Two Algorithms and the Values of D-Dbscan under Different

this model overcomes the classification limitations brought by a single learner and makes the performance of the classifier complementary. The flow of NA model is as follows: there is training data set G_j ; The number of iterations is N , and the weight distribution of the initial training sample is shown in equation 4.5.

$$\begin{cases} W_1 = (w_{1,1}, w_{1,2}, \ldots, w_{1,j}) \\ w_{1,j} = \frac{1}{N}, 1, 2, \ldots, N \end{cases} \tag{4.5}$$

Then construct learning algorithm Γ, which is composed of j classifiers $C_j(u)$. In the algorithm, $C_j(u)$ training data are used for classification prediction, and the classification results are counted. Return the final classification result to Γ through a simple voting algorithm. For $n = 1, 2, \ldots, N$, use the training data set with weight W_n to train the base Γ. Input the integrated classifier model $C(u)$ to get the weak classifier $R_n(u)$, see equation 4.6.

$$R_n(u) = \Gamma(G_j, W_n, C(u)) \tag{4.6}$$

Calculate the classification error rate of $R_n(u)$ for the training data set, and then calculate the weight of $R_n(u)$ in the strong classifier according to the classification error rate, and update the weight distribution of the training sample set. After iteration N of the above process, the final classifier result can be obtained, as shown in equation 4.7.

$$F(u) = \text{sign}\left(\sum_{j=1}^{N} \theta_n R_n(u)\right) \tag{4.7}$$

In equation 4.7, θ_n is the proportion of $R_n(u)$ in the strong classifier. According to the prediction results of the algorithm, students' abnormal grades are filtered.

5. Result Analysis of NA Model. To evaluate the clustering effect of the D-DBSCAN algorithm proposed in the study, and determine the optimal number of clusters. Silhouette Coefficent (SC) was selected for evaluation. SC can evaluate the cohesion and separation between sample data points at the same time, and can evaluate the clustering effect when the formal sample data set category is unknown. The research selects DBSCAN algorithm and D-DBSCAN algorithm for comparison, and obtains the results of the contour coefficients of the two algorithms and the values of D-DBSCAN under different *minPts* conditions, as shown in Figure 5.2. It can be seen from Figure 5.1(a) - that DBSCAN algorithm needs to constantly find the optimal

Table 5.1: Comparison Results of Operation Time of Two Algorithms

Algorithm	Total running time (s)	Total operation times/time	Total operation times/time
DBSCAN algorithm	2617.00	100	26.17
D-DBSCAN algorithm	2904.60	10	290.46

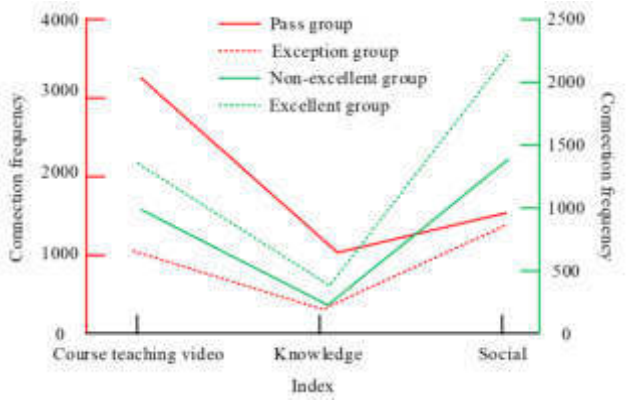


Fig. 5.3: The use of different groups in sports online education

solution of $minPts$ and χ. When $\chi = 0.1$, $minPts$ is any value, and the corresponding SC value is negative. It shows that the clustering effect is the worst at this time. When $\chi = 0.9$ and $minPts = 6$, the SC value is 0.643, and the parameter is the optimal solution. DBSCAN++represents that the algorithm only needs one parameter, so it only needs to find the optimal solution under different $minPts$. D-DBSCAN algorithm has better clustering effect than DBSCAN algorithm for other parameter values except . From Figure 5.1(b), when $minPts = 6$, the SC value is 0.711, which is the optimal solution of the D-DBSCAN algorithm. Compared with DBSCAN algorithm, D-DBSCAN clustering performance is improved by 10.6%, and the corresponding number of clusters is 4. The results show that when clustering students' behavior characteristics, the number of clusters is 4, and the corresponding SC value is 0.711, which is the best clustering effect.

Table 5.1 shows the comparison results of the operation time of the two algorithms. It can be seen from Table 2 that DBSCAN algorithm performs 100 operations on the two parameters and ; D-DBSCAN algorithm performs 10 operations on to obtain better clustering results. The average single run time of D-DBSCAN algorithm is 11.10 times that of DBSCAN algorithm, but its total run time is 1.11 times that of DBSCAN algorithm, and the calculation time is 9% longer than DBSCAN algorithm. The above results are due to the fact that compared to other algorithms, the D-DBSCAN algorithm proposed in the study can discover clusters with different shapes and adaptively select appropriate neighborhood radii, making it more suitable for analyzing complex academic behavior of students in universities.

See Figure 5.3 for the results of the use of different groups in sports online education. In terms of connection frequency between the abnormal group and the passing group, the passing students have more access to the network physical education curriculum resources than the abnormal students; In terms of social behavior in online classroom, the two groups of students visited the same. In the comparison between the excellent group and the non-excellent group, the excellent students have the highest connection frequency in the social aspects of sports online education class, and the connection frequency in sports knowledge is 1.7 times of the non-excellent students.

Figure 5.4 shows the results of online duration of online physical education courses for different groups of students each month. The students in the abnormal group generally spend less time in physical education courses than the students in the passing group. During the four months of learning, the online duration of

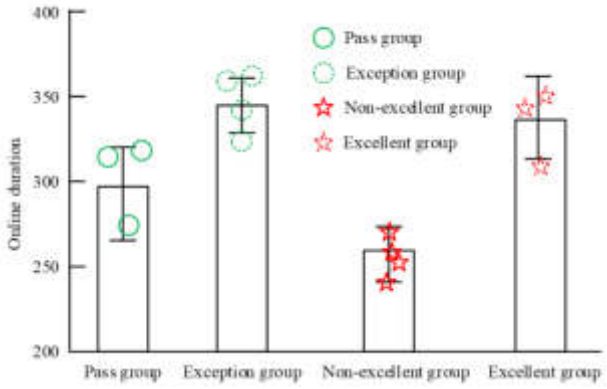


Fig. 5.4: Monthly Online Duration Results of Online Physical Education Courses For Students in Different Groups

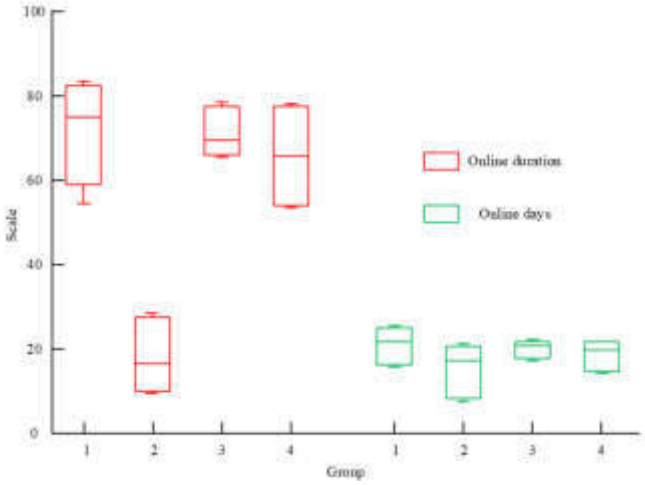
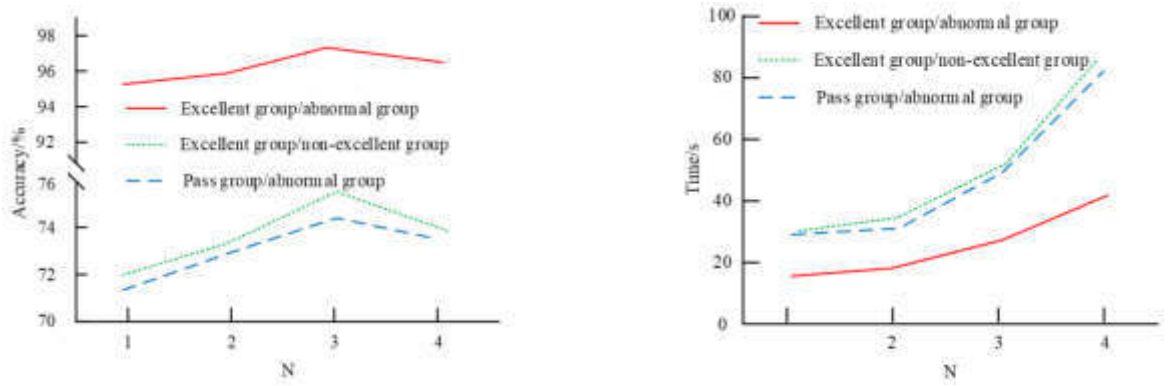


Fig. 5.5: Index Chart of Network Self-Study Of Different Groups Of Students in the Library

students in the abnormal group was more scattered, and the online duration of students in the abnormal group was significantly reduced near the test. The average online duration of the excellent group is the highest, and the distribution is more concentrated, and the overall performance of the semester is also more stable.

Figure 5.5 shows the results of relevant indicators of the self-study network of different groups of students in the library every month. 1-4 corresponds to the passing group, abnormal group, excellent group and non-excellent group respectively. The students in the abnormal group have significantly lower online duration and online days of self-study, which indicates that the students in the abnormal group will not have less time to self-study. The students in the excellent group showed lower variance in both indicators. It shows that the data of the excellent group converges, indicating that the behavior pattern of the excellent students is more fixed. From the overall performance analysis, the difference between the abnormal group and the qualified group is more obvious. Therefore, it is easier to find students with abnormal results in the test in the subsequent prediction.

To verify the performance of the model proposed in the study, the study used the tenfold cross-validation method for training evaluation. The experiment selects the average of accuracy and F1 value as the final result.



((a)) ((b))

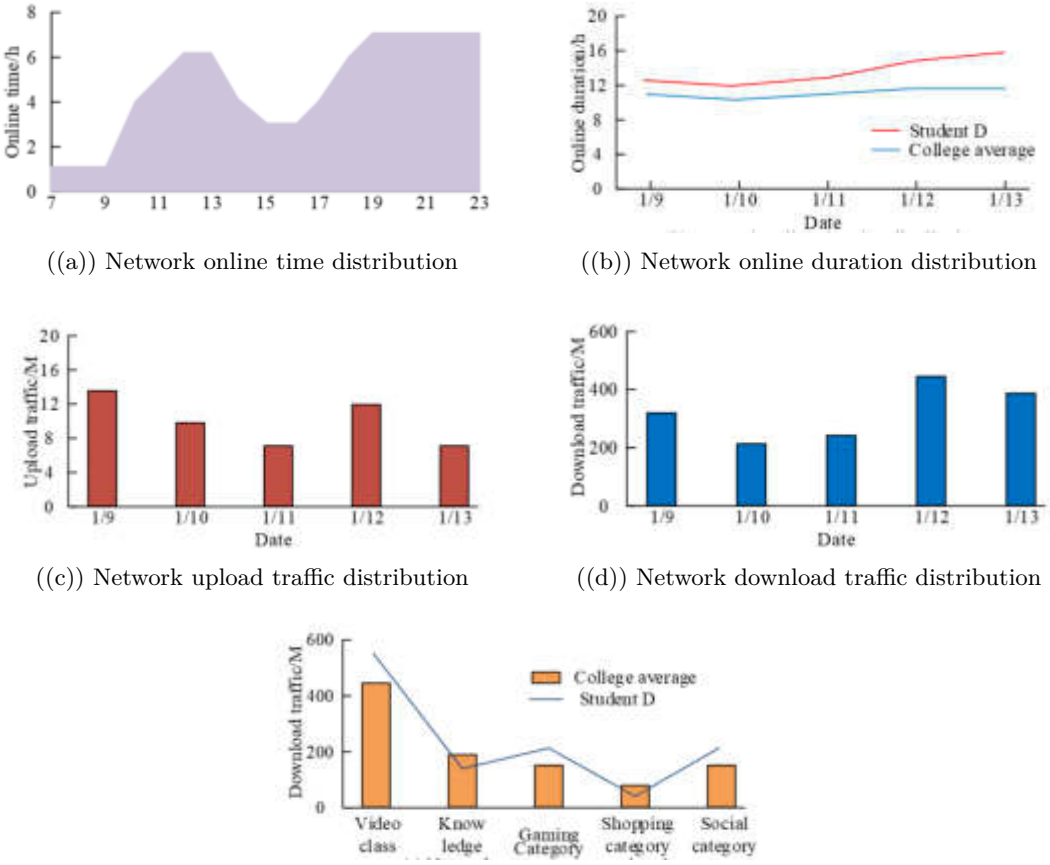
Fig. 5.7: Prediction Accuracy and Consumption Time Results of Each Group of Samples at Different N

Table 5.2: Prediction Results of Different Models for Each Group in Physical Education Network Teaching Test

Group	Model	Accuracy (%)	Precision (%)	Recall (%)	F1
Abnormal group/pass group	C4.5	66.34	79.89	68.30	0.743
	SVM	67.57	83.20	69.06	0.756
	IL	60.18	77.03	63.20	0.691
	NA	97.93	98.63	83.27	0.801
Abnormal group/excellent group	C4.5	72.83	70.42	75.63	0.743
	SVM	73.35	73.86	77.49	0.757
	IL	65.40	66.34	69.73	0.672
	NA	98.16	98.58	97.26	0.958

Figure 9 shows the prediction accuracy and consumption time results of each group of samples at different N. Figure 5.6(a)- shows that in the abnormal group and the qualified group, the excellent group and the non-excellent group, the prediction accuracy rate is the highest when N=3, but the integration rate is lower when compared with N=2. In the excellent group and the abnormal group, the prediction accuracy is the highest when N=3. Figure 5.6(b) shows that when N=3, it takes 50 seconds. However, combined with the accuracy curve, when N=2, the accuracy can be effectively improved in the case of similar time consumption. Therefore, select N=2, that is, the base classifier is composed of C4.5 model and SVM model, as the NA model. By comparing the results of multiple experiments mentioned above, the effectiveness of the NA model in predicting academic performance in online physical education teaching in universities can be determined. By studying the impact of the value of N in the NA model, a solid data foundation has been laid for predicting academic performance in a wider range of related universities in the future.

Table 5.2 shows the prediction results of different models of each group in the physical education network teaching test. Compared with the other three models, the accuracy of NA model was 97.93% for the students in the abnormal/passing group. The recall rate of SVM represents that 69.06% of the positive samples are accurately predicted; The NA model iteratively trains the base learner, so its accuracy is 31.59% and 30.36% higher than C4.5 model and SVM model. The F1 value of NA model is also the highest, 0.801. In the abnormal group/excellent group, the accuracy rate of the model was 98.16%; The recall rate was 97.26%; F1 value is 0.958. The research results show that the integrated learning algorithm performs better than the single prediction model in the prediction of sports network teaching results, and can accurately predict students' abnormal results. The above results are due to the fact that compared to a single prediction model, the NA



((a)) Network online time distribution

((b)) Network online duration distribution

((c)) Network upload traffic distribution

((d)) Network download traffic distribution

((e)) Network resource access situation

Fig. 5.9: Data Analysis and Score Risk Prediction Results for Student D

model proposed in the study is no longer limited to the use of a single algorithm. Instead, based on the characteristics exhibited by behavioral data, a NA model composed of multiple classifiers is used for prediction analysis, complementing the advantages of different classification models, improving the shortcomings of the model, and thereby improving the prediction accuracy of the model, which is better used for predicting grades in online physical education teaching in universities. To further validate the effectiveness of the NA model proposed in the study in practical applications, the study randomly selected student D from a certain university for analysis and score risk prediction.

The data analysis and performance risk prediction results of student D are shown in Figure 5.9. From Figure 5.9, it can be observed that the time distribution of student D's online presence during the day is maintained for a long time from 11:00 to 13:00 and from 19:00 to 23:00. And through its network distribution in the Xi'an market within a week, it was found that the student's average online duration was significantly higher than that of the college. In the online upload and download traffic of student D, it was found that the usage was highest on January 12th and lower on January 11th. Through the access to network resources, it was found that video accounted for the most at 23%, while shopping and knowledge accounted for the least at 2% and 7%, respectively. Based on the above analysis of results, the network usage of student D is high, the network viscosity is severe, and their life is irregular. The predicted result is a failure. Through the above analysis and risk warning, administrators can save the risk list of failing students for further offline communication and communication, improve their non-standard life behavior, and enhance their academic level. In summary, the

application of NA model in online teaching and testing of physical education in universities can more efficiently manage and communicate with different students, improve their academic performance, and provide higher quality students to society.

6. Conclusion. The development of education informatization has made a huge collection of education data. How to mine valuable information from a large number of data and accurately classify and predict the students with abnormal results in the physical education network teaching test is an important means to help managers make scientific decisions. This research was based on the feature library of student behavior portrait, and used D-DBSCAN algorithm for clustering analysis. The experiment constructed a multi-classification based NA model to predict the abnormal scores of students in college physical education network test. The experimental results showed that when, the SC value was 0.711, which was the optimal solution of D-DBSCAN algorithm. At this time, compared with DBSCAN algorithm, D-DBSCAN clustering performance was improved by 10.6%, and the corresponding number of clusters was 4. When N=2, that was, the base classifier of NA model was composed of C4.5 model and SVM model, the prediction accuracy and time consumption were the most appropriate. Compared with C4.5 model, SVM model and IL model, the accuracy, recall rate and F1 value of NA model were 98.16%, 97.26% and 0.958 respectively. To sum up, the NA model based on classifiers proposed in the study had better performance and could accurately predict the abnormal performance of students in college physical education network teaching test. However, there are still shortcomings in this study. For example, the amount of research data is not very large. With the development of high-performance computing technology, high-performance platforms can be used to build a distributed cluster environment in future research. And then realize the parallelization of student behavior data processing and calculation, and improve the operation efficiency of the overall model.

REFERENCES

[1] Vatsalan, D., Rakotoarivelo, T., Tyler, P., Ladjalet, D. & Bhaskar, R. Privacy risk quantification in education data using Markov model. *British Journal Of Educational Technology*. **53**, 804-821 (2022)

[2] Dhanalakshmi, R., Muthukumar, B. & Canessane, R. Analysis of Special Children Education Using Data Mining Approach. *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*. **30** pp. 125-140 (2022)

[3] Banerjee, D., Gidwani, C. & Rao, T. The role of "Attributions" in social psychology and their relevance in psychosocial health: A narrative review[J]. *Indian Journal Of Social Psychiatry*. **36**, 277-283 (2021)

[4] Qi, S. Approaches to Information Service and Management Construction in University Libraries. *International Journal Of Social Science And Education Research*. **2**, 41-45 (2019)

[5] Gj, A., Dw, A., Dy, A., Dha, B., Qd, A., Role, W. & Domain-Based, O. Access Control Model for Graduate Education Information System. *Procedia Computer Science*. **176** pp. 1241-1250 (2020)

[6] Wang, S., Zhang, F., Gong, Q., Bolati, D. & Ding, J. Research on PBL teaching of immunology based on network teaching platform. *Procedia Computer Science*. **183**, 750-753 (2021)

[7] Grin, N. INFORMATIZATION IN EDUCATION. *Scientific Papers Collection Of The Angarsk State Technical University.*, 348-351

[8] Sang, H. Analysis and Research of Psychological Education Based on Data Mining Technology. *Security And Communication Networks*. **2021**, 1-8 (2021)

[9] Yahya, A. & Osman, A. data-mining-based approach to informed decision-making in engineering education. *Computer Applications In Engineering Education*. **27**, 1402-1418 (2019)

[10] Kerzic, D., Aristovnik, A., Tomazevic, N. & Lan, U. Assessing the impact of students' activities in e-courses on learning outcomes: a data mining approach. *Interactive Technology And Smart Education*. **16**, 117-129 (2019)

[11] Doleck, T., Lemay, D., Basnet, R. & Bazelais, P. Predictive analytics in education: a comparison of deep learning frameworks. *Education And Information Technologies*. **25**, 1951-1963 (2020)

[12] Fan, J., Zhang, M., Sharma, A. & Kukkar, A. Data mining applications in university information management system development. *Journal Of Intelligent Systems*. **31**, 207-220 (2022)

[13] Ramanathan, K. & Thangavel, B. Minkowski Sommon Feature Map-based Densely Connected Deep Convolution Network with LSTM for academic performance prediction. *Concurrency And Computation Practice And Experience*. **33** pp. 4 (2021)

[14] Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D. & Khanna, A. CatBoost - An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance. *Advances In Data Science And Adaptive Analysis: Theory And Applications*. **13**, 1-21410 (2021)

[15] Ade, R. Students performance prediction using hybrid classifier technique in incremental learning. *International Journal Of Business Intelligence And Data Mining*. **15**, 173-189 (2019)

[16] Deepika, K., Relief-F, S. & Tree, B. Random Forest Based Feature Selection for Student Academic Performance Prediction. *International Journal Of Intelligent Engineering And Systems*. **12**, 30-39 (2019)

[17] Yusuf, A. & And, J. and synthetic minority oversampling techniques for academic performance prediction. *International Journal Of Informatics And Communication Technology (IJ-ICT)*. **8**, 122-127 (2019)
[18] Ns, A., Maa, B., Rb, C. & Hh, D. The effect of the virtual social network-based psycho-education on the hope of family caregivers of clients with severe mental disorders. *Archives Of Psychiatric Nursing*. **35**, 290-295 (2021)
[19] Envelope, D., Akbar, M., Bagaskara, A. & Vinarti, R. Improving classification algorithm on education dataset using hyper-parameter tuning - ScienceDirect. *Procedia Computer Science*. **197** pp. 538-544 (2022)
[20] An, Y. & Zhou, H. Short term effect evaluation model of rural energy construction revitalization based on ID3 decision tree algorithm. *Energy Reports*. **8** pp. 1004-1012 (2022)
[21] Sun, F. & Shi, G. Study on the application of big data techniques for the third-party logistics using novel support vector machine algorithm. *Journal Of Enterprise Information Management*. **35**, 1168-1184 (2022)
[22] Tan, F. & Xie, X. Recognition Technology of Athlete's Limb Movement Combined Based on the Integrated Learning Algorithm. *Journal Of Sensors*. **7557**, 1-30575 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Nov 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 299–311, DOI 10.12694/scpe.v25i1.2296

RESEARCH ON THE APPLICATION OF SPEECH DATABASE BASED ON EMOTIONAL FEATURE EXTRACTION IN INTERNATIONAL CHINESE EDUCATION AND TEACHING

XIANGLI ZHANG*

Abstract. The advanced analysis of the relationship between acoustic and emotional characteristics of speech signals can effectively improve the interactivity and intelligence of computers. Given the current status of speech recognition and the problems encountered in international Chinese education, the study proposes to extract emotional characteristics to achieve speech construction of the database. Based on considering the emotional characteristics of speech, a hybrid algorithm based on spectral sequence context features is proposed. The DBN-BP algorithm is used to process emotional data of different dimensions, and a speech database is constructed. After testing and analyzing the algorithm model, it is found that the dynamic recognition accuracy of the DBN-BP model fused with emotional features is over 90%, and the negative emotion recognition rates in the three databases are all above 60%. At the same time, the accuracy rate of the model in the algorithm comparison experiment remains above 85%, the data information extraction is relatively complete, and the average test time of less than 1s is less than 3%. The speech database based on multi-emotional feature extraction can effectively provide a new reference for the improvement of the quality of Chinese international education and the improvement of the speech recognition system.

Key words: Emotional characteristics; Speech database; Chinese education; International teaching; DBN-BP algorithm

1. Introduction. With the continuous development of artificial intelligence technology, speech recognition technology has become increasingly mature, and it has been applied in various fields such as medical services, remote education, transportation, etc. However, due to the complexity and differences of emotional activities, the current development of speech emotion recognition still faces significant limitations. At the same time, internal and external factors such as environmental noise interference, emotional fluctuations in the speaker, channel distortion in the speech library, native language habits, individual differences in the speaker, and language learning environment can have a significant impact on speech recognition and application effectiveness, and inevitably lead to language understanding ambiguity between the communication parties. In current international Chinese education and teaching, the communicative function of language teaching has not been given sufficient attention, and the breadth and depth of Chinese language and culture have further increased the difficulty of learning teaching aids. Murvey B scholar found that some foreign college students who went to China to study had different cognitive attitudes and life paths before and after graduation. They were often in ambivalence in the learning of cultural knowledge and language, and had poor initiative [1]. One of the key aspects of increasing the quality of international Chinese education is to make this cultural dissemination method more vivid. Building a voice database can effectively reduce teaching difficulty. The Kaur J scholar team described the automatic spectral speech recognition technology and studied the current development of tonal language [2]. Zehra W scholars introduced ensemble learning into cross corpus and multilingual emotion recognition, and proved that the mutual application of different corpus data can improve the accuracy of corpus data training [3]. Emotional features refer to a series of features in human speech that can reflect emotional states, including tone, volume, speed, intonation, etc. These features can identify the speaker's emotional state in speech, such as anger, joy, sadness, etc. And a speech database refers to a resource library that collects and stores a large amount of speech data, which contains pronunciation samples from different populations. Analyzing different speech samples in the speech database can extract features related to emotions, allowing computers to automatically recognize the speaker's emotional state. There are differences in acoustic feature patterns under different emotional states. Research suggests that there is a close relationship between emotional

*School of Chinese Language and Literature, Panzhihua University, Panzhihua 617000, China (xianglidream@sina.com)

features and speech databases. Therefore, to further improve the quality and effectiveness of Chinese education and teaching, the study starts from the dimension of emotional feature extraction to achieve the construction of speech databases.

2. Literature review. The emotional feature is a kind of information feature with diversity and complexity. Most scholars have studied its feature recognition. Among them, Fan X and other scholars have used the singular value non-solution algorithm after the subband division of the speech signal with the help of wavelet packet change. The dictionary training was performed on two different feature sets, real and false. The results denoted that the classification module integrating cepstral features and sparse decomposition expressed a recognition rate of more than 75% during the experiment, which further improved the ability of the Chinese deception detection system [4]. The machine team of Pan L scholars used a machine learning algorithm to construct a model of the English part of speech and eliminates the ambiguity in the recognition based on relevant rules and phrase structure. The experimental outcomes indicated that the classification algorithm had a good application effect [5]. Scholars such as Mnassri A used a genetic algorithm to optimize the parameters of the support vector machine, thereby improving the accuracy of speech recognition. The test findings indicated that the selected Arabic words could be effectively input after cepstrum processing, and could be used in the comparison. A high recognition rate could be achieved in a short training time [6]. The RNN transducer greatly simplified the automatic speech recognition system, but the realization of its training process was quite difficult. Based on this, scholars such as Wang S have used the learning rate decay strategy and added convolutional layers to improve the ability to understand Chinese [7]. Koduru A believed that paying attention to the extraction of speech signals could effectively understand its speech emotions. MFCC coefficients, zero-crossing algorithms, and global features were used to achieve feature extraction and information screening. Simulation findings expressed that the extraction algorithm could effectively improve happiness, etc. The general sentiment was extracted [8].

Scholars such as Kumaran U proposed to use deep C-RNN to realize the change of emotions in the classification stage, that is, to distinguish emotional features of different natures through the extraction of high-level spectral features and the learning of contextual features. The data loss value was small, and the accuracy rate of emotion recognition of speech signals was more than 80% [9]. Kerkeni L and his research team used the empirical pattern decomposition system to realize speech emotion recognition, that is, the signal was decomposed into feature modulation and emotion recognition to improve its classification performance. And the research outcomes illustrated that the system was supported by machine algorithms, in database verification. It had a recognition rate of more than 85% [10]. Aiming at the problem that the prosody and sound quality features were greatly affected by the signal-to-noise ratio (SNR) in the speech emotion recognition, Huang Y and other scholars proposed to use weighted ideas and deep belief networks to realize feature processing and fusion operations. The results showed that the feature learning structure could better reduce the interference problem of noisy environments and improve the accuracy and application performance of emotion recognition [11].

At the same time, Daneshfar F and other scholars used the QPSO algorithm to perform dimension reduction projection processing on the extracted high-dimensional rich features and improve the algorithm considering the classifier parameters. The research outcomes proved that the accuracy of the research system in the emotional speech database was better than other comparison algorithms [12]. Chen M and other scholars proposed a three-dimensional attention convolutional recurrent neural network to distinguish SER features, which reduced the interference of other irrelevant factors based on retaining information and emotional features. The experimental findings indicated that the method had high application effectiveness, and the recall rate was high [14]. Scholars such as Kwon S proposed that the deep convolutional network of INCA performed the most characteristic prediction, and collected and processed the data from the extraction of spectral and spatial domain features. The experiment outcomes expressed that the prediction system under the classifier performed more than 80% recognition rate, with good application effectiveness [15]. There were many types of research on feature algorithms for emotion recognition, but they were rarely applied in international Chinese teaching. Scholars such as Widodo HP believed that under the current globalization trend, Chinese teachers should pay attention to the construction and negotiation of their professional identity [13]. Yuan R and other scholars analyzed the cognition of college students' international courses with a new perspective of identity. The experimental findings illustrated that participants' positioning of themselves often fell into the paradox

of personal roles and social roles [17]. Xinhan N scholar started with the research on student management teaching systems, constructed an intelligent analysis system based on neural network technology and emotion feature recognition algorithms, and designed a relevant scale evaluation index system using machine learning methods. The results denoted that the proposed model had good classroom application effectiveness [18]. Hu Jingchao, a scholar, combined deep learning with HMM feature algorithms to design a teaching state detection system, and completed the construction of recognition models through the collection and processing of subjective evaluation data and feature discretization. The outcomes expressed that this algorithm could effectively recognize student state features, with a recognition accuracy of over 90% [19]. Byun S W scholars used recursive neural network models to extract emotional recognition features and classify emotions from different aspects of acoustic features. The findings indicated that the accuracy of the designed system exceeded 85%, and its applicability was good [20]. Shah V scholars believed that introducing machine learning algorithms into text data analysis could effectively identify emotional states contained in information data [21].

3. Research on the construction of a speech database based on emotional features.

3.1. Extraction algorithm based on spectral sequence context mixed features. Feature extraction is an important step in speech emotion recognition and database establishment. The acoustic parameters contained in the speech signal are the main distinguishing points of different speech features. Generally, the validity and difference of the feature set are processed with generation and evaluation modules. The emotional acoustic features of speech are less likely to fluctuate due to differences in expression methods, which are largely related to the emotional attitude and emotional fluctuations presented by the speech. When different people express the same language meaning, they may unconsciously reveal individual emotional tendencies due to their own language habits and personal preferences, which are reflected in speech acoustic features such as time-frequency domain and cepstrum features. Time domain features refer to the features exhibited by speech signals within a certain time range after windowing processing. When the time domain waveform of a single frame signal crosses the time axis and causes different changes in adjacent sampling values, the speech signal exhibits high and low-frequency features. The number of changes is positively correlated with the frequency. The common time-domain features include short-time energy, average amplitude, autocorrelation, and so on. It is difficult to estimate the period of short-time autocorrelation due to large amount of calculation and long time consumption, and it is difficult to determine the appropriate size of the window length. Therefore, the study uses the short-term average amplitude difference function to calculate the period, and the calculation formula is shown in equation 3.1 [22].

$$F_n(k) = \sum_{m=N-1-k}^{N-1} |x_n(m) - x_n(m+k)| \tag{3.1}$$

In equation 3.1, $x_n(m)$ is the voice signal; $x_n(m+k)$ is the maximum delay point; N means the time; m represents frame shift. The frequency domain feature reflects the eployment of signal energy in different frequency bands, and can reflect the overall periodic performance of the signal. Part of the formula is shown in equation 3.2.

$$\begin{array}{c} S_f = \sum_{n-1}^{N}(A_i(n) - A_{i-1}(n))^2 \\ \sum_{n-1}^{S_r} A(n) = \frac{17}{20}\sum_{n-1}^{N} A(n) \end{array} \tag{3.2}$$

In equation 3.2, S_f, S_r is the spectrum transition parameter and the spectrum cutoff parameter; $(A_i(n) - A_{i-1}(n))$ denotes the current amplitude spectrum of the frame number and the previous amplitude spectrum; n is the number of spectral lines [23].

$$C(n) = \mathcal{F}^{-1}\left(\ln\left|\mathcal{F}\left(x_n(m)\right)\right|\right) \tag{3.3}$$

Equation 3.3 is the cepstrum characteristic parameter, in which $F(), F^{-1()}$ respectively represent the forward and inverse changes of Fourier, and $|\mathcal{F}(x_n(m))|$ denotes the real part of the complex number [24]. Speech signals are often continuous and whole, and the intonation and emotion between the previous frame and the

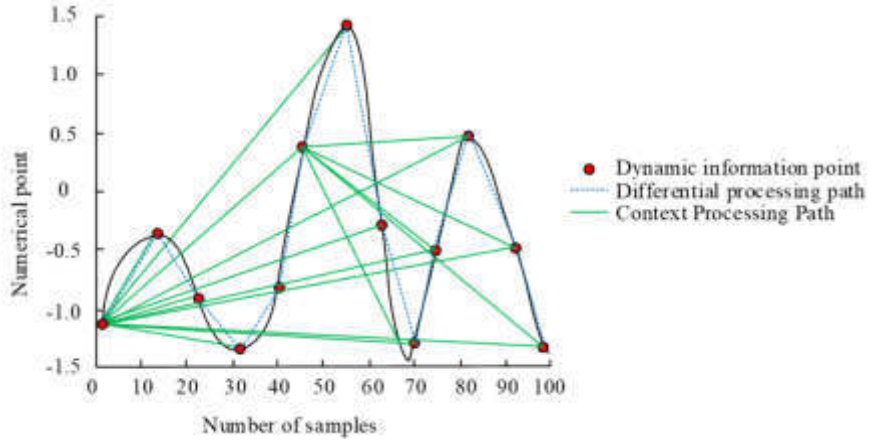


Fig. 3.1: Dynamic information processing mode

next frame are mutually influenced and run through the whole speech sequence. Therefore, the study proposes a feature extraction based on the spectrum sequence context feature (SSC). Algorithms are utilized to strengthen the grasp of dynamic correlation information between all frames. Figure 3.1 shows two information processing modes.

The contextual processing in Figure 3.1 can get the relevant dynamic information between all the frames better, and it ensures the dynamic spectral information and reduces the loss of information compared with the traditional differential processing of time series. The difference distance between the spectral sequence frames are calculated, as shown in equation 3.4.

$$\begin{gathered} D_{pq} = c_q - c_p, p, q = 1, 2, \ldots, M \\ Q = \begin{bmatrix} 0 & D_{12} & D_{13} & \ldots & D_{1M} \\ D_{12} & 0 & D_{23} & \ldots & D_{2M} \\ D_{13} & D_{32} & 0 & \ldots & D_{3M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ D_{1N} & D_{2M} & D_{3M} & \ldots & 0 \end{bmatrix} \end{gathered} \tag{3.4}$$

In equation 3.4, N means the frame index; p, q are the spectral frame index; c_q, c_p express the spectral sequence of the corresponding frame; $D_{i,j}$ denotes the difference value between two frames; Q refers to the order vector matrix. Then, the average value of the feature set and the distance between the feature spectrum and the average spectrum are obtained, as shown in the equation 3.5.

$$\begin{aligned} C_{\text{avs}} &= \textstyle\sum_{i=1}^{N} \frac{c_p}{N} \\ \text{diag}_p &= C_p - C_{\text{avs}}, \quad p = 1, 2, \ldots, n \\ F_m &= \begin{cases} S_{pq} + \text{diag}_p, & \text{if } p = q \\ S_{pq}, & \text{if } p \neq q \end{cases} \end{aligned} \tag{3.5}$$

In equation 3.5, C_{avs} indicates the average value; *diag* expresses the distance; S stands for the difference matrix; *diag* refers to the spectral center difference; F_m means the fused feature matrix. Figure 3.2 is a schematic diagram of a spectral context feature extraction process.

In Figure 3.2, the input speech signal is firstly processed by adding windows and splitting frames. The speech signal data is collected in one segment. To ensure the batch processing of the data by the programme, it needs to be transformed into the programmed data structure according to the specified length, i.e., subframe. At the same time, the signal processing requirements for continuous conditions. If the signal is disconnected during the subframe processing, it is necessary to add windows to the subframe data to better ensure the continuity of

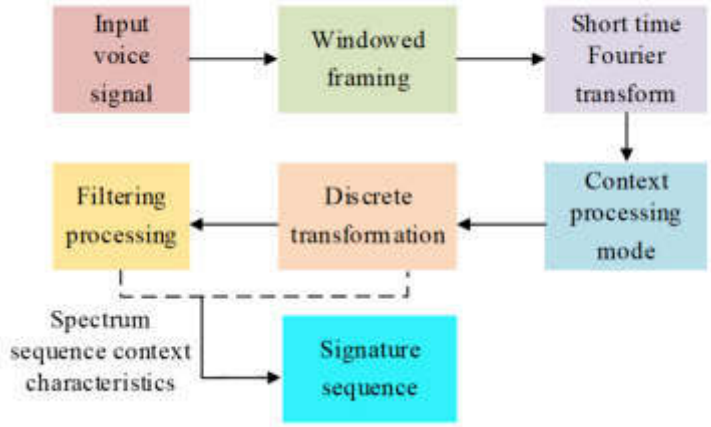


Fig. 3.2: Schematic diagram of spectrum context feature extraction process

the signal data. Subsequently, the processed speech signal data is subjected to short-time Fourier transform to better obtain the relationship between the time domain and the frequency domain. And the transformed signal is subjected to frame Gammotone filtering results and power-law compression, DCT transform, and then the set of spectral features is obtained. The set of spectral features is subjected to discrete cosine transform and context processing to obtain the context features. The different distances between the context features can be used as the basis for their differential processing, and the distance between each frame and the average value can be calculated to obtain the feature sequence needed for the study.

The rotation-invariant performance of contextual features can reduce the impact of the complexity of the vocal system and the diversity of speech content on speech recognition in complex background environments. The dimensional feature data extraction is performed on the context features by spatiotemporal Gabor filtering, that is, the temporal modulation filter is used as a row vector, and the frequency domain modulation filter is represented as a column vector, and is convolved with the feature channel and frame, respectively. The calculation method is shown as equation 3.6.

$$\text{filtcrcdSSC}(g,z) = \sum_{i,j} \text{SSC}(g-i, z-j) \cdot \text{filtcrfunction}(i', j') \tag{3.6}$$

In equation 3.6, g and z are the spectral and time indices of i', j', respectively, indicating the relative center offset of the frequency spectrum and time. Figure 3.3 shows the framework of the speech recognition system.

In Figure 3.3, the recognition system for speech signals mainly includes three parts: preprocessing, feature extraction and classification and recognition, in which the similarity comparison index is needed to differentiate in signal recognition. The high and low-frequency speech signals can reflect the high and low interest in the emotional content of the speech, and the long-term variation of prosodic speech can also represent the emotional difference of speech. The study introduces the prosodic feature into the statistical function to realize the transformation of the feature vector while ensuring its usability in the classifier while reducing the recognition complexity. By mixing all feature combinations and generalizing the acoustic properties of emotional speech, the most robust speech emotion feature representation set can be extracted, as shown in Figure 3.4.

Figure 3.4 shows the hybrid feature combination of speech emotion. MFCC, rhythmic and SSC features in the speech data have their own unique feature data, among which SSC features can better extract the differences in speech emotion, thus avoiding the extraction errors caused by the differences in language styles and sentence lengths.

3.2. Design of emotional speech database based on DBN - BP algorithm. Emotional features in complex environments will be affected by subjective emotional styles and relatively vague emotional demarcation

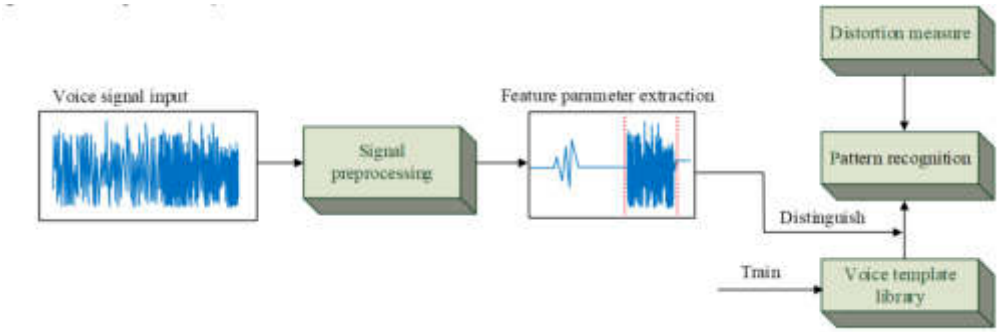


Fig. 3.3: Voice recognition system framework

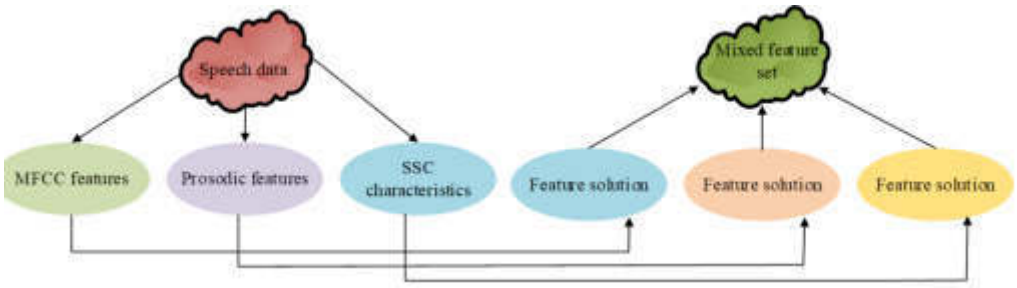


Fig. 3.4: Schematic diagram of mixed features for speech emotion extraction

points, which makes it difficult for classifiers to identify. Contextual and acoustic prosody features are highly subjective in algorithm testing. Therefore, on this basis, the deep learning algorithm is introduced to represent the speech signal at multiple levels, and the characteristic parameter factors with high robustness are extracted. Deep belief networks (DBN) are based on restricted Boltzmann machines and are trained to deal with the correlation between different hidden layers with a constructed joint distribution function. DBN includes an explicit layer responsible for data transmission and a hidden layer that adjusts the weight assignment of the data and is often trained with the " contrast divergence " algorithm, whose mathematical expression is shown in equation 3.7.

$$\begin{aligned} P(v|h) &= \prod_{i=1}^{d} P(v_i|h) \\ P(h|v) &= \prod_{j=1}^{q} P(h_j|v) \end{aligned} \tag{3.7}$$

In equation 3.7, d, q are the neurons in the explicit and hidden layers; v and h are the state vectors corresponding to the visible and hidden layers, respectively. The updated formula of the connection weight is shown in equation 3.8.

$$\Delta w = \eta(vh^T - v'h'^T) \tag{3.8}$$

In equation 3.8, T means transposition and η denotes connection parameters. DBN is stacked and connected by multiple Boltzmann machines and effectively trained with layers. After the pre-training of each layer is completed, the whole network is trained with back propagation (BP) neural network algorithm, and a deep network model is obtained. BP algorithm realizes nonlinear transformation and learning of sample data using gradient and iterative algorithms, and its mathematical expression is shown in equation 3.9.

$$\begin{aligned} u_j &= \textstyle\sum_{i=1}^{M}(\omega_{ij}x_i - \theta_j) \\ y_j &= f(u_j) = \frac{1}{1+e^{-u_j}} \end{aligned} \tag{3.9}$$

The input value and the corresponding input signal are classified. If the corresponding input signal belongs to the input value category, it is expressed as ± 1, and if not, it is 0. In equation 3.9, $x(x = 1, 2, ..., N$ is the input vector; j denotes the network node; u_j is the weighted sum of the node j threshold and the input value θ_j . The network node weights and thresholds are corrected to obtain the equation 3.10.

$$\begin{aligned} \omega_{ij}(t+1) &= \omega_{ij}(t) + \lambda \sigma_j x_i \\ \theta_j(t+1) &= \theta_j(t) + \lambda \sigma_j \end{aligned} \tag{3.10}$$

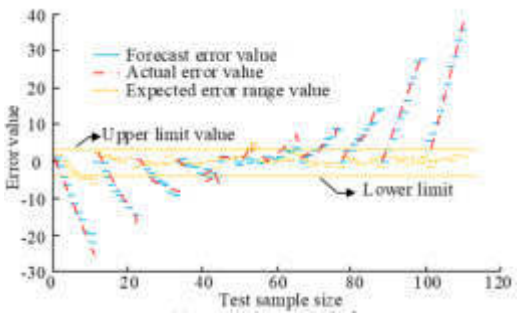
In equation 3.10, ω_{ij} is the weight x_i from node x_i to node i at a time t ; j indicates the input of the $i-th$ node; λ expresses the gain factor. Depending on whether the ideal value is clear, the value can be expressed as equation 3.11.

$$\begin{aligned} \sigma_j &= y_j(1-y_j)(d_j - y_j) \\ \sigma_j &= x_i(1-x_i)\textstyle\sum_k \sigma_k W_{jl} \end{aligned} \tag{3.11}$$

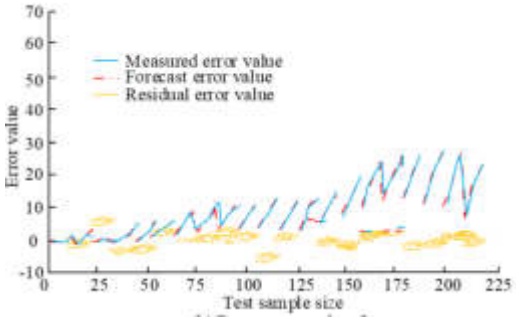
In equation 3.11, (d_j, y_j) mean the ideal output and actual output of the output node j ; l is the total number of nodes in the upper layer of the hidden layer node . When $\omega_{ij}\theta_j$ are in a steady state, the algorithm ends. Emotion recognition, as a unique feature recognition of human beings, is highly subjective, social, and cultural. Only when the two communicating parties show roughly the same emotional ups and downs, can they have the same voice characteristics. Strengthening the establishment of a voice database can effectively promote research on the characteristics of emotional data. The establishment of a voice database needs to adhere to the principles of authenticity, interactivity, continuity, and richness. Research is focused on the construction of a corpus database using specific sentence recordings and editing of related emotional video data. Taking into account the different types of emotional speech, database construction is implemented through database creation, data table definition, and data import calculation.

4. Analysis of the application results of the speech database for emotional feature extraction. The study proposed the construction of a speech database supported by hybrid algorithm based on SSC features and DBN-BP algorithm to realise the revelation of the effect of Chinese teaching on the basis of considering the characteristics of the phonological sentiment. The most important thing in Chinese teaching was to master the semantic expression between different information, in which emotion was an important acoustic feature. SSC features could be collected on the dynamic mutual information of speech signal data, while the DBN-BP algorithm extracted emotional features from speech using deep belief network and BP algorithm. This was achieved through emotional speech sample data collection, emotional feature extraction, labelling and classification of emotional categories and data storage of emotional feature information. In this study, the corpus database construction was carried out with the clips of specific utterance recordings and related emotional video data. The database construction was realized from the database creation-data table definition and data import calculation, taking into account the different types of emotional speech. The construction of emotional feature database could provide rich speech resources for international Chinese teaching, including practice materials for pronunciation, intonation and emotional expression. And through the speech database, emotional features could be used to assess and correct learners' pronunciation, helping learners to pronounce more accurately. For example, by comparing the learner's pronunciation with the standard pronunciation in the database, the learner could understand and improve his/her own pronunciation for the pronunciation characteristics of a particular emotional state. At the same time, the emotional speech samples in the database could be used to demonstrate and practice the characteristics of intonation and emotional expression in international Chinese language teaching, and learners could improve their intonation and emotional expression by imitating the emotional speech samples in the database.

4.1. Performance test of algorithm model based on emotional features. The experimental environment was designed as follows: the central processor was Intel core i5-6500; the deep learning framework was Caffe; the interface was MATLAB; the computer memory size was 12GB; the programming language was Python. The setting of iteration times was determined based on the test and training datasets in the voice database. When testing the algorithm, it set the learning rate and the maximum number of iteration steps to 0.001 and 600, respectively, and conducted training and test analysis on the data in the speech database to

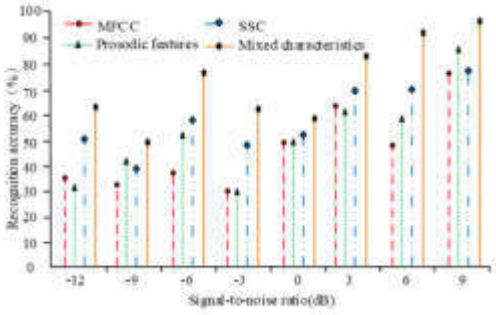


((a)) Data error results before neural network compensation

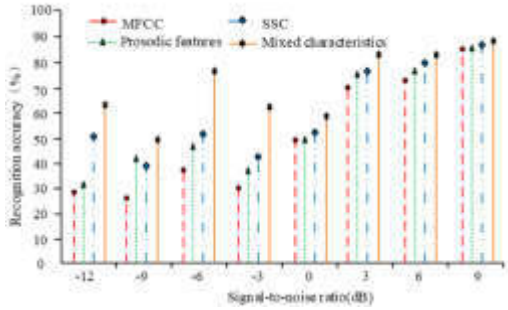


((b)) Data error results after neural network compensation

Fig. 4.2: Error results of data extraction before and after BP neural network compensation



((a)) Differential recognition result of dynamic information



((b)) Results of feature recognition accuracy under different signal-to-noise ratios

Fig. 4.4: Dynamic information recognition of different speech features and comparison of recognition accuracy results under SNR

better test the feasibility and applicability of the proposed algorithm. Figure 4.2 is an analysis of the error results of data extraction before and after adding BP neural network.

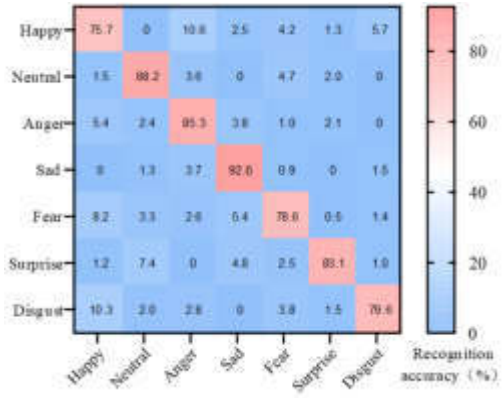
Figure 4.2 shows that the differences in the data error results were obvious before and after the data compensation with the aid of the neural network. The specific performance was that in Figure 4.2.a, the predicted and the actual error value curves shown by the proposed algorithm when extracting data features were roughly the same. The curve error fluctuation range under a small sample size was 0.38%, and the positive and negative error values were divided into two parts with the sample data volume 40 as the dividing point. The maximum prediction error value was -22 and 34, and the expected error range was between (-4, 2). The characteristic curve in Figure 4.2.b showed that the variation range between the measurement error and the prediction error was 0.042%, and the residual error value was lower than 0. The overall value was less affected by the change in the sample size, which effectively realized that the extraction of speech features ensured the accuracy of the algorithm to a certain extent. At the same time, considering that the extraction of speech signal features was more likely to be influenced by the external and objective environment and other factors, resulting in the generation of noise, the test data of the proposed feature fusion method and single speech feature extraction were compared, and the results are shown in Figure 4.4.

From Figure 4.4.a, the information accuracy rates of different speech features in differential dynamic recognition were different. When the SNR was negative, the information recognition rates from low to high were MFCC> Prosodic features>SSC>Mixed characteristics, and the smaller the SNR value, the more obvious the difference in feature information. Among them, the speech information extraction effect of fusion features was significantly higher than that of cepstral coefficient, prosodic, and SSC features. The accuracy differences at 6dB were 41.2%, 32.6%, and 19.%, respectively. And its average recognition accuracy was higher than the other three feature algorithms, and with the increase of the SNR value, the difference amplitude value decreased. The average recognition accuracy of the hybrid feature algorithm was above 94%, which was higher than that of MFCC (70.8%), Prosodic (81.3%), and SSC features (88.9%), and the maximum improvement rate exceeded 20%. The above results showed that the study of speech signal recognition from the perspective of fusion features could effectively improve its anti-noise interference ability, and had better accuracy and stability. A certain number of datasets were selected from emotional corpora in three different languages. The emotional tags expressed in the datasets were extracted, and 7 types of tags with different emotional attributes were obtained, namely happy, neutral, angry, sad, fear, surprise, and disgust were compared for recognition rates. In Figure 4.6.a - Figure 4.6.c, the languages of the three datasets were German, Chinese, and English respectively. The learning rate of the DBN network was set to 0.08 and the number of hidden layer nodes was 8000 to obtain the emotional feature recognition under different datasets, as shown in Figure 4.6.

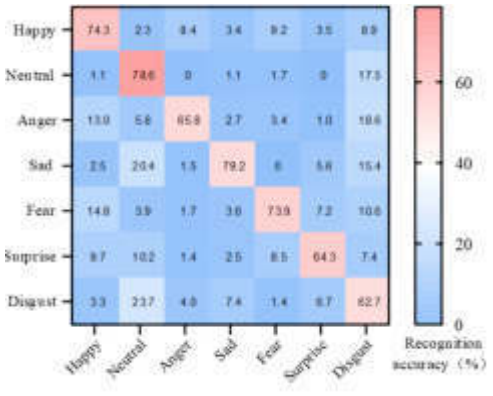
Figure 4.6 is the accuracy confusion matrix of different sentiment label classifications, in which the row and column represent the actual and the predicted values, respectively. In the case of Figure 4.6.a dataset 1, the algorithm's recognition accuracy for 7 different sentiment attributes was 75.7%, 88.2%, 85.3%, 92.6%, 78.6%, 83.1%, 79.6%, respectively. The difference between the recognition rates of angry and happy features exceeded 15%, and the system's recognition effect of emotional features on negative attributes was always high in positive emotions. In Figure 4.6.b, under the Chinese data set, the emotion recognition rates of the algorithms were all above 60%, among which the extreme values of anger and surprise were 79.2% and 64.3%, respectively. Compared with dataset 1, the rate was worse, but its overall control over the recognition of emotional features was better. In Figure 4.6.c, the algorithm showed a recognition rate of 75% on the anger emotion feature, and the difference in recognition accuracy between the fear emotion attribute and the disgust attribute was 17.8% and 17.1%, respectively. The recognition rate of the algorithm in different datasets was not the same. The reason was that the emotional characteristics presented by different databases were related to a certain cultural background, so the recognition effects were not the same. It had good performance in emotion recognition, especially in the recognition and classification of negative emotion features.

4.2. Applicability test of algorithms based on emotional features. Paying attention to the accuracy of language information transmission and the sufficient performance of emotional expression in international Chinese teaching has effectively helped teachers improve the quality of teaching management, and to a certain extent has greatly improved teaching effectiveness. The speech data for emotion recognition proposed by the research was used to study the effect of application recognition and was combined with the BP algorithm, support vector machine algorithm (SVM), long short-term memory network (LSTM), bidirectional long short-term memory network (BILSTM). The joint attention mechanism, bidirectional long short-term memory-attention (BILSTM-Attention), was compared, and the results are shown in Table 4.1.

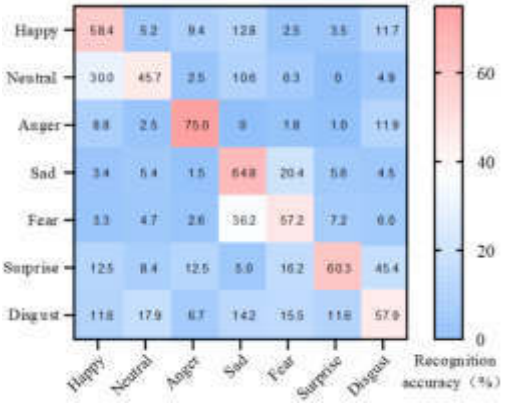
As shown in Table 4.1, the accuracy and recall data changes of different model algorithms under different datasets were different. The specific performance was as follows: the accuracy of the BP model under the three data sets was lower than 75%. The accuracy of the BP model decreased with the increase of the difficulty of the information covered by the data, and its performance was inferior to other comparison algorithms. The accuracy rates of the single model SVM and LSTM on the simple and medium difficulty datasets were 76.37%, 81.15% and 71.24%, 78.97%, respectively, and they also showed a certain drop in accuracy. The reason was that some informationwas missing, which in turn led to a decrease in the recognition accuracy of emotional information. The accuracy rates of the BILSTM model with the attention mechanism were 83.26% and 82.15% under the medium difficulty dataset, and the corresponding F1 values were 83.28 and 82.38. The overall performance of the data recognition was better, and its accuracy rate was only decreased by only 0.57%, significantly less than the other two single models. However, there was still a certain gap in the recognition accuracy of the DBN-BP model combined with the emotional feature analysis proposed in the study. The accuracy and recall rate of



((a)) Recognition results of emotion classification under dataset 1



((b)) Recognition results of emotion classification under dataset 2



((c)) Recognition results of emotion classification under dataset 3

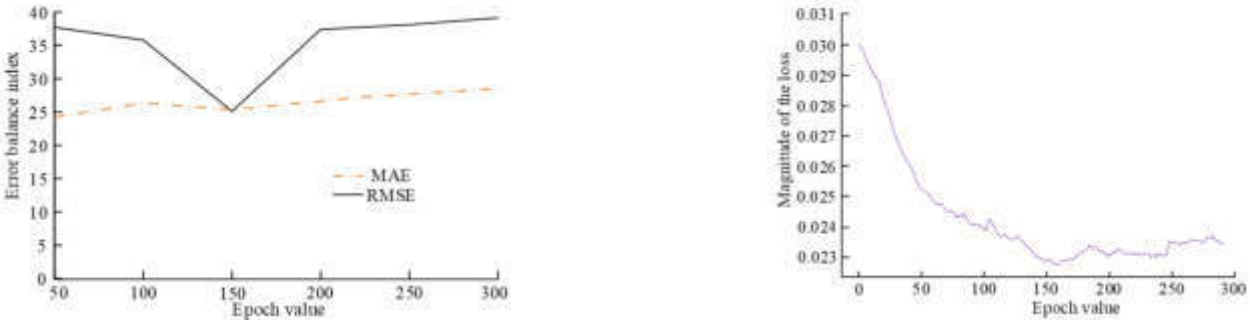
Fig. 4.6: Accuracy matrix results of emotional features on different datasets

the DBN-BP hybrid model was less affected by the difficulty of the sample data and remained at 85%. % and 86%. Its F1 values under the three data sets were 88.02, 87.65, and 86.53, respectively, and the performance of the algorithm was relatively stable. The above results indicated that the fusion DBN-BP model could effectively identify and process the features of emotional data, and its performance and application accuracy were good. The reason was that the model performed multi-dimensional processing and recognition based on the characteristics of emotional information, which overcame the problem of missing information data by a single algorithm. Then, the error analysis was carried out with the more difficult data set. Each training time was equivalent to one batch (Epoch). Multiple groups of Epochs were set to perform data statistics on the loss function results of the fusion algorithm. The results are shown in Figure 4.8.

In Figure 4.8, the evaluation index of the fusion model showed a trend of first decreasing and then increasing, and its root mean square error (RMSE) and mean absolute error (MAE) curve values tended to converge and stabilize with the increase of training batches. At the same time, the loss function value of the model algorithm showed a downward trend, and its value was lower than 0.024 in the later stage of training. The loss of data was small, and the feature retention of information data was better. The application results of the proposed model integrating emotional features and its running and test time in multiple experiments were analyzed. The average value multiple times was taken as the final result, and its emotional information extraction was

Table 4.1: Statistical results of data processing performance under different algorithms

Dataset classification	Model	Accuracy (%)	Recall rate (%)	F1
Low-difficulty data set	BP	73.25	77.13	74.38
	SVM	76.37	79.42	78.11
	LSTM	81.15	80.39	80.61
	BILSTM-Attention	83.26	84.22	83.28
	DBN-BP under emotional characteristics	85.33	87.15	88.02
Medium difficulty dataset	BP	68.22	69.13	68.34
	SVM	71.24	74.26	73.21
	LSTM	78.97	78.22	79.01
	BILSTM-Attention	82.15	81.33	82.38
	DBN-BP under emotional characteristics	86.02	88.14	87.65
Difficult data set	BP	65.23	64.31	64.08
	SVM	71.35	73.29	72.24
	LSTM	75.32	78.16	79.37
	BILSTM-Attention	79.20	81.24	80.09
	DBN-BP under emotional characteristics	85.16	86.32	86.53



((a)) Training batch experiment results

((b)) Experimental results of loss function

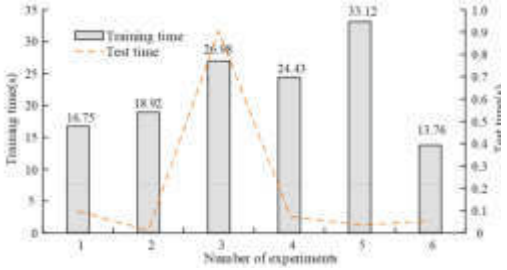
Fig. 4.8: Statistics of training batches and loss functions of mixed DBN-BP model

analyzed. The results are shown in Figure 4.10.

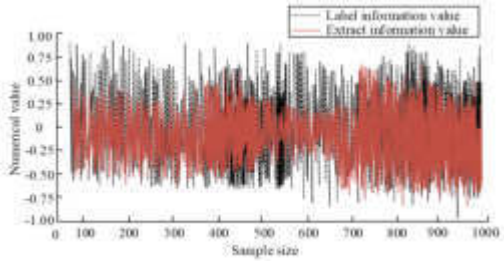
From Figure 4.10.a, the training time of the hybrid model under a different number of experiments was less than 1 minute, the average training time was 22.32 s. The test time of the model was less than 1s, and the maximum value was 0.93 s. The average test time was 0.64 s, and the overall test time was relatively low. It was stable when the number of experiments was greater than 4. The results in Figure 4.10.b showed that the number of samples would not cause a great interference with the performance of the model to extract information, and the information data contained in the information extraction value was basically in the label value data, and to a certain extent, the extreme value was reduced. The accuracy of identification information exceeded 86%, and the performance was good. At the same time, the validity of the algorithm was tested with the recording data of an international Chinese classroom in a university, and compared with the actual classroom emotional performance. The results are shown in Figure 4.11.

As shown in Figure 4.11, the difference between the predicted value of the model application and the real value in the four dimensions of positive, negative, neutral, and extreme emotions was small, which were 2.24%, 0.27%, 0.56%, and 0.24% respectively. The information prediction effect of emotional traits was better.

5. Conclusion. Strengthening speech emotion recognition is an important means to accelerate the promotion of intelligent human-computer interaction, which focuses on the emotional characteristics of speech data



((a)) Training time and testing time of fusion model under different sample sizes



((b)) Extraction of emotional feature information by fusion model

Fig. 4.10: Application time consumption of fusion model and feature extraction of emotional information

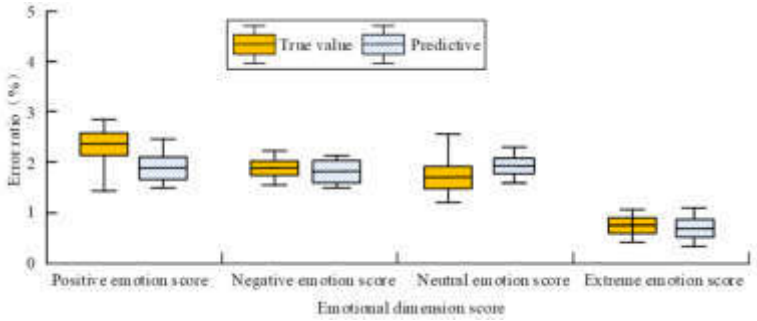


Fig. 4.11: Error comparison results between the predicted value and real value of the fusion model

in the current international Chinese teaching and effectively improves its teaching quality. The main idea of the research was to use the DBN-BP algorithm to extract emotional features, and build a speech database. After testing, the proposed algorithm showed a difference between measurement error and prediction error when extracting data features. The range of change was 0.042%, and the accuracy of dynamic identification information was higher than that of cepstral coefficient (70.8%), prosodic (81.3%) and SSC features (88.9%). At the same time, the recognition rate of different emotions of the DBN-BP model in the Chinese database was above 60%, the overall recognition control was good, and the accuracy and recall rate shown in the comparison with other algorithms were less affected by samples. The influence of the difficulty of the data remained above 85% and 86%, and the F1 values of the corresponding data sets were 88.02, 87.65, and 86.53, respectively, which were higher than the performance of other algorithms of the same dimension. The value of the model proposed was lower than 0.024 in the later stage of training, and the feature retention of information data was better. The average training time and test time were 22.32s and 0.64s, and the accuracy rate of emotional label recognition information exceeded 86%. The difference in scores on all four emotional dimensions was less than 3 percent. Focusing on emotional feature extraction can effectively improve the accuracy and applicability of speech recognition, and strengthening the multi-dimensional inspection of data is one of the ideas for future research and improvement.

Fundings. The research is supported by the Key Research Base of Humanities and Social Sciences of Sichuan Province-Sichuan International Education Development Research Center“Research on University Alliance Serving the Construction of the ‘the Belt and Road’(SCGJ2022-20); Foundation project: Panzhihua University 2022 school-level research project: Ideas and opportunities for the development of Chinese international education in sino-foreign cultural exchanges.

REFERENCES

[1] Mulvey, B. International higher education and public diplomacy: A case study of Ugandan graduates from Chinese universities. *Higher Education Policy*. **33**, 459-477 (2020)

[2] Kaur, J., Singh, A. & Kadyan, V. Automatic speech recognition system for tonal languages: State-of-the-art survey. *Archives Of Computational Methods In Engineering*. **28**, 1039-1068 (2021)

[3] Zehra, W., Javed, A., Jalil, Z. & Others Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*. **7**, 1845-1854 (2021)

[4] Fan, X., Zhao, H., Chen, X. & Others Deceptive Chinese speech detection based on sparse decomposition of cepstral feature. *Chinese Journal Of Acoustics*. **38** pp. 01 (2019)

[5] Pan, L., Hu, L. & Li, Z. Simulation of English part-of-speech recognition based on machine learning prediction algorithm. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **40**, 2409-2419 (2021)

[6] Mnassri, A., Cherif, A. & Bennasr, M. Algorithm Optimizing SVM Multi-Class Kernel Parameters Applied in Arabic Speech Recognition. *International Journal Of Systems Signal Control & Engineering Applications*. **12**, 85-92 (2019)

[7] Wang, S., Zhou, P., Chen, W. & Others Exploring run-transducer for Chinese speech recognition//2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). *IEEE*. pp. 1364-1369 (2019)

[8] Koduru, A., Valiveti, H. & Budati, A. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal Of Speech Technology*. **23**, 45-55 (2020)

[9] Kumaran, U., Radha Rammohan, S., Nagarajan, S. & Others Fusion of Mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. *International Journal Of Speech Technology*. **24**, 303-314 (2021)

[10] Kerkeni, L., Serrestou, Y., Raoof, K. & Others Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication*. **114** pp. 22-35 (2019)

[11] Huang, Y., Tian, K., Wu, A. & Others Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *Journal Of Ambient Intelligence And Humanized Computing*. **10** pp. 1787-1798 (2019)

[12] Daneshfar, F. & Kabudian, S. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimedia Tools And Applications*. **79**, 1261-1289 (2020)

[13] Widodo, H., Fang, F. & Elyas, T. The construction of language teacher professional identity in the Global Englishes territory:'we are legitimate language teachers'. *Asian Englishes*. **22**, 309-316 (2020)

[14] Chen, M., He, X., Yang, J. & Others 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*. **25**, 1440-1444 (2018)

[15] Kwon, S. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *International Journal Of Intelligent Systems*. **36**, 5116-5135 (2021)

[16] Widodo, H., Fang, F. & Elyas, T. The construction of language teacher professional identity in the Global Englishes territory: 'we are legitimate language teachers'. *Asian Englishes*. **22**, 309-316 (2020)

[17] Yuan, R., Li, S. & Yu, B. Neither "local" nor "global": Chinese university students' identity paradoxes in the internationalization of higher education. *Higher Education*. **77**, 963-978 (2019)

[18] Xinhan, N. Intelligent analysis of classroom student state based on neural network algorithm and emotional feature recognition. *Journal Of Intelligent And Fuzzy Systems*. **40**, 1-12 (2020)

[19] Hu, J. & Zhang, H. Recognition of classroom student state features based on deep learning algorithms and machine learning. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **40** pp. 2 (2021)

[20] Byun, S. Lee S P . *Study On A Speech Emotion Recognition System With Effective Acoustic Features Using Deep Learning Algorithms*. **2021** pp. 4 (0)

[21] Shah, V. Mehta M .Emotional state recognition from text data using machine learning and deep learning algorithm. *Concurrency And Computation: Practice And Experience*. **2022** pp. 17 (0)

[22] Atmaja, B., Sasou, A. & Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*. **140** pp. 11-28 (2022)

[23] Pham, N. Dang N M D, Nguyen S D. A Method upon Deep Learning for Speech Emotion Recognition. *Journal Of Advanced Engineering And Computation*. **4** pp. 4 (2021)

[24] Long, L. & Liang, T. Multi-Distributed Speech Emotion Recognition Based on Mel Frequency Cepstogram and Parameter Transfer. (Chinese Journal,0)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Nov 16, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 313–326, DOI 10.12694/scpe.v25i1.2297

THE EVALUATION OF ETHNIC COSTUME COURSES BASED ON FP-GROWTH ALGORITHM

RUI XU*

Abstract. In order to make full use of the accumulated curriculum data of Folk costume and dig out useful information from it, so as to provide useful information for curriculum teaching, the article proposes three general functions based on the requirement analysis, and pre-processes the completed grade data of ethnic costume students in 4 academic years, analyzes these data by FP-growth algorithm to understand the situation of association rules between different courses, and through K-means++ algorithm The clustering analysis of students with different levels of achievement is carried out and the results are validated by examples. In the algorithm performance analysis, the performance of FP-growth algorithm is better, the average absolute error of FP-growth algorithm is always smaller than that of Apriori algorithm; When the support degree is 20%, the running time of FP-growth algorithm is 0.4s, which is 0.4s less than that of Apriori algorithm. when the number of calculation nodes is 5, the running time of FP-growth algorithm and the accuracy of the K-means++ algorithm were higher than that of the K-means algorithm. In the Iris dataset, the accuracy of the K-means++ algorithm was 91.05%, which was 8.94% higher than that of the K-means algorithm. When mining the course grade data, the confidence level of the obtained association rules was even higher, even up to 97.15%. The standardized test score for the second group of students was 0.960. The course evaluation method used in the article was more objective and the accuracy of the data analysis was higher, providing valuable reference information for teachers' teaching.

Key words: Ethnic dress; Course evaluation; FP-growth; K-means++ algorithm; Association rules

1. Introduction. The development of technology has not only brought convenience to people's lives, but has also facilitated the development of education. Computer technology is used to improve the quality of teaching and learning, allowing teachers to gain a deeper understanding of their students and to deepen their knowledge with the help of computer technology. Education-related systems have been developed to make the work of education more convenient. And by running the system for a long time, a lot of data is accumulated in the database. This data is not fully utilized. How to make the best use of the data and uncover the latent useful information is the problem that needs to be solved. By applying improved association rules, the running time of the recommender system is reduced and better quality rules are obtained [1]. The efficiency of tyre quality data analysis is improved by mining and analyzing abnormal data through improved Frequent-Pattern growth (FP-growth) algorithm [2]. The K-means algorithm based on principal component analysis was used for clustering analysis of handwritten digital datasets, and the algorithm had better performance [3]. The FP growth algorithm, as an association analysis method, has good performance in frequent itemset data mining. Therefore, facing the problem of data mining of folk costume curriculum, in order to understand the learning situation of students, this paper analyzes the association rules of folk costume curriculum performance data, adopts FP growth algorithm, and uses improved K-means algorithm to cluster the performance data, hoping to mine useful information. The research is divided into four parts. The first part is a literature review, which introduces the research status of domestic and foreign scholars on curriculum teaching, FP growth algorithm, and K-means++algorithm. The second part constructs the teaching evaluation system of folk costume course through FP growth algorithm and K-means++algorithm, and preprocesses the data. The third part analyzes the algorithm performance and application effectiveness. The fourth part summarizes the research methods and points out the research prospects, shortcomings, and future research directions.

2. Related Work. The development of technology has not only brought convenience to people's lives, but has also facilitated the development of education. Computer technology is used to improve the quality of teaching and learning, allowing teachers to gain a deeper understanding of their students and to deepen their

*Academy of Arts Jinling Institute of Technology Nanjing 211169, China (lele1220mao@163.com)

knowledge with the help of computer technology. Education-related systems have been developed to make the work of education more convenient. And by running the system for a long time, a lot of data is accumulated in the database. This data is not fully utilized. How to make the best use of the data and uncover the latent useful information is the problem that needs to be solved. By applying improved association rules, the running time of the recommender system is reduced and better quality rules are obtained [1]. The efficiency of tyre quality data analysis is improved by mining and analyzing abnormal data through improved Frequent-Pattern growth (FP-growth) algorithm [2]. The K-means algorithm based on principal component analysis was used for clustering analysis of handwritten digital datasets, and the algorithm had better performance [3]. The FP growth algorithm, as an association analysis method, has good performance in frequent itemset data mining. Therefore, facing the problem of data mining of folk costume curriculum, in order to understand the learning situation of students, this paper analyzes the association rules of folk costume curriculum performance data, adopts FP growth algorithm, and uses improved K-means algorithm to cluster the performance data, hoping to mine useful information. The research is divided into four parts. The first part is a literature review, which introduces the research status of domestic and foreign scholars on curriculum teaching, FP growth algorithm, and K-means++algorithm. The second part constructs the teaching evaluation system of folk costume course through FP growth algorithm and K-means++algorithm, and preprocesses the data. The third part analyzes the algorithm performance and application effectiveness. The fourth part summarizes the research methods and points out the research prospects, shortcomings, and future research directions.

Liu T et al. optimized the error back propagation (BP) algorithm by improving the particle swarm algorithm and applied the improved algorithm to multimedia courseware evaluation, which has higher prediction accuracy than the algorithm before the improvement [4]. Hideya et al. used topic modeling to collect course evaluation-related data for analysis. After validation, the labels used were found to be valid and able to display the thematic proportions and distribution of course information through visual information [5]. Kazanidis I et.al. applied the overlay algorithm to e-course learning assessment to better assess course quality and facilitate teachers' understanding of teaching and learning [6]. Kazanidis I Bz A et al. analyzed the impact of collaborative learning on students outside the classroom by having them form their own learning groups for a pharmacogenomics course and showed that collaborative learning had a greater impact on students' abilities [7]. Wu X addresses the problem that the "online and offline" teaching mode can improve teaching efficiency, mixes English teaching and civics teaching, and integrates "online and offline" for integrated teaching, analyzes the teaching method of hybrid teaching and proposes reform [8]. The experimental results show that the model has higher recommendation accuracy and lower prediction error [9] on foreign language teaching [10]. Yang Z designed a data mining system based on the Apriori algorithm in order to mine accurate information from the database, and the experiments showed that the system could quickly get effective information, which could effectively improve the quality of education teaching and promote the modernization process of education [11].

Novita R et al. designed a data mining system with the FP-Growth algorithm to determine the relationship pattern of the related topics of the Quran, and the experimental results showed that the system can quickly implement the relationship pattern of the Quran's associated topics and promote the understanding of the Quran [12]. Jiye et al. designed a data mining system with the FP-Growth algorithm in order to improve the detection of irregular behaviors in power management systems, an FP-Growth algorithm detection model was designed, and the experimental results showed that the model can effectively detect behaviors that cause security problems in regular user behaviors and can ensure the security of power association information systems [13]. liu Y et al. In order to improve the dynamic of intelligent rail shuttle trolley scheduling accuracy, a model based on genetic algorithm and K-Means++ algorithm was designed, and the experimental results showed that the model can reduce the system imbalance caused by the CNC system failure and meet the greater production demand [4]. Ye J et al. designed an improved cluster analysis algorithm in order to improve the detection accuracy of UAV FM models by constructing an experimental environment, and the results showed that the algorithm proposed in the study can achieve the detection of UAVs and guarantee a certain accuracy rate, which has good application prospects [15]. Cheng S et al. designed a K-means++ algorithm model in order to improve the performance of neural network models, and evaluated the learning results of the network using the confusion matrix, and the experimental results showed that the model determined a more compact number of nodes in the hidden layer [16]. Zhang G et al., in order to study the impedance adjustment techniques of artificial

Table 2.1: Comparison of Research Contents

Reference	Research Contents
Ref. [4]	Multimedia courseware evaluation
Ref. [5]	Visualization of course information topics
Ref. [6]	Electronic Course Learning Assessment
Ref. [7]	Influence of extracurricular learning cooperation on students of Pharmacogenomics
Ref. [8]	Strategies for Combining Ideological and Political Education with Online and Offline Teaching in English Curriculum
Ref. [9]	Propose a recommendation model for online offline courses
Ref. [10]	Intelligent Online Video Corpus in Teaching
Ref. [11]	University teaching management evaluation combining Big data and data mining
Research	The Application of Data Mining and Clustering Based on Students' Course Score Data in Folk costume Course Evaluation

belt grinding, obtained arm impedance adjustment data based on its relevant data, and obtained the relevant controller through core parameter learning and other processes. During this period, the K-means++algorithm was used to estimate the manual grinding impedance. The results show that the proposed method has a good application effect and can effectively estimate the impedance of manual grinding [17].

To sum up, there is little research on the analysis of students' course scores in the course evaluation, and the data of course scores are not fully utilized. Therefore, the article explores the data of folk costume course scores to understand the students' learning situation. In view of the good performance of FP growth algorithm and K-means++algorithm in data mining, it is applied to the data mining analysis of folk costume courses.

Table 2.1 presents several directions of instructional research. From the data used in the research, there is relatively little literature on conducting research based on student course performance data. Therefore, the unlike cited literature, researching and exploring the relationship between student course performance data and course evaluation can achieve the automation of course evaluation, improve evaluation efficiency and objectivity.

3. Evaluation of Ethnic Costume Courses based on FP-growth Algorithm.

3.1. Course evaluation data pre-processing and correlation analysis. The folk costume course is a study of the design and production of folk costumes. The assessment of the course is carried out to understand the students' learning so that teachers can understand the students' situation and analyze whether they have reached the criteria for graduation. The evaluation is carried out through quantitative and qualitative analysis , which changes the previous method of evaluation which used only grades. In the ethnic dress course evaluation system, a needs analysis is first required to collect course data and grade data. In the first category of data, the course name of the ethnic costume course, the number of students in the relevant course, the basis of assessment and evaluation that each course has and other relevant information need to be collected; in the latter category of data, the students' names, student numbers and other personal information need to be collected, the completion of homework in the usual study, in the examination and practical operation experiments, the performance of students in the data related to students' performance will be collected. Data collection can be done by manual input or batch import. Users can choose different data collection methods according to their own habits, thus enhancing the user experience. The report enquiry and download function allows users to understand how students are learning in the course. Through the data analysis and decision making function, you can understand the students' performance in learning, as well as the achievement of the index points and graduation requirements, and display these data through the information of graphs and charts, so that teachers can make adjustments to the teaching methods based on these visual data, thus improving the quality of teaching. Based on the requirements analysis: three general functions were identified, namely data collection, data analysis, and report management. The use case diagram for data collection and data analysis is shown in Figure 3.1.

Prior to data analysis, data pre-processing is required. The article uses the grade data of students who

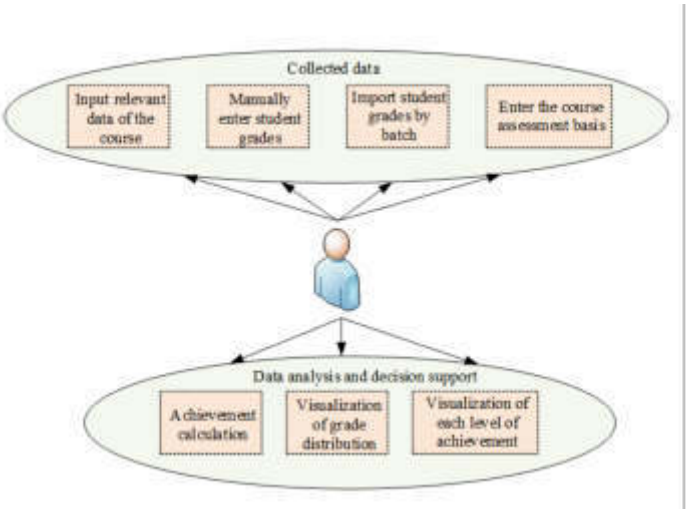


Fig. 3.1: Relevant Use Case Diagrams

completed studies in ethnic dress during the academic year 2016-2020, which is subjected to data cleaning and integration processes, followed by data transformation so that students' exam grades are standardized to lie between [0,1], and finally data statute processing, in which discrete processing results in greater than the course mean data correlation and clustering analysis was then performed, in which the association rule algorithm was applied. In the association rules, the set of items is set as shown in equation 3.1.

$$I = \{i_1, i_2....., i_n\} \tag{3.1}$$

In equation 3.1, represents the term and then sets the transaction term , whose mathematical expression is shown in equation 3.2.

$$D = \{T_1, T_2...., T_p\} \tag{3.2}$$

In equation 3.2,T_p represents a transaction; T_p is the set of i_n and $T \epsilon I$. If $X \epsilon I, X\epsilon \neq \theta, Y \epsilon I, Y \neq \theta$ and $X \cap Y \neq \theta$, there are association rules in the non-empty set X and Y consisting of certain items, and , are the preconditions, outcomes of the relationship. The mathematical expression for the support of this relation is shown in equation (3).

$$sup(X => Y) = \frac{|\{T : X \cup Y \epsilon T, T \epsilon D\}|}{|D|} \tag{3.3}$$

The mathematical expression for the confidence level of the association rule is shown in equation 3.4.

$$Confidence(X => Y) = \frac{|\{T : XUY \epsilon T, T \epsilon D\}|}{|\{T : X \epsilon T, T \epsilon D\}|} \tag{3.4}$$

In equation 3.4, the confidence level represents the probability of X and Y occurring at the same time, and the higher the confidence level, the more reliable the corresponding association rule is. In the case of the item set support $sup(X)$, the formula is shown in equation 3.5.

$$sup(X) = \frac{|\{T : X \epsilon T, T \epsilon D\}|}{|D|} \tag{3.5}$$

In equation 3.5, if $sup(X) \geq min_sup$,min_sup represents the minimum support, then represents the frequent itemset. Since the Apriori algorithm has a high computational overhead and is not conducive to data analysis,

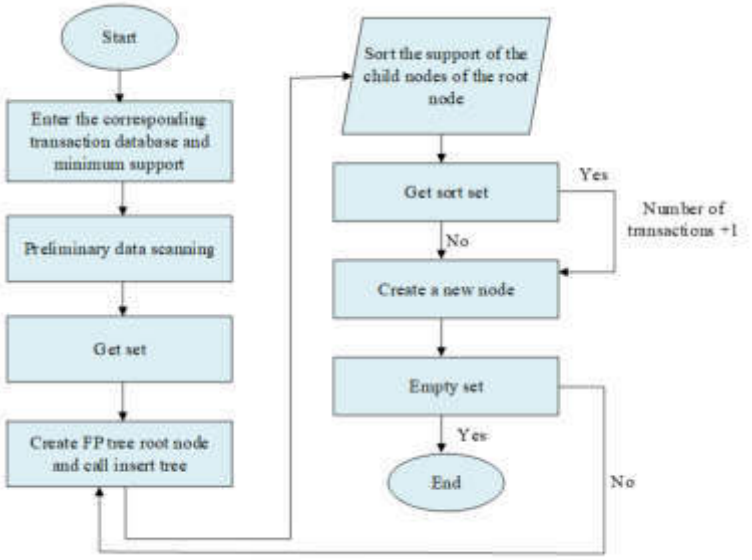


Fig. 3.2: Relevant Use Case Diagrams

the article uses the FP growth algorithm, which mines the entire set of frequent items without generating a candidate set of items. This algorithm can overcome the flaw of multiple scans of transactional databases in the Apriori algorithm. It is worth mentioning that the data processed by the algorithm cannot include subjects that have little relevance to the folk costume course, which will affect the accuracy of the results. Therefore, certain processing of the data is necessary. At runtime, this algorithm first constructs a Frequent Pattern Tree (FP Tree), as shown in Figure 3.2.

In the figure 3.2, the scan of D, which is known through several experiments, and $minsup = 50$, with a minimum confidence level of 0.85, are used to calculate the set of frequent terms, as shown in equation 3.6.

$$F = \{j_1, j_2, ..., j_m\} \tag{3.6}$$

In equation 3.6, F represents the set of frequent items and j_n represents the frequent items. The support number of frequent items, which represents the number of times the item appears in D , is calculated and sorted in descending order to obtain the frequent items table L . The FP-Tree root node is then created and marked by null. It performs a second scan of D and in T and removes the non-frequent items to obtain the frequent items, which are sorted according to L . It creates a branch of T to represent the nodes of the FP-Tree by item herding dust and supports number, where T_n indicates that there is a n transaction T . When linking T_1 frequent items, the root node links the first frequent set, and then links the frequent items in the same way that the latter frequent necklace is linked to the previous frequent item. When a shared path exists in the T_2 branch, the number of all nodes in which the path of that segment is located is added by one and the remaining different paths need to be recreated. When all T branches are inserted, the FP-Tree is obtained. Then frequent item mining is performed, as shown in Figure 3.3.

The association rules are mined by calling FP-growth from FP-Tree. In the recursive process, FP-growth calls the first layer of the FP-Tree with a null value for the node, thus obtaining the frequent 1-item set. Then, the recursive call to FP-growth is made on all the resulting itemsets, resulting in a multivariate frequent itemset. As can be seen from the algorithm's operation, the algorithm is divided into two main parts, building the FP-Tree to mine the frequent itemsets. The FP-growth algorithm is then used to analyze the course performance data of ethnic dress students and to mine the association rules that exist in them. According to the association rules mined, teachers can reasonably formulate the professional training plan, so that the teaching quality of this major can be improved.

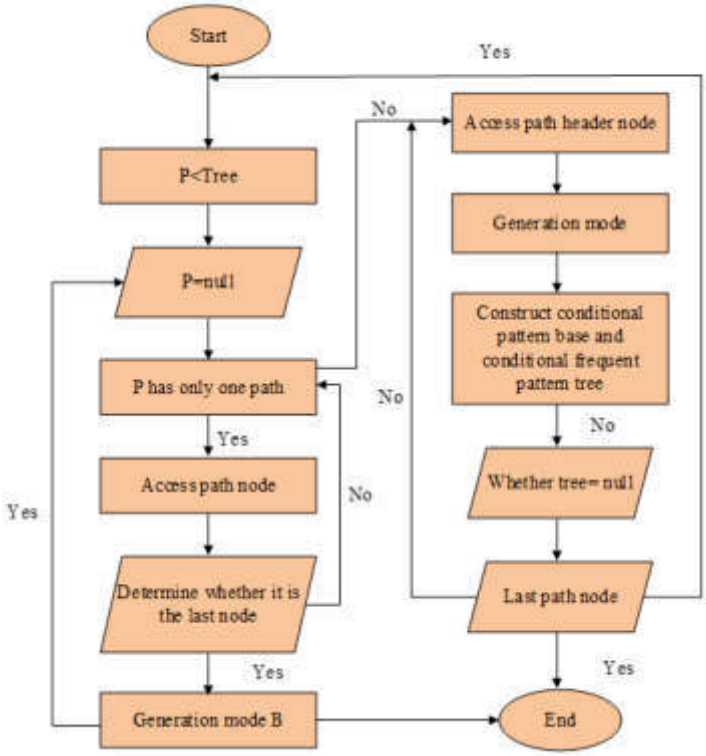


Fig. 3.3: Operation process

3.2. Classification of course performance data based on cluster analysis. The K-means algorithm is also a clustering algorithm that is generally suitable for large-scale data mining, but the number of clusters can affect the results of the clustering analysis. The K-means++ algorithm can make up for the shortcomings of this algorithm, and the flow of the K-means++ algorithm is shown in Figure 3.4.

In Figure 3.4, the first cluster centroid c_1 is randomly selected from the set C and the shortest distance from each point in C to the cluster $d(x_i)$ is calculated using the Euclidean distance formulax_i. represents the first vector$i$ in C . The sum-of-squares operation is performed dx_ion to obtain the sum-of-squares value of x_i $sum(d_(^2x_i))$. The probability of the midpoint of C becoming the next cluster centre is calculated and the relevant formula is shown in equation 3.7.

$$P_i = \frac{d^2(x_i)}{sum(d^2(x_i))} \tag{3.7}$$

In equation 3.7,$P(i)$ represents the probability of being the next cluster centre. It chooses a random number R which lies in the interval [0,1], lets R subtract $P(1)$; When $R - P(1) > 0$, it continues to subtract $P(2)$, before again determining the magnitude of the difference between R and $P(2)$, when it is greater than 0, use the same method, do the subtraction operation until the obtained difference is not greater than 0, stop down, the second clustering centre is the point corresponding to $P(i)$. Starting with the operation to calculate the shortest distance $d(x_i)$, the subsequent steps are repeated to find the optimal K initial clustering centre ck . The distance between the points of C and ck is then calculated, and the nearest ck point to the set C is found and then grouped into the corresponding subset. The mean value of the set is calculated and the calculated mean value is used to update the clustering centres and to perform the error averaging and calculation, the

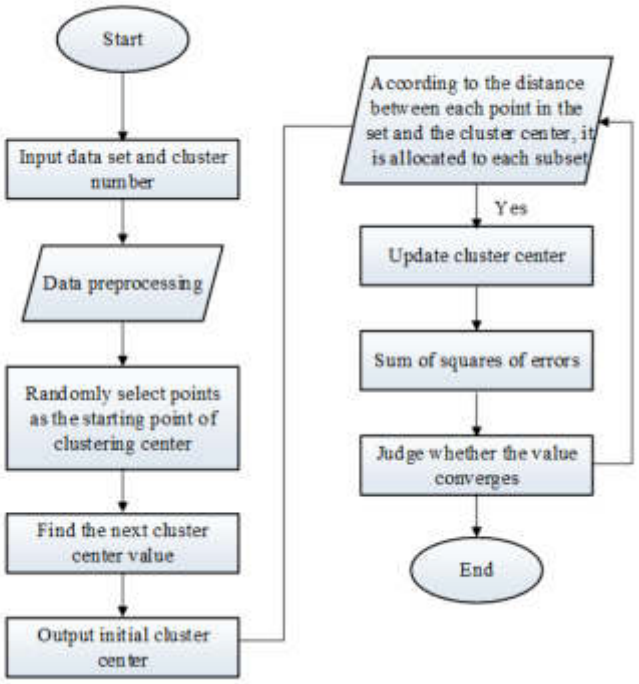


Fig. 3.4: Operation process

corresponding formula is shown in equation 3.8. mi denotes the mean value of C.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C} |q - mi|^2 \tag{3.8}$$

In equation 3.8,E represents the error sum of squares and the set of points is set to q . This step is repeated from calculating the distance between the points at C and ck and stops when the error sum of squares converges. Applying the K-means++ algorithm to the analysis of student performance data, the numerical class of performance data needs to be discretized before use. After this type of processing, the type of student performance data is transformed into a binary data type, the specific mathematical expression of which is shown in equation 3.9 .

$$\begin{cases} i = (x_{i1}, x_{i2}, ..., x_{ip}) \\ j = (x_{j1}, x_{j2}, ..., x_{jp}) \end{cases} \tag{3.9}$$

In equation 3.9, i and j are the student's performance vectors and the number of attributes of the values is p . The Euclidean distance d(i,j) is one of the commonly used distance metrics and is shown in equation 3.10.

$$d_{ij} = \sqrt{\sum_{k}^{n} (x_{ik} - x_{jk})^2} \tag{3.10}$$

In equation 3.10, d_{ij} represents the Euclidean distance. The Manhattan distance is also a distance metric and its mathematical expression is shown in equation 3.11.

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \tag{3.11}$$

In addition to this, there is the Minkowski distance method, which is calculated using the formula shown in equation 3.12.

$$d(i,j) = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} \tag{3.12}$$

After discretizing the data related to test scores, this data exists in two states, when the data is 0, it means that the student's score is not greater than the average score; If it is 1, it is greater than the average score. With the help of a binary data column table can reflect the dissimilarity of these two data, while for $d(i,j)$ symmetric binary data, the symmetric dissimilarity $d(i,j)$ is calculated as shown in equation 3.13.

$$d(i,j) = (b+c)/(a+b+c) \tag{3.13}$$

In Equation 3.13, when the values of j and i are both 1 or 0, the corresponding state numbers are a or$d$ respectively; When $i = 0$,$j = 1$ the state number is c ; For asymmetric binary data, the two data states of 0 or 1 are of different importance, and since $i = 1, j = 1$ is more significant than , $i = 0, j = 0$ d is omitted, resulting in the corresponding mathematical expression for the phase difference as shown in equation 3.14.

$$d(i,j) = (b+c)/(a+b+c+d) \tag{3.14}$$

The K-means++ algorithm was then applied to cluster the results of student performance in a course and the target number of clusters was determined $k = 4$. In order to test the performance of the algorithm, an acceleration ratio can be used, which is calculated as shown in equation 3.15.

$$s_m = T_1/T_m \tag{3.15}$$

In equation 3.15, s_m represents the speedup ratio, and the Central Processing Unit (CPU) time to complete the task for m and 1 processor are T_m and T_1 respectively.

4. Analysis of Experimental Results. In order to study the performance of the FP-growth algorithm, the article uses the Apriori algorithm as the comparison algorithm, the number of transactions included in the test data is 8428, by mining test data to reflect the performance of the two algorithms; The CPU used is 2GHZ, to study the average absolute error of the algorithm with different number of iterations, and the running time of the two algorithms with different support, the specific results are shown in Figure 4.1.

As can be seen in Figure 4.1, the mean absolute error of both the FP-growth algorithm and the Apriori algorithm decreases as the number of iterations increases. At 700 iterations, the average absolute error of the FP-growth algorithm is 0.665, while the average absolute error of the Apriori algorithm is 0.725, with the average absolute error of the former algorithm being 0.06 smaller than the average absolute error of the latter. Overall, the average absolute error of the FP-growth algorithm is always smaller than that of the Apriori algorithm. When the support degree was 10%, the running time of the FP-growth algorithm was 1.3s, while the running time of Apriori algorithm was 6.5s. The time difference between the latter algorithm and the former algorithm was 5.2s. With the increase of support degree, the running time of both algorithms decreased continuously, and when the support degree was before 16%, the running time of the Apriori algorithm decreased, the running time of the Apriori algorithm decreased faster than that of the FP-growth algorithm. At 20% support, the FP-growth algorithm runs in 0.4s, which is 0.4s less than the Apriori algorithm, and at 30% and 40% support, the FP-growth algorithm runs in 0.3s and 0.2s respectively. Overall, the FP growth algorithm always has a shorter runtime than the Apriori algorithm at the same level of support. It can be seen that the FP-growth algorithm has a better performance compared to the Apriori algorithm. The performance of the FP growth algorithm was further analysed with the test dataset of webdocs, and study the running time of FP growth algorithm, the Apriori algorithm, the decision tree and the random forest under different numbers of computing nodes, and the running time of the four algorithms under different data set sizes, as shown in Figure 4.3.

In Figure 4.3, the running time of the FP growth algorithm and Apriori algorithm varies with the number of computing nodes. With the increase of the number of computing nodes, the running time of the two algorithms

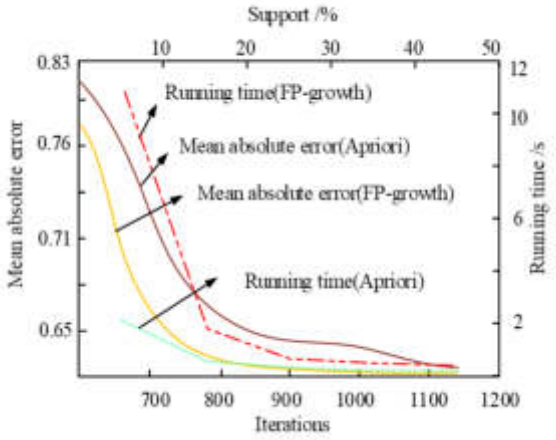
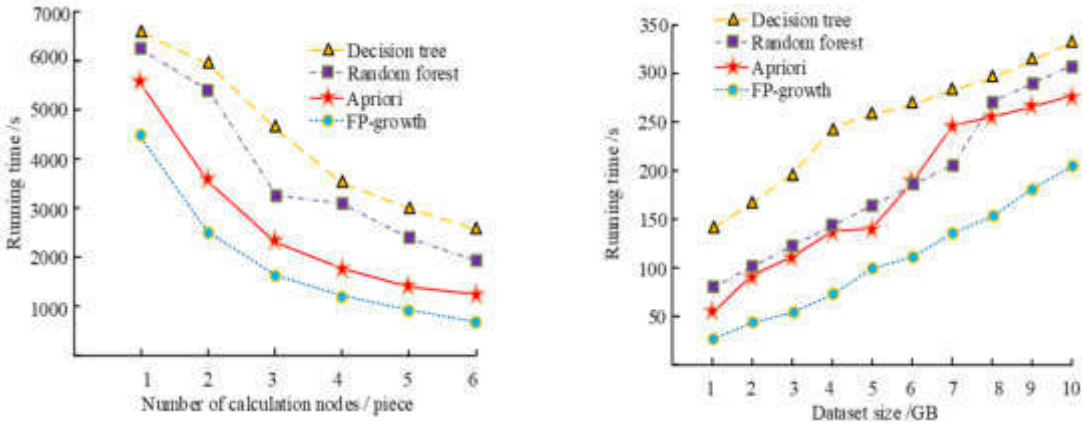


Fig. 4.1: Mean Absolute Error and Running Time of the Two Algorithms



((a)) Calculate the running time under the node

((b)) Run time at dataset size

Fig. 4.3: Running time of four algorithms

changes, and both are constantly decreasing. When the number of computing nodes is 1, the running time of Apriori algorithm is 5534s, while that of FP growth algorithm is 4476s. When the number of computing nodes is 2, the running time of FP growth algorithm is 2602s, 1082s less than that of the Apriori algorithm. When the number of computing nodes is less than 3, the running time of the two algorithms decreases rapidly; When the number of calculation nodes is greater than 3, the running time of the FP growth algorithm and the Apriori algorithm decreases slowly. When the number of calculation nodes is 4, the corresponding running time of the FP growth algorithm is 1235s, while the running times of the Apriori algorithm is 1727s, which is 492s more than the former. When the number of computing nodes is 5, the running time of the FP growth algorithm and Apriori algorithm is 1000s and 1451s respectively, and the time difference between them is 451s. From the time difference between the two algorithms when the number of computing nodes is 4 and 5, it demonstrates that the time difference between the two algorithms is decreasing. As the data set increases, the running time of the four algorithms increases, and the running time of the FP growth algorithm is relatively minimum. It tests the data sets webdoc1 and webdoc4, and studies the running time and acceleration ratio of the FP growth

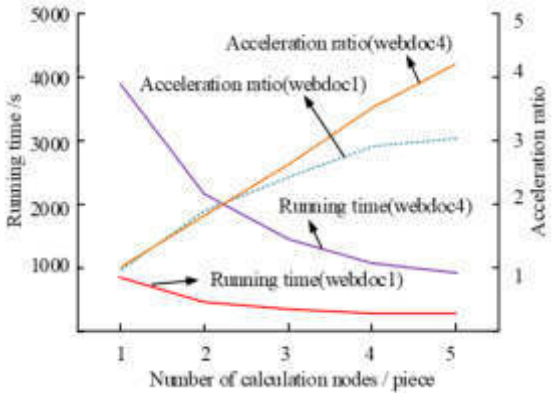


Fig. 4.4: Number of calculation nodes /piece

algorithm under different data sets and different number of computing nodes, as shown in Figure 4.4.

In Figure 4.4, there are differences in the runtime of the FP growth algorithm under the same number of computing nodes in different datasets. As the number of computing nodes increases, the corresponding runtime decreases, and the time difference for mining different datasets becomes smaller and smaller. Overall, the FP growth algorithm has the least runtime when mining dataset webdoc1. When the number of nodes is calculated as 1, the FP growth algorithm takes 854 seconds to mine dataset webdoc1, while it takes 3817 seconds to mine dataset webdoc4. The FP growth algorithm takes much less time to mine dataset webdoc1. When calculating the number of nodes to 2, the FP growth algorithm's mining time for the dataset webdoc4 rapidly decreases, with a running time of 2143 seconds, which is 1732 seconds longer than the algorithm's mining time for dataset webdoc1, while the algorithm's mining time for dataset webdoc1 is 411 seconds. When mining the dataset webdoc1, the FP growth algorithm ran for 264 seconds and 253 seconds when calculating the number of nodes as 4 and 5, respectively. The time difference between the two is relatively small, with a time difference of 11 seconds. In the acceleration ratio of the FP growth algorithm, when the number of computing nodes is not greater than 2 in different datasets mining, the acceleration ratio of the FP growth algorithm in datasets webdoc1 and webdoc4 mining is basically the same. When the number of computing nodes is not less than 3, as the number of computing nodes increases, the acceleration ratio of the algorithm running on the two datasets continues to increase, and the difference in acceleration ratios between the two continues to widen. When the number of computing nodes is 5, the FP growth algorithm has an acceleration ratio of 3.1 on dataset webdoc1, while the algorithm has the acceleration ratio of 4.3 on dataset webdoc4. The former has an acceleration ratio 1.2 less than the latter. Overall, from the performance analysis of the FP growth algorithm, it can be seen that its performance is significantly better than the Apriori algorithm. It studies the accuracy and runtime of the K-means++algorithm and the K-means algorithm under different datasets, as shown in Figure 4.5.

In Figure 4.5, in the accuracy of the algorithms, the accuracy of the different algorithms varied across the datasets, and overall the accuracy of the K-means++ algorithm was higher than the accuracy of the K-means algorithm. In the Iris dataset, the accuracy of the K-means++ algorithm is 91.05%, which is 8.94% higher than that of the K-means algorithm, while the accuracy of the K-means algorithm is 82.11%. The accuracy of the K-means++ algorithm in the datasets Glass and Flame is 83.00% and 95.07% respectively. Overall it can be seen that the two algorithms have the highest accuracy in the dataset Wine, with the K-means++ algorithm having an accuracy of 96.21% and the K-means algorithm having an accuracy of 90.00%. In the running time of the algorithms, the K-means algorithm had the longest running time in the dataset Wine with a time of 6.49s, 1.67s more than the K-means++ algorithm, while the K-means++ algorithm had a running time of 4.82s. In the dataset Flame, the K-means++ algorithm had a running time of 3.12s, 1.67s less than the K-means. The running times of the K-means++ algorithm were 3.32s and 4.01s in the datasets Iris and Glass respectively. It can be seen that the running time of the K-means++ algorithm was less than that of the K-means algorithm;

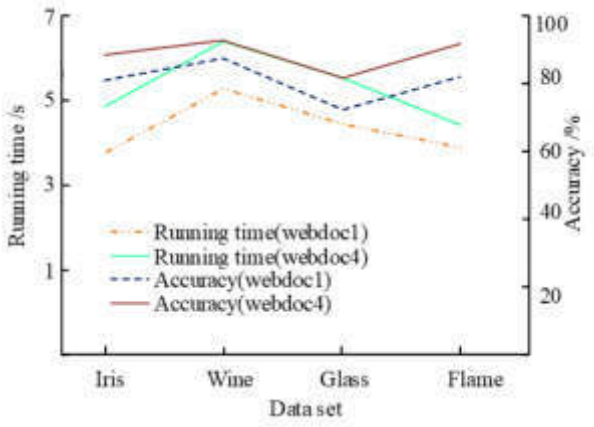


Fig. 4.5: Accuracy and runing time under different data sets

Table 4.1: Association Rules of Some Courses of National Costume Major

Precondition	Result	Confidence Level (%)
National costume culture, northern national folk culture	Appreciation of national costumes	94.47
Foundation of creation, sketch, colour, photography	Art design practice	97.15
Foreign language, computer-aided design	English for art design	89.16

both algorithms had the most running time in the dataset Wine. In addition, through experiments, in the data set Iris, compared with the K-means algorithm, the K-means++algorithm has a smaller Mean absolute error and a larger recall rate. It can be seen that the K-means++ algorithm performs better than the K-means algorithm. The FP-growth algorithm was used to mine the course grade data and the relevant results were obtained as shown in Table 4.1.

As can be seen from Table 4.1, the confidence levels of the association rules vary from course to course. Based on the relationship between the confidence levels, it is known that with the two courses, Ethnic Costume Culture and Northern Ethnic Folk Culture, as prerequisites, the confidence level between them and the course on Ethnic Costume Appreciation as an outcome is 94.47%, indicating that the inference between them is more reliable. Similarly, it can be seen that inferences from the other two course association rules are also more reliable. With the prerequisites of Fundamentals of Creative Writing, Drawing, Colour and Photography, the confidence level of the association rule with the Art and Design Practice course as the outcome was 97.15%. The usual and examination results of the eight students were processed to obtain the corresponding dataset, and sample A6 was used as the initial clustering centre, with the relevant data shown in Table 4.2.

From Table 4.2, it can be seen that there is a difference between the usual and examination results of these students. is the distance from the sample to sample A6, from which the cumulative values of and are known $P(x)$,$P(x)$ as shown in Figure 4.6.

As can be seen in Figure 4.6, the probability of a sample becoming the next cluster centre differs from sample to sample. The probability of sample A1 becoming the next cluster centre is 20.00% and the cumulative value of this probability is also 20.00%; the probability of sample A2 becoming the next cluster centre is greater than that of sample A1 with a probability of 32.50% and the corresponding cumulative value is 52.5%. Among the different samples, sample A6 has the smallest probability of being the next cluster centre of 0, while sample A7 has a probability of being the next cluster centre of 5.00%, which is 2.50% greater than that of sample A5. The probability and cumulative probability of sample A3 being the next cluster centre are 12.50% and 65.00% respectively. The cumulative value of for the A8 sample is 100.00%. The cumulative value of for the first 4

Table 4.2: Association Rules of Some Courses of National Costume Major

Sample serial number	A1	A2	A3	A4
Usual performance	3	4	3	4
Examination results	4	4	3	3
$d(x)^2$	8	13	5	10

Sample serial number	A5	A6	A7	A8
Usual performance	0	1	0	1
Examination results	2	2	1	1
$d(x)^2$	1	0	2	1

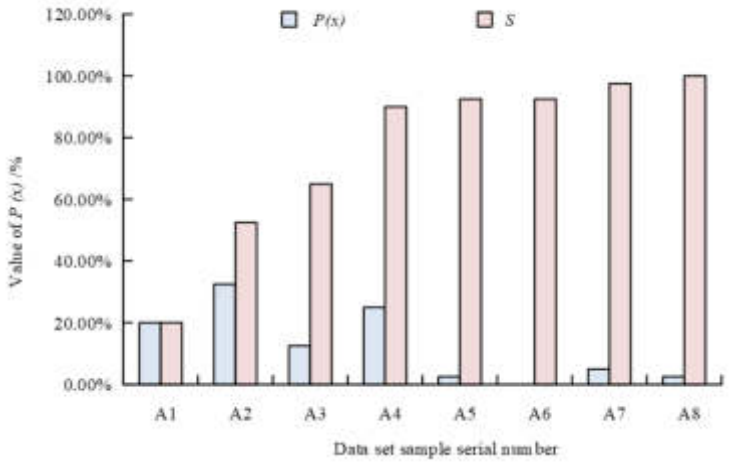


Fig. 4.6: Data set sample serial number

samples was 90.00% and it just so happened that the first 4 samples were farther away from A6 and had a higher compared to the other sample points, validating the theory of the algorithm involved in the study to find different types of students at different test difficulty levels. Using the K-means++ algorithm to cluster discrete art and design practice courses, the relevant results were obtained as shown in Figure 4.7.

In Figure 4.7, students are divided into four categories, with standardized test scores and usual scores for each category. Comparing the standardized scores, it can be seen that students in the first category had good scores in both categories, with a standardized test score of 0.986, while students in the second category had a standardized usual score of 0.715, which was 0.277 higher than the standardized usual score of students in the third category. It can be seen that the application effect of the research method is good.

In the course evaluation of folk costume, when mining the association rules between students' scores, the research did not choose the traditional Apriori algorithm, but carried out the related mining work through FP growth algorithm. By using this algorithm, good mining results were achieved. Some face the problem of frequent pattern mining in large databases and propose Apriori algorithm and FP growth algorithm. After testing the open-source data of Istacart, it was found that compared to the Apriori algorithm, the FP growth algorithm runs faster and has better application effects [20]. It can be seen that selecting an appropriate association rule analysis algorithm is beneficial for improving mining efficiency. The choice of clustering algorithm will affect the clustering effect. Some scholars, in order to improve the effectiveness of hot spot mining for taxi passengers, use K-means++algorithm to carry out relevant clustering work based on the concept of secondary segmentation. By comparing this algorithm with methods such as the K-means algorithm, it was found that

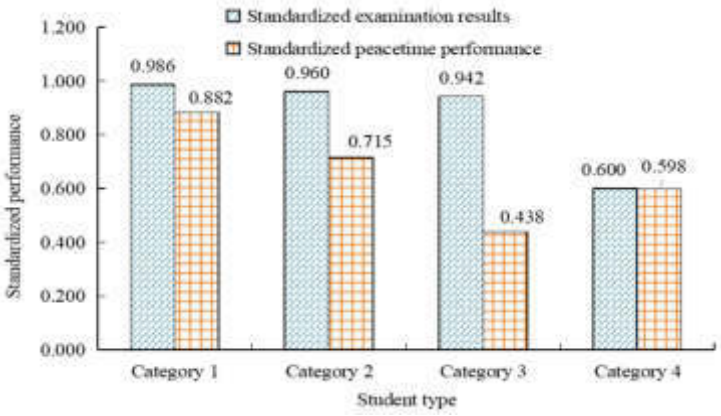


Fig. 4.7: Standardized grades of each category of students

the K-means++algorithm has better clustering accuracy. From this, it can be seen that selecting a suitable clustering algorithm based on the actual situation is beneficial for achieving better clustering results. In the course evaluation of Folk costume, when conducting association rule analysis and cluster analysis on students' course performance data, selecting appropriate algorithms is conducive to improving the effect of folk costume course evaluation, so as to help teachers make better teaching decisions.

5. Conclusion. In order to explore the correlation between ethnic costume courses and understand the learning situation of different types of students, the article uses the FP-growth algorithm to mine the results of courses in ethnic costume, find the hidden useful information in them, and realize the cluster analysis of students with different learning results through the K-means++ algorithm, so that teachers can adjust the learning situation of different types of students according to This allows teachers to adjust their teaching methods according to the learning situation of different types of students, thus realizing personalized teaching. The performance analysis of the algorithm shows that the FP-growth algorithm outperforms the Apriori algorithm, with a running time of 0.4s when the support degree is 20%, which is 0.4s less than the Apriori algorithm. The FP-growth algorithm took the least amount of time to mine the webdoc1 dataset among the different datasets. In the accuracy of the algorithms, the accuracy of different algorithms varied across datasets, with the K-means++ algorithm having a higher accuracy than the K-means algorithm overall. The confidence level of the association rule with the art and design practice course as an outcome was 97.15% with the prerequisites of creative fundamentals, drawing, color and photography. The cumulative value of for the first four samples was 90.00%, while it just so happened that the first four samples were further away from A6 and had higher compared to the other sample points. The first group of students had good scores in both categories, with a standardized test score of 0.986, while the second group of students had a standardized usual score of 0.715. The application effect of the research method is good, and it can be applied to other course evaluations, which is conducive to improving teaching quality. There are still some shortcomings in the research. In the data, there are fewer types of data for student grades, and more types of data should be added to make the analysis more comprehensive. In the data mining analysis, the amount of data involved in mining is relatively small, and it is necessary to increase the amount of data analyzed to further demonstrate the feasibility of the research method. In the future, research can be conducted to add data types and data volumes to improve the applicability of the method.

Funding. The research is supported by: High-level Scientific Research Initiation Project of Jinling Institute of Technology "Study on the Motif of Ethnic Minority Dress Decoration in Northwest China under the Background of One Belt and One Road, Project No. jit-b-201920; General Research Project on Philosophy and Social Sciences in Jiangsu University "A Comparative Study of Chinese Kazakh Clothing on the Silk Road and Central Asian Kazakh Clothing" Project No.2020SJA0533; and National first-class disciplines program funding.

REFERENCES

[1] Hamidi, H. & Hashemzadeh, E. An Approach to Improve Generation of Association Rules in Order to Be Used in Recommenders. *International Journal Of Data Warehousing & Mining*. **13**, 1-18 (2017)

[2] Li, M., Ding, D. & Yi, Y. Data Analysis of Tyre Quality Based on Improved FP-Growth Algorithm. *Zhongguo Jixie Gongcheng/China Mechanical Engineering*. **30**, 244-251 (2019)

[3] Li, B. An Experiment of K-Means Initialization Strategies on Handwritten Digits Dataset. *Intelligent Information Management*. **10**, 43-48 (2018)

[4] Liu, T. & Yin, S. An improved particle swarm optimization algorithm used for BP neural network and multimedia course-ware evaluation. *Multimedia Tools & Applications*. **76**, 11961-11974 (2017)

[5] Hideya, M. Makiko, et al. Analysis of the free descriptions obtained through course evaluation questionnaires using topic modeling. *Educational Technology Research*. **41**, 125-137 (2019)

[6] Kazanidis, I., Valsamidis, S., Gounopoulos, E. & Others Proposed S-Algo+data mining algorithm for web platforms course content and usage evaluation. *Soft Computing*. **24**, 14861-14883 (2020)

[7] Bz, A., Sk, A., Yx, B. & Others A student initiative to implement peer-led study groups for a pharmacogenomics course: Evaluation of student performance and perceptions. *Currents In Pharmacy Teaching And Learning*. **12**, 549-557 (2020)

[8] Wu, X. Research on the Reform of Ideological and Political Teaching Evaluation Method of College English Course Based on "Online and Offline" Teaching. *Journal Of Higher Education Research*. **3**, 87-90 (2022)

[9] Yz, A., Hao, L., Ping, Q. & Others Heterogeneous teaching evaluation network based offline course recommendation with graph learning and tensor factorization - ScienceDirect[J]. *Neurocomputing*. **415**, 84-95 (2020)

[10] Liu, Y., Han, L., Bo, J. & Others The application and teaching evaluation of Japanese films and TV series corpus in JFL classroom. *The Electronic Library*. **36**, 721-732 (2018)

[11] Zhao, Y. Research on the application of university teaching management evaluation system based on Apriori algorithm. *Journal Of Physics: Conference Series*. **1883**, 012033 (2021)

[12] Novita, R. & Mustakim, S. Determination of the relationship pattern of association topic on Al-Qur'an using FP-Growth Algorithms. *IOP Conference Series: Materials Science And Engineering*. **1088**, 12020-01202 (2021)

[13] Jiye, W. Zhihua, et al. *FP-Growth Based Regular Behaviors Auditing In Electric Management Information System - ScienceDirect. Procedia Computer Science*. **139**, 275-279 (2018)

[14] Liu, Y. & Ling, P. Intelligent RGV Dynamic Scheduling Strategy Based on Greedy Algorithm. *World Scientific Research Journal*. **5**, 278-287 (2019)

[15] Ye, J., Zou, J., Gao, J. & Others A New Frequency Hopping Signal Detection of Civil UAV Based on Improved K-Means Clustering Algorithm. *IEEE Access PP (*. **99** pp. 1-1 (2021)

[16] Cheng, S., Wu, Y., Li, Y. & Others TWD-SFNN: Three-way decisions with a single hidden layer feedforward neural network. *Information Sciences*. **579**, 15-32 (2021)

[17] Zhang, G., Zhang, G., Ni, F. & Others Learning impedance regulation skills for robot belt grinding from human demonstrations. *Assembly Automation*. **41**, 431-440 (2021)

[18] Alcan, D., Ozdemir, K., Ozkan, B., Mucan, A. & Ozcan, T. A comparative analysis of apriori and fp-growth algorithms for market basket analysis using multi-level association rule mining. *Global Joint Conference On Industrial Engineering And Its Application Areas*. pp. 128-137 (2022)

[19] Wang, Y. & Ren, J. Taxi Passenger Hot Spot Mining Based on a Refined K-Means++ Algorithm. *IEEE Access*. **9** pp. 66587-66598 (2021)

[20] Usmani, S., Kamran, S., Zeeshan, M., Islam, N. & Khan, Z. Z."A comparative analysis of apriori and FP-growth algorithms for frequent pattern mining using apache spark". *Proceedings Of International Scientific Research Conference*. (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Sep 11, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 327–339, DOI 10.12694/scpe.v25i1.2298

CONSTRUCTION OF SEMANTIC COHERENCE DIAGNOSIS MODEL OF ENGLISH TEXT BASED ON SENTENCE SEMANTIC MAP

PENG GUO*

Abstract. The current English composition automatic correction system rarely involves coherence quality analysis of compositions. Therefore, a semantic coherence diagnosis model for English texts was constructed based on sentence semantic maps, and its effectiveness was verified through experiments. Experimental results show that sub Graphical model 10, 12 and 13 exceed 200 on the first two test texts and 1 on the last two test texts. However, the differences between these three subgraphs on different test texts are not significant, with differences below 30 and 0.3. In addition, when extracting incoherent sentences, the F1 value reaches the optimal value at a threshold of 0.34, which is 87.54%. When the threshold is fixed at 0.34, the accuracy of extracting non coherent sentences also increases with the number of articles, reaching a maximum of 88.43%. At the same time, there was no significant difference in accuracy, recall, and F1 values among different English composition numbers, maintaining between 83% and 89%. The Pearson coefficient calculated in the actual comparison with the teacher's manual composition score is 0.6025, indicating a strong correlation between the two, indicating that the diagnostic results are reliable. Overall, the diagnostic model constructed in the study has strong accuracy and effectiveness, and is practical in the diagnosis of semantic coherence in actual English texts.

Key words: Sentence semantic map; English text; Preprocessing; Diagnostic model

1. Introduction. The coherence of the text is highly subjective, and in the language field, it is regarded as the fluency of the text in the text [1]. The coherence of the article is an important characteristic of good text, and it is also an important criterion for measuring the readability of the article. In a coherent work, the arrangement of sentences and words in it is not random, but has certain logic and consistency [2]. The development of artificial intelligence has prompted the emergence of many English text correction systems, which not only reduces the pressure on teachers to correct, but also improves the writing ability of students. Based on this, scholars at home and abroad have conducted in-depth research on it. Hu L et al. constructed a model of English text grammar error correction by using neural network, so as to effectively realize the error correction of English grammar [3]. Gondaliya Y et al. achieved English grammar and spelling check by using rules, thus effectively improving the coherence of sentences [4]. By building a spelling error correction system, Chaabi Y et al. realized the error correction of English vocabulary, thereby improving the quality of the text [5]. Currently, there is very little research on automatic grading of English compositions to design the indicator of coherence quality, and the performance of the actual core construction system still needs to be improved. Therefore, a diagnostic model for semantic coherence in English texts was constructed based on sentence semantic maps, with the aim of achieving effective analysis of the coherence quality of English texts and accurately evaluating the overall quality of English function.

2. Related Work. With the rapid development of artificial intelligence and its related fields, computers can gradually independently evaluate the comprehensive quality of English texts [6]. However, the coherence quality is seldom included in the evaluation by computer, which leads to the unreasonable scoring results [7]. Based on this, scholars at home and abroad have conducted in-depth research on it. Srivastava K et al. ensured the semantic coherence between different parts of the article by using some sentences of the article to test the semantic coherence [8]. Yang X et al. ensured the semantic coherence of the original sentence by using context-aware word replacement [9]. Aminovna ensures the coherence of text writing by proposing the most effective methods in writing and emphasizing the importance of coherence in academics [10]. Saleh M et al. provided help for the coherence of people's text creation by analyzing the means of grammatical cohesion [11]. Dassanayake N

*College of Finance and Management, Guangzhou Institute of Technology, Guangzhou, 510630, China (wayne007008@163.com)

Table 2.1: Summary and Discussion of Previous Research

Author	**Research questions**	**Research results**
Srivastava K et al.	The issue of semantic coherence in different parts of the article	Strengthening semantic coherence through the use of partial sentences in test articles
Yang X et al.	Semantic Incoherence of Original Sentences	Ensure semantic coherence of the original sentence by using context aware vocabulary substitution
Aminovna B D et al.	Discontinuity in text writing	Emphasizing the importance of coherence in academia, ensuring coherence in text writing
Saleh M et al.	Discontinuity in text creation	The analysis of grammatical cohesive devices provideassistance for the coherence of text creation
Dassanayake N et al.	Lexical Incoherence in Chinese Translation of Sri Lankan Language	Utilizing relevant texts translated by Sri Lankan learners to ensure lexical coherence in Sri Lankan language translation into Chinese
Akmilia P M et al.	Discontinuity of text	Utilizing cohesive devices to ensure text coherence
Abdusalomovna K Y et al.	Semantic Incoherence in Discourse	Analyzed the relationship between cohesion and coherence to provide assistance in ensuring semantic coherence in discourse
Zhang K et al.	Discontinuity in semantic themes	The Topic model of short text is constructed by using two related knowledge to ensure the coherence of semantic topics
Linnik A et al.	Discontinuity in overall semantics	And utilizing combination rating operations such as information content to ensure overall semantic coherence
Rebuffel C et al.	Discontinuity and fluency of text	Propose a multi branch decoder to enhance text coherence and fluency
Gaur M et al.	The problem of weak semantic logicality	The use of knowledge graphs to generate information ensures the logic of semantics
Ruzikulovich A T et al.	The Semantic Incoherence of Poetry	The analysis of the functional semantics of imperative structures ensures semantic coherence in poetry

ensures the lexical coherence of Sri Lankan language translation into Chinese by using relevant texts translated by Sri Lankan learners [12]. Akmilia PM effectively enhances persuasiveness at research conferences by using cohesive devices to ensure the coherence of the text [13]. Abdusalomovna provides help in ensuring semantic coherence in discourse by analyzing the relationship between cohesion and coherence [14].

In addition, Zhang et al. constructed a short text topic model by using two-item related knowledge to ensure the coherence of semantic topics [15]. Linnik et al. compared the differences between aphasia and non-aphasia, and used combined rating operations such as information content to ensure the coherence of the overall semantics [16]. Rebuffel et al. proposed a multi-branch decoder to effectively train word-level labels, thereby improving text coherence and fluency [17]. Gaur M et al. used dynamic meta-information retrieval to analyze the coherence of contextual sentences, and used knowledge graphs to generate information to ensure semantic logic [18]. Ruzikulovich analyzes the functional semantics of imperative structures to ensure that the poetic language is both poetic and semantically coherent [19]. Judging from the research of scholars at home and abroad, most scholars do not start from the overall coherence of the text to distinguish different degrees of sentence coherence. Therefore, the research on the construction of an English text semantic coherence diagnosis model based on sentence semantic maps is innovative. It can not only effectively improve the current lack of coherence quality assessment in English composition evaluation, but also provide assistance for students' English composition grades. At the same time, it also lays the foundation for the development of an English composition evaluation system while improving the overall evaluation quality of the role of English. In addition, a summary of specific research literature is shown in Table 2.1.

3. Construction of a semantic coherence diagnosis model of English text based on sentence semantic graph.

Fig. 3.1: The Model Structure of English Text Semantic Coherence Diagnosis

3.1. Analysis of semantic coherence diagnosis model structure and semantic space correlation algorithm. To realize the effective diagnosis of the semantic coherence quality of English composition texts, the research constructs a semantic coherence diagnosis model of English text based on the sentence semantic graph, and evaluates it through experiments. In English text scoring systems, semantic coherence is rarely involved as an evaluation metric. Semantic coherence refers to paying attention to the combination and cohesion of sentences in written writing. To achieve discourse coherence, attention should be paid to not only topic, word order, language use (cohesion, echo), but also context and sentence pattern. coordination [20]. Semantic coherence theory is a widely accepted theory in coherence analysis. In this theory, there are two very important concepts, namely macrostructure and connection. The macro structure represents the semantic relationship between the parts of the whole text, while the connection is the semantic relationship between the various elements in the idiom.

Research on the application of semantic coherence theory, Entitative graph model, related algorithms of vector semantic space and related knowledge of sub graph matching in the construction of semantic coherence diagnosis model of English text. Among them, the Entitative graph model is fundamentally an improvement on the grid model of test questions. It uses the form of sentence graph to make the connection between sentences more three-dimensional, and can analyze the semantic relationship within the text from a macro perspective. The related algorithms in Vector space are mainly used to map the text to the semantic space of the vector, and express the semantic similarity between text units by using the association between vectors. Therefore, the model structure of the semantic coherence diagnosis of English text constructed by the research is shown in Figure 3.1.

From Figure 3.1, for the input English text, the research first uses the preprocessing module of this article to perform a series of preprocessing, such as word segmentation, part of speech tagging, dependency analysis, etc; Secondly, through the results of preprocessing, entity words in English text are identified, and the boundaries of entity word phrases are determined based on the form of syntactic trees. On this basis, the co referential phenomenon between entity words is resolved. Then, the entity words are marked with grammatical roles, and the Entitative graph structure of English text is established by combining relevant information. Then, under the guidance of semantic coherence theory, the semantic similarity information between sentences is combined with Entitative graph to generate the studied sentence semantic map model. Then, the improved matching technology is used to mine the frequent sub Graphical model in the sentence semantic map to capture the coherence patterns in the text, and the coherence features in the sentence semantic map are extracted according to the coherence quality analysis module built by the research to conduct the corresponding coherence quality analysis, finally forming the coherence quality analysis results of the English text to be approved required by the research. In addition, in natural language processing, in order to obtain a large amount of lexical information and semantic information, it is necessary to perform a series of preprocessing on the text. The researched preprocessing module consists of three parts, whose contents are shown in Figure 3.2.

From Figure 3.2, the model preprocessing module constructed by the research includes text segmentation

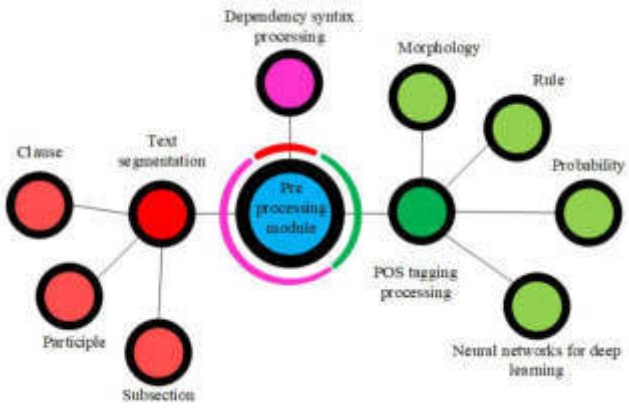


Fig. 3.2: Specific Contents of Text Preprocessing Module

processing, part-of-speech tagging processing and dependency syntax analysis processing. The text segmentation process includes segmentation, sentence segmentation and word segmentation. On paragraphs, since two consecutive line breaks appear between each paragraph, this feature is needed to divide the text. In clauses, the period is the only sign of the end of the sentence, and in English, the period can not only be used as the sign of the end of the sentence, but also can be used as the abbreviation of the first letter and surname. Therefore, the segmentation of Chinese should also be considered on a case-by-case basis. In word segmentation, there is a space between each word in English, so it can be used for word segmentation. In order to segment English text effectively and accurately, this paper studies the segmentation processing of English text using regular expressions. The part-of-speech tagging process includes part-of-speech tagging methods using lexical, rules, probability, and deep learning neural networks. In addition, dependency parsing is an important basic technique that can directly analyze words in text sentences. Dependency syntax refers to the semantic dependencies between words. On this basis, it is centered on the tree-like node in the root of the tree, and the rest are used as modifiers.

It is worth noting that in the English semantic coherence diagnosis model, text analysis occupies a large proportion. In natural language research, it is usually necessary to understand the semantics and internal connections of each component of the text, and then analyze it as a whole [21]. The study uses a group of embedding models used to generate word vectors (Word to vector, Word2vec). The focus of this model is to learn the weights of the neural network to obtain a distributed lexical representation. It includes two neural network models, namely skip mode and continuous bag of words (CBOW). Among them, the skip word mode mainly predicts the usage of the current word, while CBOW predicts the current word through the context-related vocabulary [22]. Therefore, the research uses CBOW for word vector training, and combines it with negative value sampling, so as to improve the expression efficiency and accuracy of English text, and achieve distributed semantic representation related to English text.

When calculating the output of the relevant hidden layer in the CBOW model, usually in the case of using the weight matrix of the hidden layer and the average vector, the average value of the input context word is calculated. The relevant output formula of the hidden layer is shown in equation 3.1.

$$h = \frac{1}{A} W^T (x_1 + x_2 + \ldots + x_A) = \frac{1}{A} (v_{\omega 1} + v_{\omega 2} + \ldots + v_{\omega A})^T \tag{3.1}$$

In equation 3.1, h represents the output value of the hidden layer; A represents the number of words in the context; W represents the weighting matrix; x_A represents the input vocabulary; v represents the number of dimensions of the vocabulary vector; ω_A represents the words of the context T ; There is a different weighting matrix in the process from the hidden layer to the output layer W'. According to the weights in the weighting

matrix, the calculation formula of the word score in the vocabulary is shown in equation 3.2.

$$p_j = v_{\omega j}^{'} h \tag{3.2}$$

In equation 3.2, it p_j represents the score value of each word; $v_{\omega j}^{'}$ represents the degree of the j-th column dimension in different weighted matrices W' . On this basis, the study uses the correlation log-linear classification model to obtain the posterior distribution of English words, and outputs the multinomial distribution. The relevant calculation formula is shown in equation 3.3.

$$O\left(\omega_j \mid \omega_I\right) = y_j = \frac{\exp\left(p_j\right)}{\sum_{j'=1}^{v} \exp\left(p_{j'}\right)} = \frac{\exp\left(v_{\omega_j}^{T} v_{\omega_I}\right)}{\sum_{j'=1}^{v} \exp\left(v'^{T}_{\omega_{j'}} v_{\omega_I}\right)} \tag{3.3}$$

In equation 3.3, $O\left(\omega_j \mid \omega_I\right)$ represents the multinomial distribution value; y_j represents the output value of the th unit of the v_ω output layer ; j represents W the row of the v'_ω weighting matrix; represents W' the column of the weighting matrix. For the input of context words, the purpose of research and training is to maximize the probability in equation 3.3, and according to the weight of the input context words = words, the conditional probability of the actual output words is obtained, and the final loss function is calculated. The formula is shown in equation 3.4.

$$\begin{aligned} E &= -\log\left(\omega_O \mid \omega_I, 1, \ldots, \omega_I, A\right) \\ &= -p_{j*} + \log \sum_{j'=1}^{v} \exp\left(p_{j'}\right) \\ &= -\left(v'\right)_{\omega_O}^{T} h + \log \sum_{j'=1}^{v} \exp\left(\left(v'\right)_{\omega_O}^{T} h\right) \end{aligned} \tag{3.4}$$

In equation 3.4, it E represents the loss function; j^* it represents the relevant index of the actual output word of the output layer. After the corresponding loss function is obtained, the correction formula for the weights of the hidden layer and the output layer can be derived. Among them, for the reciprocal of the net input of the first unit of E the output layer j, the relevant calculation formula is shown in equation 3.5.

$$\frac{\partial E}{\partial r_j} y_j - t_j := e_j \tag{3.5}$$

In equation 3.5, r_j represents the net input value; t_j represents y_j the label; e_j represents the reciprocal value. Similarly, the formula for calculating the reciprocal E of the weighting matrix from the hidden layer to the output layer is shown in W' equation 3.6.

$$\frac{\partial E}{\partial \omega'_{ij}} = \frac{\partial E}{\partial p_j} \frac{\partial p_j}{\partial \omega'_{ij}} = e_j h_i \tag{3.6}$$

In equation 3.6, it represents the h_ith unit in the hidden layer. i Combining equation 3.5 and equation 3.6, the modified formula of the weight from the hidden layer to the output layer is obtained as shown in equation 3.7.

$$v_{\omega j}^{'new} = v_{\omega j}^{'old} - \eta e_j h \quad for\ j = 1, 2, \ldots, v \tag{3.7}$$

In equation 3.7, it η represents the learning rate; it $v'_{\omega,j}$ represents the relevant output vector of the word . ω_j Similarly, the correction formula of the weights from the output layer to the hidden layer is shown in equation 3.8.

$$v_{\omega_{I,a}}^{(new)} = v_{\omega_{I,a}}^{(old)} - \frac{1}{A} \eta E H^T \quad for \quad a = 1, 2, \ldots, A, \ldots \tag{3.8}$$

In equation 3.8, it represents $v_{\omega_{I,a}}$ the input vector of the th word in the RH input context; a it is a vector of dimension N, which represents the sum of the output vectors of all words in a vocabulary. If the weight is calculated according to its expected error, the definition expression of each component is shown in equation 3.9.

$$EH_i = \frac{\partial E}{\partial h_i} = \sum_{j=1}^{v} \frac{\partial E}{\partial p_j} \frac{\partial p_j}{\partial h_i} = \sum_{j=1}^{v} e_j \omega'_{ij} \tag{3.9}$$

On this basis, in order to improve the efficiency of the Word2 vec embedding model, a training method using negative sampling is used to optimize the output vector update of the model. When sampling, a better probability distribution is selected, and it is called the noise distribution. By simplifying the corresponding training objective, high-quality word embeddings can be produced. The relevant calculation formula is shown in equation 3.10.

$$E = -\log \sigma\left(v_{w_k}^T h\right) - \sum_{w_j \in W_{neg}} \log \sigma\left(-v_{w_k}^T h\right) \tag{3.10}$$

In equation 3.10, w_k represents the output word; v_{w_j} represents W_k the output vector of the word; represents W_{neg} the set of negative samples of the noise distribution. In order to obtain the correction formula of the relevant words under the condition of negative sampling, the net input sum of the output unit needs to be derived, and E the formula is shown in equation 3.11.

$$\frac{\partial E}{\partial v_{w_j}'^T h} = \begin{cases} \sigma(v_{w_j}'^T h - 1), & \text{if } w_j = w_k \\ \sigma(v_{w_j}'^T h), & \text{if } w_j \in W_{\text{neg}} \end{cases} = \sigma(v_{w_j}'^T h) - t_j \tag{3.11}$$

In equation 3.11, W_j when it is a positive sample, the t value is 1; w_j when it is not a positive sample, the t value is 0. At this point, the relative derivative of the output vector of the word w_j can be obtained. E The calculation formula is shown in equation 3.12.

$$\frac{\partial E}{\partial v'_{w_j}} = \frac{\partial E}{\partial v_{w_j}'^T h} \frac{\partial v'^T{}_{w_j} h}{\partial v'_{w_j}} = \left(\sigma(v_{w_j}'^T h) - t_j\right) h \tag{3.12}$$

According to equation 3.12, the correction formula of the output vector can be obtained. The relevant calculation formula is shown in equation 3.13.

$$v_{w_j}'^{(new)} = v_{w_j}'^{(old)} - \eta(\sigma(v_{w_j}'^T h) - t_j)h \ldots \tag{3.13}$$

Therefore, the relevant iterative calculation of the embedding model only needs to w_j update the words belonging to the words in the vocabulary accordingly, so as to improve the calculation efficiency. In order to pass the prediction error to the hidden layer, take the E derivative of the output sum of the hidden layer. The calculation formula is shown in equation 3.14.

$$\frac{\partial E}{\partial h} \sum_{w_j \in \{w_k\} \cup W_{neg}} \frac{\partial E}{\partial v_{w_j}'^T h} \frac{\partial v_{w_j}'^T h}{\partial h} = \sum_{w_j \in \{w_k\} \cup W_{neg}} \left(\sigma\left(v_{w_j}'^T h\right) - t_j\right) v'_{w_j} := EH \tag{3.14}$$

Output vector can be updated accordingly by substituting the EH value calculated by equation 3.14) into equation 3.8. In addition, in order to verify the validity of the diagnostic model, the study introduces the Pearson correlation coefficient, and the calculation formula is shown in equation 3.15.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3.15}$$

In equation 3.15, $\rho_{X,Y}$ represents the Pearson coefficient; n represents the number of samples; X and Y represents the value of a certain sample in the sample; $\overline{X}$ and $\overline{Y}$ represents the mean value of the sample. $\rho_{X,Y}$ A value greater than zero indicates a positive correlation between the two samples, and a value smaller than zero indicates a negative correlation. And the larger the absolute value, the stronger the correlation between samples.

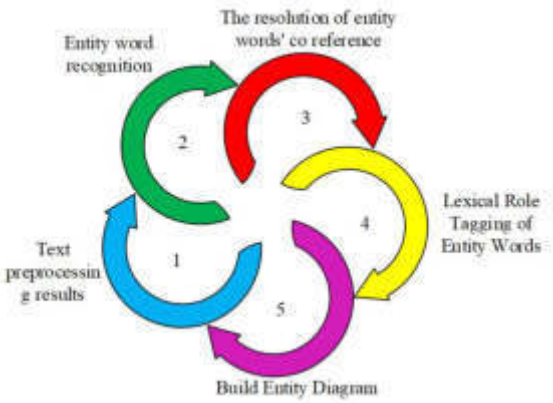


Fig. 3.3: Construction Process of Entity Diagram

3.2. Research on entity graph and sentence semantic graph based on model preprocessing results. After the English text preprocessing results are obtained, the relevant entity words in the text can be identified accordingly, and the co-referential phenomenon between the entity words can be eliminated at the same time. And combine the above information to build a text entity graphic model. The corresponding process of entity graph construction is shown in Figure 3.3.

From Figure 3.3, in the construction process, the first step is to preprocess the text and recognize entity words based on its results. In English text, entity words are usually used as morphological attributes of nouns or pronouns, so they can be identified according to this feature. Since most of the entity words in the English language are nouns or pronouns, the accuracy of part-of-speech tags is critical when identifying entity words. By marking the obtained parts of speech, nouns and pronouns can be extracted from the text. However, in some English languages, words such as numbers and symbols are sometimes marked as nouns. However, these words not only fail to explain the coherence of the article, but also act as a distraction. Therefore, after extracting all the entity words, in order to reduce the noise, the research also filters the entity words. In English texts, the research only focuses on the subject, predicate and existence of three grammatical roles played by entity words for the time being. Therefore, when marking the function of the entity word, it is only necessary to determine whether the entity word in the sentence is the subject or object, and the rest are marked as "existence". In the entity graph model constructed by the research, the process of the grammar role labeling module is shown in Figure 3.4.

From Figure 3.4, the process of the grammar role labeling module first reads the relevant syntactic analysis results of English discourses, and then classifies them. The specific operation steps first traverse the nodes in the dependency syntax tree of the statement. When traversing the entity word, look up the synonym of the entity word, and look up the entity word and its sibling nodes in the corresponding word tree library, and if it exists, it is set to mark it. Then look up all the dependencies of the current entity word, if it contains a noun subject relationship or a clause component subject relationship, mark the entity word as the subject. If there is a relationship with a direct object or an indirect object, mark the entity word as an object, if not, mark it as "existing", and save the marking result. Finally, the syntactic tree traversal is completed, and the result of the entity lexical role tagging of the English text is output.

After the sentence semantic map of the English text is generated, its features need to be analyzed, and then the entire coherence of the English text is analyzed. The English discourse has good coherence, and the discourse must have logic inside. This structure is reflected in the sentence semantic graph as the different connection modes between sentences, that is, the subgraph pattern. On this basis, the research first extracts features such as frequent subgraph frequencies, graph signatures and subgraph semantic values from the graph, and uses these feature values to further study the consistency of the text. On this basis, a coherent quality analysis can be performed. The specific analysis process is shown in Figure 3.5.

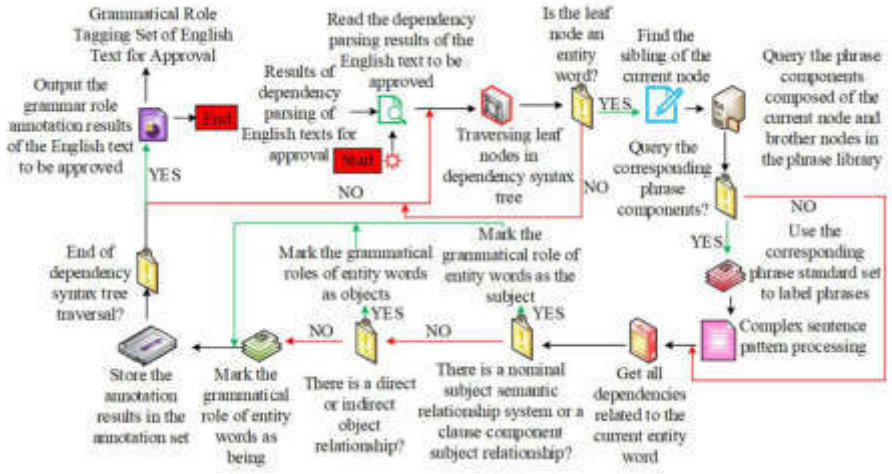


Fig. 3.4: Process of Syntax Role Annotation Module

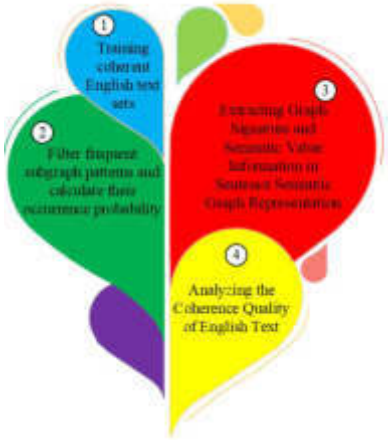
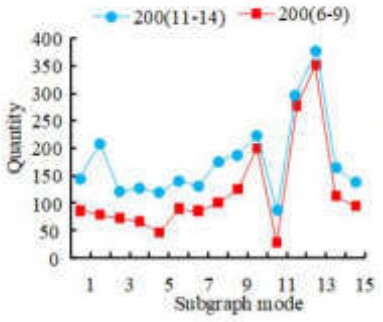


Fig. 3.5: Process of Coherent Quality Analysis

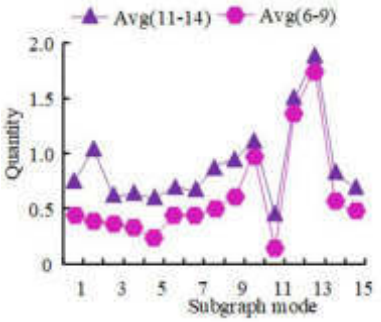
From Figure 3.5, the coherence quality analysis first uses a large number of English texts with good continuity as the training set, and counts the dot subgraphs of all three nodes and four nodes; Concentrate the subgraphs that appear frequently, and calculate the occurrence probability of each frequency subgraph to generate a frequency subgraph model and use it as the continuous subgraph distribution feature of the English text; then extract the graph features and the subgraphs in the English text. Semantic value information; finally, an algorithm is designed to analyze the consistency of English discourse by using the distribution characteristics of frequent subgraphs in the semantic graph of sentences. In addition, in the experimental setup of the English text semantic coherence diagnosis model, the research experimental environment is divided into hardware and software environments. The hardware environment includes model number 880@3.07GHz Intel processor with a memory setting of 16.00GB; The software environment is Microsoft's 64 bit Bitwise operation system. Eclipse development tools, Java programming language and Excel 2016 data analysis tools are selected. The experimental data set contains the International Corpus for Asian English Learners, which has 1.2 million words and a single text length of 200 400 words; The college English textbook corpus contains 930 English articles, all of which are from college English textbooks; The Chinese English Learner Corpus, which contains 1

Table 4.1: Relevant data set before experiment

Hardware environment		**Software environment**					
Memory	16 GB	Operating system	Windows 10 64 bit operating system	Programming language	Java Java		
Corpus of model experiment data							
-	Data _	-	Data _	-	Data _	-	Data _
ICNALE	1000 articles	COLEN	100 texts	CELC	400 compositions	TECCL	500 test sets
Grade (Points)	Criterion of comments						
0-7	The whole passage is abrupt, the semantics between sentences are not necessarily connected, the language is fragmented, and the coherence is poor						
7-13	The overall transition of the essay is poor, the semantic connection between sentences is not close enough, and the coherence is poor						
13-20	The overall transition of the essay is more natural and smooth, the semantic connection between sentences is closer, and the coherence is better						
20-25	The overall transition of the article is natural, smooth, and coherent						



((a)) Number of frequent subsets in CELC



((b)) Average number of frequent subsets in CELC

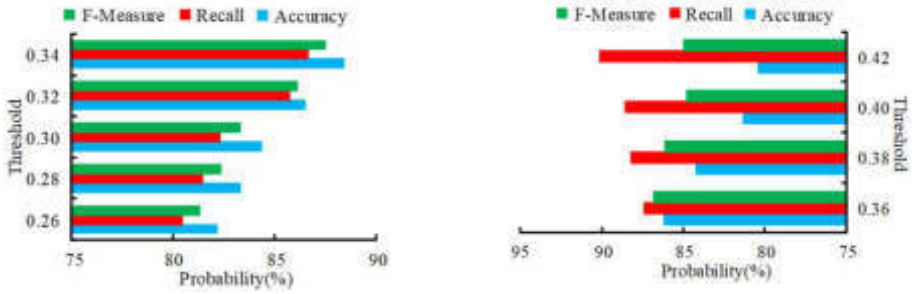
Fig. 4.2: Number of Frequent Subgraphs in Different Test Texts

million words, is a collection of essays written by college English majors and non major students; The Chinese Student English Composition Corpus, which contains 1817335 words, contains 10000 compositions written by Chinese middle school and college students for English reasons.

4. Evaluation and analysis of the semantic coherence diagnostic model of English texts. To verify the actual effect of the English text semantic coherence diagnosis model constructed by the research, the research analyzes it through three experiments, namely subgraph screening, incoherent sentence extraction, and sentence sorting. and compared it with the teacher's rating. Before the experiment, the research set up the experimental environment, experimental data and evaluation indicators of the diagnostic model, the contents of which are shown in Table 4.1.

From Table 4.1, the model experimental data selected for the study are 4 known corpora, namely the International Corpus of Asian English Learners (ICNALE) and the Corpus of College English Textbooks. English Textbooks (COLEN), Chinese English Learner Corpus (CELC) and Chinese Students 'English Composition Corpus (TECCL). The study selected 1000 articles in ICNALE as a test set for incoherent sentence extraction, 100 English texts in COLEN as a test set for sentence ordering, 200 essays in CELC with scores of 11-14, and 6 -9 points 200 essays as a test set for frequent sub-atlas and 500 essays in TECCL for comparison experiments with teacher ratings. On this basis, the research first conducts an experimental analysis on the screening of subgraphs, and the results are shown in Figure 4.2.

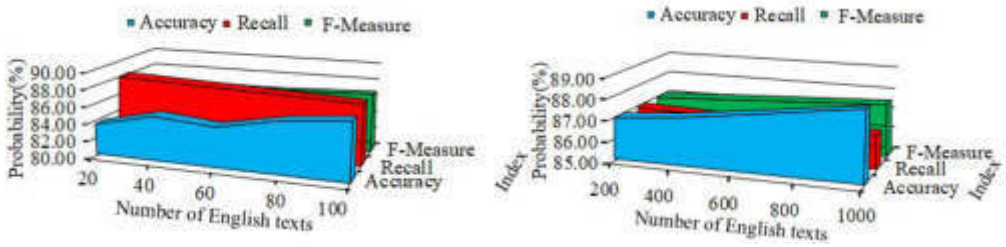
From Figure 4.2, it can be found that the subgraph patterns 10, 12, and 13 all exceed 200 on the first two



((a)) Probability%

((b)) Probability%

Fig. 4.4: Extraction of Discontinuous Sentences under Different Thresholds



((a)) Results of Discontinuous Sentences Extraction in the First Five Different English Texts

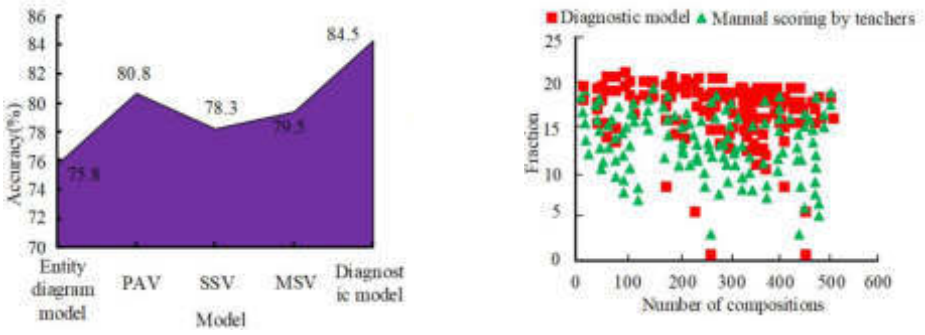
((b)) Results of Discontinuous Sentence Extraction in the Last Five Different English Texts

Fig. 4.6: Extraction Results of Discontinuous Sentences in the Number of Non English Compositions

test texts, and exceed 1 on the last two test texts. However, the differences between these three subgraphs on different test texts are not large, and the gaps are all lower than 30 and 0.3. Taken together, most of the spectrograms show large differences in both types of test text, and can distinguish between coherent text and incoherent text very well. But not all subgraphs can capture the coherent information of the text well, such as subgraph patterns 10, 12, 13, etc. There is not much difference between the average number of occurrences of these subgraphs in the test text with high degree of continuity and the test text with discontinuity, and it cannot capture the coherence of the text very well. If it is placed in the frequent subgraph, it will affect the frequency distribution of the frequent subgraph. In addition, the study extracted incoherent sentences to evaluate the constructed coherent diagnosis model, and the experimental results are shown in Figure 4.4.

From Figure 4.4, the precision rate and recall rate are different under different thresholds, and the F1 value (F-Measure) value under the combination of the two is also different. Among them, the highest precision appears at 0.34, which is 88.43%, and the highest recall rate appears at 0.42, which is 90.15%. Taken together, the optimal value of the diagnostic model occurs when the threshold is 0.34, and the F1 value is 87.54%. Too high or too low a threshold will decrease the probability of both indicators, so the optimal threshold for diagnosing the model is 0.34. On this basis, in order to verify the performance of the diagnostic model in extracting incoherent sentences, the study sets the extraction threshold to 0.34. And randomly select a certain amount of essays from the test set, and divide them into ten groups for experiments according to the number. The results are shown in Figure 4.6.

From Figure 4.6, the precision rate, recall rate and F1 value are not very different under different numbers of English compositions, which are maintained between 83% and 89%. Among them, the highest precision



((a)) Sentence Sorting Result Under Different Model

((b)) Comparison results between diagnostic model and teacher's manual scoring

Fig. 4.8: Sentence Sorting Experiment and Actual Composition Scoring Results

rate is 88.43%, the highest recall rate is 88.20%, and the highest F1 value is 87.54%. On the whole, when the number of English writings is used for testing, the accuracy of the discontinuous sentences extracted by the model still maintains a good stability, and the extraction accuracy of non-coherent sentences also increases with the increase of the number of articles. Therefore, the diagnostic model shows better performance in the actual English incoherent sentence extraction experiment. Finally, the research conducted an experiment on English sentence ordering, and compared the diagnostic model score with the teacher's manual score, and the results are shown in Figure 4.8.

In Figure 4.8(a), the study introduces an entity graph model and three semantic similarity models. The semantic similarity models are a single similar vertex (Single S imilar Vertex, SSV), multiple similar vertices (Multiple S imilar Vertices)) and the preceding adjacent vertex (P receding Adjacent Vertex). As can be seen from the figure, the diagnostic model given by the study is significantly better than the other four models, and the sentence sorting accuracy rate reaches 84.50%. As can be seen from Figure 4.8(b), in the actual English composition scoring, the diagnostic model constructed by the research is generally consistent with the teacher's manual scoring, roughly in the range of 5-25 points. Although there is a big difference in the scores of English compositions, this is a subjective task and will be affected by many subjective factors. For example, there are many mistakes in students' compositions, and teachers' understanding and correction requirements are also different. So some gaps are understandable. In addition, the Pearson correlation coefficient between the calculated diagnostic model and the manual score was 0.6025, indicating a strong correlation, indicating high effectiveness. On the whole, the diagnostic model constructed by the research has better performance in English text recognition, and it also shows high accuracy in the actual composition correction and scoring, and has strong practicability.

5. Discussion. The coherence of a discourse is a major indicator of its readability. Excellent discourse does not have a random combination of sentences and vocabulary, but rather has certain logical and coherent rules. In English articles, the brain naturally searches for special ways of logic and coherence. If this way of logic and coherence can be found, then the content of the article can be understood. However, in some large English exams in China, candidates often use texts containing a large number of advanced vocabulary and complex sentence structures in English articles to please the examiners in order to pass. However, when expressing the theme and content of the article, its logic is very chaotic and incoherent. At the same time, current research on automatic grading of English compositions has rarely designed the indicator of coherence quality, and the performance of the actual kernel construction system still needs to be improved. Therefore, an English text semantic coherence diagnosis model was constructed on the basis of sentence semantic graph, and its accuracy and effectiveness were verified through experiments.

The experimental results show that sub Graphical model 10, 12 and 13 exceed 200 in the first two test

texts and 1 in the last two test texts; The accuracy and recall rates are different under different thresholds, and the F1 value (F-Measure) value under the combination of the two is also different. Among them, the highest accuracy occurs at 0.34, which is 88.43%, and the highest recall rate occurs at 0.42, which is 90.15%. This result is consistent with the results of Farkhodovna M S [23]. At the same time, there was no significant difference in accuracy, recall, and F1 values among different English composition numbers, maintaining between 83% and 89%. Among them, the highest accuracy rate is 88.43%, the highest recall rate is 88.20%, and the highest F1 value is 87.54%. This result is superior to Uru OB et al.'s 87.65% [24]. The diagnostic model provided by the study is significantly superior to the other four models, with a sentence sorting accuracy of 84.50%. This result has advantages compared to Keskin D et al [25].

Overall, the diagnostic model constructed in the study has better performance in English text recognition, and it also shows high accuracy in actual essay grading.

6. Conclusion. To effectively diagnose the semantic coherence quality of English texts, a diagnostic model is constructed by using the sentence semantic graph, and its performance is experimentally verified. The experimental results show that in the frequent subgraph screening experiment, the subgraph modes 10, 12 and 13 have little difference in different text tests, and the rest all show large differences. Therefore, in order to improve the performance of the diagnosis model, it is necessary to delete these Subgraph; in the incoherent sentence extraction experiment, the highest precision reached 88.43, the highest recall rate reached 90.15%, and the best threshold for the comprehensively calculated F1 value was 0.34, and the probability reached 87.54% at this time; the threshold was fixed at 0.34. In the experiment of extracting incoherent sentences, the precision rate and recall rate were 88.43% and 88.20% respectively under different numbers of English compositions; in the comparison experiment between sentence sorting and actual scoring, the diagnostic model had the highest accuracy rate of 84.50%, and in practice There is not much difference between the scores and the manual scores of teachers, showing high accuracy. In addition, the calculated Pearson correlation coefficient was 0.6025. On the whole, the diagnostic model constructed by the research shows good performance, high accuracy and high reliability in the actual English incoherent sentence extraction experiment and English composition correction. It is worth noting that although the research has introduced semantic information between sentences into the Entitative graph model, the semantic relationship between sentences cannot be well expressed, so its ability to represent the semantic information of the whole text has limitations, which needs further improvement in the future. At the same time, the frequency of nodes that the research mainly focuses on in coherence quality analysis is not comprehensive enough, so it is easy to solve the problem of sparse data. In the future, it is necessary to increase the focus to solve this problem.

REFERENCES

[1] Najafi, E., Valizadeh, A. & Darooneh, A. The Effect of Translation on Text Coherence: A Quantitative Study. *Journal Of Quantitative Linguistics*. **29**, 151-164 (2022)
[2] Akmilia, P., Faridi, A. & Sakhiyya, Z. The Use of Cohesive Devices in. (Conference to Achieve Texts Coherence. English Education Journal, 12(1): 66-74,2022)
[3] Hu, L., Tang, Y., Wu, X. & Zeng, J. Considering optimization of English grammar error correction based on neural network. *Neural Computing And Applications*. **34**, 3323-3335 (2022)
[4] Gondaliya, Y., Kalariya, P., Panchal, B. & Nayak, A. A Rule-based Grammar and Spell Checking. *SAMRIDDHI: A Journal Of Physical Sciences, Engineering And Technology*. **14** pp. 01 (2022)
[5] Chaabi, Y. & Allah, F. Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. *Journal Of King Saud University-Computer And Information Sciences*. **34**, 6116-6124 (2022)
[6] Li, X., Li, X., Chen, S., Ma, S. & Xie, F. Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. *Connection Science*. **34**, 1638-1653 (2022)
[7] Chen, J., Zhang, L. & Parr, J. Improving EFL students' text revision with the self-regulated strategy development (SRSD) model. *Metacognition And Learning*. **17**, 191-211 (2022)
[8] Srivastava, K., Dhanda, N. & Shrivastava, A. Optimization of Window Size for Calculating Semantic Coherence Within an Essay. *ADCAIJ: Advances In Distributed Computing And Artificial Intelligence Journal*. **11**, 147-158 (2022)
[9] Yang, X., Zhang, J., Chen, K., Zhang, W., Ma, Z., Wang, F. & Yu, N. . *Tracing Text Provenance Via Context-aware Lexical Substitution//Proceedings Of The AAAI Conference On Artificial Intelligence*. pp. 11613-11621 (2022)
[10] Aminovna, B. Importance of coherence and cohesion in writing. *Eurasian Research Bulletin*. **4**, 83-89 (2022)
[11] Saleh, M. & Bharati, D. The Use of Cohesive Devices in Descriptive Text by English Training Participants at PST". *English Education Journal*. **12**, 95-102 (2022)

[12] Dassanayake, N. Exploring Coherence among Sri Lankan CFL Learners in Chinese-English Translation: Decoding and Interpreting of Culture-loaded Content. *International Journal Of Language And Literary Studies*. **4**, 350-363 (2022)
[13] Akmilia, P. & Faridi, A. Sakhiyya. (Z. The Use of Cohesive Devices in,2022)
[14] Abdusalomovna, K. Theoretical background of using cohesion in discourse. *Web Of Scientist: International Scientific Research Journal*. **3**, 687-693 (2022)
[15] Zhang, K., Zhou, Y., Chen, Z. & Others (2022) Incorporating Biterm Correlation Knowledge into Topic Modeling for Short Texts. *The Computer Journal*. **65**, 537-553 (0)
[16] Linnik, A., Bastiaanse, R., Stede, M. & Khudyakova, M. Linguistic mechanisms of coherence in aphasic and non-aphasic discourse. *Aphasiology*. **36**, 123-146 (2022)
[17] Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R. & Gallinari, P. Controlling hallucinations at word level in data-to-text generation. *Data Mining And Knowledge Discovery*. **36**, 318-354 (2022)
[18] Gaur, M., Gunaratna, K., Srinivasan, V. & Jin, H. . *Iseeq: Information Seeking Question Generation Using Dynamic Meta-information Retrieval And Knowledge Graphs//Proceedings Of The AAAI Conference On Artificial Intelligence*. pp. 10672-10680 (2022)
[19] Ruzikulovich, A. Functional-semantic and linguo-poetic capabilities of imperative structures. *EPRA International Journal Of Multidisciplinary Research (IJMR)*. **8**, 219-221 (2022)
[20] Zhao, L. & Xu, J. A Study on the Translation Strategies of The Nine Songs from the Perspective of Cognitive Construal: A Comparative Analysis of the Yangs' and Waley's Versions. *International Journal Of Linguistics, Literature And Translation*. **5**, 56-62 (2022)
[21] Wang, Z. & Wang, J. The Grammatical and Semantic Functions of "with" Structure in Chinese-English Translation. *International Journal Of Linguistics, Literature And Translation*. **5**, 109-116 (2022)
[22] Feng, Y., Hu, C., Kamigaito, H. & Others (2022) A simple and effective usage of word clusters for CBOW model[J]. *Journal Of Natural Language Processing*. **29**, 785-806 (0)
[23] Farkhodovna, M. THE STRUCTURE OF PROVERBS AND PHRASEOLOGICAL UNITS IN ENGLISH. *IJTIMOIY FANLARDA INNOVASIYA ONLAYN ILMIY JURNALI*. **3**, 83-87 (2023)
[24] Uru, O., Sudirman, A. & Nugroho, A. Exploring cohesions in EFL academic writing: A state of the art on the study of cohesions. *Elsya: Journal Of English Language Studies*. **3**, 141-149 (2021)
[25] Keskin, D. & DEMİR, B. The role of theme and rheme in thematic progression patterns in English argumentative essays by Turkish University students. *Edu*. **10**, 64-82 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Nov 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 341–354, DOI 10.12694/scpe.v25i1.2299

TRADITIONAL CULTURAL NETWORK ONLINE EDUCATION INTEGRATING DEEP LEARNING AND KNOWLEDGE TRACKING

HENG ZHAO* AND ZHIYUAN SUN†

Abstract. With the popularization of the computer network and the development of artificial intelligence technology, the traditional education industry has been reformed. In the past two years, online education has developed rapidly. The combination of the Internet and education enables students to study online at any time, no longer relying on the time and place requirements in traditional education. However, with the rapid development of online education, many problems have gradually emerged. In online education, with a large amount of knowledge and question banks, students are faced with a large number of choices. Therefore, positioning and tracking the knowledge level of students and realizing personalized online education have become the main problems facing the moment. Based on this, this study integrates deep learning and knowledge tracking technology to build a traditional cultural network online education model, aiming at accurately positioning students' knowledge levels and recommending personalized question banks. The experimental results show that the average AUC of the model proposed in this study is 0.781, and the average accuracy rate is 0.886, which is significantly better than other online education models. Through the combination of deep learning and knowledge tracking technology, the research successfully provides a new and efficient model for personalized learning in the field of online education, which is of great guiding value for promoting further innovation and development of online education. In addition, the research also provides practical solution strategies for related fields, which have obvious practical significance and popularization value.

Key words: Online education; Personalized recommendation; Deep learning, Knowledge tracking cost

1. Introduction. As the popularization of the Internet and the growth of artificial intelligence technology, online education has gradually become one of the trends of modern education [1]. With the advantages of the Internet and big data, online education has changed the demand for time and location in traditional education [2]. And it can provide users with rich, real-time updated learning materials, reducing the gap between the education level of different development levels in the region [3]. However, with the gradual development of online education, more problems are gradually revealed. These problems are as follows: teachers can not follow up the learning status of students in real time; students in the knowledge base is difficult to choose; the increase of meaningless learning and other problems. The current online education industry practitioners are plagued by these problems [4]. In the traditional teaching mode, teachers are difficult to rely on manual modeling of learning resources for each student to achieve the screening of learning, and they are also unable to accurately perceive the learning changes of learners. As the advancement of artificial intelligence technology, each learner can easily build a unique learning resource model. However, how the huge database of teaching resources can be accurately recommended to students to achieve their personalized development is still a major problem that needs to be solved [5]. To address this problem, this study proposes a traditional culture network online education model that integrates deep learning (DL) and knowledge tracking (KT), aiming to discover learning resources suitable for learners from a large amount of data through DL. And through KT technology, students' learning data are mined and used for online education, aiming to improve the correct rate of predicting students' correct answers to questions, thus improving the effectiveness of online teaching. In addition, the research innovatively combines knowledge mapping to obtain potential connections between items through the description of semantic associations. The research also expands users' interests through various types of associative relationships, which improves the accuracy and diversity of traditional online education recommendation systems. The research aims to combine modern science and technology to provide more effective aids for online teaching and to help learners learn more efficiently. The research also provides a

*School of Tourism and Hotel Management, The Open University of Shaanxi, Xi'an, 710119, China (hengzhaohz@outlook.com)
†School of Marxism, The Open University of Shaanxi, Xi'an, 710119, China (xuegen2021@163.com)

reference for the application of DL and KT in different fields. This research is composed of five main parts. The first part is an overview of the research as a whole; the second part is a summary of related work at home and abroad; the third part is divided into two subsections; the first one describes the model improvement of KT algorithms based on DL, and the second one describes the construction of KT online education models based on DL; the fourth part is an experimental validation of the performance of the KT models proposed in the study; the fifth part is a summary of this study, and it has been an outlook on the future research.

2. Related works. Recently, research on DL has attracted wide attention in various fields. Shen L et al. believed that the short-term forecast of passenger flow had an important significance to the scheduling of subway systems and other aspects. An improved gravity model with DL was proposed, and a short-term forecasting method was upgraded to balance model interpretability and forecasting accuracy. This method had good performance in the experimental results, which was better than other models [6]. This method provided a more reliable solution for improving the accuracy of short-term forecasting of passenger flow in subway systems. In the data-driven decision support system, Garg et al. designed a new data processing framework based on a DL network by using the combination of density-based clustering and variable-length sequence decoding method. The concept of data association was used to track the occurrence and evolution of such events to understand the operating environment. Simulation experiment denoted that this method could perceive the sensitivity of performance and parameters better than others [7]. The method provided a new reliable reference for DL in data processing. Tang et al. found that the linear defect detection of a photovoltaic module was a key link in the health assessment of photovoltaic devices, but the traditional defect diagnosis was actually inefficient manually. Therefore, by distributing computing tasks among edge devices, edge servers and cloud servers, a defect detection algorithm was proposed based on DL. The effectiveness and accuracy of the method were verified by experiments [8]. The method simplified the task of allocating computation between edge and cloud servers for big data processing. Bernardini et al. introduced a DL framework to predict baryon fields. The new model included two network parts, U-Net and Wasserstein, and combined the universe volume and amplified fluids from the real-life environmental feedback project. The dynamical simulation was used to represent a large range of scales. The experimental results indicated that the accuracy was within 10%, which was in good agreement with the cosmological simulation [9]. The research introduced DL to cosmological tasks, providing a viable solution for baryon field prediction. Liu C et al. developed a hierarchical approach based on DL network embeddings to identify patient subtypes from large-scale patient somatic mutation profiles. A network embedding approach encoded genes on the protein interaction set to construct patient vectors. The high classification accuracy indicated that the web-embedded based patient features were reliable in classifying patients [10]. This study provided a powerful DL approach for personalized cancer treatment.

As online education has become a popular research field, the research on data mining technology in online education has also attracted the attention of scholars, and the KT model has become a key technology for simulating the state of students. Song X et al. believed that there were problems in KT research that considered the single relationship and the lack of interpretability, so they proposed a deep KT framework based on a joint graph convolutional network, which modeled multi-dimensional relationships as graphs for relationship fusion. Experiments proved that the method performed better than others. And the interpretability of learning analysis was proved through a case study [11]. The study provided a more effective solution to the problem of considering a single relationship with a lack of interpretability. Pavlik PI et al. introduced a formal learner modeling method - logical knowledge tracing (LKT), which integrated many existing learner modeling methods. The advantage was that the logistic regression model of choice provided the specification of a symbolic system that could specify many existing models in the literature and many new ones. Experimental outcomes indicated that the approach considering multiple learner model characteristics and learning environments was correct [12]. This study integrated existing learner modeling approaches and provided an easier reference for future research. To alleviate the sparsity problem in the summary KT online learning system, Gan W et al. proposed an enhanced learning model with an attention mechanism for graph representation, and used the model to skill embeddings. Experiments from three datasets verified the superiority and interpretability of this model [13]. This research provided a reliable solution to the sparsity that exists in online learning systems. Huang Y et al. found that it was difficult for students to find suitable exercises from a large number of topics provided by many online systems in programming training. This study found a new model for KT, which added additional information

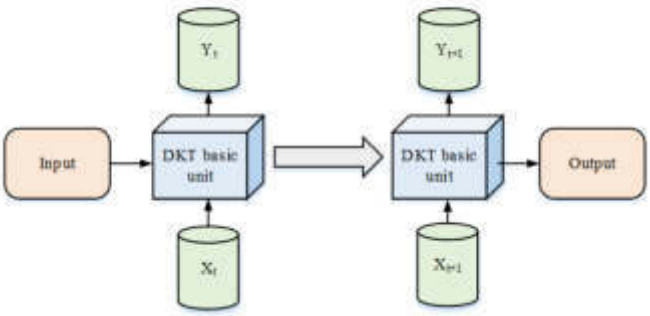


Fig. 3.1: DKT model structure

representing the relationship between exercises to the input data and compressed the input vector to solve the dimensionality problem. Research outcomes proved that the area of the proposed deep KT model was 0.7761, which was better than other KT models and ran faster [14]. This research addressed the dimensionality problem in KT for online teaching and learning systems. Liu S et al. found that the current KT model only attributed the learner's feedback, while ignoring the influence of mental ability factors. Therefore, a new ability-enhanced KT model has been found. It introduced the ability factor into the attribution of feedback, and designed a continuous matrix factorization model. The results proved that the proposed method was better than other KT algorithms in terms of quantitatively evaluated predictive accuracy [15]. This study innovatively introduced the influence of mental ability factors to optimize the KT system more comprehensively.

To sum up, DL is widely used in computer algorithms to improve the accuracy of algorithm models. At the same time, KT models have also been widely used and developed in online education. Therefore, this study innovatively integrates DL and KT, and proposes a new type of traditional cultural online education system. To further improve the interpretability of the deep KT model, the study is based on a dynamic key-value memory neural network, which incorporates two main memory modules for storing knowledge conceptual information and students' state information. In addition, the study recodes the data with the embedding method to avoid the waste of memory resources caused by too high dimensionality. The study aims to provide a more efficient method for accurate recommendation of online learning resources.

3. DL combined with KT online education model.

3.1. Improvement of KT algorithm model based on DL. With the popularity of online education, KT has become one of the best hot research directions. KT uses students' historical learning information to mine students' learning data for online education, with the purpose of improving the correct rate of predicting students' correct answers to questions [16]. The traditional DL-based KT algorithm model, which is named deep knowledge tracing (DKT), is constructed on the long-short-term memory network (LSTM) model. The LSTM model is based on the recurrent neural network (RNN), adding forgetting, input and output gates, so that the network can remember a longer input process [17]. The DKT model is shown in Figure3.1. The input data type of the model is a $0, 1^2 N$ sequence, in which N is the number of items. Only one position in the sequence is 1, and the others are 0. For example, if the record "the answer to the first a question is correct", the first element a in the code is 1, and the others are 0. If it is recorded that "the answer to the first a question is wrong", the first element $N + a$ is 1, and the others are 0. The model is an encoder-decoder type model whose output is a sequence of length. During training, the one-hot code $t + 1$ at $q^t + 1$ moment contains both the topic information and the correct and incorrect information. The topic concept scalar at moment corresponds to the prediction information of the next moment in the output sequence. The input gate formula is shown in equation 3.1.

$$i_t = \sigma(W_i[h_t - 1, x_i]) + \sigma b_i \tag{3.1}$$

In equation 3.1, i_t represents the input gate at t moment; W_i means the weight matrix of the input gate; $[h_t - 1, x_t]$ denotes the connection of two vectors into a longer vector; $h_t - 1$ and x_t expresses the output items in the two matrices, respectively; b_i refers to the bias of the input gate; the set item *sigma* means the Sigmoid function. The formula of the forget gate is shown in equation 3.2.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t]) + \sigma(b_f) \tag{3.2}$$

In equation 3.2, f_t stands for the forget gate at t time; Wf refers to the weight matrix of the forget gate; b_f represents the bias term of the forget gate;σ expresses the Sigmoid function. The forget gate memory formula is shown in equation 3.3.

$$\widetilde{C_t} = tanh(W_c \cdot [h_{t-1}, x_t]) + tanh(b_c) \tag{3.3}$$

In equation 3.3,$\widetilde{C_t}$ refers to the forget gate memory at t moment; W_c denotes the weight matrix of the forget gate memory; b_c indicates the bias item of the forget gate memory; $tanh$ means the hyperbolic tangent function. The output gate formula is shown in equation 3.4.

$$O_t = \sigma(W_\circ \cdot [h_{t-1}, x_t]) + \sigma(b_\circ) \tag{3.4}$$

In equation 3.4, O_t indicates the output gate at t moment; W_t is the weight matrix of the output gate; $b\circ$ stands for the bias item of the output gate; σ expresses the Sigmoid function. Long memory and short memory are shown in equation 3.5 .

$$\begin{cases} h_t = o_t * tanh(C_t) \\ C_t = f_t * C_{t-1} + i * \widetilde{C_t} \end{cases} \tag{3.5}$$

In equation 3.5, h_t and C_t denote long memory and short memory, respectively. f_t means forget gate, $\widetilde{C_t}$ denote forget gate memory; O_t indicates output gate; $tanh$ refers to hyperbolic tangent function. The above formula shows that the output in the model is determined by the long memory, and the long memory is affected by the short memory, so the output is affected by the current input and the long-term input. In the LSTM model, long and short memory vectors are passed between each unit [18]. The improved model of this study is based on the dynamic key value memory neural network with two main memory modules. The memory module composed of M and d vectors stores the knowledge concept information and the student's state information, respectively. M is a parameter that can be set, and the dimensions in the module d can also be different. Another improvement is to re-encode the data using the embedding method. Through the embedding matrix, the length of the data $2N * d$ is reduced to the length of the data $2N$, avoiding the waste of memory resources caused by excessive dimensionality. The improved model is shown in Figure 3.2.

As shown in Figure 3.2, the improved algorithm model mainly consists of the embedding process, the graph process, and the knowledge concept encoding process, which correspond to the blue, green, and yellow labels, respectively. In the figure, the embedding process converts high-dimensional, sparse data (e.g., words, student IDs, topic IDs, etc.) into low-dimensional, continuous vector representations, so that data with similarities are closer to each other in this low-dimensional space. In the education domain, the embedding of students and topics captures their implicit attributes or properties. The purpose of the graph process is to construct a network of relationships between knowledge points or between topics to capture and convey information through graph structures. In each iteration of the DKVMN model, the representation of each node is updated according to its neighboring nodes. In this way, the relationships and hierarchies between knowledge points can be explicitly encoded in the model. The purpose of the knowledge concept encoding process is to encode each knowledge point or concept as a continuous vector that captures the dependencies and complexity between knowledge points. The implementation of knowledge concept encoding first obtains an initial representation of each knowledge point through the embedding process. Then, these representations are refined and optimized through the graph process. Eventually, each knowledge point will have a vector representation that adequately captures the essence of the knowledge and its relationship with other knowledge. The calculation of the weight is processed by the embedding matrix, and the dot product operation is performed on the input data with

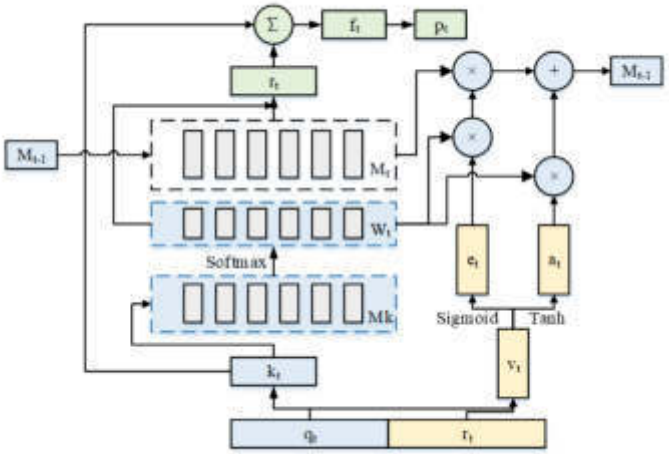


Fig. 3.2: DKVMN model

only the concept of the problem and the vector in the knowledge memory matrix. A vector whose length is equal to the number of vectors in the memory matrix is got, and the weight vector after going through the Softmax layer is got. Softmax is used as an activation function, and its calculation process expression is shown in equation 3.6.

$$w_t(i) = Softmax(k_t^T M^k(i)) \tag{3.6}$$

As shown in the formula, the activation function Softmax will output the sum of the weight vector elements equal to 1. The improved algorithm model uses knowledge memory and state memory modules. The state memory module is used to record the state at a certain moment, and the weight vector is used to weight and add the vectors in the state matrix during reading M . Embedding matrices are enabled to further correlate inputs in global gradient descent optimization. The reading process is shown in equation 3.7.

$$\begin{cases} r_t = \sum_{i=1}^{N} w_t(i) M_t^v(i) \\ f_t = tanh(W_1^T[r_t, k_t] + b_1) \\ p_t = Sigmoid(W_2^T f_t + b_2) \end{cases} \tag{3.7}$$

In equation 3.7, r indicates the product sum of the weight and the state matrix, which is the value obtained after the connection of r and k; p indicates the output of the Sigmoid function. The writing part is to adjust the matrix according to the knowledge concept at t moment and the information of whether the answer is correct. The input information generates vectors that control forgetting and memory after embedding. In the writing, both the control memory vector and the previous vector are affected by the association weight, and this improvement makes the process of writing affected by the relation of knowledge concepts. The process of updating the state matrix is shown in equation 3.8.

$$\begin{cases} e_t = sigmoid(E^T v_t + b_e) \\ \widetilde{M_t^v}(t) = M_{t-1}^v(t)[1 - w_t(i)e_t]) \\ a_t = Tanh(D_{vt}^T + b_a)^T \\ M_t^v(i = \overline{M}_{i-1}^v(i) + w_t(i)a_t \end{cases} \tag{3.8}$$

In the analysis of the improved model, the study find that the three main modules of the model interact with each other, which will complicate the interaction between knowledge concepts and states, thereby reducing

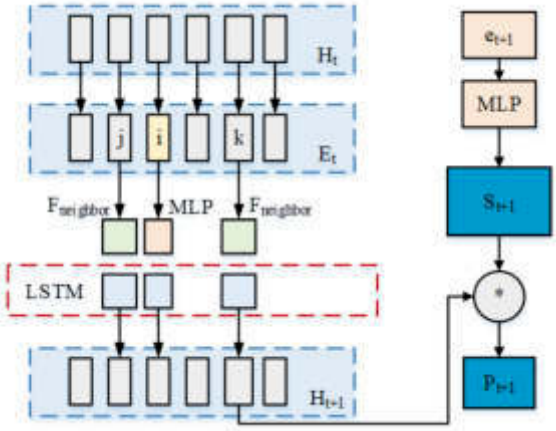


Fig. 3.3: Optimized DL model

the interpretability of knowledge concepts. Therefore, this study introduces deep reinforcement learning for personalized recommendation. The inputs of reinforcement learning include states, actions, and rewards. The key to implementing reinforcement learning is to output the optimal action in the current state to maximize the reward. In reinforcement learning, the long-term rewards obtained through training are optimal, so the summation of rewards is very important. The study introduces a discount factor γ for weighted summation. The expression of weighted summation is shown in equation 3.9.

$$U(S_\circ) = E[\sum_{t=1}^{\infty} \gamma^t R(S_t)] \tag{3.9}$$

In equation 3.9, U stands for the sum of rewards, that is, the average expectation of the weighted sum of rewards for all future action choices. R indicates the reward, that is, the system determines the value of an action S. represents the state, which is the only description of the current environment. This study exploits a recurrence relation when calculating total value to obtain a policy function that maximizes long-term returns. Ideas in reinforcement learning are utilized to divide the long sequence of action decisions, and the mathematical expression is shown in equation 3.10.

$$U^*(S) = max_a R(a) + U^*(S') \tag{3.10}$$

In equation 3.10, S expresses all the best paths; a denotes the current moment selection action; S' means the remaining paths. After optimization, a small part of the end of the optimally selected path can be truncated, and the remaining path will also become the optimal path.

3.2. Construction of KT online education model based on DL. Some studies have introduced two memory modules to store knowledge status and conceptual relationships, increasing the interpretability of the traditional cultural KT [19]. There are some classic theoretical models in the KT problem, including the combination of dynamic key-value storage network for KT and item response theory of classical cognitive diagnosis model. On this basis, some scholars proposed the knowledge query network (KQN) structure, which increases the interaction between knowledge status and knowledge concepts, to clarify the principle of the specific operation of the model, and further enhance the interpretability of the KT model [20]. The optimized DL model is shown in Figure 3.3. In the Figure 3.3, the improved DL model includes three parts. The first part is the embedding process, in which H is a memory matrix, representing the matrix that stores each knowledge state. H contains N independent vectors, in which N is the total number of problem concepts [21]. H_t

denotes each vector dimension, means the time-memory matrix. E_t indicates the result matrix after inputting embedding processing. The moment matrix E_t is a matrix containing $N-e$ dimensional vectors, and the expression is shown in equation 3.11.

$$E_k^t = \begin{cases} x^t E(k=i) \\ E_c(k)(k \neq i) \end{cases} \tag{3.11}$$

As shown in equation 3.11, the second part is the graph process. The input at t time is (q_t, a_t) . q_t denotes the concept of the problem at t time and the adjacent nodes are recorded as j, k . At this time, only the vectors j and k in the memory matrix are updated. The update process goes through a fully connected part and a LSTM network [22]. The result is recorded as the vector i in the memory matrix, as shown in equation 3.12.

$$H_k^{t+1} \begin{cases} RNN(f_{MLP}(h_k^{'t}))(k_i) \\ RNN(f_{neighbor}(h_i^{'t}))(k \neq i) \end{cases} \tag{3.12}$$

In equation 3.12, the core of this part is the function of k vector. In this study, the setting threshold $f_{neighbor}$ of the function is improved to simplify the model. When the moment is t, k is the title.k and n vectors are not considered, only the adjacent vectors need to be updated. It assumes that the concept input at $t+1$ time is e_{t+1}, after the feed-forward network, the output is a code of length S_{t+1}, the expression is shown in equation 3.13

$$S_{t+1} = \sigma(W_1) \cdot (\sigma(W \circ \cdot e_{t+1} + b_\circ)) + b_1) \tag{3.13}$$

The proposed model uses DKT network and the deep reinforcement learning network DQN to solve the problem of large memory requirements and long calculation time of the DKT network. It solves the problem that the DQN network cannot further study the relationship between clustering internal concepts with the help of directed graphs relationship problem [23]. In the application of recommendation algorithms, this study innovatively combines knowledge graph (KG) into an online education system that integrates DL and KT, and establishes a complete traditional cultural online education system [24]. KG obtains potential links between items through the description of semantic associations, expands users' interests through various types of associations, improves the accuracy and diversity of recommendations. KG can also maintain users' historical learning data well, increasing interpretability of recommendation results. In the improved recommendation algorithm, the conceptual calculation of the user's operation on the item is shown in equation 3.14.

$$P_i = \frac{exp(v^t R_i h_i)}{\sum_{(h,t,r)\epsilon S_u^1} exp(v^T R_i h_i)} \tag{3.14}$$

In equation 3.14, v^T indicates the candidate item vector; R_i mean the edge of the i node; h_i represents the feature vector; S_u^1 refers to propagation process set. The user representation is obtained according to the probability, and the calculation process is shown in equation 3.15.

$$\begin{cases} O_u^1 = \sum_{h_{i,r_i,t_i \in S_u^1}} p_i t_i \\ u = o_u^1 + o_u^2 + ...o_u^H \end{cases} \tag{3.15}$$

In equation 3.15, t_i stands for the feature vector of the node's tail node. The similarity between the user and the item is calculated to complete the recommendation at last. The expression is shown in equation 3.16.

$$\widehat{y_{uv}} = \frac{1}{1 + exp(-u^T v)} \tag{3.16}$$

In equation 3.16, $\widehat{y_{uv}}$ indicates the user u 's click probability for the item v . The research and design of the learning system of traditional cultural network online education divides users into students and teachers

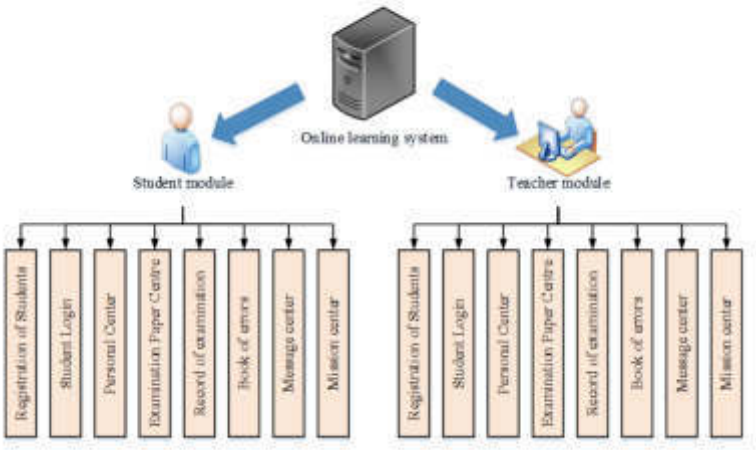


Fig. 3.4: System function overview diagram

Table 3.1: Data set information

Data Set	**Algebra_2006_2007**	**Assistments2009-2010**
Number of Logs	52324	58851
Individual Users	511	976
Number of Test Questions	1734	12461
Points of Knowledge	216	110

[25]. Teachers are complex in student management, knowledge management, topic preparation, task release, etc.; students are responsible for completing tasks, taking exams, and answering questions online. An overview of system functions is shown in the Figure 3.4.

In the overall process of traditional cultural online education, the teacher first logs in to the system, checks the students' learning and answering conditions, changes the topics and publishes learning tasks; then the students log in to the system, checks the tasks and notifications, and conducts answering training and knowledge learning. System statistics record students' answers, and input the knowledge structure relationship hidden behind the questions into the model of research design to obtain the sequence of topic recommendations. Finally, students conduct feedback exercises through recommended topics, conduct targeted training, and improve learning efficiency. The overall flow chart of traditional cultural online education is shown in the Figure 3.5.

3.3. Performance comparison results of KT models based on fusion DL. To verify the effect of the knowledge recommendation model based on KT proposed in this study, the experimental data set used algebra_2006_2007 and Assistment2009-2010 data sets that were often used in the field of KT. The algebra_2006_2007 data set was the interaction record between the user and the computer-aided system, and the Assistment2009-2010 data set recorded the practice log of elementary school students' mathematics exercises, including 62955 records. The ratio of the training, test, and verification sets used in the research was 8:1:1. After preprocessing the data, the relevant information of the data is shown in Table 3.1.

The study first explored the impact of different discount factors and different lengths of question sequences on the system, and then established evaluation indicators including area under curve (AUC), accuracy and knowledge mastery fluctuations to verify the improved model's performance proposed in this study. Under different discount factors, the changes in the cumulative rewards of the algorithm model during training iterations

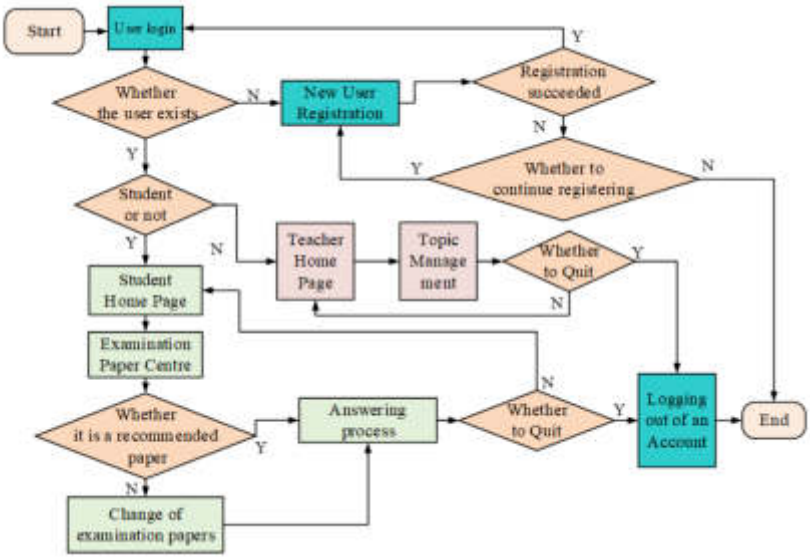
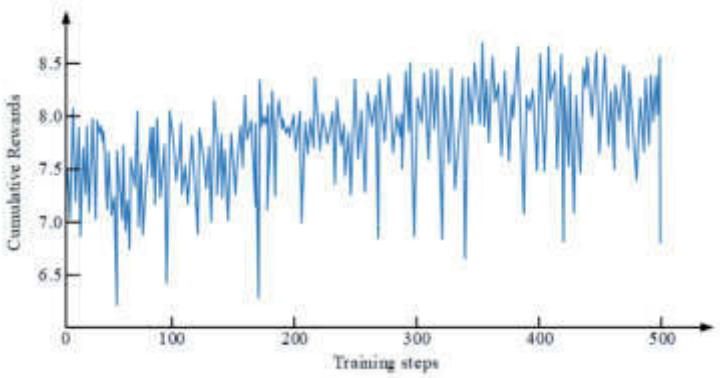


Fig. 3.5: Traditional culture online education flow chart

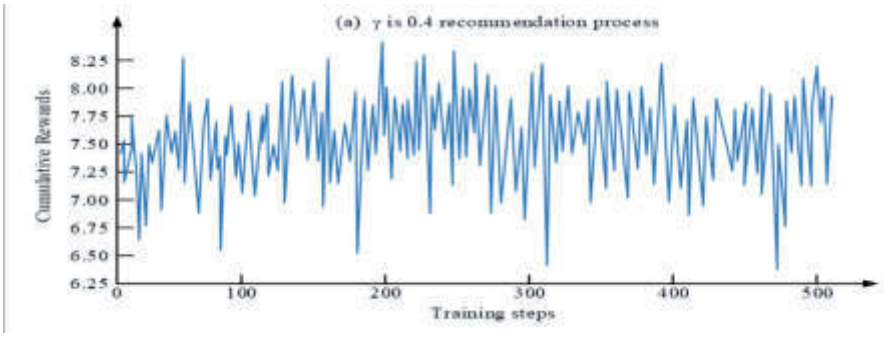
are shown in Figure 3.7.

From Figure 3.6(a), a clear trend can be observed with a discount factor set at 0.4. The cumulative reward of the improved algorithm model exhibited a consistent increase as the number of training iterations escalated. This incremental pattern suggested a potential correlation between the training iterations and the model's efficacy. Conversely, Figure 6(b) portrayed a different scenario when the discount factor was zeroed. The cumulative reward oscillated between 6.32 and 8.47 throughout the training iterations. Notably, for up to 500 training iterations, the reward's trajectory appeared relatively static, devoid of any pronounced pattern. A deeper dive into Figure 3.7 revealed an enlightening insight: the model harnessing a 0.4 discount factor outperformed its counterpart, especially as iterations proliferated. This underpinned the assertion that a long-term recommendation strategy held an advantage over a heuristic approach. To further underscore this, it was essential to test the improved model's capability across various topic sequence lengths. Topic sequences of different lengths were recommended to verify the recommendation ability of the improved model. The result is shown in the Figure 3.9.

Figure 3.9 provided an in-depth analysis of the fusion DL and KT recommendation model across varying application settings. Specifically, Figure 3.8(a) demonstrated the model's performance when recommending 3 out of 5 questions. By approximately the 150th training step, the model verged on full convergence, maintaining a steady trajectory with an overall reward range between 0.38 and 1.53. Meanwhile, Figure 3.8(b), which depicted the recommendation of 4 out of 8 questions, revealed a similar pattern of stability by around the 275th training step, having a reward range of 1.78 to 2.25. However, when ventured into more complex scenarios, the dynamics shifted. Figure 3.8(c), representing 5 out of 15 recommendations, suggested that the model struggled to converge even after 500 training iterations, with a reward spectrum of 3.05 to 4.03. Similarly, Figure 3.8(d) which tackled 10 out of a hefty 50 questions confirmed this challenge. Despite 500 training cycles, convergence remained elusive, and rewards oscillated between 6.26 and 8.55. Drawing insights from these findings, a pattern emerged: the model's convergence and stability were inversely proportional to the number of topics to be recommended. For compact topic sets, the model converged swiftly. However, as the breadth of topics expanded, the convergence rate dwindled, culminating in the model's inability to adequately cater to personalized learning needs across a diverse student body. AUC represents the area enclosed by the ROC and the X coordinate, which can better reflect the pros and cons of the classifier. The closer the AUC value was to 1, the better the classifier effect. Common algorithms KT, DKT, DKVMN, GKT and DL-KT algorithm



((a)) γ is 0.4 recommendation process



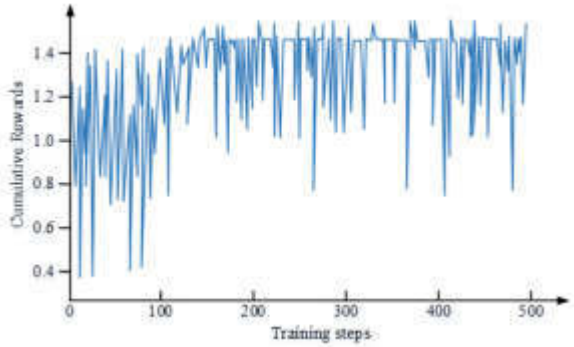
((b)) γ is 0 recommendation process

Fig. 3.7: Training performance of the algorithm model under different discount factors

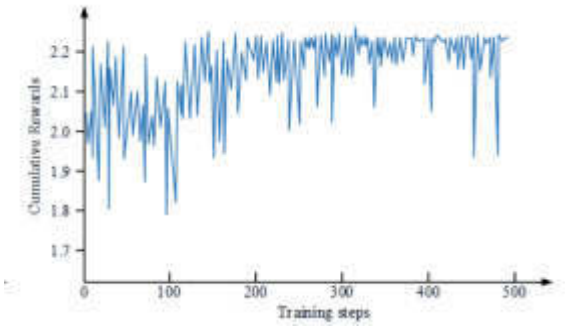
model proposed in this study were compared. The experimental results of AUC and accuracy are shown in Figure 3.11.

Figure 3.10(a) shows the AUC comparison of different algorithm models on different training sets. From the figure, the DKT algorithm performed the worst, with an average AUC of 0.762 in the two training sets, followed by the DKVMN, with an average AUC of 0.769, and then the GKT algorithm, with an average AUC of 0.774. The best performance was the algorithm model proposed in this study. The average AUC of the DL-KT algorithm was 0.781. Figure 3.10(b) shows the comparison of the accuracy of different algorithm models in different training sets. From the figure, the DKVMN performed the worst, with an average accuracy rate of 0.846, followed by the DKT algorithm, with an average accuracy rate of 0.857. The GKT algorithm performed better, with an average accuracy rate of 0.863. The best performer DL-KT was with an average accuracy of 0.886. From the comparison in Figure 8 the DL-KT model proposed in this study performed better than other models in AUC and accuracy. It was verified that the construction mode of the DL-KT model that combined the knowledge display relationship with the potential relationship of knowledge points would make the effect of KT better. The fluctuation of knowledge mastery value (KMV) refers to the fluctuation value between the current knowledge mastery and the previous knowledge mastery after completing the recommended knowledge. KMV is the knowledge fluctuation numerical index. A comparative experiment was carried out on the degree of mastery of knowledge points, and the results are shown in the Figure 3.12.

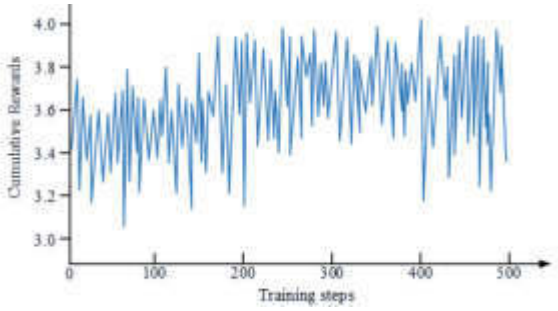
From Figure 3.12, in the Assistment2009-2010 data set, the KMV was the highest when the student's knowledge mastery in the algebra_2006_2007 data set was 0.63. On the whole, the change trend of KMV in the two training sets was the same, and the overall KMV fluctuation range was within 0 to 0.16. The results



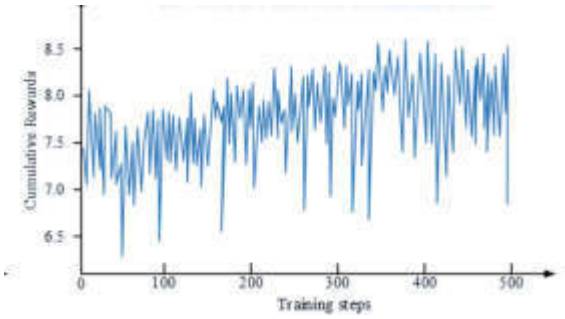
((a)) A total of 3 out of 5 are recommended in turn



((b)) A total of 4 out of 8 are recommended in turn

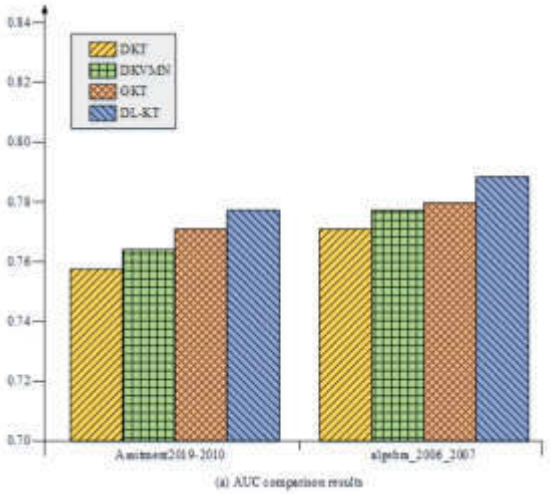


((c)) A total of 5 out of 15 are recommended in turn

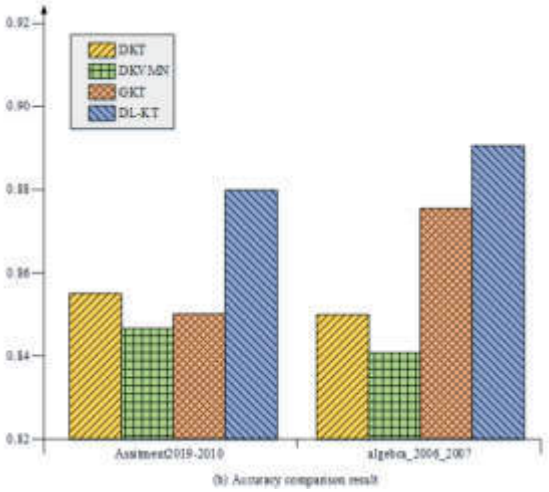


((d)) A total of 10 out of 50 are recommended in turn

Fig. 3.9: Model effect performance in different application environments



((a)) AUC comparison results



((b)) Accuracy comparison result

Fig. 3.11: Comparison of AUC and accuracy of different algorithm models

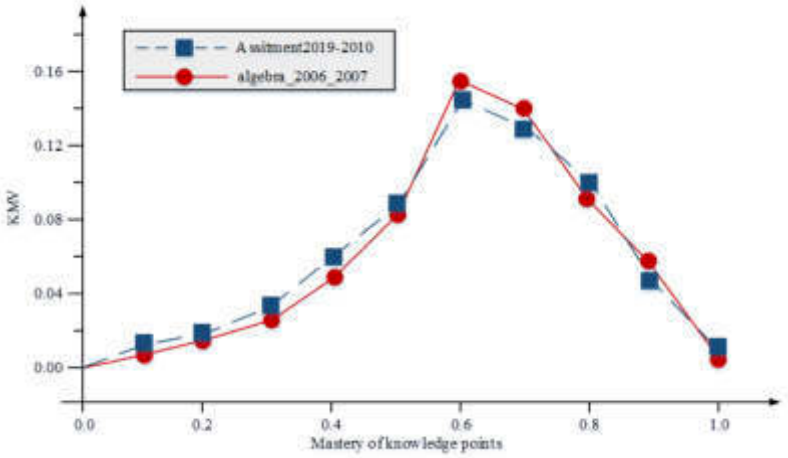
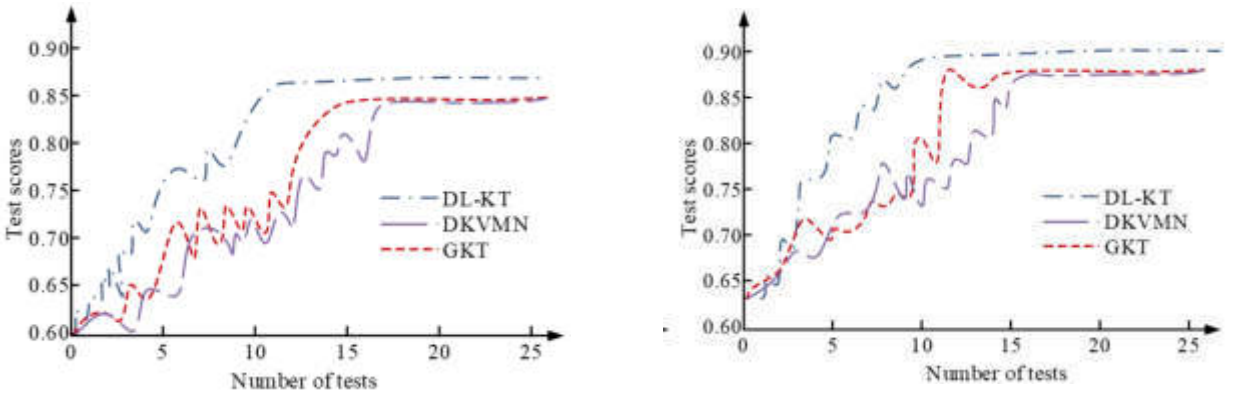


Fig. 3.12: Comparison of knowledge points



((a)) Changes in test scores for senior students

((b)) Changes in test scores for lower grades students

Fig. 3.14: Comparison of application of different models in mental health education

showed that recommending knowledge within the range of 0.6-0.7 to users improved the user's knowledge mastery the fastest, which was obviously better than other algorithm models. It was verified the performance of traditional cultural network online education model proposed in this study that integrated DL and KT. To get the effect of the practical application of DL-KT in online education, the study selected senior and junior students to conduct a comparative experiment, respectively, and there was no significant difference in the performance of each group before the experiment. The results of the effect of different models on students' performance in traditional culture online learning were obtained, as shown in Figure3.14.

Figure 3.13(a) shows the changes in the grades of the students in the higher grades. With the assistance of the DL-KT model, students' grades increased faster and stabilized at about 86 after the 10th quiz. The difference in students' final grades with the assistance of the DINA and HO-DINA models was not significant. Figure 3.13(b) shows the changes in students' grades in the lower grades. With the assistance of DL-KT model, students' grades increased at a faster rate and their final grades stabilized around 89, which was better than the DINA and HO-DINA models. From Figure 3.14, the DL-KT model proposed in the study was more effective in improving the online learning performance of students in lower grades and was more suitable for application

in online education for students in lower grades.

4. Conclusion. The online learning system is a new learning platform that uses the Internet and artificial intelligence technology to enable students to learn the required knowledge more quickly and conveniently. It is difficult to pass the actual level of students in the current online education system and accurately recommend the knowledge question bank. Therefore this research built an online education model that integrated DL and KT. The experimental outcomes denoted that the algorithm model with a discount factor of 0.4 had better performance, which verified that the long-term recommendation effect used in this study was better than the previous heuristic recommendation effect. The average AUC of the model proposed in this study was 0.781, and the average accuracy rate was 0.886. The results showed that recommending knowledge within the range of 0.6-0.7 to users improved the user's knowledge mastery the fastest, which was obviously better than other online education models. However, the shortcoming of this study was that relatively little research has been conducted on the other key subject in online education, the teacher. Teachers play a crucial role in the learning process of students, especially how to track and adapt to the learning progress and needs of different students in real time. In addition, although the model outperformed other models in some metrics, whether it can be generalized across different educational cultures and curricula still requires further research. In future work, the study plans to delve into the role and function of teachers in online learning systems, especially how to enable real-time tracking of student learning by teachers and how to better integrate teachers and technology to provide personalized learning advice to students. As the fields of technology and education converge further, the study hopes to bring about more far-reaching and broader implications for online learning systems.

Fundings. The research is supported by Scientific Research Program Funded by Education Department of Shaanxi Provincial Government (Program No.22JZ016).

REFERENCES

[1] Dooly, M. & Sadler, R. If you don. *T Improve, What's The Point?' Investigating The Impact Of A 'flipped' Online Exchange In Teacher Education*. **32**, 4-24 (2020)

[2] Adiego, J. & Martín-Cruz, N. Training competences in smart cities: an online program for higher education students. *International Journal Of Sustainability In Higher Education*. **22**, 1630-1645 (2021)

[3] Griffith, P., Doherty, C., Smeltzer, S. & Mariani, B. Education initiatives in cognitive debiasing to improve diagnostic accuracy in student providers: A scoping review. *Journal Of The American Association Of Nurse Practitioners*. **33**, 862-871 (2021)

[4] Root, W. & Rehfeldt, R. Towards a Modern-Day Teaching Machine: The Synthesis of Programmed Instruction and Online Education. *The Psychological Record*. **71**, 85-94 (2021)

[5] Chowdhury, T., Hoque, M., Wanke, P., Raihan, M. & Azad, M. Antecedents of Perceived Service Quality of Online Education During a Pandemic: Configuration Analysis Based on Fuzzy-Set Qualitative Comparative Analysis. *Evaluation Review*. **46**, 235-265 (2022)

[6] Shen, L., Shao, Z., Yu, Y. & Chen, X. Hybrid Approach Combining Modified Gravity Model and Deep Learning for Short-Term Forecasting of Metro Transit Passenger Flows: Transportation Research Record. (2021)

[7] Garg, V. & Wickramarathne, T. Deep Learning Approach for Enhancing Situational Awareness in Surveillance Applications with Ubiquitous High-Dimensional Sensing. *IEEE Journal Of Selected Topics In Signal Processing*. **16**, 869-878 (2022)

[8] Tang, W., Yang, Q., Hu, X. & Yan, W. Deep learning-based linear defects detection system for large-scale photovoltaic plants based on an edge-cloud computing infrastructure. *Solar Energy*. **231**, 527-535 (2022)

[9] Bernardini, M., Feldmann, R., Anglés-Alcázar, D., Boylan-Kolchin, M., Bullock, J., Mayer, L. & Stadel, J. From EMBER to FIRE: predicting high resolution baryon fields from dark matter simulations with deep learning. *Monthly Notices Of The Royal Astronomical Society*. **509**, 1323-1341 (2022)

[10] Liu, C., Han, Z., Zhang, Z., Nussinov, R. & Cheng, F. network-based deep learning methodology for stratification of tumor mutations. *Bioinformatics*. **37**, 82-88 (2021)

[11] Song, X., Li, J., Tang, Y., Zhao, T., Chen, Y. & Guan, Z. Joint Graph Convolutional Network based Deep Knowledge Tracing. *Information Sciences*. **580** pp. 510-523 (2021)

[12] Pavlik, P., Eglington, L. & Harrell-Williams, L. Constrained Framework for Learner Modeling. *Institute Of Electrical And Electronics Engineers (IEEE)*. **14**, 624-639 (2021)

[13] Gan, W., Sun, Y. & Sun, Y. Knowledge structure enhanced graph representation learning model for attentive knowledge tracing. *International Journal Of Intelligent Systems*. **37**, 2012-2045 (2022)

[14] Huang, Y. & Cheng, Y. Prediction of Online Judge Practice Passing Rate Based on Knowledge Tracing. 2021. *0*. **38** pp. 003 (0)

[15] Liu, S., Yu, J., Li, Q., Liang, R., Zhang, Y., Shen, X. & Sun, J. Ability boosted knowledge tracing. *Information Sciences: An International Journal*. **596** pp. 567-587 (2022)

[16] Pawan, Y., Prakash, K., B., C., S., H. & C., Y. Particle swarm optimization performance improvement using deep learning techniques. *Multimedia Tools And Applications*. **81**, 27949-27968 (2022)

[17] Lai, Z., Wang, L. & Ling, Q. Recurrent knowledge tracing machine based on the knowledge state of students. *Expert Systems*. **38**, 1-12782 (2021)
[18] Gan, W., Sun, Y., Peng, X. & Sun, Y. Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing. *Applied Intelligence*. **50**, 3894-3912 (2020)
[19] Prakash, A., Patro, K., Hammad, M., Tadeusiewicz, R. & Pławiak, P. secured biometric authentication system using ECG signal based on deep learning techniques. *Biocybernetics And Biomedical Engineering*. **42**, 1081-1093 (2022)
[20] Huo, Y., Wong, D., Ni, L., Chao, L. & Zhang, J. Knowledge modeling via contextualized representations for LSTM-based personalized exercise recommendation. *Information Sciences*. **523** pp. 266-278 (2020)
[21] He, J. & Huang, S. Full-length de novo protein structure determination from cryo-EM maps using deep learning. *Bioinformatics*. **37**, 3480-3490 (2021)
[22] Rakhshani, H., Idoumghar, L., Ghambari, S., Lepagnot, J. & Brevilliers, M. On the performance of deep learning for numerical optimization: An application to protein structure prediction. *Applied Soft Computing*. **110**, 7596-15 (2021)
[23] Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S., Altuwaijri, G., Abdul, W. & Faisal, M. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Computing And Applications*. **35**, 14681-14722 (2023)
[24] Li, Y., Bao, T., Gao, Z., Shu, X., Zhang, K., Xie, L. & Zhang, Z. new dam structural response estimation paradigm powered by deep learning and transfer learning techniques. *Structural Health Monitoring*. **21**, 770-787 (2022)
[25] Rastogi, K., Bodani, P. & Sharma, S. Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto International*. **37**, 1501-1513 (2022)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 17, 2023
Accepted: Sep 11, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 355–368, DOI 10.12694/scpe.v25i1.2300

APPLICATION ANALYSIS OF ENGLISH PERSONALIZED LEARNING BASED ON LARGE-SCALE OPEN NETWORK COURSES

HAINI YANG*

Abstract. In the context of Big data, large-scale open online courses increase learning paths for learners, but in the face of countless high-quality curriculum resources, it is easy for derivative learners to face the dilemma of rich curriculum resources but difficult to choose resources, which leads to information maze for learners. How to help learners quickly and accurately find their own learning resources in the explosive growth of MOOC resources is an urgent problem in the field of education Big data. However, the traditional Collaborative filtering recommendation technology does not perform well when dealing with sparse data and cold start. The recommendation content is repeated and can not effectively deal with high-dimensional and nonlinear data of online learning users, resulting in low efficiency of resource recommendation. Therefore, the study adopts a deep belief network (DBN) to construct a personalized resource recommendation model. The model combines the learner behavior characteristics with the curriculum resource content attribute characteristics to form the learner feature vector. The parameters of the model are adjusted according to the characteristics of learners. Through experiments, the proposed model has shown good performance. The experiment explored the effects of training set size, learner characteristics, and GPU on model performance. The experimental results show that when the training set proportion is 100%, the RMSE, Accuracy, Recall, and F1 values of the model are 0.76, 0.946, 0.957, and 0.951, respectively. When the model is trained using a training set containing learner features, the RMSE, Accuracy, Recall, and F1 values of the model are 0.75, 0.962, 0.908, and 0.958, respectively. After using GPU to accelerate the model, the running time of the model decreased from 360 minutes to 90 minutes. The results indicate that the model cannot effectively mine data information when the degree of correlation between sample information is low. The richer the relationships between samples, the better the performance of the model. Simultaneously learning hunger feature vectors and learner behavior feature vectors for training can significantly improve the recommendation accuracy of the model. The main contribution of this study is to propose a recommendation method based on DBN classification to replace traditional similarity calculation methods, using DBN's efficient feature abstraction and feature extraction capabilities to fully explore learners' interest and preference for course resources. In addition, in view of the common problems of cold start and data sparsity in traditional Collaborative filtering recommendation methods, the research deeply mines the characteristics of learners' Demographics and curriculum resources' content attributes, and constructs a learner interest model based on DBN combined with learners' behavior characteristics, which effectively solves the problems of cold start and data sparsity, as well as the inaccurate expression of learners' interest preferences for curriculum resources.

Key words: Open network courses; Deep belief network; Restricted Boltzmann machine; Back propagation neural network; Personalized recommendation

1. Introduction.

There are many teaching methods in the current society, and the emerging personalized learning method of English based on large-scale open network courses has gradually become the focus of social attention. Due to the differences in students' knowledge structure and individual learning needs, they cannot quickly and accurately find their own curriculum resources. Therefore, it has caused a huge waste of manpower, material resources and time. In this case, the recommendation system can solve this problem well. The Personal Recommendation System (PRS), as the name implies, is to establish a personalized user model based on the characteristics and interests of users. After systematic calculation, the model finally realizes the user's requirements [1, 2]. Unlike search engines, users can obtain personalized suggestions by artificial means without consciously. However, due to the exponential growth of learners and teaching resources, the traditional collaborative filtering algorithm is not efficient in processing sparse data and cold start. In addition, the duplication of recommendation content and the inability to effectively process high-dimensional and non-linear data will lead to inefficient resource recommendation. On the basis of the DBN model, the research conducted an in-depth discussion on the RBM and BP of each component of the DBN [3, 4]. In the process of in-depth modeling, the feature vector based on "learner-course resources" and the relevant course scores

*School of Humanities and Arts, Shaanxi Technical College of Finance and Economics, Xianyang 712000, China (yhn202320230@163.com)

provide a training example for students. On this basis, the feature vector of "learner curriculum resources" is introduced. This method mainly regards learners' evaluation of the course as a supervised marker, and fine-tune the whole network using unsupervised pre-training and supervised feedback, and finally form a complete DBN personalized recommendation model. The classification result is learners' preference for curriculum resources.

The main theoretical contribution of the study is to improve the accuracy of DBN-MCPR scoring prediction by optimizing the initial values of model parameters and parameter setting rules based on DBN-MCPR. Verify the classification accuracy of DBN-MCPR using public datasets and compare it with several traditional recommendation methods to highlight its excellent recommendation accuracy; Then, the DBN-MCPR is used to train the real dataset of the M0OC platform of the Teacher's College, mining recommendation rules between learners and course resources, further completing learners' predictive scoring of courses, and verifying the effects of the training dataset, learner feature vectors, and GPU acceleration on the performance of DBN-MCPR. The practical significance and potential contributions of the study have four aspects. Firstly, enhancing personalized learning experience. By using the DBN-MCPR model, tailored course recommendations can be provided for each learner based on their interests and needs. This will enable learners to more efficiently choose suitable learning resources, improve their learning experience and effectiveness. Secondly, to enhance learner engagement, personalized recommendation models can help improve learner engagement and motivation. By providing relevant and challenging course resources based on learners' interests and learning needs, it can stimulate learners' learning motivation and encourage them to participate more deeply in the learning process. Thirdly, the improvement of automated evaluation systems. The DBN-MCPR model can provide more accurate and personalized suggestions for automated evaluation systems, thereby helping learners better evaluate their learning progress and understanding. This will have a positive impact on the improvement and optimization of automated evaluation systems. Fourthly, educational decision-making support, utilizing the DBN-MCPR model, can analyze learners' learning preferences and behavioral characteristics, and provide valuable data and insights for educational decision-making. This information can be used for the formulation of educational policies, allocation of teaching resources, and other aspects, helping educational institutions make more intelligent and effective decisions. This method solves the Hard problem of consciousness problem of resource selection faced by learners in the context of large-scale open online courses, and improves the shortcomings of traditional Collaborative filtering recommendation technology. The use of deep belief networks to construct feature fusion and recommendation models is a relatively new method in the field of education Big data.

The research mainly consists of four parts. The first part is to summarize the research results of domestic and foreign scholars on personalized recommendation algorithm and DBN model; The second part is to build a personalized recommendation model based on DBN, and introduce the DBN model and the operation of the personalized recommendation model in detail; The third part is to test the proposed model and analyze its performance; The last part is the summary of the full text, the analysis of the deficiencies in the study, and the recommendations for the follow-up study.

2. Related Work. The essence of personalized English learning based on large-scale open online courses is personalized recommendation based on learners' characteristics. Many scholars at home and abroad have relatively mature research on personalized recommendation model. Zhong L et al. proposed a personalized news recommendation algorithm based on the topic model and the finite Boltzmann machine. Based on the LDA2vec topic model, the topic information and audience rating data are used as the condition layer and visual layer to achieve the purpose of news recommendation. Experiments had proved that taking the news theme of LDA2vec as a prerequisite layer can better predict and enhance the effect of news recommendation [5]. Chen Y team constructed a personalized recommendation model based on attention flow network. It proposed a weighted attention flow net, and then recommended products according to the transfer probability of the attention flow net. The algorithm was tested with several sets of actual data, which proved its superiority [6]. A location identification and customized suggestion technique for tourism sites based on image processing was suggested by Zhang Q and other academics. This method was based on hashing and consists of offline and online phases. In addition, a personalized recommendation model based on geographical location and time was also presented in the experiment. It was obviously that the proposed method has high precision and efficiency [7]. Naserian E and other researchers proposed a local personalized recommendation model based on classification. The model was divided into several local models, each of which had several characteristics. Other strategies

Table 2.1: Overview Table

Reference	Research direction	Research results
Reference [5]	Personalized news recommendation algorithm	Enhance news prediction and recommendation effectiveness
Reference [6]	Personalized recommendation algorithm based on attention flow network	The algorithm has superiority in experiments
Reference [7]	Image Processing Based Location Recognition and Personalized Recommendation Method for Tourist Attractions	The proposed method has high accuracy and efficiency
Reference [8]	Classification based local personalized recommendation mode	Tested using data on Foursquare and achieved better recommendation results
Reference [9]	Proposed an intelligent agent model that supports deep learning	This algorithm has good economy and effectiveness in use
Reference [10]	Taylor based T-DBN classifier	The specificity of this model reaches 90.757%, sensitivity is 92.225%, and accuracy is 92.122%
Reference [11]	A New Fuzzy Fusion Method Based on FG-SMOTE	The performance of this method was evaluated using a DBN classifier, and the AUC using Fuzzy SMOTE technology reached 93.7%; The predicted value of F1 score is 94.2%; The geometric average score is 93.6%

were also implemented to support the development of individual local patterns. On this basis, a new method of extracting hidden patterns from geographical clusters was proposed. On Foursquare, the data on Foursquare were used for testing, it was verified that the model performed better effect than the existing algorithm [8]. The DBN model used in the study is also the focus of current social research. Su T et al. proposed an agent model that supports deep learning, which could effectively improve the computational efficiency while ensuring high accuracy. For this reason, the experiment combined the deep belief network (DBN) with the reference frame based non-dominant sequence genetic algorithm (NSGA-III) to establish a new prevention and control system. Through a large number of simulations of IEEE experimental system, the effectiveness and economy of the algorithm were verified[9]. Vijay G and other researchers proposed a Taylor based T-DBN classifier, which could accurately locate the lesions and features in retinal fundus images. The specificity of the model reached 90.757%, the sensitivity was 92.225%, and the accuracy was 92.122% [9]. Hemalatha P and other scholars proposed a new fuzzy fusion method based on FG-SMOTE for processing unbalanced data. The algorithm was composed of a few oversampling based on fuzzy Gaussian synthesis and a deep belief network classifier. DBN classifier was applied to evaluate the function of this method. The results show that the AUC using Fuzzy SMOTE technology reaches 93.7%; The predicted value of F1 achievement is 94.2%; The geometric average score is 93.6% [11].

In summary, many scholars have conducted research and analysis on the DBN model and achieved good classification results. The key directions in each study are shown in Table 2.1. Due to the limited application of the DBN model in personalized recommendation in the field of education, the study is the first to combine the two for exploration, aiming to improve the learning quality of learners in a large-scale open online learning environment.

3. Construction of personalized learning recommendation model based on DBN in MOOC environment.

3.1. Construction of deep belief network model. The research optimizes the personalized English learning of the current large-scale open online courses through the deep belief network (DBN) [12, 13, 14]. DBN is a probability generation model, which consists of multiple hidden layers and a display layer to form a deep hybrid neural network. The DBN used in the study is composed of the unsupervised restricted Boltzmann machine (RBM) and the back propagation neural network (BPNN). Figure 3.1 indicates its structure [5, 16, 17].

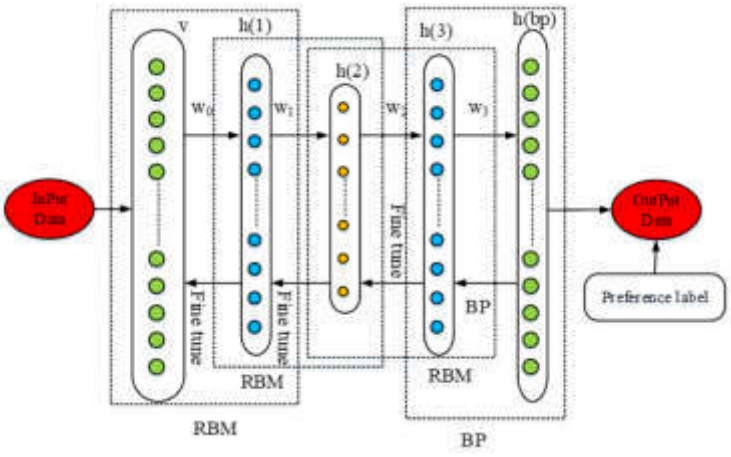


Fig. 3.1: DBN model structure

Figure 3.1 shows the training process of the DBN model, where v represents the state vector of the visible layer; b represents the visible layer bias vector; h represents the hidden layer state vector; c represents the hidden layer bias vector; W represents the connection weight. The DBN training process mainly involves unsupervised pre training and supervised parameter tuning. Among them, unsupervised pre training uses unlabeled training sets to train each layer's RBM layer by layer, independently, and unsupervised from bottom to top. After Input Data, initialize the first layer RBM visible layer unit using the original sample data, and ensure that the features of the input data are well mapped onto the first layer RBM hidden layer unit through abstraction, enabling the model to extract the most essential original features of the data. Then, the hidden layer of the first layer RBM is used as the visible layer input of the second layer RBM. After training to obtain the nth layer, the output of the nth layer is used as the input of the n+1 layer, and the RBM parameter sets of each layer are obtained repeatedly. Supervised parameter adjustment involves fine-tuning network parameters using labeled data from top to bottom, globally, and with supervision. By comparing the expected values of Oup Daa with Preene label, the error is fed back to the network for weight adjustment. BP weights and thresholds are corrected by the quickest descent method, which is more suitable for global optimization of DBN network parameters and can avoid the optimal characteristics of a single RBM network. The learning process of this model is comparable to the startup process of the BP weight, which compensates the defect that the weight matrix and deviation in BP network are random initialization. The model formalizes the traditional BP neural network into an optimized BP neural network, fine-tuning the network using BP back-propagation algorithm to achieve the best classification effect, effectively solving the issue that BP network takes a long time. RBM is the core architecture of DBN, mainly composed of visible layer and hidden layer. Its feature is that neurons are not interconnected, but are interconnected between the visible layer and the hidden layer. In the visual layer, each neuron is used to describe a feature or attribute of the training sample. The hidden layer neurons are used to extract the corresponding features. The core of RBM learning is to correctly describe the characteristics of the observable stratum by adjusting the parameters. The RBM network structure is shown in Figure 3.2.

In Figure 3.2, there are m neurons and hidden neurons in RBM. The energy based model (EBM) is the core of RBM. For a group of known states, the energy of this state can be described by equation 3.1.

$$E(v,h|\theta) = -\sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i w_{ij} h_j \tag{3.1}$$

The matrix vector form of the known state is shown in equation 3.2.

$$E(v,h) = -b^T v - c^T h - h^T W v \tag{3.2}$$

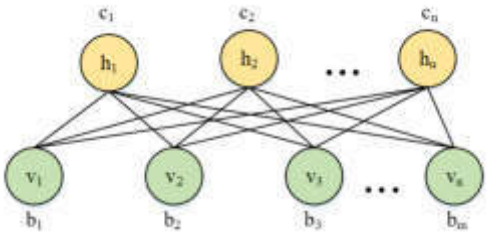


Fig. 3.2: RBM structure diagram

In equation 3.1, (v, h) is the description of the known state; v indicates the visible layer state vector; b indicates the visible layer offset vector; h indicates the hidden layer state vector; c indicates the hidden layer offset vector; W indicates the connection weight. Because RBM is the connection between visible layer and hidden layer, the weight of W is the connection weight between visible layer unit and hidden layer unit; $E(v, h|\theta$ Indicates the energy of the current state of the system; And θ is the parameter set of W, b, c. When the value of θ is determined, the joint probability distribution of RBM at a certain time is shown in equation 3.3.

$$\begin{cases} p(v, h|\theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)} \\ z(\theta)) = \sum_{v,h} e^{-E(v,h)} \end{cases} \tag{3.3}$$

In equation 3.4, p represents the joint probability of a given parameter model; $Z(\theta)$ represents the partition function. If the state vector v of the visible layer is known, the neuron activation possibility of the concealed layer is shown in equation 3.4.

$$p(h_j = 1|v) = \text{sigmoid}\left(c_j + \sum_i v_i w_{ij}\right) \tag{3.4}$$

Similarly, if the hidden layer state vector h is known, the neuron activation possibility of the visible layer is shown in equation 3.5.

$$p(v_j = 1|h) = \text{sigmoid}\left(b_j + \sum_i h_i w_{ij}\right) \tag{3.5}$$

In equation 3.4 and equation 3.5, *sigmoid* represents the activation function, and its value range is from 0 to 1. When analyzing a practical problem, the essence of solving the problem is to solve the probability allocation of the visible layer v. The edge allocation of visible layer v is shown in equation 3.6.

$$p(v) = \frac{1}{Z(\theta)} \sum_h e^{-E(v,h|\theta)} \tag{3.6}$$

In equation 3.6, $p(v)$ represents the edge distribution of the visible layer. Similarly, the edge distribution formula of hidden layer h can be gotten, as shown in equation 3.7.

$$p(h) = \frac{1}{Z(\theta)} \sum_v e^{-E(v,h|\theta)} \tag{3.7}$$

It can be seen from the above formula that the main goal of training the RBM model is to modify θ to make the probability distribution represented by the model as consistent as feasible with the probability distribution of the training sample. Because the model is difficult to calculate $Z(\theta)$, the result of $p(v, h|\theta)$ is difficult to obtain. Traditional solving methods have seriously reduced the training efficiency of the model, and

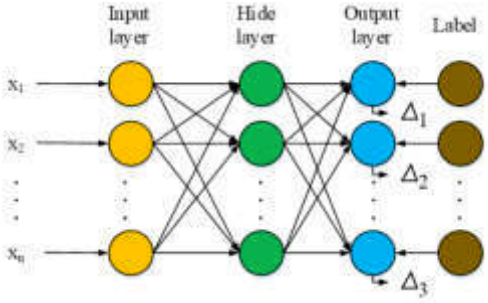


Fig. 3.3: Structure diagram of BPNN

the convergence velocity is also slow, so the learning efficiency of RBM model is low. To handle the above issues, the research applies the Contrast Divergence (CD) algorithm to initialize the neuron states of all visible layers [18, 19, 20]. CD algorithm only needs one Gibbs sampling to get a better approximation, so it is popular in RBM model. As a supervised learning classifier in DBN, BPNN uses a combination of forward propagation of signals and reverse fine tuning of errors to conduct autonomous learning [3, 4, 12]. Figure 3.3 demonstrates the structure of BPNN.

BPNN model consists of forward propagation and back propagation. Forward propagation is the training data from the input to the output layer through the middle layer to get the output Y; At this time, Y is a function of X and W. Back propagation is that when the output layer fails to get the expected value, W is modified by the minimum principle to reduce the model error. As long as appropriate hidden layer neurons are set, BPNN can approximate the nonlinear function of any discontinuous point. Calculate the error gradient between the actual output and the ideal output of the output node as shown in equation 3.8.

$$\delta_k = v_i'(1 - v_i')(v_i - v_i') \tag{3.8}$$

In equation 3.8, v_i' represents the actual output; v_i indicates the desired output. The gradient error of h is shown in equation 3.9.

$$\delta_h = v_h'(1 - v_h')\theta_{hk}\delta_k \tag{3.9}$$

In equation 3.9, θ represents the connection weight value. The update method of calculation weight is shown in equation 3.10.

$$\theta_{ij} = \theta_{ij} + \Delta\theta_{ij} = \theta_{ij} + \epsilon o_i \delta_j \tag{3.10}$$

In equation 3.10, ϵ represents the learning rate; o_i represents node output. The minimum value of output error is calculated as shown in equation 3.11.

$$E = \sum_S \left(\sum_Z (d_{sz} - o_{SZ}) \right)^2 \tag{3.11}$$

4. English personalized learning recommendation based on DBN classification. Study the application of the constructed DBN model to personalized resource recommendation in the education big data environment. In the experiment, a personalized recommendation model on the bias of DBN classification (DBN-MCPR) was established in the MOOC environment. Personalized recommendation are seen as a categorization prediction issue, and DBN-MCPR is the key to personalized recommendation. Figure 4.1 is the structure of the DBN-MCPR.

The data preprocessing module shown in Figure 4.1 refers to the data acquired by the data collection component, which cannot be directly imported into the DBN classification model for data feature extraction.

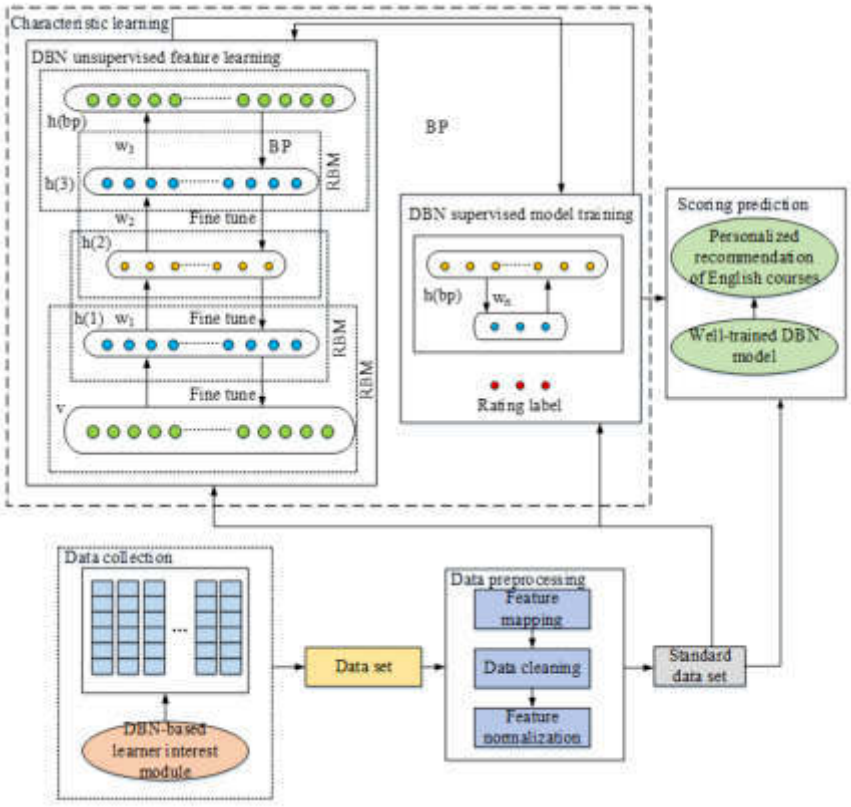


Fig. 4.1: Structure of DBN-MCPR model

Because there are many bad data in the data collected in the early stage, it is necessary to carry out analysis and categorization, characteristic mapping, noise processing, pseudo-data or null value cleaning, characteristic normalization and other operations. Data preprocessing usually affects the learning effect of the model. After the preliminary statistics, analysis and collection of data, the data with large deviation will be eliminated, and then the data will be digitized to eliminate the data that is meaningless for model training. The data will be standardized with equation 4.1.

$$x^* = \frac{x - x_{\min}}{x \ \min_{\max}} \tag{4.1}$$

In equation 4.1, x^* represents the normalized value; x indicates raw data; x_{min} and x_{max} represent the extreme value of a single attribute. The performance of DBN-MCPR model mainly depends on the training of DBN classification model, and its goal is to fit the data extracted from DBN depth feature to the maximum extent. The training of the model consists of two types: supervised and unsupervised.

$$W = W + \epsilon \left(\frac{1}{\text{dataset}} \Delta W \right) \tag{4.2}$$

Figure 4.2 demonstrates the training of DBN-MCPR.

In formula (2.13), ϵ represents the learning rate. Similarly, as indicated in formula (2.14), both the concealed layer offset vector and the apparent layer offset vector are changed.

$$\begin{cases} b = b + \epsilon \left(\frac{1}{\text{dataset}} \triangle b \right) \\ c = c + \epsilon \left(\frac{1}{\text{dataset}} \triangle c \right) \end{cases} \tag{4.3}$$

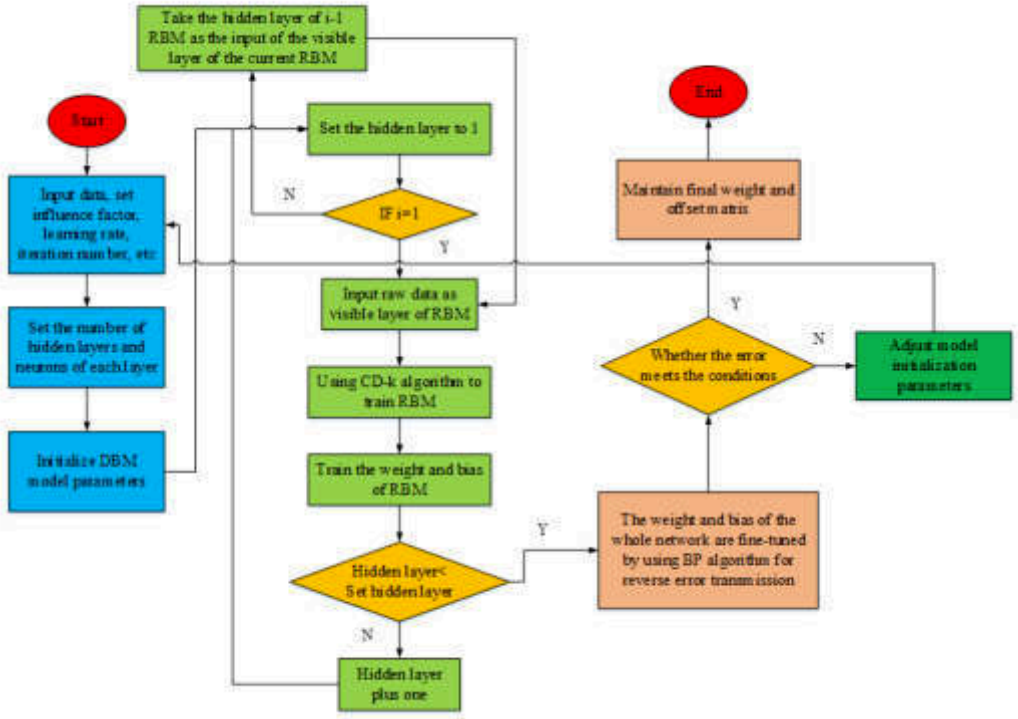


Fig. 4.2: Training the DBN-MCPR

To improve the recommended effect of DBN-MCPR, all parameters must be set effectively before the model runs. These settings usually include initialization and parameter settings. The model implicit weight matrix is initialized using the activation function. The main goal of model weight initialization is to ensure that the initialization neurons can work under the activation function's influence in the initial state, so as to ensure the transmission of information. Because the reverse propagation of gradient is invalid for the offset of hidden layer, the offset of this layer is 0, and the initialization of this value is also 0. In the training of DBN classification model, learning rate is very important. A good learning rate strategy can efficiently shorten the training time of the model and accelerate the convergence of the algorithm. Momentum is introduced into parameter updating, and the direction of parameter updating after two iterations is combined to avoid local extremum problem. In addition, the algorithm not only accelerates the convergence speed, but also greatly enhances the robustness of the model. The formula after adding momentum is shown in equation 4.4.

$$\theta := \varphi\theta + \epsilon\frac{\partial \ln L_s}{\partial\theta} \tag{4.4}$$

In equation 4.4, φ represents the learning rate of momentum. The research first sets φ to 0.5, and then adjusts it according to the reconstruction error during the training process. The number of iterations is the same as the learning rate setting strategy. According to the situation of training data,The quantity of initial pre-training iteration is 100, and the quantity of iterations in the fine-tuning stage is 200. Theoretically, increasing the quantity of hidden layers and hidden layer nodes can effectively improve the efficiency of feature extraction of the model, but there is no clear theoretical basis for specific parameter settings. And it needs to be set according to different models. The smaller quantity of hidden layers and nodes will weaken the ability of feature extraction, resulting in under-fitting; The quantity of hidden layers and nodes in the network is large, which makes the network structure more complex and vulnerable to local minimization, resulting in over-fitting.

Table 5.1: File Information of MovieLens 1M Dataset

Filename	Description	Field
Users.dat	User information files that include demographic characteristics	UserID, gender, age, occupation, zip-code
Movies.dat	A movie information file that includes movie content properties	MovieID, movietitle, genres
Ratings.dat	A file that contains information about the user's rating of the movie	UserID, movied, rating, timestamp

Therefore, according to Kolmogorov's law, the hidden layer is decided to be 3 layers and the neurons meet the $(2n+1)/n$ condition.

5. Effect analysis of deep belief network in personalized English learning under MOOC environment.

5.1. Analysis of personalized recommendation effect of DBN-MCPR model. Through the construction and analysis of the above model, detailed settings for experiments, materials, and data collection will be studied. The experiment was conducted using the MovieLens 1M dataset; Compare the performance of DBN based recommendation method, matrix factorization based Collaborative filtering recommendation method, BP network based recommendation method, and RBM based hybrid recommendation algorithm; Set the top-level feature dimension to 5; Examining the variation of RMSE values of different algorithms with the number of iterations; For the DBN-MCPR model, further parameter analysis is carried out, including the number of iterations, Learning rate, number of hidden layers, hierarchical settings, and training data set size. The materials and data were collected from the MovieLens 1M dataset, including users. dat, movies. dat, and ratings. dat files; Parsing dataset files to extract information about users, movies, and ratings; Collect demographic characteristics data such as gender, age, occupation, etc. based on user information; Collect film content attribute data for film information, such as categories, directors, actors, etc; Associate user rating data for movies with demographic characteristics and movie content attribute data to construct a dataset for model training; Organize and preprocess the dataset, such as removing duplicate data, handling missing values, etc. MovieLens 1 M was chosen as the test data to simulate the test data to the maximum extent. The experiment uses the demographic characteristics of users and the film content attributes for model training. This data set includes 6040 users of MovieLens, who evaluated 3952 films with 5-star rating and provided free text annotation. The data set mainly includes three parts: users.dat, movies data and ratings data. The details of each file are shown in Table 5.1 below.

The recommendation methods based on DBN and the collaborative algorithm of matrix decomposition are compared; The recommendation methods based on BP network and matrix decomposition are compared. Figure 5.1 shows the performance of different recommended algorithms in the MovieLens1M data set.

The top feature dimension of each recommended method is set to 5 in the experiment. Figure 5.1 illustrates how, when the upper dimension is the same, the RMSE values of various suggestion algorithms vary with the number of repetitions. In Figure 5.1, the number of repetitions has no effect on the User-CF and Item-suggestion CF's precision. With a rise in iterations, the RMSE graph greatly changes. However, in this data set, it is estimated that it will take more than 90 iterations to get better results. The more iterations of the recommended algorithm of BP network, the lower the RMSE; The smaller the RMSE is after 12 iterations. The more BP repeats, the lower RMSE. The RMSE value of BP fluctuates around 87.4%. From this point, we can see that the recommendation quality of BP is not high. After more than 60 iterations, the RBM hybrid recommendation algorithm has achieved good recommendation results; After 100 repetitions, the RMSE of the RBM recommended algorithm can reach 0.83. The RMSE value decreases and the recommendation's precision increases with the number of repetitions. From the experimental results, when there are more than 200 rounds, the precision of the recommendations is unaffected by the number of repetitions, and the recommendations outperform the RBM technique.When there are a set amount of repeats, the hybrid recommendation algorithm based on RBM is much better than the DBN classification method. In the hybrid recommendation of RBM, the result of recommendation is much worse than that of DBN classification because there is no supervised

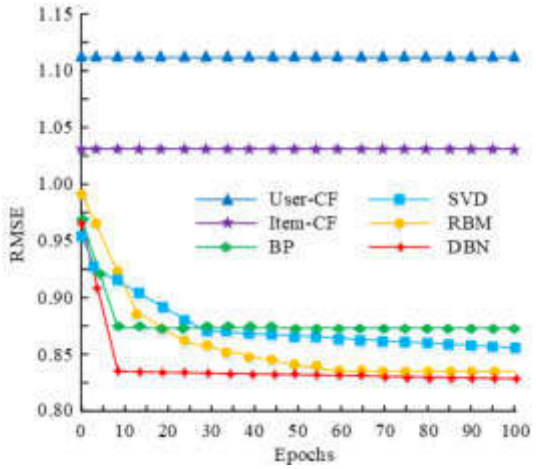


Fig. 5.1: Recommended performance of different algorithms

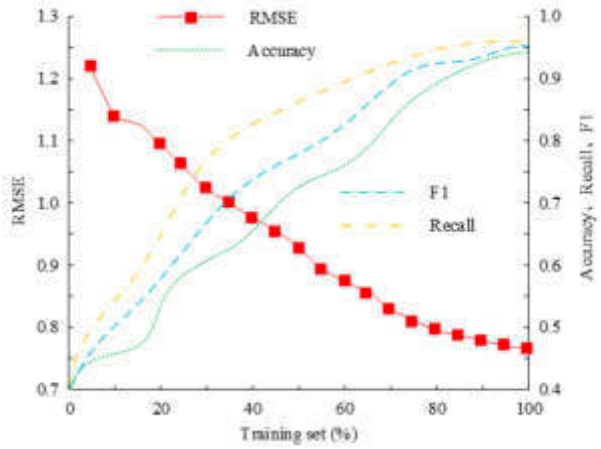
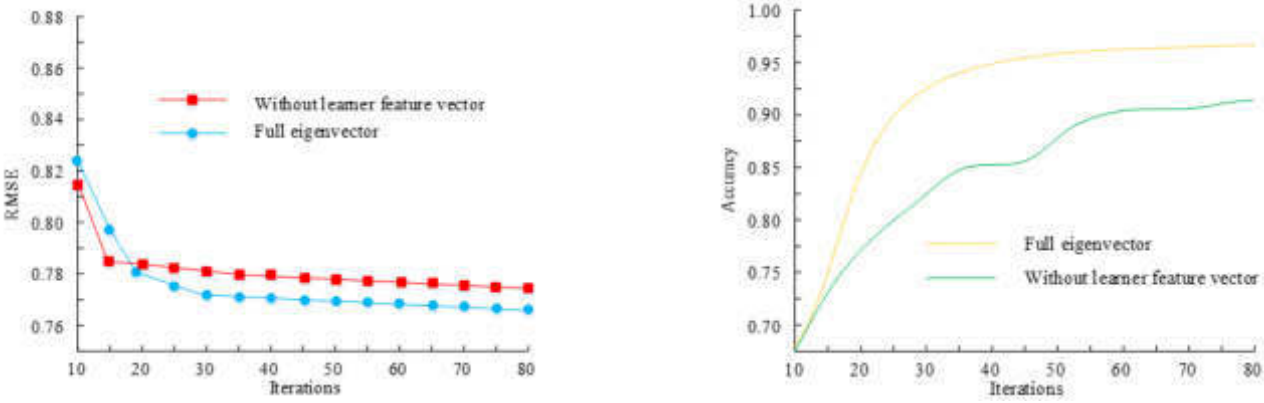


Fig. 5.2: Change of DBN-MCPR model performance with training set size

back-propagation algorithm. Moreover, the recommendation method based on DBN classification is superior to other recommendation methods in terms of convergence speed.

5.2. Performance analysis of parameters on DBN-MCPR model. In the experiment, the correctness of the recommendation algorithm of DBN classification is tested. The results show that the correct rate of the model is different with the adjustment of each parameter. It contains the quantity of training rounds, the quantity of feedback iterations for fine-tuning, the setting of learning rate, the quantity of concealed layers, and the layer setting. The size of the training data set will also have a great effect on the recommendation precision of the model. The data set of MovieLens1M is more about the demographic characteristics of users and the attributes of films. So in the evaluation of the film, there is no more about the user's behavior. This is very disadvantageous to deeply mining the user preference relationship model. For this reason, the research uses the actual student course selection data of the normal university to test the DBN-MCPR to achieve higher recommendation accuracy. The parameters of the DBN model are initialized by the proposed method.The batch size is 10; The pre-training learning rate was 0.01; Fine tuning learning rate is 0.1; The number of pre-training



((a)) Effect of training set size on model RMSE

((b)) Effect of training set size on model Accuracy

Fig. 5.4: Effect of learner feature vector on RMSE and Accuracy of DBN-MCPR model

iterations is 100; The quantity of fine-tuning iterations is 200; The characteristic dimension is 40; The output neuron is 5; The median neurons were 81, 163 and 327, respectively. The experiment will verify the influence of training data set, learner feature vector and GPU accelerator on the performance of the model. The RMSE, Accuracy, Recall and F1 values were selected as the criteria for judging the performance of the model.

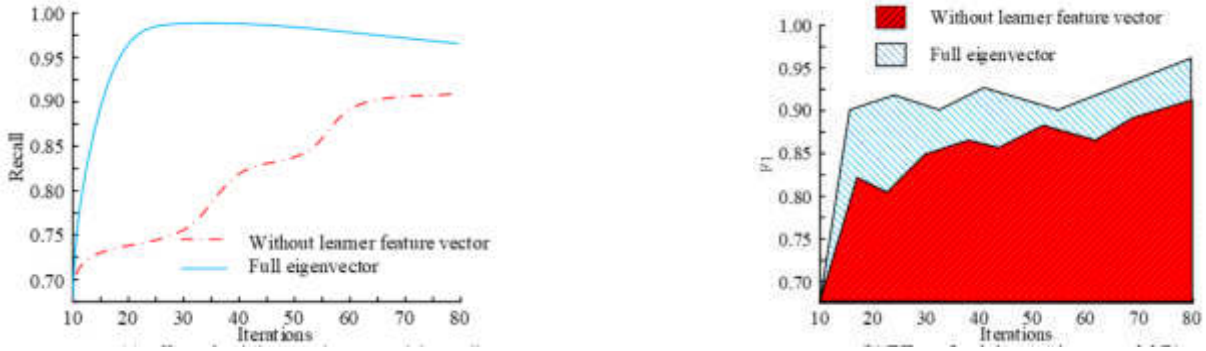
Figure 5.2 shows the impact of training set size on DBN-MCPR model. In Figure 5.2, the RMSE of the model is directly influenced by the percentage of the training sample. The higher the proportion of the training set, the lower the RMSE value. When the training set proportion reaches 100%, the RMSE value of the model is 0.76. From the analysis of the accuracy results, the higher the training set proportion, the higher the precision value. When the training set proportion reaches 100%, the accuracy value of the model is 0.946. From the analysis of Recall results, the higher the proportion of training set, the higher the Recall value of the model. When the training set proportion reaches 100%, the Recall value of the model is 0.957. From the analysis of F1 value results, the higher the proportion of training set, the higher the F1 value. When the training set proportion reaches 100%, the F1 value is 0.951.

Figure 5.4 shows the effect of learner feature vectors on the model. In Figure 5.3(a), when the iterations of the model achieves to 80, the RMSE value is 0.79 when the training data set without the learner feature vector is trained. When the training data set of learner feature vector is used for training, the RMSE value of the model is 0.75. In Figure 5.3(b), when the iterations of the model achieves to 80, the precision value is 0.917 when the training data set without the learner feature vector is trained. When the training data set of learner feature vector is applied for training, the accuracy value of the model is 0.962.

Figure 5.6 shows the effect of learner feature vectors on the Recall and F1 values of the model. In Figure 5.5(a), the Recall value of the model is 0.908 when the training data set without the learner's feature vector is trained. The Recall value of the model is 0.966 when the training data set of the learner feature vector is used for training. In Figure 5.5(b), the F1 value is 0.913 when the training data set without the learner feature vector is trained. The F1 value is 0.958 when the training data set of the learner feature vector is used for training.

Figure 5.7 depicts how the Processor affects how effectively the model runs.. In Figure 5.7, the running time of the model increases with iterations. When iterations reaches 100, the running time is about 360 minutes. The RMSE value is also increasing, and the final RMSE value is 0.92.

Figure 5.8 depicts how the Processor affects how effectively the model runs. In Figure 5.8, the running time increases with the iterations. However, when iterations reaches 100, the running time is only about 90 minutes. The RMSE value is also increasing, and the final RMSE value is 0.75.



((a)) Effect of training set size on model Recall

((b)) Effect of training set size on model F1

Fig. 5.6: Effect of learner's feature vector on Recall and F1 values of DBN-MCPR model

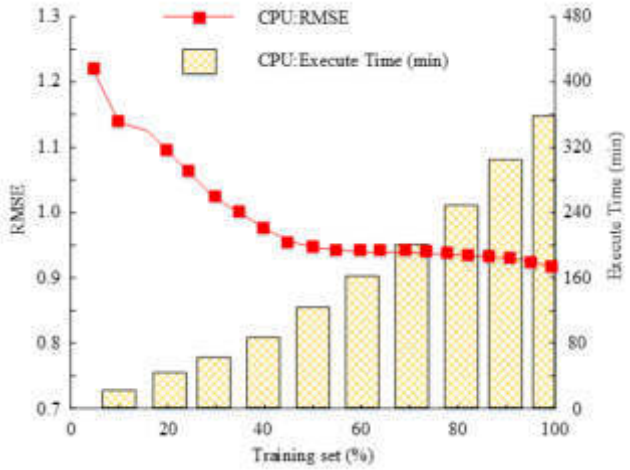


Fig. 5.7: Effect of CPU on the running efficiency of DBN-MCPR model

6. Discussion. Based on the experimental results, we conducted a comprehensive analysis and explanation of the performance of the DBN-MCPR model on the MovieLens 1M dataset, and linked it to research questions and existing literature. First, by comparing the performance of different recommendation methods, it is found that the recommendation method of DBN classification is superior to other recommendation methods in Rate of convergence. Specifically, when the number of repetitions is fixed, the hybrid recommendation algorithm based on RBM performs slightly worse than the recommendation algorithm based on DBN classification. This shows that although the hybrid recommendation algorithm can improve the recommendation accuracy to a certain extent, the DBN based classification method performs better in Rate of convergence and recommendation quality. Secondly, a detailed analysis was conducted on the parameters of the DBN-MCPR model. It is found that the recommendation accuracy of the model is affected by several parameters, including the number of iterations during training, the number of iterations during feedback fine-tuning, the setting of Learning rate, the number of hidden layers and layers, and the size of the training dataset. In the experiment, gradually debug the combination of these parameters and evaluate the performance of the model. The experimental results show that a higher proportion of training sets, the use of learner feature vectors for training, and GPU acceleration can significantly improve the recommendation accuracy and performance of the model. The

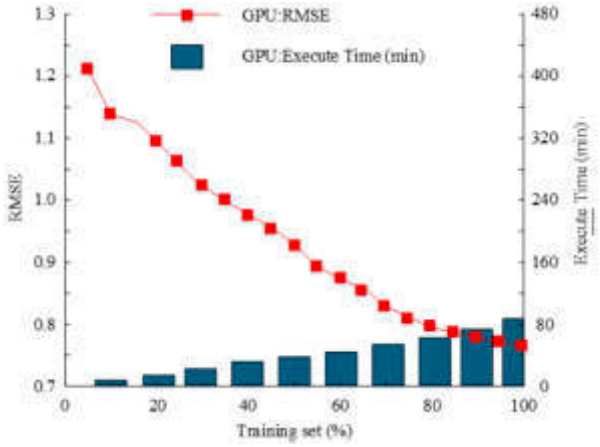


Fig. 5.8: Effect of GPU on the operation efficiency of DBN-MCPR model

experimental results show that the DBN based recommendation method has advantages over other methods in Rate of convergence and recommendation quality. This is consistent with the expected advantages of DBN in the research question.

7. Conclusion. The study explores the application effect of personalized English learning based on large-scale open online courses, and uses DBN to construct a personalized resource recommendation model. This model integrates learner behavior features with course resource content attribute features to form learner feature vectors. The parameters of the model are adjusted based on the learner's characteristics. Through experiments, the DBN-MCPR model showed the best performance when compared with other models, with an RMSE value of 0.83. At the same time, the effects of the size of training set, learner characteristics and GPU on the function of the model are also discussed. The experimental results indicate that the RMSE, Accuracy, Recall and F1 values of the model are 0.76, 0.946, 0.957 and 0.951 respectively when the training set is 100%. When the model uses the training set containing learner characteristics for training, the RMSE, Accuracy, Recall and F1 values of the model are 0.75, 0.962, 0.908 and 0.958, respectively. When GPU is used to accelerate the model, the running time of the model is reduced from 360 min to 90 min. The results show that the model cannot well mine data information when the degree of sample information association is low. The richer the relationship between samples, the better the performance of the model. Meanwhile, the learner characteristic vector and the learner behavior characteristic vector are trained together, which can significantly improve the recommendation accuracy of the model. There are still some deficiencies in the experiment carried out in the study. At present, the score data of learners on the MOOC platform on English curriculum resources are few. The DBN model has some difficulties in mining information, which leads to the improvement of model performance. Therefore, the follow-up research can explore how to effectively use the learner feedback information as the original training data set, so as to train a more efficient personalized recommendation model.

REFERENCES

[1] Xie, J., Zhu, F., Guan, H. & Zheng, L. Personalized query recommendation using semantic factor model. *China Communications*. **18**, 169-182 (2021)

[2] Li, G., Zhuo, J., Li, C., Hua, J., Yuan, T., Niu, Z., Ji, D., Wu, R. & Zhang, H. Multi-modal Visual Adversarial Bayesian Personalized Ranking Model for Recommendation. *Information Sciences*. **572**, 378-403 (2021)

[3] Li, X., Shao, H., Jiang, H. & Xiang, J. Modified Gaussian convolutional deep belief network and infrared thermal imaging for intelligent fault diagnosis of rotor-bearing system under time-varying speeds. *Structural Health Monitoring*. **21**, 339-353 (2022)

[4] Li, H., Wang, H., Xie, Z. & He, M. Fault diagnosis of railway freight car wheelset based on deep belief network and cuckoo search algorithm. *Proceedings Of The Institution Of Mechanical Engineers*. pp. 501-510 (2022)

[5] Zhong, L., Wei, W. & Li, S. Personalized news recommendation based on an improved conditional restricted Boltzmann ma-

chine. *The Electronic Library: The International Journal For Minicomputer, Microcomputer, And Software Applications In Libraries*. **39**, 553-571 (2021)
[6] Chen, Y., Dai, Y., Han, X., Ge, Y. & Li, P. Dig users' intentions via attention flow network for personalized recommendation. *Information Sciences*. **547**, 1122-1135 (2021)
[7] Zhang, Q., Liu, Y., Liu, L., Lu, S., Feng, Y., Identification, Y. & Recommendation, P. of Tourist Attractions Based on Image Processing. *Traitement Du Signal: Signal Image Parole*. **38**, 197-205 (2021)
[8] Naserian, E., Wang, X., Dahal, K., Alcaraz-Calero, J. & Gao, H. Partition-Based Partial Personalized Model for Points of Interest Recommendations. *IEEE Transactions On Computational Social Systems*. **8**, 1223-1237 (2021)
[9] Su, T., Liu, Y., Zhao, J. & Liu, J. Deep Belief Network Enabled Surrogate Modeling for Fast Preventive Control of Power System Transient Stability. *IEEE Transactions On Industrial Informatics*. **18**, 315-326 (2022)
[10] Vijay, G. & Athalye, S. Taylor series-based deep belief network for automatic classification of diabetic retinopathy using retinal fundus images. *International Journal Of Imaging Systems And Technology*. **32**, 882-901 (2022)
[11] Hemalatha, P. & Fg-smote, A. Fuzzy-based Gaussian synthetic minority oversampling with deep belief networks classifier for skewed class distribution. *International Journal Of Intelligent Computing And Cybernetics*. **14**, 270-286 (2021)
[12] Shirke, S. & Udayakumar, R. Hybrid optimisation dependent deep belief network for lane detection. *Journal Of Experimental And Theoretical Artificial Intelligence*. **34**, 175-187 (2022)
[13] Moholkar, K. & Patil, S. Lioness Adapted GWO-Based Deep Belief Network Enabled with Multiple Features for a Novel Question Answering System. *International Journal Of Uncertainty, Fuzziness And Knowledge-based Systems: IJUFKS*. **30**, 93-114 (2022)
[14] Et., A. Breast Cancer Detection Using Deep Belief Network by Applying Feature Extraction on Various Classifiers. *Turkish Journal Of Computer And Mathematics Education (TURCOMAT)*. **12**, 471-487 (2021)
[15] Zhong, L., Wei, W. & Li, S. Personalized news recommendation based on an improved conditional restricted Boltzmann machine. *The Electronic Library: The International Journal For Minicomputer, Microcomputer, And Software Applications In Libraries*. **39**, 553-571 (2021)
[16] Kondratyev, A. Non-Differentiable Leaning of Quantum Circuit Born Machine with Genetic Algorithm. *Wilmott*. **2021**, 50-61 (2021)
[17] Biswal, A., Borah, M. & Hussain, Z. Music recommender system using restricted Boltzmann machine with implicit feedback. *Advances In Computers*. **122**, 367-402 (2021)
[18] Kurup, A., Ajith, M. & Ramón, M. Semi-supervised facial expression recognition using reduced spatial features and Deep Belief Networks. *Neurocomputing*. **3672** pp. 188-197 (2019)
[19] Chu, J., Wang, H., Liu, J., Gong, Z. & Li, T. Unsupervised Feature Learning Architecture with Multi-clustering Integration RBM. *IEEE Transactions On Knowledge And Data Engineering*. **34**, 3002-3015 (2020)
[20] Liang, H., Liu, Y., Sheng, G. & Jiang, X. Fault-Cause Identification Method Based on Adaptive Deep Belief Network and Time-Frequency Characteristics of Traveling Wave. *IET Generation Transmission & Distribution*. **13**, 724-732 (2019)
[21] Li, X., Shao, H., Jiang, H. & Xiang, J. Modified Gaussian convolutional deep belief network and infrared thermal imaging for intelligent fault diagnosis of rotor-bearing system under time-varying speeds. *Structural Health Monitoring*. **21**, 339-353 (2022)
[22] Li, H., Wang, H., Xie, Z. & He, M. Fault diagnosis of railway freight car wheelset based on deep belief network and cuckoo search algorithm. *Proceedings Of The Institution Of Mechanical Engineers*. pp. 501-510 (2022)
[23] Shirke, S. & Udayakumar, R. Hybrid optimisation dependent deep belief network for lane detection. *Journal Of Experimental And Theoretical Artificial Intelligence*. **34**, 175-187 (2022)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 17, 2023
Accepted: Nov 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 369–381, DOI 10.12694/scpe.v25i1.2303

CONSTRUCTION AND APPLICATION OF MOOC+FLIPPED CLASSROOM MIXED TEACHING MODEL IN VOLLEYBALL TEACHING IN COLLEGES AND UNIVERSITIES

JUBIN ZHANG* AND JIE YU†

Abstract. "MOOC+Flipped Classroom" (MFC) is a new model that make the online learning and offline learning integrated. And it also shows a new way in developing the teaching reform. Based on MFC, a mixed teaching mode suitable for college volleyball was studied and established. And a reasonable evaluation index system was developed. It fosters educational fairness and improvements in volleyball learning performance. To be applicable to many kinds of problems, the support vector machine (SVM) is improved to obtain the DBT-SVM based quality evaluation model of college volleyball mixed teaching. The average score of students in the mixed teaching mode was higher than that in the traditional teaching mode. The comprehensive performance scores of the former and the latter after the experiment are 9.12 ± 0.75 and 7.63 ± 0.56 respectively. The mixed teaching mode results in higher student attendance. Among the prediction results of DBT-SVM, BT-SVM and SVM, DBT-SVM has higher accuracy, less training and testing time, and smaller error. This shows that DBT-SVM can accurately evaluate the teaching effect and quality.

Key words: MOOC; Flipped classroom; Mixed teaching; College volleyball teaching; Quality evaluation; DBT-SVM

1. Introduction. As the network develops rapidly, resource sharing concept in college education has been deepened and the awareness of informatization has been strengthened. This has facilitated the growth of Massive Open Online Course (MOOC) and flipped classroom [1]. MOOC's emerging has changed teaching methods and learning methods. This not only promotes educational fairness, but also improve the teaching method. Flipped classroom is a teaching method that emphasizes "people-oriented". It overlaps with the teaching methods of blended learning to some extent, aiming at making students learn more actively and flexibly. In the teaching process, flipped classroom has no preconditions. However, in the case of uneven students' foundation, the implementation of flipped classroom must have some basic prerequisites and use the appropriate teaching platform [2]. The combination of MFC can promote the smooth completion of teaching objectives and improve students' knowledge level. To a certain extent, it can improve students' motivation to learn and the classroom teaching quality (TQ) [3]. Volleyball, as one of the ball games, can improve people's physical coordination and is conducive to physical health. But at present, the problems of single teaching mode and low interest of students still exists in the volleyball teaching in higher education. This does not bode well for college volleyball education in the long run. It should make effective use of the network tools and platforms in the information age to build a new teaching model and enhance students' enthusiasm for volleyball courses. At the same time, the construction and implementation of the classroom TQ evaluation system will highly facilitate teaching theory development and ensure teachers' classroom TQ evaluation. After SVM is introduced into TQ evaluation, all school members will be evaluated. Not only can human errors be avoided, but the teacher's teaching process can also be fully demonstrated. In view of this, the research will build a mixed mode for volleyball teaching based on MFC. At the same time, reasonable indicators are selected to establish a TQ evaluation model to verify whether the mixed teaching mode is effective.

2. Related Works. In response to the impact of informatization on teaching models, many scholars studied online teaching models, including MOOC and flipped classroom, etc. The evaluation model based on BP neural network and SVM has been established for TQ, and many research results have been achieved.

*Department of Physical Education, North China Institute of Aerospace Engineering, Langfang 065000, China
†Department of Physical Education, North China Institute of Aerospace Engineering, Langfang 065000, China (yujie221108@163.com)

For more effective educational data mining, Qian Y and other scholars used MOOCs based flipped classroom and back-propagation neural network to predict students' grades and analyze the impact of teaching. Under the new model, students' performance had improved [4]. For more autonomous learning ability of college students' English, Chang H proposed a teaching model for flipped college English classes with big data and deep neural networks to investigate changes in student performance. This model provided a reference for cultivating college students' autonomous learning ability in English [5]. To overcome the low efficiency of college English evaluation, Zhang Y et al. proposed a method for evaluating inputs of distance education and determined a reasonable input evaluation index. The effectiveness of the keyboard rack results was more than 90%, and the students' English scores had improved [6]. In view of the vigorous development of online and offline teaching models, Wu X proposed a mixed teaching evaluation method of ideological and political education in college English teaching, and put forward suggestions on reforming the evaluation system for the two parts [7]. To improve BP neural network's low efficiency for evaluating the TQ, Jiang L et al. proposed a model based on AHP and particle swarm optimization BP neural network to evaluate TQ. The model parameters were adjusted and the results were verified by ANOVA. This method can effectively overcome the shortcomings of BP neural network and improve the evaluation accuracy [8].

To improve the evaluation efficiency of classroom TQ, Yuan T proposed a modified model to evaluate mixed TQ with Markov chain and designed a comparative experiment to select reasonable evaluation indicators. This model could improve the efficiency of evaluating the classroom TQ [9]. Addressing the inefficiency of traditional English TQ evaluation, Huang W proposed an improved Gaussian algorithm evaluation model, and combined with machine learning technology, improved the correlation vector machine model. The model had good quality evaluation effect and can be applied to English intelligent teaching [10]. To make reasonable efficiency use of the results to evaluate TQ, Yu H improved the Apriori Tid algorithm and proposed a model for evaluating online TQ on the basis of teaching needs. TQ evaluation using data mining proved that the model had good performance [11]. For addressing that online education evaluation model is inefficient, Hou J proposed a deep neural network-based online education quality evaluation model. The current BP neural network was improved by adaptive learning rate. This model could process large-scale data sets and improve the efficiency of the evaluation for TQ [12].

The above is the research of different researchers on the mixed teaching model and quality evaluation model. The mixed teaching model based on MFC can enhance the learning performance of students with certain effectiveness in evaluating. Therefore, this study will build a college volleyball mixed teaching mode based on MFC, and evaluate its TQ to promote college volleyball teaching.

3. Construction of mixed teaching mode and quality evaluation model of college volleyball based on MFC.

3.1. Mixed teaching mode of college volleyball based on MFC. MOOC is an online course with a large number of participants and no access conditions. In short, MOOC is to transform "teaching" into a new and open learning mode. Teachers use the Internet for teaching, students learn online, and conduct a higher level of knowledge exchange during this period. The "online+offline" teaching mode can be realized by MOOC, which is mixed with teachers' teaching, students' independent learning, quality curriculum of foreign and domestic schools [13]. During the teaching process, students can learn at their own pace based on their preferences and use various technologies to change the traditional way of teaching and learning in the classroom. Figure 3.1 shows MOOC teaching model.

The MOOC teaching model in Figure 3.1 mainly includes student activities and teacher activities. Learning and assessment are carried out through platforms and announcements, and teachers mainly play the role of supervision and assistance. In this process, we will achieve a higher level of knowledge exchange. Before the teaching activities are carried out, the teacher will issue a notice in the class group to guide students to complete the registration on the MOOC platform, and publish the selected textbooks, grading standards, and other requirements in the form of a notice to remind students to read. Then, the teacher fulfills the leading responsibility of supervising students' autonomous learning, while students control their dominant position and solve problems encountered during the learning process through independent research or communication and discussion. The teacher only needs to assist from the sidelines. Finally, the assessment is organized by teachers, with students actively participating in the assessment and actively verifying learning outcomes.

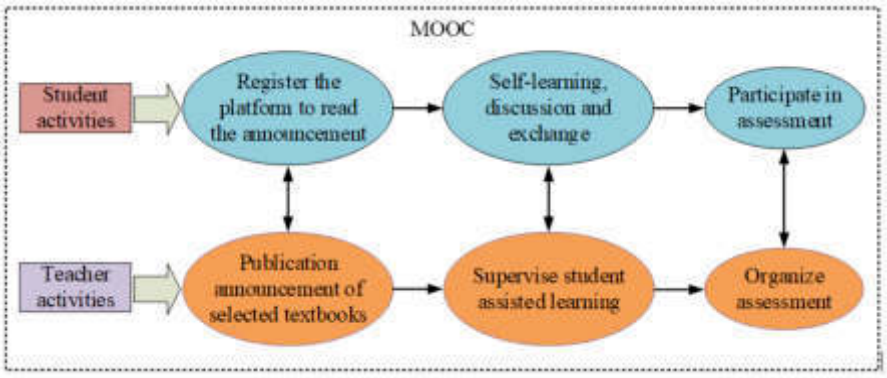


Fig. 3.1: MOOC teaching model

Flipped classroom (FCM), also known as "Flipped Class", refers to readjusting the time inside and outside the classroom, transferring the decision-making power of learning from teachers to students. It is a teaching method of reverse innovation through classroom discussion, practice, internalization and strengthening knowledge. In this teaching mode, the valuable time in the classroom allows students to focus more on proactive project-based learning, work together to research and solve problems, and thus gain a deeper level of understanding. Also reallocate time both inside and outside the classroom, transferring the learning power of teachers to students. Flipped classroom means that the physical classroom of "teachers' teaching and students' listening" has become a way to stimulate students' active and active learning. In this teaching mode, students can use their valuable spare time to engage in theme based learning, explore problem-solving together, and deepen their understanding of the problem. Teachers no longer occupy classroom time to impart information, which requires students to complete self-learning before class. They can watch video lectures, listen to podcasts, read enhanced e-books, and discuss with other classmates online. They can access the necessary materials at any time. Teachers can also have more time to communicate with everyone. After class, students independently plan their learning content, pace, style, and presentation of knowledge, while teachers use teaching and collaborative methods to meet students' needs and facilitate personalized learning. The goal is to enable students to achieve more authentic learning through practice. The Flipped classroom model is a part of the big education movement. It overlaps with blended learning, inquiry learning, and other teaching methods and tools in meaning, all of which are designed to make learning more flexible, active, and enable students to participate more. In the The Internet Age, students learn rich online courses through the Internet, and do not have to go to schools to receive teachers' lectures. The Internet, especially the mobile Internet, gave birth to the "Flipped classroom" teaching model. "Flipped classroom" is a complete subversion of the traditional classroom teaching structure and teaching process based on printing, which will lead to a series of changes in teacher roles, curriculum models, management models, etc.

Teaching in a mixed way is proposed before. Narrowly speaking, mixed teaching is a teaching context in which there is both online and offline teaching. Broadly speaking, mixed teaching is a combination of traditional teaching and digital teaching under the condition of information technology. The broad mixed teaching is applied in this study. Hybrid teaching is to organically combine the advantages of network and physical classroom teaching based on teacher-led and student-centered. Improvement for the teaching effect can be achieved by integrating seven links including teaching objectives, teaching environment, resources, content, time, assessment standards and teaching evaluation. Figure 3.2 shows the mixed teaching mode of "MFC".

The mixed teaching mode is mainly composed of classroom online activities and classroom offline activities, the two parts are conducted separately in the classroom and outside the classroom. In implementing "MFC" mixed teaching mode construction, teaching design and activity implementation, the teacher-led and student-centered principle should be strictly followed. At the same time, based on the needs of college sports volleyball students and teachers, it can grasp the support of its mixed teaching activities. Then, a rigorous and detailed

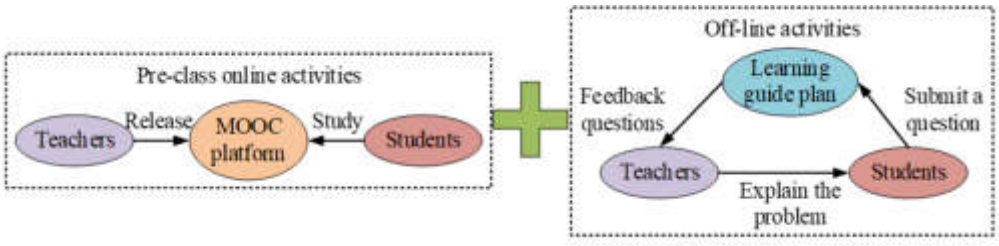


Fig. 3.2: Mixed teaching mode of "MFC"

teaching design is carried out in accordance with the characteristics of college mixed volleyball teaching. The "MFC"-based mixed teaching activities are adjusted in a timely and flexible manner to meet the actual teaching situation of the school for continuous efficiency improvement. Among them, in online activities, learning guidance plans assign tasks, manage platforms, and collect information into the course. Students also watch videos, provide feedback, and discuss online during the course. Teachers are mainly responsible for supervising and guiding. In classroom activities, the learning guidance plan involves teaching design practice for students. Students and teachers respectively provide feedback on teaching issues and guide teaching activities into the learning guidance plan. Teachers are also responsible for teaching and consolidating knowledge for students.

At the same time, in the theory and practice of the "MFC" mixed teaching mode of college volleyball, we must always adhere to the guidance of constructivism learning and relevance. Constructivist learning theory believes that learning is the process in which learners actively construct internal psychological representations in the process of interacting with the environment. Related scholars believe that the practical significance of constructivism refers to the connections between things, their properties, and laws. During the learning process, it helps learners understand the properties, laws, and connections between current things and other things. Due to the fact that current learners are in a certain background environment, with the help of other help, such as teachers, peers, etc., by establishing interaction with each other The process of actively constructing knowledge meaning through collaborative activities. Therefore, in the "mixed" teaching mode of "MFC", we will organically combine relevant teaching theories with practice, realize the "mixed" teaching mode of "MFC", and maximize the organizational activities of college volleyball. And the constructed model basically conforms to the principles of feasibility, systematicness and subjectivity. In the development of college volleyball mixed teaching based on "MFC", the process evaluation and the final evaluation are combined. This overcomes the disadvantages of the single evaluation information obtained only by the final score in the past, and makes its conclusions more authentic and credible.

3.2. Improved binary tree SVM multi-class classification algorithm. The evaluation of classroom TQ is important for college classroom teaching activities. Traditional evaluation methods include analytic hierarchy process and BP neural network evaluation. In a sense, both evaluation methods have yielded some results. However, some problems stay remained. The former accounts for a large proportion of the subjective factors, which leads to the objective and reliability of the final evaluation result is not ideal. There is a huge amount of data in training BP neural network model. There are problems such as learning and memory instability in the network, which leads to the failure of the model [14]. SVM can effectively solve the shortcomings of the above two methods, and analyze them on this basis to obtain the relevant dependencies, so as to correctly evaluate them.

SVM is a supervised learning method by learning statistically. Its main feature is that the network's generalization rate can be effectively increased with minimal structural risk. This method has obvious advantages in solving nonlinear and small sample recognition problems, and has obtained good application prospects in pattern recognition and data mining.

SVM should first be solved from the linearly separable classification surface, assuming the linearly separable training sample set $S = [(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)] \in (X \times Y)^l$. Among them, $i = 1, 2, ..., l$, $x_i \in R^n$, $y_i \, in 1, -1$.

SVM is to construct a classifier through the information provided by training samples, as shown in equation 3.1.

$$f : R^n \to \pm 1 \tag{3.1}$$

Among them, the optimal classification surface can be converted into an optimization problem, and equation 3.2 is the expression.

$$\begin{cases} \min\limits_{w,b} & \frac{1}{2}\|w\|^2 \\ \text{s.t} & y_i\left((w \cdot x) + b\right) - 1 \geq 0, \quad \text{for } i = 1, 2, ..., l \end{cases} \tag{3.2}$$

(w, b) is a parameter pair in equation 3.2. In the process of solving, the optimal solutions w and b of parameter pair (w, b) can be obtained. In equation 3.3, the decision function expression is constructed.

$$f(x) = \text{sign}(w \cdot x + b) \tag{3.3}$$

The Lagrange function can be used to solve it, and equation 3.4 is the expression of Lagrange function.

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i [y_i(w \cdot x_i + b) - 1] \tag{3.4}$$

In equation 3.4, $\alpha = (\alpha_i \ldots, \alpha_l)$. The solution of the quadratic optimization problem can be obtained through the Lagrange function, and the decision function can also be deformed in equation 3.5.

$$f(x) = \text{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i'(x \cdot x_i) + b'\right) \tag{3.5}$$

But because many problems in practice are nonlinear problems, it is necessary to introduce relaxation variables to soften the constraints, or use the idea of kernel function. In the case of nonlinearity, equation 3.6 shows the expression of final optimization problem.

$$\textbf{Maximize } F(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \left[\phi(x_i) \cdot \phi(x_j)\right] \tag{3.6}$$

In equation 3.6, $\phi(x)$ is used to describe the training samples mapped to the samples in the high-dimensional space. Assume $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, then equation 3.6 can be transformed to equation 3.7.

$$\textbf{Maximize } F(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3.7}$$

$K(x_i, x_j)$ is the kernel function in equation 3.7. At present, there are mainly four kinds of kernel functions in common use. Equation 3.8 is the first linear kernel function.

$$K(x_i, x_j) = x_i^T x_j \tag{3.8}$$

The second kernel function is Gaussian kernel function, also called radial basis RBF kernel function. The expression is shown in equation 3.9.

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \quad \gamma > 0 \tag{3.9}$$

The third kernel function is polynomial kernel function, and the expression is shown in equation 3.10.

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0 \tag{3.10}$$

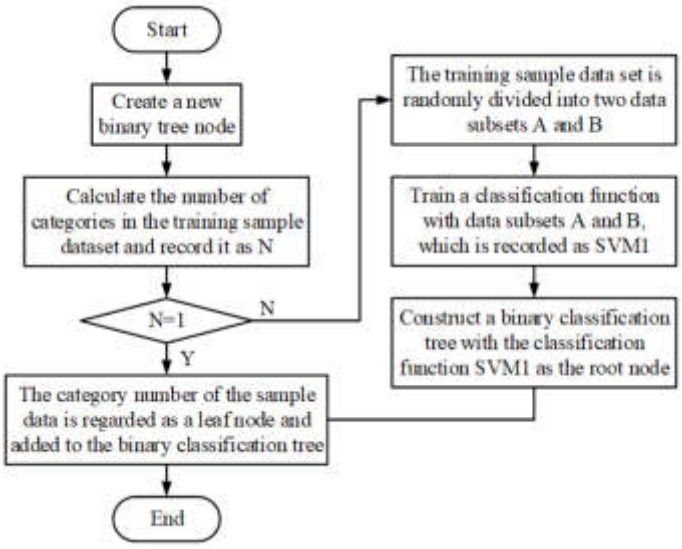


Fig. 3.3: Steps in BT-SVM algorithm training stage

The fourth kernel function is a two-layer neural network kernel function, which is also called sigmoid kernel function. The expression is shown in equation 3.11.

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \tag{3.11}$$

For the nonlinear separable problem, SVM can use RBF to avoid the complex calculation process, and most armies use RBF kernel function to solve [15]. At the same time, the main goal of SVM algorithm is to solve the second class classification problem. But in real life, most problems are multi-class problems. For multi-class problems, it is to decompose the multi-class problems into two classes, and then use the two-class classification function to solve them. Among them, Binary Tree SVM (BT-SVM) is a very popular multi-class classification algorithm. BT-SVM algorithm generally consists of training phase and test phase, and the steps of training phase are shown in Figure 3.3.

In the training phase, if the output category number is consistent with the category number when the data to be tested enters the node, the training is successfully completed in Figure 3.3. After that, it can test the data. In Figure 3.4, the steps in the BT-SVM algorithm test phase are shown.

The test phase is mainly to input the data to be tested into the root node in the training phase in Figure 3.4. Similarly, when the category label of the leaf node is the same as the category of the input test data, it also means that the test phase is successfully completed. However, BT-SVM will spend more time and economic costs in the process of training and testing, so it needs to be improved.

The Euclidean-based distance binary tree SVM (DBT-SVM) is improved. Assume that X is a sample set including k categories, and X_i is the training sample set of class i. In equation 3.12, the Euclidean distance of the nearest sample between class i and j is defined.

$$d_{i,j} = \min\{\|x_i - x_j\|\} \tag{3.12}$$

In equation 3.12, $i = 1, 2, \ldots, k$ means $d - i, j = 0$ and $d_{i,j} = d_{j,i}$. At the same time, the sample center expression of class i is shown as follows.

$$c_i = \frac{1}{n} \sum_{x \in x_i} x \tag{3.13}$$

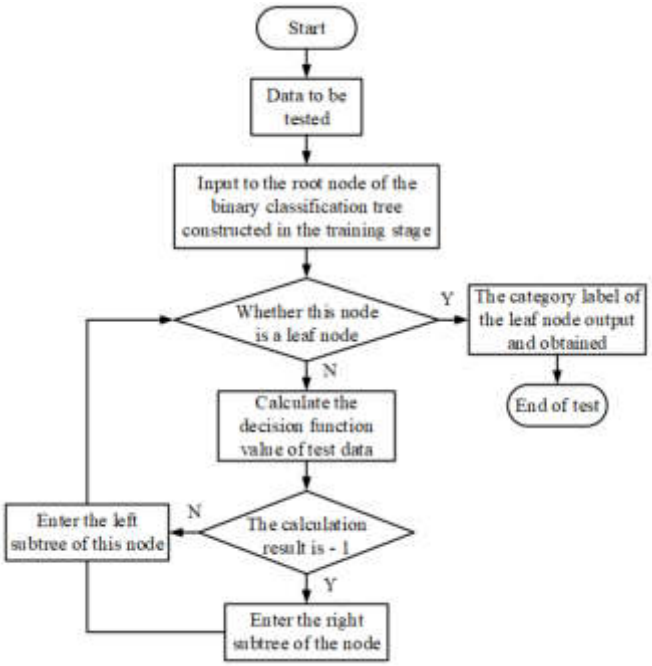


Fig. 3.4: Steps in the BT-SVM algorithm test phase

n represents samples number of class i in equation 3.13. If C_i is the sample centers of Class i, c_j is the sample centers of Class j, equation 3.14 shows the Euclidean distance between the centers of Class i and Class j.

$$d'_{i,j} = \|c_i - c_j\| \tag{3.14}$$

Like BT-SVM, DBT-SVM also includes training stage and test stage. First, all the existing category numbers are sorted from small to large in the experiment, and placed in set C. If a classification problem has k categories, equation 3.15 is the matrix D expression of Euclidean distance between category i and category j .

$$D = \begin{bmatrix} d_{1,2} & d'_{1,2} \\ d_{1,3} & d'_{1,3} \\ \vdots & \vdots \\ d_{k,(k-1)} & d'_{k,(k-1)} \end{bmatrix} \tag{3.15}$$

Then, in set C from matrix D, it should find out class i and j with the largest distance between samples. According to the order from small to large, it should put them in C_1 and C_2. An SVM classifier is also created at the corresponding node of the binomial tree. When each subtree has only one category, it can end the training. Similarly, when the node contains only one category, the test can also be ended.

3.3. Model for evaluating the DBT-SVM based college volleyball mixed TQ . Based on the principles of guidance, objectivity, fairness, conciseness and efficiency, this paper selects the evaluation indicators of the college volleyball mixed TQ. And combining the actual situation of teaching, it should formulate a reasonable index evaluation system to improve the accuracy and effectiveness of teaching evaluation. The evaluation index system adopts tower structure, and the appropriate first-level index is selected first. Then it is decomposed into second or third level indicators. Each indicator is given a score for evaluation and calculation [15]. In Table 3.1, the index system established by the study for evaluating volleyball mixed TQ is shown. The generation strategy of nearly complete binary tree and the related definition of class distance in clustering are

Table 3.1: Mixed TQ Evaluation Index System

First-level evaluation index	Index No	Secondary evaluation index
Learning attitude	X_1	Take classes on time, don't be late and leave early
	X_2	Actively participate in training
	X_3	Teamwork ability
Technological achievements	X_4	Serve
	X_5	Catch the ball
	X_6	Spike
	X_7	Learning duration
Network resource learning	X_8	Homework after class
	X_9	Online discussion

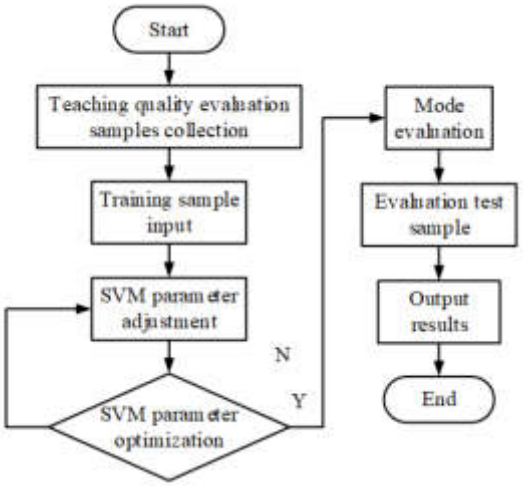


Fig. 3.5: Flow chart of university DBT-SVM based volleyball model for evaluating mixed TQ

used. Then, multiple binary tree SVM are combined to build a model for evaluating the DBT-SVM based college volleyball mixed TQ. Figure 3.5 shows the model flow chart.

The TQ evaluation samples need to be collected first, the samples preprocessed, the training samples input, and the SVM parameters adjusted in Figure 5. When the SVM parameters are in the optimal state, the mode can be evaluated, and the test samples can be evaluated to obtain the output results. On the contrary, if the SVM parameters do not reach the optimal state, it needs to continue to be adjusted until the parameters reach the optimal state before evaluation.

4. Analysis of colleges volleyball mixed teaching effect and quality evaluation results.

4.1. Analysis on teaching effect. The mixed teaching mode of university volleyball based on MOOC+ Flipped classroom and the evaluation model of university volleyball mixed teaching quality based on DBT-SVM will be handed over to experts for quality evaluation. The results show that the mixed teaching mode and teaching model are scientific and comprehensive. Since the focus of this research is the construction and application of the mixed teaching mode of MOOC+Flipped classroom in college volleyball teaching, the results will not be repeated. 50 students in the volleyball class selected by the 2020 level of a university are set as the experimental subjects. They are divided into two groups for observation and experiment, with 25 students in each group. Before the experiment, the scores of all subjects were recorded, including serving, receiving,

Table 4.1: Average score of the two groups of students

Group	Stage	Serve	Catch the ball	Spike	Comprehensive performance
Experimental group	Before experiment	5.29	6.11	4.99	5.41 ± 0.78
	After experiment	9.36	9.85	9.01	9.12 ± 0.75
Observation group	Before experiment	5.29	6.11	5.13	5.45 ± 0.71
	After experiment	6.77	8.65	7.59	7.63 ± 0.56

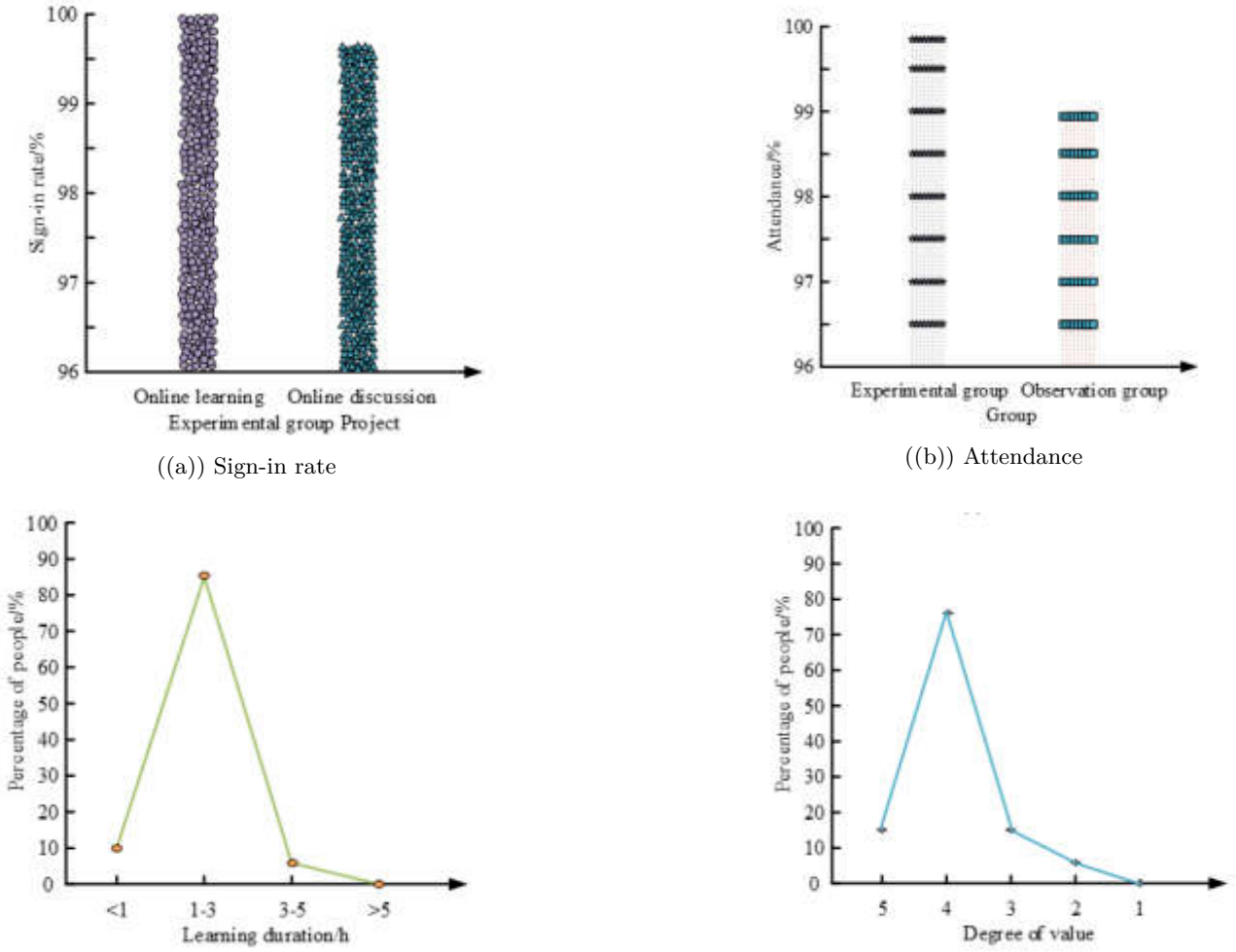
spiking and comprehensive performance. And the average value of each item of the two groups of students was calculated. After that, a semester was selected for the experiment, which lasted for 4 months, 4 weeks each month, 2 classes a week, and 90 minutes each time. The experimental group adopted the volleyball mixed teaching mode of "MFC", while the observation group adopted the traditional volleyball teaching mode with other conditions remained unchanged. The results were statistically analyzed. In Table 4.1, to facilitate the analysis, a ten-point system was adopted.

The results shown above are very close before the experiment in Table 4.1. After a semester of different teaching modes, the experimental group have greatly improved in serving, receiving, spiking and comprehensive performance. And the average score was significantly higher than that of the observation group. The comprehensive performance scores of the experimental group and the observation group before the experiment were 5.41 ± 0.78 and 5.45 ± 0.71 respectively. The two groups' comprehensive performance scores after the experiment are 9.12 ± 0.75 and 7.63 ± 0.56 respectively. In the volleyball mixed teaching mode of "MFC", the experimental group also have an online learning class than observation group. To further investigate students' learning attitudes under different teaching modes, in Figure 4.2, comparative analysis is made on students' attendance and learning duration.

In Figure 4.2, sub-figure (a) shows the attendance rate of online learning and online discussion of the experimental group. Subfigure (b) shows the attendance rate of two groups. Subfigure (c) shows the average network course learning time of the experimental group every week. Subfigure (d) shows the network resources value evaluation by the students in the experimental group. Students in the experimental group have an ideal check-in situation for online resources, and the attendance rate of the experimental group performs better than that in the observation group. Students are more interested in volleyball courses under the mixed teaching mode of "MFC". From an online resource learning perspective, most students' weekly learning time of online resources is controlled within 1-3 hours. A small number of students study within 1 hour and 3-5 hours. No student's study time is more than 5 hours. In evaluating the students' network resources value, the higher the score, the higher the value that students think network resources have. From sub-graph (d), most students believe that network resources have certain value for volleyball teaching. In volleyball learning and improve students' performance, "MFC" can be used to build a mixed teaching mode of volleyball to enhance students' interest.

4.2. Evaluation of TQ. The research uses the volleyball teaching data of a university as the experimental data, and uses the evaluation scale filled by teachers and students as the data source. Finally, it selects three suitable data sets and divide them into data1, data2 and data3. In Figure 4.3, it uses SVM, BT-SVM and DBT-SVM to classify the data, and calculates the accuracy of the three algorithms. On different data sets, the three algorithms' accuracy is DBT-SVM, BT-SVM and SVM from high to low, that is, DBT-SVM has higher prediction accuracy in Figure 4.3. In dataset data1, the accuracy rates of DBT-SVM, BT-SVM and SVM are 97.01%, 96.67% and 94.56% respectively. In dataset data2, the accuracy rates of DBT-SVM, BT-SVM and SVM are 98.03%, 97.33% and 93.87% respectively. In dataset data3, the accuracy of DBT-SVM, BT-SVM and SVM are 97.34%, 96.78% and 93.96% respectively. After that, eight experiments were conducted on the dataset using three algorithms. The average time consumption of the three algorithms in Figure 8 is obtained by statistics of the time consumption during the experiment.

Sub-graph (a) shows the average training time and sub-graph (b) shows the average test time in Figure 4.5.



((a)) Sign-in rate

((b)) Attendance

((c)) Learning duration

((d)) Value evaluation of network resources

Fig. 4.2: Learning situation analysis

The training and testing time of SVM is significantly more than that of BT-SVM and DBT-SVM. The latter two both take less time, and the gap is not very large. This is because the amount of data and categories in this experiment is small, so the amount of calculation is small, making the gap not obvious. However, the training time and test time of DBT-SVM are less than that of BT-SVM, that is to say, it is effective to use DBT-SVM for mixed college volleyball TQ evaluation. At the same time, DBT-SVM, BT-SVM and SVM can also be used to predict the results, and 15 groups of test data sets can be selected for prediction. The results are shown in Figure 4.7.

Sub-graph (a), sub-graph (b) and sub-graph (c) respectively show the prediction of DBT-SVM, BT-SVM and SVM on performance in Figure 4.7. The predicted results of DBT-SVM are closer to the original results, and the predicted results are significantly better than those of BT-SVM and SVM. To better reflect the evaluation effect of DBT-SVM, the prediction result error is compared to obtain the evaluation results and error comparison under the three algorithms, as shown in Figure 4.8.

Among the three algorithms, the error of prediction results is DBT-SVM, BT-SVM and SVM from small

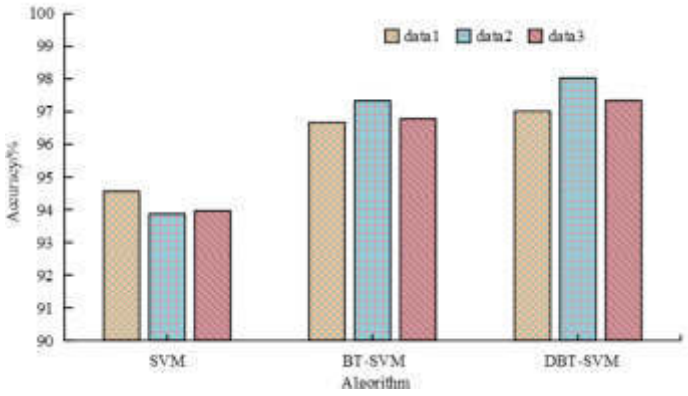
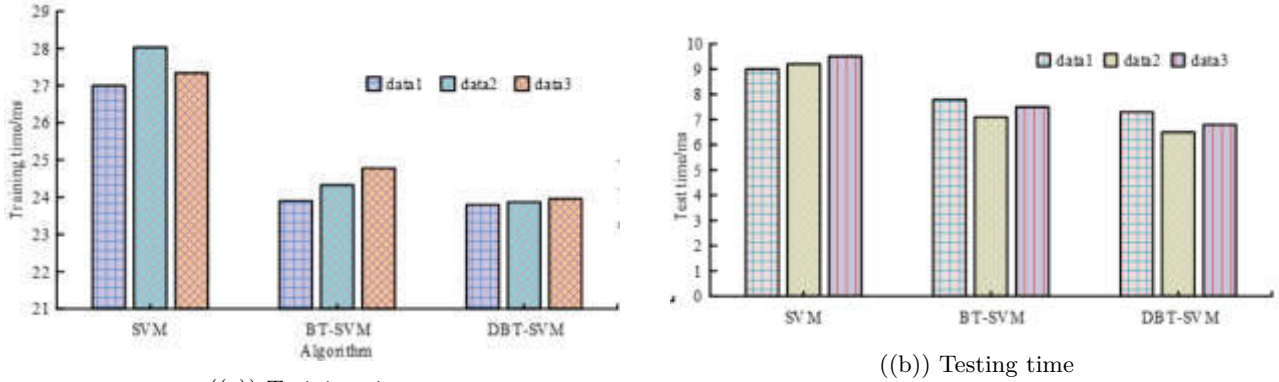


Fig. 4.3: Accuracy of three algorithms



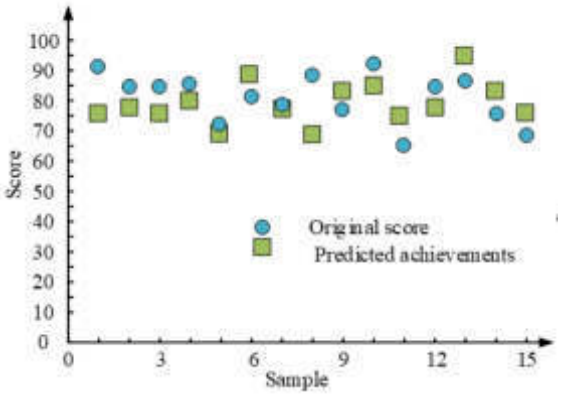
((a)) Training time

((b)) Testing time

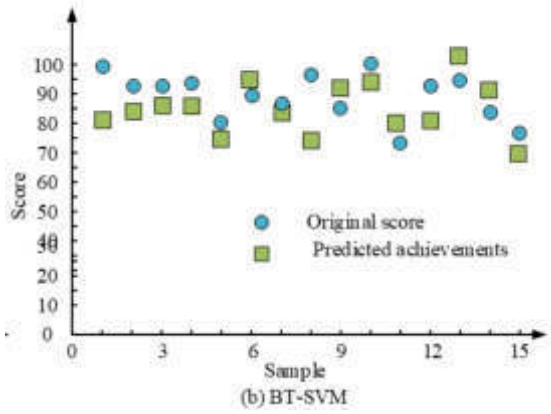
Fig. 4.5: Evaluation time of three algorithms

to large in Figure 4.8. This further shows that DBT-SVM can be used to build a mixed TQ evaluation model for college volleyball, which can achieve more accurate evaluation of teaching effect and quality. Then, it can monitor the teaching mode for a good college volleyball teaching environment creation.

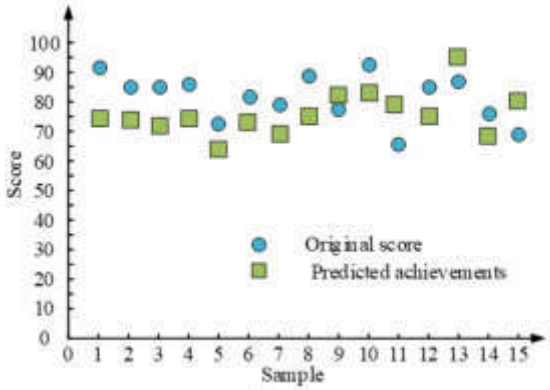
5. Conclusion. The information society has triggered the information wave of education. The development of MOOC had promoted the education equity, and provided sufficient high-quality resources to improve the flipped classroom. This provided a good choice for many universities to carry out teaching reform. "MFC" can form a student-centered teaching model. And it can make students' learning initiative and enthusiasm stimulated. To promote college volleyball teaching, it needs build a volleyball mixed teaching mode of "MFC". To further evaluate the effectiveness of this model, a model for evaluating the TQ was established based on DBT-SVM. Under the volleyball mixed teaching mode of "MFC", students' performance had been significantly improved. Students were also highly motivated to use online resources. The attendance rates of online learning and online discussion were 100% and 99.58% respectively. The attendance rate of the experimental group was higher than that of the observation group, which was 99.79% and 98.96% respectively. In the experiment, the TQ was evaluated by DBT-SVM. The accuracy of DBT-SVM in dataset data1, data2 and data3 was 97.01%, 98.03% and 97.34% respectively, and its comprehensive performance was better than BT-SVM and SVM. This showed that the TQ evaluation based on DBT-SVM is effective. In the future, more objective and scientific indicators will be selected to deeply evaluate the college volleyball mixed TQ.



((a)) DPT-SVM



((b)) BT-SVM



((c)) SVM

Fig. 4.7: Prediction results of three algorithms

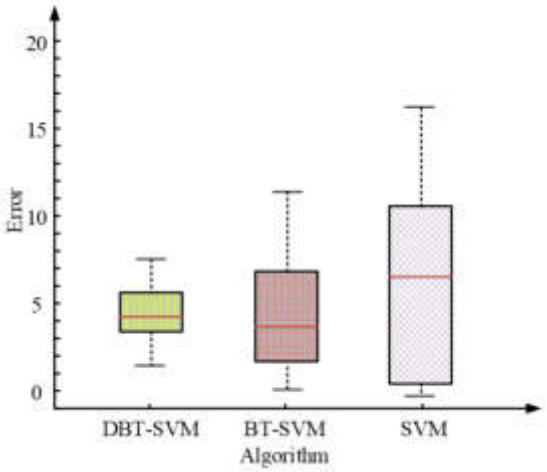


Fig. 4.8: Comparison of evaluation results and errors under three algorithms

Fundings. Hebei Provincial Department of Education 2023 "Humanities and Social Science Research Project of Hebei Education Department" (No.SZ2023123); Scientific Research Fund of North China space industry College in 2022 (Youth Fund Project)-Research on the Branding Construction of Student Sports Clubs in Our School in the New Era" (No.KY-2022-06).

REFERENCES

[1] Constantinescu, A. & Moore, N. Applying Adult Learning Principles, Technology, and Agile Methodology to a Course Redesign Project. *Journal For Quality And Participation*. **41**, 26-29 (2019)
[2] Li, X. & Cao, H. Research on VR-Supported Flipped Classroom Based on Blended Learning — A Case Study in "Learning English through News". *International Journal Of Information And Education Technology*. **10**, 104-109 (2020)
[3] Burkhart, S. & Craven, D. P161 Digital Workbooks to Develop and Evidence Learning in a Flipped Nutrition Classroom in Higher Education. *Journal Of Nutrition Education And Behavior*. **52**, S92-S93 (2020)
[4] Qian, Y., Li, C., Zou, X., Feng, X., Xiao, M. & Ding, Y. Research on predicting learning achievement in a flipped classroom based on MOOCs by big data analysis. *Computer Applications In Engineering Education*. **30**, 222-234 (2022)
[5] Chang, H. College English Flipped Classroom Teaching Model Based on Big Data and Deep Neural Networks. *Scientific Programming*. **8433**, 1-99184 (2021)
[6] Zhang, Y. & Yang, Y. The evaluation method for distance learning engagement of college English under the mixed teaching mode. *International Journal Of Continuing Engineering Education And Life-long Learning*. **32**, 159-175 (2022)
[7] Wu, X. Research on the Reform of Ideological and Political Teaching Evaluation Method of College English Course Based on "Online and Offline" Teaching. *Journal Of Higher Education Research*. **3**, 87-90 (2022)
[8] Jiang, L. & Wang, X. Optimization of Online TQ Evaluation Model Based on Hierarchical PSO-BP Neural Network. *Complexity*. **2020**, 1-12 (2020)
[9] Yuan, T. Algorithm of Classroom TQ Evaluation Based on Markov Chain. *Complexity*. **2021**, 1-12 (2021)
[10] Huang, W. Simulation of English TQ evaluation model based on gaussian process machine learning. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **40**, 2373-2383 (2021)
[11] Yu, H. Online TQ evaluation based on emotion recognition and improved AprioriTid algorithm. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **40**, 7037-7047 (2021)
[12] Hou, J. Online TQ evaluation model based on SVM and decision tree. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology*. **40**, 2193-2203 (2021)
[13] An, L., Yu, S., Fan, Q. & Others Exploration of the Mixed Teaching Mode of "Three Classes" under the Intelligent Teaching-based on Computer Programming Course. *Education Study*. **2**, 33-42 (2020)
[14] Li, Y., Wang, Y., Li, Y. & Yang, X. Practical Experience of Mixed Teaching Mode Based on SPOC in Physiology Experiment Course for International Students. *Education Study*. **2**, 304-312 (2020)
[15] Guo, M. & Huang, S. Exploration of "Flipped Classroom" Teaching Mode in College Chinese Course in the Era of "Internet plus ". *Basic & Clinical Pharmacology & Toxicology*. **124** pp. 310-311 (2019)
[16] An, L., Yu, S., Fan, Q. & Zhao, Y. Exploration of the Mixed Teaching Mode of "Three Classes" under the Intelligent Teaching-based on Computer Programming Course. *Education Study*. **2**, 33-42 (2020)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 17, 2023
Accepted: Nov 6, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 383–394, DOI 10.12694/scpe.v25i1.2304

RESEARCH ON THE APPLICATION OF APRIORI ALGORITHM IN THE TEACHING OF BALL SPORTS TECHNIQUES AND TACTICS

JIANQUN AN* AND YONGFENG ZHAO†

Abstract. To address the problem that technical and tactical aspects of ball games have a large impact on match-winning, association rules are used to mine and analyse the data. Item constraint terms are added to the original Apriori algorithm to reduce the generation of redundant data, and the operational efficiency of the improved association rules is improved with the aid of the GSA-PSO algorithm to give full play to the fast convergence advantage of the hybrid algorithm. The experimental results show that the algorithm proposed in the study is more convergent than the Apriori algorithm, and the former has fewer iterations on average in the single-peaked function; the overall search accuracy is significantly improved. The algorithm is also less disturbed by the dataset than the Apriori algorithm, with an average running time of 19.23s, 21.54s and 25.61s for dataset sizes of 200, 500 and 1000 respectively. The deviation rate between the predicted and actual coaching strategies was less than 5% and the scoring tolerance rate was between 0.04 and 0.10. It indicates that the algorithm proposed in the study can improve the efficiency of data analysis, which is conducive to the development of a perfect ball sport technical and tactical scheme and promote the stability of ball sport technical and tactical development.

Key words: Apriori Algorithm, GSA-PSO, Association Rules, Sports Teaching, Tactics, Data Mining

1. Introduction. Since the founding of New China, the development of sports in China has grown from small to large and from weak to strong, and our country has become an important force with strong competitiveness in the international sports arena. Ball games are an integral part of sport, and many ball games are popular with our nationals. At the same time, in the process of modernising sports, the rules of ball games are constantly being updated, bringing certain changes and challenges to ball game techniques and tactics [1]. Techniques and tactics have a significant impact on the performance of athletes in the game, and the rapid and accurate development of ball sport techniques and tactics to guide the game is one of the key factors in winning the game. At the same time, with the strong development of computer technology, computer science is able to improve the efficiency and accuracy of data analysis, so information science techniques are being applied more and more frequently in sports [2]. Among them, association rule-based data mining is superior for the analysis of sports-related match data, and the Apriori algorithm is the most classical algorithm for association rules, but it is easy to fall into local optimal solutions and the number of scans of the most databases makes the efficiency of the algorithm greatly reduced [3]. Regarding the issue of the significant impact of technical and tactical skills on winning games in ball games, association rules are used to mine and analyze data. Optimize the Apriori algorithm to improve the efficiency of data analysis, facilitate the development of comprehensive ball sports technical and tactical plans, provide auxiliary support for improving the level of ball sports technical and tactical skills, and promote the stable development of ball sports technical and tactical skills.This study will optimize the Apriori algorithm by adding project constraints to the original Apriori algorithm, and utilize the GSA-PSO algorithm to improve the efficiency of association rules, fully leveraging the fast convergence advantage of hybrid algorithms. This is also the innovation of this study. To improve the efficiency of data analysis, promote the development of comprehensive ball sports technical and tactical plans, provide auxiliary support for improving the level of ball sports technical and tactical skills, and promote the stable development of ball sports technical and tactical skills.

2. Related Works. Data mining based on association rules can extract key data from a large amount of data and analyse it with high accuracy. The Apriori algorithm, as the most classical algorithm in association

*College of Sports Science, Lingnan Normal University, Zhanjiang, 524048, China
†College of Sports Science, Lingnan Normal University, Zhanjiang, 524048, China (yongfengzhaolnnu@163.com)

rules, is widely used in many fields. Many scholars have combined association rules to explore a large number of data mining, and have achieved a lot of research results.

Yang et al. constructed a data mining model based on the Apriori algorithm in order to solve the problem of quickly finding effective information from data materials, and the experimental results proved that the model can quickly mine effective information and promote the practicality and modernization of the university teaching management evaluation system [4].

Findawati et al. constructed a data mining model based on the Apriori algorithm in order to To uncover the potential value in the sales process of HPAI products, a data mining model based on an improved Apriori algorithm was designed, which can extract valuable information from the sales information of HPAI products and help sellers analyse the potential consumption demand for purchasing HPAI products [5].

In order to mine the potential behaviour of customers from the database, an association rule based on Apriori algorithm was established, which can more accurately mine the potential consumption behaviour of users and facilitate companies to develop more appropriate marketing strategies [6].

Luo et al. proposed an Apriori algorithm-based association rule in order to solve the problem of unreasonable medication use caused by redundant information in the medical industry. a data mining system based on Apriori algorithm, which can analyze and mine the correlation information in the data, so that the laws of rational drug use can be analyzed from the laws of drug use, dosage studies, etc. [7].

Liu et al. proposed a method and means based on Apriori algorithm and machine learning for the problem that the methods and means of volleyball technical prediction in China are relatively lagging behind Apriori algorithm and machine learning data mining algorithm, and the experiment proved that the algorithm can predict the score of volleyball matches with high accuracy based on the training data [8].

Lu et al. established an association rule analysis identification method based on the Apriori algorithm in order to find effective combinations of acupoints for the treatment of diabetic gastroparesis, and the experimental results proved that the method was able to link acupoints and establish effective combinations of acupoints that could play a better role in the treatment of diabetic gastroparesis [9].

Li et al. proposed an Apriori algorithm data mining model in order to study the the effect of course sequencing on student performance, an Apriori algorithm data mining model was designed, and experimental results proved that the model was able to reveal the internal connections between various course clusters and provide guidance to students on course selection [10].

In order to establish the reliability of a new type of badminton interval training, a simulated match and evaluation game mechanism, and experimental results showed that the mechanism was able to effectively match athletes' training and rest intervals based on their physiological responses and time to exhaustion [11].

Wang et al. structured a mathematical model of the vertical height and horizontal speed of the basketball offensive line in order to improve the hitting rate of basketball players, and experimental results showed that athletes using the dominant hand is a higher hit rate, and the model improves the teaching results of teachers explaining basketball training techniques and tactics [12].

Zhao used complexity computer simulation to simulate a football field, combined with mapping software to draw the football and players to develop real-time tactics, and the experimental results showed that using complexity computer simulation for early school football overall play with high accuracy of passing and running, which can provide implications for the overall tactics of football [13].

Yuki et al. designed a data mining model combining clustering and association rule analysis in order to perform effective data mining in multi-objective topologies, and the experimental results showed that the model was able to obtain visualization results of the target space and promote optimization of multi-objective topologies [14]. Malik et al. structured a data mining scheme to generate vehicle paths in abnormal road event maps, a safety-authentication-based data collection scheme was designed to mine and analyse the data using association rules and VANET techniques, and experimental results showed that the scheme was able to predict the vehicle paths in abnormal situations and save time for emergency rescue [15].

As can be seen in the above, Apriori algorithms have important applications in various industries and data mining with Apriori algorithms can provide intrinsic connections between behaviours. Given the importance of tactical instruction in ball games to the outcome of the game, the Apriori algorithm will be used to mine data in ball games with a view to providing better tactical instruction in ball games.

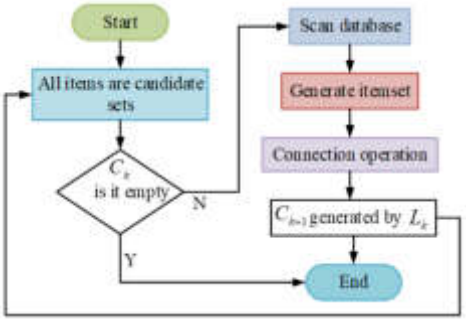


Fig. 3.1: Apriori algorithm flow chart

3. Application of Association Rule Mining Algorithm based on Apriori Algorithm in Sports Teaching.

3.1. Optimal Apriori algorithm under constraints. Data mining is the process of exploring large amounts of fuzzy random data for useful knowledge that can be used for information queries and management decisions, and predicting potential knowledge patterns with the help of relevant intelligence techniques and computer science to provide users with sound decisions. Data mining can be divided into descriptive and predictive according to its functions, of which association rule mining, as one of the important branches, can grasp the relationship between data in the database and find the set of frequent and satisfying support attributes, and subsequently, according to the preset confidence level, the eligible rule relationship can be derived [16]. Specifically, the association rule uses Support and Confidence to filter and filter the set of items preceding and following for the attributes of the data set, and then sets a threshold to achieve the data mining purpose [17]. Where the support and confidence of $(A \rightarrow B)$ is calculated as shown in equation ref3.1.

$$\begin{cases} Support(A \rightarrow B) = P(A \cup B) \\ Confidence(A \rightarrow B) = P(B|A) \end{cases} \tag{3.1}$$

Equation. 3.1 represents the proportion of$A$ and B contained in the set of instances Y and the probability of occurrence of Y in the instances containing A, where the mathematical expression for satisfying the threshold set by the association rule is shown in Eq. (3.2).

$$\begin{cases} \text{Support}(A) \geq \text{Min\_Support} \\ \text{Confidence}(A) \geq \text{Min\_Confidence} \end{cases} \tag{3.2}$$

In equation 3.2, A is a frequent item set when the support of the item set A is greater than or equal to the minimum support; the lift in the association rule is shown in equation 3.3.

$$Lift(A \Rightarrow B) = \frac{P(B|A)}{P(A)} \tag{3.3}$$

Equation 3.3 is defined as the ratio of the confidence of the rule to its support. When the confidence level of the item set is greater than or equal to the minimum confidence level, the association rule obtained is a strong association rule [18]. Association rules can include numerical, unidimensional, and single-layer depending on the problem situation, and the Apriori algorithm is the most widely used among numerical rules [19]. The Apriori algorithm is implemented as an iterative algorithm with layer-by-layer search to react to the correlation between transactions, and the implementation flow of this algorithm is shown in Figure 3.1.

$S_{min}C_k$ If all the items belong to the candidate set C_k, the algorithm is judged to be empty. If all the items are empty, the algorithm is terminated; otherwise, the database is scanned and the frequent $(k+1)$ items are

selected from the candidate set L_{k+1} , and the frequent items are generated from L_k . C_{k+1} is judged again. The Apriori algorithm can effectively improve its efficiency in processing the candidate frequent itemset C_k , but it is more susceptible to the impact of data redundancy which makes some phenomena not meet the user's expectation [20].

The teaching of ball technology is limited by the level of information technology, and some of the important match-related data is only used in statistical analysis, and less consideration is given to the deeper analysis of the data and the correlation between winning and losing matches. In addition, the traditional Apriori algorithm has the disadvantages of being time-consuming and having a lot of rules, which makes it difficult to improve the efficiency of data analysis and cope with large matches. Therefore, we propose an improved algorithm that uses user interest as a constraint term to improve the problem of generating a large candidate set. Algorithm of Constrained Association Rule Mining based on Item, ACARMI) [21]. The ACARMI algorithm introduces thresholds to avoid the problem of the Reorder algorithm having a large number of candidate itemsets in the initial stage and the Direct algorithm generating too many in the later stage, achieving the accuracy and efficiency of data mining, thereby improving the efficiency of data analysis, facilitating the development of comprehensive ball sports technical and tactical plans, and promoting the stable development of ball sports technical and tactical plans.Firstly, by introducing the constraint , the mathematical expression is shown in equation (3.4), so that the support and confidence of the association rule contain a certain weight of the item set X, Y.

$$\begin{cases} C = e_1 \vee e_2 \vee e_3 \vee e_4 \vee e_n \\ e = a_1 \vee a_2 \vee a_3 \vee a_4 \vee a_m \end{cases} \tag{3.4}$$

In equation 3.4, $e_1 \vee e_2 \vee e_3 \vee e_4 \vee e_n$ is the Boolean expression for the constraint and is the set form. Unlike the original association algorithm which frequently computes the database, the constraint only scans the subset associated with the constraint and generates the set of transactions that meet the constraint, the computation of which consumes the time process shown in equation 3.5.

$$\begin{cases} t = |D| * (\lambda + 1) * \tau + |D| * \tau \\ t' = |D| * \tau + |D'| * (\lambda + 1) * \tau + |D| * \tau \end{cases} \tag{3.5}$$

In equation 3.5, D' is the set that meets the constraint, τ is the time consumed by a single processing transaction, λ is the maximum length of the frequent item set, and τ is the direct scan time and the filtered time. When $t' < t$ is used, then equation 3.6 is satisfied.

$$|D|/|D'| < \lambda/(\lambda + 1) \tag{3.6}$$

The data is collated to obtain the approximate rate of data reduction, as shown in equation3.7.

$$Fil\_Ratio = (\frac{|D| - |D'|}{D}) > \frac{1}{\lambda + 1} \tag{3.7}$$

The ACARMI algorithm avoids the problems of the Reorder algorithm having too many candidate items in the initial stage and the Direct algorithm having too many in the later stage by introducing a threshold to achieve accuracy and efficiency in data mining. The computational flow of the ACARMI algorithm is shown in Figure 3.2.

In Figure 3.2, the ACARMI algorithm generates the candidate item set with the minimum length frequent item set in the initial stage of the operation, and then during the scanning of the database, the iterative generated item sets are processed separately, i.e. iterations less than the maximum length are processed with the Direct algorithm and vice versa, the item sets generated in the previous cycle are sorted and subsequently new candidate item sets are generated with the Reorder algorithm [22]. This keeps its candidate item set small overall, thus achieving a guarantee on the efficiency of the algorithm.

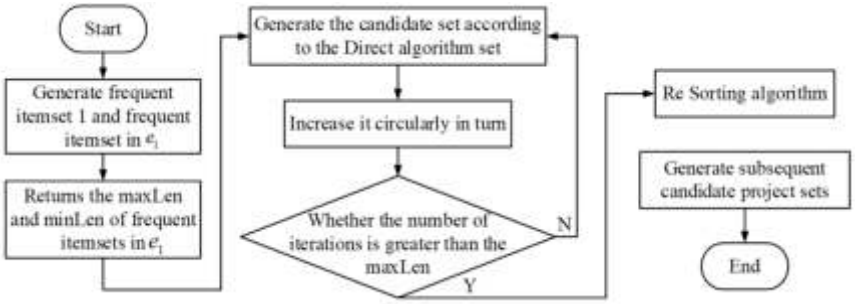


Fig. 3.2: Calculation flow of acami algorithm

3.2. Construction of an Application System for Teaching Sports Techniques and Tactics under the GSA-PSO-Apriori Algorithm. Ball sports competitions require high timeliness and therefore require good processing efficiency when using association rules for tactics development and related data analysis requirements. In order to further improve the efficiency of the Apriori algorithm in scanning data, the study relies on the Gravitational search algorithm-Particle swarm optimization (GSA-PSO) for the extraction of association rules, optimising the fast convergence characteristics of the hybrid algorithm Apriori algorithm, i.e. converting the original data into a binary format type and calculating the adaptation values for each particle. The rules are then mined with the aid of the GSA-PSO algorithm in order to update the particles and optimise the conditions [23]. The GSA algorithm considers that the mutual attraction between particles is proportional to their own mass and inversely proportional to the distance, which can improve the adaptability and applicability of the algorithm. inversely proportional to the distance, and can obtain the particle with the largest inertial mass after iteration, whose mathematical expression is shown in equation 3.8.

$$F_{ij}^{d}(t) = G(t)\frac{M_i(t)M_j(t)}{R_{ij(t)} + \epsilon}(xj^{d}(t) - xi^{d}(t)) \tag{3.8}$$

Equation 3.8 is the mathematical expression for the particle i acting on the particle j in the same dimensional space d and time t, where x_i^d is the position of the particle i in d dimensional space, M_i, M_j is the particle and its gravitational mass, ϵ is a constant, and $R_{ij}(t)$ is the Euclidean distance of the particle i, j in time. The mathematical expression for the Euclidean distance there is shown in equation 3.9.

$$R_{ij}(t) = ||X_i(t), X_j(t)||^2 \tag{3.9}$$

In Equation 3.8, $G(t)$ is the gravitational constant at the moment of t. To ensure the randomness of the gravitational search algorithm, by setting a random number taking values in the range [0,1], as shown in Equation 3.10.

$$F_i^d(t) = \sum_{j=i,j\neq i}^{N} rand\dot{F}_{ij}^{d}(t) \tag{3.10}$$

If the gravitational mass value is set equal to the inertial mass value in GSA, the sum of the current velocity of the particle and its acceleration is the velocity of the particle at the next moment, i.e. the mathematical formula for the position and velocity of the particle is shown in equation 3.11.

$$\begin{aligned} v_{id}(t+1) &= randv_{id}(t) + a_{id(t)} \\ x_{id}(t+1) &= x_{id}(t) + v_{id(t)} \end{aligned} \tag{3.11}$$

In Equation 3.11, $a_{id}(t)$ is the acceleration of the mass in the dimensional space [24][24]. In the traditional particle swarm algorithm, each particle in the population represents a feasible solution in the optimization

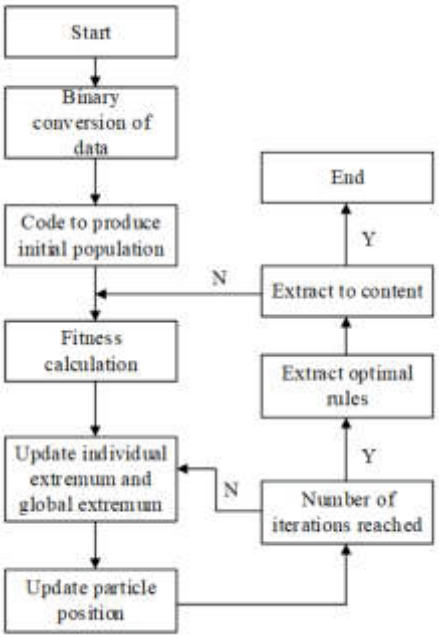


Fig. 3.3: Flow chart of gsa-pso algorithm mining association rules

problem, and the dimensionality and population size of each particle are D and N, respectively, which are calculated as shown in Equation 3.12.

$$\begin{cases} v_i(t+1) = w\dot{v}_i(t) + c_1 r_1(t)[pbest_i(t) - x_i(t)]c_2 r_2(t)[pbest_i(t) - x_i(t)] \\ x_i(t+1) = x_i(t) + v_i(t+1) \end{cases} \tag{3.12}$$

In equation 3.12, v_{t+1} and $x_i(t+1)$ are the velocities of the particles and their positions, $pbest_i$ is the local optimal solution of the particles. t is the number of iterations, w is the inertia weight, c_1, c_2 is the individual learning factor and social learning factor, and r_1, r_2 is a uniform random number in the range of [0, 1]. The combination of the GSA algorithm and the PSO algorithm can ensure the continuous improvement of particle memory while enhancing its ability to exchange information, so that the particles can enhance the ability of group information exchange while maintaining the original laws of motion, and its improved hybrid algorithm formula is shown in equation (3.13).

$$\begin{gathered} v_{id}(t+1) = r_1 v_{id}(t) + C_1 r_2 a_{id}(t) + c_2 r_3 (gbest_{id} - x_{id}) \\ x_{id}(t+1) = x_{id}(t) + v_{id(t+1)} \end{gathered} \tag{3.13}$$

In equation 3.13, v_{id} is the position of the particle at the moment, c_1, c_2 is the learning factor, r_1, r_2, r_3 is a random number in the range of [0,1], and $gbest_{id}$ is the best position of the group. By adjusting the value of the learning factor, the particle's own gravitational force can be balanced with the global exchange capacity, and thus improve the particle's performance in finding the best position. Figure 3.3 shows the flow chart of the GSA-PSO algorithm for mining association rules.

Figure 3.3 shows the key process of mining association rules using the GSA-PSO algorithm, which is the core part of the improved algorithm. When extracting rules with the GSA-PSO algorithm, data types such as the total number of particle populations, learning factor, maximum number of iterations and influence factor need to be input, and the recording of each particle position and new population is achieved with the help of the fitness calculation and particle optimal value update to ensure the extraction of optimal association rules. The

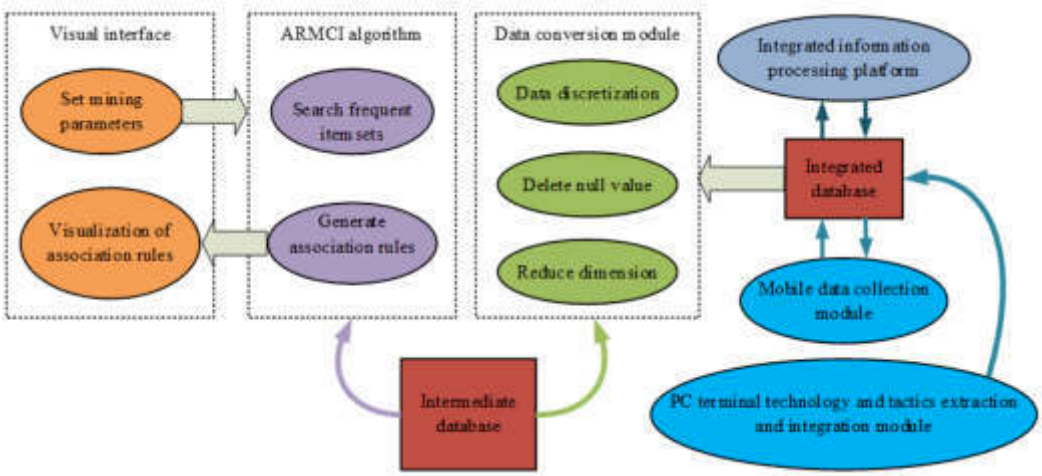


Fig. 3.4: Schematic diagram of the overall architecture flow of the system

GSA-PSO algorithm has improved the operational efficiency of the improved association rules, improved the efficiency of scanning data, achieved particle updates, and optimized conditions; The introduction of gravity search algorithm can ensure that the population particles fully exchange information, effectively improving the adaptability and applicability of the algorithm.The confidence and support degrees are used to adapt the degree function, and the influence factors are multiplied and summed to obtain the adaptation function, whose mathematical expression is shown in equation 3.14.

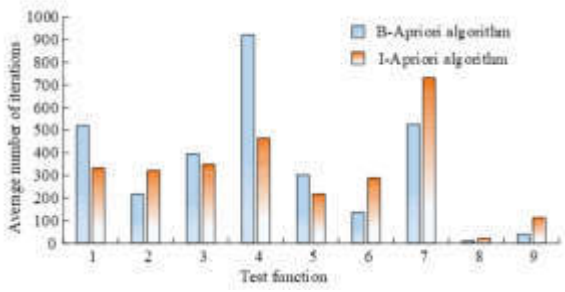
$$F(x) = aS(x) + bC(x) \tag{3.14}$$

In equation 3.14, x is the particle and a, b is the parameter of $S(x)$ support and $C(x)$ adaptation in a function that takes values in the range [0,1]. Each particle has its own velocity, position and fitness values, and the binary processing allows the particle swarm formula to be updated with the mathematical expression shown in equation 3.15.

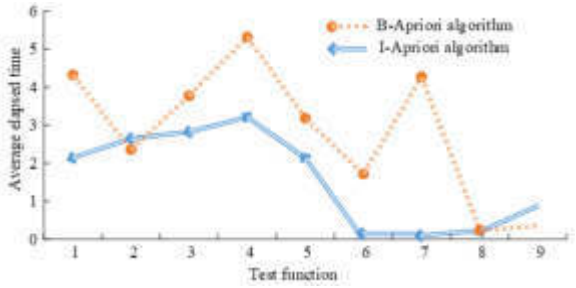
$$\begin{aligned} v_{id}(t+1) &= r_1 v_{id}(t) = c_1 r_3 (gbest_{id} - x_{id} \\ x_{id} &= \begin{cases} 1, r_4 < sig(v_{id}(t+1)) \\ 0, r_4 \geq sig(v_{id}(t+1) \end{cases} \end{aligned} \tag{3.15}$$

In equation 3.15, $i(i = 1, 2, 3, \ldots, N)$ is the population size and the sig function takes values in the range (0,1). The velocity determines the maximum distance a particle can move in an iteration, with a larger velocity indicating a closer local search to the vicinity of the optimal solution. At the same time, the computer information platform is used to conduct data statistics on the scores, duration and errors of ball players during the competition, and this data is transferred to the back-end database. After pre-processing the data, improved association rule mining is introduced with a view to providing better decision analysis for athletic training of athletes and technical teaching of coaches. Figure 3.4 illustrates the overall architecture of the analysis system.

The system design consists of data collection, data conversion, algorithm implementation and visualisation interface. The pre-processing section converts numerical statistical information into Boolean data, and dimensionality reduction, constraint restriction and data type re-conversion make it possible to perform better association rule mining. The visualisation interface enables the presentation of relevant data results and thus provides strategic guidance to the user. At the same time, there are many different types of data, too many uncertainties and too many different attributes, so similar attributes need to be merged in the data processing phase to reduce the redundancy of frequent item sets.



((a)) Average iteration times of Apriori algorithm before and after the experiment



((b)) Time consumption of Apriori algorithm before and after the experiment

Fig. 4.2: Iteration times and execution time of Apriori algorithm before and after improvement

Table 4.1: Experimental Environment Parameter Settings

Parameter	**Setting**
Platform	Win7, MATLAB
Memory	4GB, CPU 1.8GHz
Population size	50
Maximum iterations	1000
Number of function runs	20

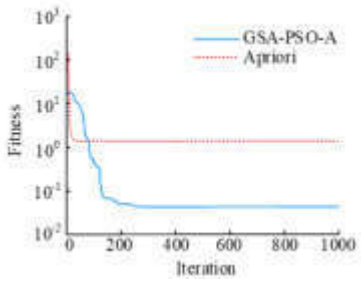
4. Analysis of the application of improved Apriori algorithm in the teaching of ball sports techniques and tactics.

4.1. Application Performance Analysis of the Improved Apriori Algorithm. After setting the maximum number of iterations, the data was counted for the number of iterations and execution time of the algorithm before and after the improvement of the Apriori algorithm. 50 experiments were conducted for each function value and the experimental results were counted as shown in Figure 4.2.

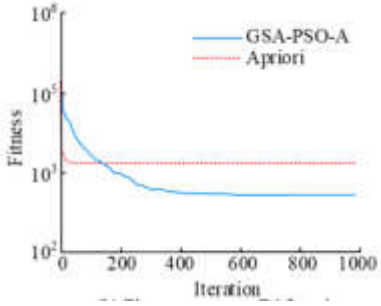
In Figure 4.2, the Apriori algorithm before the improvement is shown, and the Apriori algorithm after the improvement is shown. Test functions 1-9 are the minimisation benchmark functions that test the performance of the algorithm, where 1-4 are single-peaked functions and 5-9 are multi-peaked functions. Due to different peaks and test functions, the value of test function 8 is minimized.The results in Figure 5 show that the average number of iterations of the improved Apriori algorithm on the single-peaked function is less than that of the pre-modified algorithm, basically less than 500, but the difference in execution time is larger, indicating that the improved algorithm has better convergence on the single-peaked function and consumes less than 3 s. On the multi-peaked function, the number of iterations of the improved algorithm is higher than that of the pre-modified algorithm. This is because the improved algorithm performs particle swapping and global optimisation processes, but it takes less time to execute, essentially less than 2 s. Overall, the improved algorithm appears to be more effective for specific applications. The improved algorithm was then tested for fitness performance and the corresponding environmental parameters were set, as shown in Table 4.1.

The statistical analysis of the operational performance of the improved algorithm under different benchmark functions was carried out and the data collation results are shown in Figure 4.2.

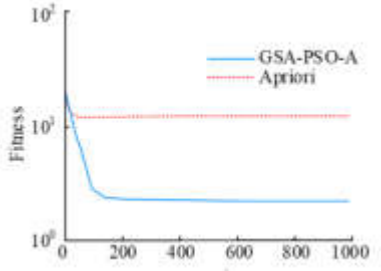
The results in Figure 4.4 show that the optimization results of the GSA-PSO-A algorithm on different benchmark functions differ, where the change in the fitness value of the GSA-PSO-A algorithm on the F2 function tends to decrease with the increase in the number of iterations, and the fitness value at 200 iterations is 0.055, which tends to be smooth at a later stage, much lower than that of the Apriori algorithm after



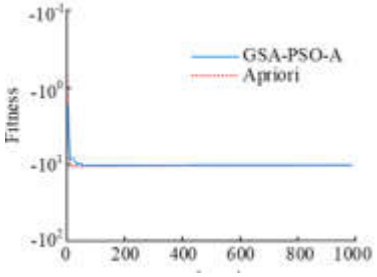
((a)) Fitness Curve on F2 Function



((b)) Fitness Curve on F6 Function



((c)) Fitness Curve on F7 Function



((d)) Fitness Curve on F9Function

Fig. 4.4: The performance of the improved algorithm under different benchmark functions

Table 4.2: Data Operation Results

/	/	Dataset Size	Average Number of Rules	Average Running Time	Support Threshold
Different Data Sets	Confidence Threshold	200	7.6	19.23	/
	Support Threshold	500	9.4	21.54	/
	Learning Factor	1000	12.7	25.61	/
Different Support Thresholds	Dataset Size	/	10.1	22.49	10%
	Confidence Threshold	/	9.1	19.87	15%
	Learning Factor	/	8.2	17.74	220%

smoothing. On the F6 function, the improved algorithm has a higher fitness value before 200 iterations than the improved algorithm, with a maximum value of over 1000, while the fitness value that plateaus at a later stage is 830. local optimum problem. Also with the help of MATLAB to achieve frequent item set and association rules mining run times 10 times, to get the run table under different data sets, the results are shown in Table 4.2.

The results in Table 4.2 show that the running time of the GSA-PSO algorithm is less affected by the number of databases, with the average running times of 19.23s, 21.54s and 25.61s for dataset sizes of 200, 500 and 1000 respectively, indicating that it can better adapt to the application effect of large datasets. While there is an inverse relationship between the change in support threshold and the number of association rules mined, which consumes more time, the improved algorithm proposed in the study can reduce the interference of some useless data during the run.

Table 4.3: Algorithm Mining Results after Introducing Constraints

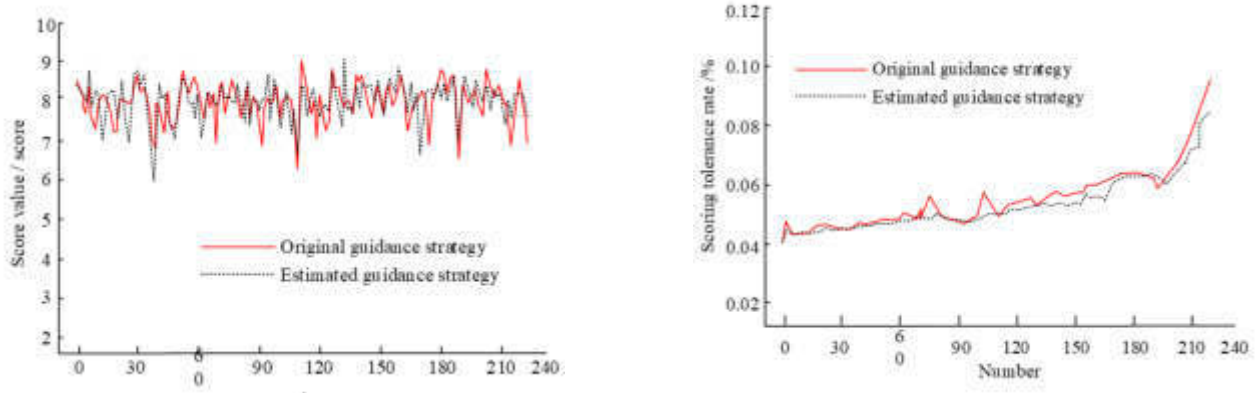
Rule	Support	Confidence Level
7-D Anti kill	15.38%	11.12%
5-J Anti high pair	8.97%	13.35%
4-J Medium high alignment	8.56%	16.27%
6-A Head block pair	8.54%	22.17%
6-A Far and straight	8.27%	30.98%
9-K Anti cross checking	8.02%	32.23%
11-J Anti blocking	7.11%	13.97%
12-L Anti blocking	6.75%	12.37%
4-L Reverse gear far	6.34%	34.56%
5-E Killing right	5.87%	13.08%
6-L Anti block	3.16%	36.27%
6-A Anti high school	16.28%	16.96%

4.2. Analysis of Application Effects. When performing data mining, the minimum support and minimum confidence were entered and constraints were set to make the results of the analysis. Also to reduce the impact of uncertainty on the results of sporting competitions and to reduce the loss of value knowledge, the minimum support and minimum confidence were set to 4% and 10% respectively and the derived results were collated as shown in Table 4.3.

Table 4.3 shows the algorithm mining results after introducing constraint conditions, which can have a certain impact on tactics and teaching when combined with rules.The association rules in Table 4.3 are represented as 'number-letter-behaviour', for example '7-D reverse kill pair' represents an event where the player hits the ball from zone 7 to zone D and the technical means is a reverse kill pair and the 'score loss "The probability of the event occurring at the same time as the 'loss of point' event was 11.12% and the support level was 15.38% indicating that the ball player was aware of the consequences of this event resulting in a loss of point. The results in Table 1 show that the probability of a player hitting the ball from zone 6 to zone L with a technical means of back-blocking being the highest for this event to occur in conjunction with a 'lost point' event was 36.27%, while only 3.16% of players were aware of this issue. The support results showed that the players' strategy of using backhand attacks was more successful, therefore, in the later stages of sport teaching, the training of targeted technical characteristics should be strengthened to improve their professional skills and abilities, taking into account the probability of losing points. To further validate the effectiveness of the application of the association rule proposed in this study, the data of ball players' participation in a sports season in a certain year was used as experimental data for application analysis and the differences were compared with the original strategy formulation, and the degree of effectiveness of its application was indicated by a rating value of 1-10, and the results are shown in Figure 4.3.

The results in Figure 4.6 show that the improved association rule algorithm proposed in the study can effectively analyse the technical movements of ball players, and the scoring results between its proposed guidance strategy and the original guidance strategy are similar, basically between 7-9 points, and the overall change is relatively smooth, with a small fluctuation range and a deviation rate of no more than 5%. And the scoring tolerance rate between the predicted strategy and the original strategy is basically between 0.04-0.10, which means that the improved algorithm can better meet the realistic requirements and has better application results.

5. Conclusion. At present, there are many statistical data of ball games. Association rules can be introduced into the statistical data of ball games to analyze and obtain the rules related to the winning factors of the game. Improve the most classic Apriori algorithm in association rules, and study the extraction of association rules with the help of gsa-pso, so as to further improve the efficiency of Apriori algorithm in scanning data. The experimental results show that the fitness of the proposed algorithm tends to decline with the increase of iteration times in the corresponding parameter environment, and the F2 function is 0.055; It is 830 in F6 function, and the overall search accuracy has been significantly improved, avoiding the algorithm



((a)) Score Value ((b)) Scoring Error

Fig. 4.6: A comparative study on the guiding strategies of ball players in competition

falling into the local optimization problem. And the average running time of the proposed algorithm when the data set size is 200, 500 and 1000 is 19.23s, 21.54s and 25.61s respectively, which is less disturbed by the data set. Through the application analysis of this algorithm, it is found that it can make a better correlation analysis of athletes' actions and loss of points in the guidance teaching, in which area 6 hits area L and the technical means is anti blocking. The probability of this event and "loss of points" event is the highest, reaching 36.27%, while only 3.16% of athletes are aware of this problem. At the same time, the deviation rate between the predicted guidance strategy and the actual guidance strategy is no more than 5%, and the scoring tolerance rate is basically between 0.04-0.10, which has good application value and effect. It shows that the proposed algorithm can help coaches make targeted training plans and assist in on-the-spot decision-making, and has very important practical value.The algorithm proposed in the study can improve the efficiency of data analysis, facilitate the development of comprehensive ball sports technical and tactical plans, and promote the stable development of ball sports technical and tactical plans.

REFERENCES

[1] Sumarly, V., Arisandi, D. & Sutrisno, T. Utilization of Apriori algorithm for book layout design in Untar library. *IOP Conference Series: Materials Science And Engineering*. **1007**, 12160-01216 (2020)

[2] Torres-Luque, G. ÁI Fernández-García, Cabello-Manrique D, et al. *Design And Validation Of An Observational Instrument For The Technical-tactical Actions In Singles Tennis*. **9**, 1-10 (2018)

[3] Daniele, C., Antonio, T., Aaron, G. & Others Investigating the game-related statistics and tactical profile in NCAA division I men's basketball games. *Biology Of Sport*. **35**, 137-143 (2018)

[4] Yang, Z. Research on the application of university teaching management evaluation system based on Apriori algorithm. *Journal Of Physics: Conference Series*. **1883**, 12033-01203 (2021)

[5] Findawati, Y., Hikmah, I., Sumarno, S. & Others HPAI consumer shopping analysis using Apriori algorithm. *IOP Conference Series: Materials Science And Engineering*. **1098**, 32087-03209 (2021)

[6] Silva, J., Varela, N., López, L. & Others Association rules extraction for customer segmentation in the SMEs sector using the Apriori algorithm. *Procedia Computer Science*. **151**, 1207-1212 (2019)

[7] Luo, Z., Cheng, B., Tian, S. & Others Review of rational drug use based on Apriori algorithm. *Journal Of Physics: Conference Series*. **1757**, 12123-01213 (2021)

[8] Liu, Q. & Liu, Q. Prediction of volleyball competition using machine learning and edge intelligence. *Mobile Information Systems*. **2021**, 1-8 (2021)

[9] Lu, P., Keng, J., Tsai, F. & Others An Apriori algorithm-based association rule analysis to identify acupoint combinations for treating diabetic gastroparesis. *Evidence-based Complementary And Alternative Medicine*. **2021**, 1-9 (2021)

[10] Li, T. & And, H. and implementation software for mining association rules (market basket analysis) to design product layout decisions. *Journal Of Physics: Conference Series*. **1869**, 12121-01212 (2021)

[11] Chia, J., Jia, Y., Barrett, L. & Others Reliability of a novel badminton intermittent exercise protocol. *Research Quarterly*

For Exercise And Sport. **90**, 2019 (0)
[12] Wang, L. Research on the application of modern computer technology in the modeling of basketball offensive line measurement and calculation. *Journal Of Physics: Conference Series*. **1952**, 42082-04208 (2021)
[13] Zhao, D. Complexity computer simulation in the study of the overall playing method of campus football. *Complexity*. **2021**, 1-9 (2021)
[14] Malik, A., Pandey, B. & Wu, C. Secure model to generate path map for vehicles in unusual road incidents using association rule based mining in VANET. *Journal Of Electronic Science And Technology*. **16** pp. 02 (2018)
[15] Yuki, S. Kazuhiro, et al. *Data Mining Based On Clustering And Association Rule Analysis For Knowledge Discovery In Multiobjective Topology Optimization – ScienceDirect. Expert Systems With Applications*. **119** pp. 247-261 (2019)
[16] Yue, T. & Zou, Y. Online teaching system of sports training based on mobile multimedia communication platform. *International Journal Of Mobile Computing And Multimedia Communications*. **10**, 32-48 (2019)
[17] Fasold, F. & Redlich, D. Foul or no foul? Effects of permitted fouls on the defence performance in team handball. *Journal Of Human Kinetics*. **63**, 53-59 (2018)
[18] Cortina, D., Miguel-Gómez, Ortega, E. & Others Effect of pitch size on technical-tactical actions of the goalkeeper in small-sided games. *Journal Of Human Kinetics*. **62**, 157-166 (2018)
[19] McDaniel R, T. Game design tactics for teaching technical communication in online courses. *Journal Of Technical Writing And Communication*. **51**, 70-92 (2021)
[20] Cantabella, M., Martínez-Espaa, R., Ayuso, B. & Others Analysis of student behavior in learning management systems through a Big Data framework. *Future Generation Computer Systems*. **90**, 262-272 (2019)
[21] Yang, D. .Application of Data Mining Technology in the Subject Tactical Teaching of Badminton. *Nternational Journal Of Emerging Technologies In Learning (iJET)*. **13**, 30-42 (2018)
[22] Rachmatika, R. & Harefa, K. Analysis of determination of strategy promotion using apriori algorithm. *Journal Of Physics: Conference Series*. **1477**, 22032-02203 (2020)
[23] Jhang, K., Chang, M., Lo, T. & Others Using the apriori algorithm to classify the care needs of patients with different types of dementia. *Patient Preference And Adherence*. **13**, 1899-1912 (2019)
[24] Care, E., Griffin, P. & Wilson, M. [Educational assessment in an information age] Assessment and teaching of 21st century skills || shifts in the assessment of problem solving. 2018. (0)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 17, 2023
Accepted: Nov 6, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 395–406, DOI 10.12694/scpe.v25i1.2305

CONSTRUCTION AND APPLICATION OF PHYSICAL EDUCATION CLASSROOM TEACHING MODEL INTEGRATING MOOC AND FLIPPED CLASSROOM

ZHENG ZHANG*

Abstract. In terms of sports itself, mastering certain skills is a reliable way to avoid injury and futile training, and the teaching process needs to focus on imparting skills. Therefore, in order to optimize the teaching mode of physical education courses, this study explores the current improvement achievements and draws on the advancement of curriculum reform. The study proposes a recommendation method based on knowledge points and student characteristics by studying the characteristics and importance of sports techniques, in order to achieve high-quality customized content recommendations. Through validation experiments, the effectiveness of this recommendation method and optimization path is demonstrated. This study adopts an empirical research design to explore the impact of recommendation methods based on knowledge points and student characteristics on the teaching effectiveness of physical education courses. The effectiveness of this recommendation method is evaluated by collecting and analyzing students' learning data. Statistical analysis methods are used to analyze experimental data, including calculating the percentage of improvement in students' technical movements under different teaching modes. The research results indicated that recommendation methods based on knowledge points and student characteristics were of great significance for optimizing the teaching mode of physical education courses. The improvement of subjects' technical movements such as hitting and swinging in the new teaching mode was significantly better than that in the traditional teaching mode, with an increase of 10-20 technical movements per unit time, with a growth rate of over 20%. Compared with traditional teaching models, this teaching model has been subjectively recognized by students and has positive reference significance for the current physical education teaching model.

Key words: MOOC; Flipped classroom; Physical education; Tennis; Recommendation model

1. Introduction. Physical exercise requires a certain level of skill preparation. Taking tennis as an example, the swing, grip, and power of the racket all affect the effectiveness of hitting the ball [1]. However, students often find it difficult to master these skills on their own, so corresponding teaching guidance is needed to help correct the lack of skills [2]. The current physical education curriculum mostly adopts a demonstration imitation teaching mechanism, which makes it difficult for students to truly understand the essence of skills. Therefore, it is very necessary to carry out the reform of the physical education curriculum. Like other subjects, the idea of transferring the dominance of the education process from teachers to students coincides with Flipped classroom [3]. In the Flipped classroom, teachers help students, not just transfer information, but students are responsible for their own learning process and must control their own learning progress [4]. At the same time, modern educational technology, represented by MOOC, not only brings convenience to physical education teachers, but also broadens students' horizons. Therefore, combining physical education with modern educational technology to explore a new classroom model that is closer to the laws of education [5]. The research will explore the feasibility of the new classroom model from the perspective of integrating Flipped classroom and MOOC, and design a recommendation model with logical consistency in consideration of students' preferences to promote the reform of physical education. Our research will provide important references for the reform of physical education curriculum.

2. Related Works.

2.1. Application of MOOC.

MOOC is widely used in various disciplines because of their massive open course resources, and their development and application are promoted in both directions by the real needs. Deng R et al. combed through the progress of MOOCs to propose optimization paths and analyzed these MOOCs through Biggs' 3P teaching and learning model framework, and they found that users mostly focused on the improvement of teaching and learning content, with little attention to learners and the interaction between the two, but rarely on the learners and the interaction between them [6]. Deng R's team then conducted

*School of Physical Education, Xinxiang University, Xinxiang, 453003, China (zheng_zhang11@163.com)

a multidimensional study of the learner profile, categorizing them into behavioral, cognitive, affective, and social engagement components of MOOC learner evaluation, and the results showed that there were significant differences in the learners' factors and learning outcomes, which could help in subsequent instructional interventions [7]. Thus, a well-established English online teaching environment was constructed with knowledge partitioning in a multimedia, multi-resource context. In the modularized online environment, students were able to learn according to their characteristics and pace, and the results showed that this attempt could promote the efficiency of conversion of students' knowledge and stimulate their desire to learn, thus improving English learning [8].

2.2. Application of Flipped classroom. The flipped classroom as a constructivist teaching strategy fit is also receiving attention at all stages of education, Wei and other scholars introduced flipped classrooms into K-12 education, building on the existing progress of flipped classrooms for mathematics-based learning, where students previewed and took notes at home before the course began and formal classroom discussions about the notes took place. The experimental results showed that this attempt significantly improved the math performance of secondary school students, with a particularly significant improvement for students at the intermediate achievement level [20]. Jdaitawi explored the self-regulation and social impact of students in the flipped classroom, tested by comparing experimental embodiments, and the results of the analysis based on ANOVA analysis showed that students in the flipped classroom showed better self-regulation and social connectedness, which is an aid to self-directed learning [10].

2.3. Research on Physical Education Teaching. In the context of increasing technological development, physical education at home and abroad is also evolving with the times, and new evaluation methods and teaching techniques are bringing new changes to the physical education classroom. Zhang N, who specialized in technical optimization of basketball programs in physical education, regarded energy management as a key pedagogical objective in sports. To achieve this, Zhang developed a movement model based on Hidden Markov Models. This model enables the identification of approximate variables and unknown challenges, allowing for efficient energy utilization by athletes. [11]. De-kun et al. proposed a scheme to reverse the confusion in the traditional evaluation of physical education. The new algorithm constructed a BiLSTM model to rank and annotate the teaching tasks, then determined the weights of each indicator according to the intelligent optimal management factors to specify the quantitative evaluation scheme, and finally used a genetic algorithm for the optimal solution of the parameters. The results showed that the new evaluation method had a better performance than traditional methods and an advantage in terms of application time [12]. Forey's experimental team targeted the discussion of sports language to compensate for the lack of discussion of the course language, expanding the sports language capacity according to the linguistic theory and pedagogical methods provided by the system, and implementing the teaching with adequate preparation of the classroom language. The results of the analysis showed that explicit teaching of explicit language had a positive impact on students and teachers, and students' performance was improved, which proved that the language of instruction also had a considerable role in physical education [13]. Wang et al. scholars proposed improvements to the problem of too many factors in the evaluation of physical education, they believed that the problems of the evaluation system of physical education in colleges and universities should be analyzed before establishing an adaptive evaluation system. For this purpose, they introduced gray correlation analysis into the evaluation model, and the evaluation results under the concept of fuzzy mathematics also provided improvement strategies for this purpose [14].

2.4. Research Review. Scholars continue to move forward in their exploration of teaching models, during which new pedagogical concepts and technologies are used to address students' learning. This has led to more rational and efficient teaching and learning activities, but most of these reforms have been directed at classroom-based subjects and have not focused enough on the physical education curriculum. The goal of a physical education classroom is to cultivate students' athletic abilities and enhance their overall quality. However, in traditional physical education teaching models, students passively accept knowledge and find it difficult to fully improve their athletic abilities. The Flipped classroom introduces preview through information technology and multimedia tools, so that students can preview relevant skills through video, text, etc. before class. Students can apply this knowledge through practice in the classroom, and teachers can provide guidance

and correction to promote the cultivation of practical sports abilities. In addition, Flipped classroom also increases the participation and interest of students. The use of multimedia tools can help improve students' understanding and mastery of skills and actions, and increase the effectiveness and interest of learning. To sum up, it is important and necessary to adopt the Flipped classroom model in the physical education class. Students' athletic abilities are cultivated through preview and practice; By increasing participation and interest, their active participation and comprehensive development are promoted. Therefore, it is very necessary to use the Flipped classroom in PE class. In light of this, the use of MOOC and flipped classrooms for physical education teaching improvement is in line with the current sense of direction of physical education curriculum reform, and the study will aim to advance this work and provide some help to this course that needs to be practiced.

3. Construction of physical education teaching model and evaluation system integrating MOOC and flipped classroom.

3.1. Construction of physical education teaching model integrating MOOC and flipped classroom. The realization path of the study is to integrate MOOC and flipped classrooms, apply this integrated teaching model to physical education teaching, and measure the feasibility of this model through effect evaluation. Therefore, MOOC and flipped classrooms need to be discussed in depth to explore their feasibility and integration nodes from the characteristics of educational psychology and physical education courses themselves. Compared with traditional cognitive learning, the Flipped classroom is favored by teachers and students at the same time, and has obtained good evaluation on teaching skills, learning flexibility, the effectiveness of teaching aids, student participation, and working environment [15]. The acquisition of sports in various dimensions is related to age and learning practice. Boys aged 8 to 9 need some practice to achieve positive growth, that is, in the process of physical education teaching, attention should be paid to the proportion of teaching methods and training [16]. Behaviorist teaching theory focuses on the construction of external stimuli, and environmental stimuli and learning methods determine the learning effect. The MOOC, with its massive and high level of external stimuli, can bring some improvement to physical education teaching. From the teaching side, such as teachers, students' learning data can be the basis for teaching according to their needs, and in the teaching design change machine can detect students' learning preferences to summarize more reasonable learning patterns. In addition, data-based learning summaries can also be used as a basis for analyzing changes in performance, visualizing and quantifying the causes of performance changes, and providing help for subsequent adaptive teaching, while sports correction based on individual differences is also one of the goals of physical education.

With the rich teaching resources, the optimization of the teaching mode can adapt to this change, while the interaction between the two can create more value-added space for learning efficiency. Traditional physical education is arranged at the beginning of the classroom, which tends to increase the time cost of students' learning and thus reduces their motivation, while the short classroom time cannot correct students' wrong exercise styles. This leads to a significant reduction in the effectiveness of physical education. The flipped classroom is an improvement of the teaching model, changing the traditional teacher-based model to a student-based model. This model makes reasonable use of time outside of class, allowing students more time and ways to prepare for what they are learning, while formal class time is spent correcting based on preparation. The teacher's identity in this model is no longer as a teaching leader but as a supporter or corrector; content such as textbooks and teaching aids is no longer the main vehicle for imparting knowledge but a reference material [17]. For physical education content, taking tennis or table tennis as examples, classroom teaching is conducted according to the model of teacher demonstration and student imitation. However, the teacher's mastery of skills is based on sufficient training and reasonable cognitive methods and levels, and simple demonstration is difficult for students to master the essentials of each movement skill, which is easy to cause a rigid situation. The flipped classroom is a constructivist approach that emphasizes that students are the main body of learning and that they accomplish their learning tasks through constructing the meaning of knowledge in a specific situation.

The study will use a constructed teaching model for the experiment, which consists of regular warm-up exercises and formal teaching tasks, with a tennis program chosen to teach forehand and backhand strokes and serves in one cycle of study. The curriculum needs to be designed holistically, with warm-up exercises related to the formal content, for example, a full-body workout to meet the running required to hit the ball,

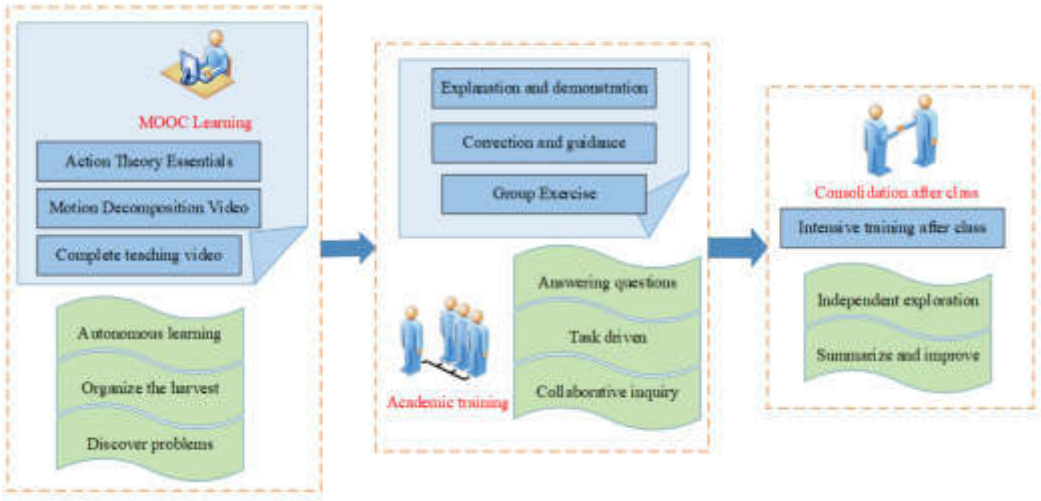


Fig. 3.1: Physical Education Teaching Mode Combining MOOC and Flipped Classroom

and upper-body warm-up as the focus during serve training [18]. At the same time, the course is conducted with the need to pay attention to the feedback of teaching information, implement the allocation of teaching percentages according to the actual situation, and advance teaching under dynamic principles. The specific teaching process is shown in Figure 3.1.

The teaching process as shown in Figure 3.1 consists of pre-course learning, classroom training, teacher instruction, and post-course consolidation. Before the start of the lesson, the teacher states the learning task and the learning points in the platform. In this part of the lesson, the teacher needs to propose the task arrangement according to the syllabus and provide the students with the learning path of each tennis movement, and then the teacher and students will agree to move to the next part of the lesson. The second half of the class will be summarized and evaluated to guide students to actively integrate what they have learned into their regular workouts. The whole teaching process will be recorded for the reference of other teachers and students on the platform.

3.2. MOOC-based knowledge sequence recommendation model. In the above-mentioned teaching process, teachers will recommend relevant courses for students to learn. Taking forehand hitting as an example, courses related to ground strokes, lead shots, and strokes will be recommended, but the sufficient number of related courses on the platform makes it difficult for students to choose, so a recommendation model based on knowledge sequences is needed to customize the recommended courses for students. The study will propose a recommendation model combining Heterogeneous Graph Attention Network (HAN) and Reinforcement Learning (RL), which can extract feature values from heterogeneous information and achieve continuous recommendations through reinforcement learning [19]. The structure of this model is shown in Figure 3.2.

The model structure as in Figure 3.2 contains three modules, which in logical order need to carry out heterogeneous network building based on user, course, and tennis knowledge after starting the operation, sampling of original paths by random wandering, and the sampling results will be used as input contents of the embedding module, after collecting students' knowledge preference feature vector by self-attention mechanism, and the contents collected from each path will be The content collected from each path will be expressed under the subsequent attention mechanism, and finally reach from user preference to knowledge point recommendation. The study will use the MOOC dataset to construct the information network, random wandering will sample all nodes for the training dataset, and the clicked knowledge points will be used as recommendation possibilities, assuming that there are A student users and a total of B sampled paths, random wandering sampling will get AxB paths, and the set of these paths will be set as N [20]. Since the node types in the network have

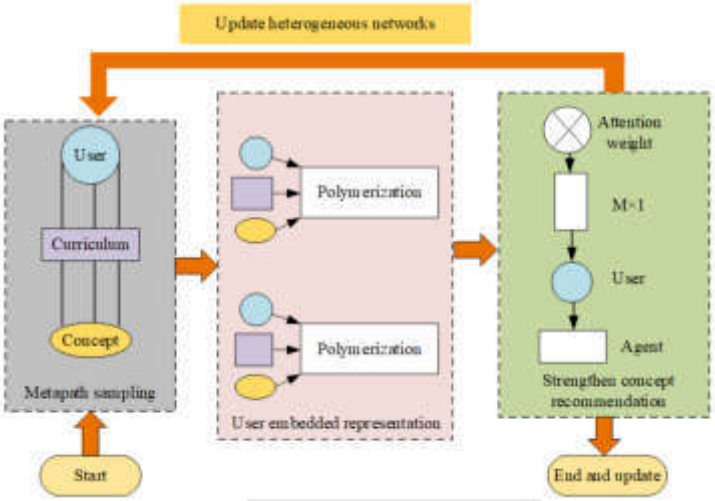


Fig. 3.2: Model structure of HAN-RL

heterogeneous characteristics, linear mapping is required for extraction. Suppose a node type is θ_i, and its type transfer matrix is M_θ, the mapping process is shown in equation 3.1.

$$E_i' = E_i.M_\theta \tag{3.1}$$

The E' and E_i of Equation 3.1 are the resultant feature values and the original feature values. Even for the same path, different nodes, their weights will be different, and then the self-attentive mechanism will be used to learn the weights of different types of nodes according to Equation 3.2.

$$\partial_{i,j}^{\phi} = \frac{exp[\sigma(a_\phi^T).[E_i'||E_j'])]}{\sum_{k\in B_i^*} exp[\sigma(a_\phi^T).[E_i'||E_j'])]} \tag{3.2}$$

The a_ϕ^T in Equation 3.2 represents the attention vector of the teacher user under the ϕ path, σ is the activation function, E_i' and E_j' represent the mapped feature vector and the mapped vector of the neighboring points after mapping, $||$ is the matrix splicing, k is the knowledge point, and $\partial_{i,j}^{\phi}$ is the weight coefficient of the two nodes i and j, whose magnitude depends on the feature because the neighboring points are different in the case of matrix splicing. i The meta-path embedding expression will be fused with the neighboring point weights as shown in equation 3.3

$$u_i^{\phi} = \sigma(\sum_{j\in N_i^{\phi}})\partial_{ij}^{\phi}.E_j' \tag{3.3}$$

In equation 3.3, u_i^{ϕ} is the meta-path embedding expression of i, according to this logic each node will be fused with its neighbors for embedding expression, and since the attention weights are generated through the unit path, it can learn all the feature information under that path. However, heterogeneous graphs have scale-free properties, so a multi-headed attention mechanism will be used for network training [21]. The node attention is replicated N times through the feature mapping and the final reflected value of the embedding expression vector as a user node is shown in equation 3.4.

$$U_i^{\phi} = \mathop{||}_{k=1}^{K} \sum_{j\in N_i^{\phi}})\partial_{ij}^{\phi}.E_j' \tag{3.4}$$

Equation 3.4 can be expressed by embedding the individual meta-path nodes of in the meta-path set $\phi_1, \phi_1, \ldots,$ ϕ_M as $U_0, U_1, \ldots, U_M$. To get the embedding expressions of different students, for the M expression vector, the corresponding weight calculation can be generalized. The attention of the path layer can learn the deep semantic information of different types of heterogeneous data, and here the weights of different paths will be learned by a single layer. Let the interlayer vector be q, then the node embedding expressions under a single path can go through a nonlinear mapping and the interlayer vector does an inner product with the result, and the normalization of the weight result is as in equation 3.5.

$$\omega_{\phi i} = \frac{1}{|V|} \sum_{i \epsilon V} q^T . tan(W.u_i^{\phi}) + b \tag{3.5}$$

The V in equation 3.5 is based on the recommended videos in the MOOC heterogeneous network, and W and b are the learnable parameters and bias values. The weight coefficients are shared among different meta-path and path layer attention implementations as in Equation 3.6 using the softmax function to normalize the overall weights.

$$\beta\phi_i = \frac{\omega\phi_i}{\sum_{i=1}^{p} exp(\omega_{\phi i})} \tag{3.6}$$

The $\beta\phi_i$ in Equation 3.6 is the normalized meta-path feature expression vector and p is the specific meta-path node. It can be seen that the results are positively correlated with the value of the meta-path. The embedding expression of the weight coefficients will show different results depending on the path, so the embedding expression of the end-user area is as in Equation 3.7.

$$U = \sum_{i=1}^{p} \beta_{\phi i} U_{\phi i} \tag{3.7}$$

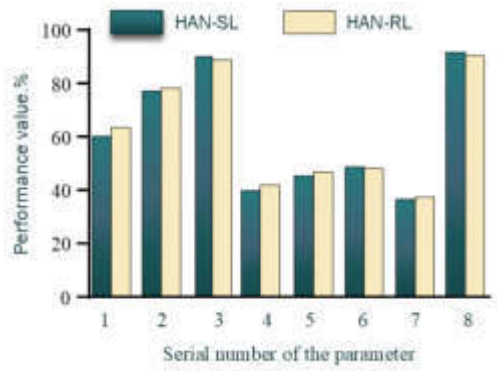
The conventional recommendation model is to establish a prediction mechanism to predict the error between the result and the true value by a loss function and to achieve the goal by reducing the error. However, this mechanism does not take into account that students' interests change over time and that the knowledge points that are extrapolated from the recommended content have a certain probability of becoming new interests of students, so the long-term interests of students need to be considered. The reinforcement approach of the study is to create an expectation reward mechanism as in equation 3.8.

$$\tau_{RL}(\theta) = E_{\pi ct|u} \sum_{t=1}^{T} r_t(ct|u) \tag{3.8}$$

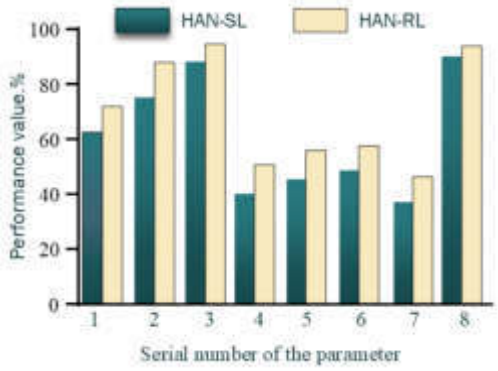
The $\pi_\theta(c|u)$ in Equation 3.8 is the recommended strategy based on the knowledge of the current user uc_t, r is the even reward, and E is the expectation function. The state of the embedded expression as reinforcement learning under the optimal recommendation strategy enables the learning process to be applied to the system, while the new state is updated by modifying the heterogeneous network. To make the model more efficient, the study will be optimized using entropy normalization as in equation 3.9.

$$H(\pi_\theta(c|u)) = -\sum_{i=1}^{T} \sum_{ct_\epsilon C} log\left[\pi_\theta(ct|u_t).\pi_\theta(ct|ut)\right] \tag{3.9}$$

At this point, the model is constructed and optimized as described above. After considering and optimizing all aspects of the three modules, the model is able to achieve recommendations based on a sequence of knowledge points. This recommendation model allows students to extract their preferences over time as they learn each knowledge point of the tennis course, avoiding that the recommended content for subsequent MOOC courses is always related to the initial data sample.



((a)) Overall knowledge recommendation performance



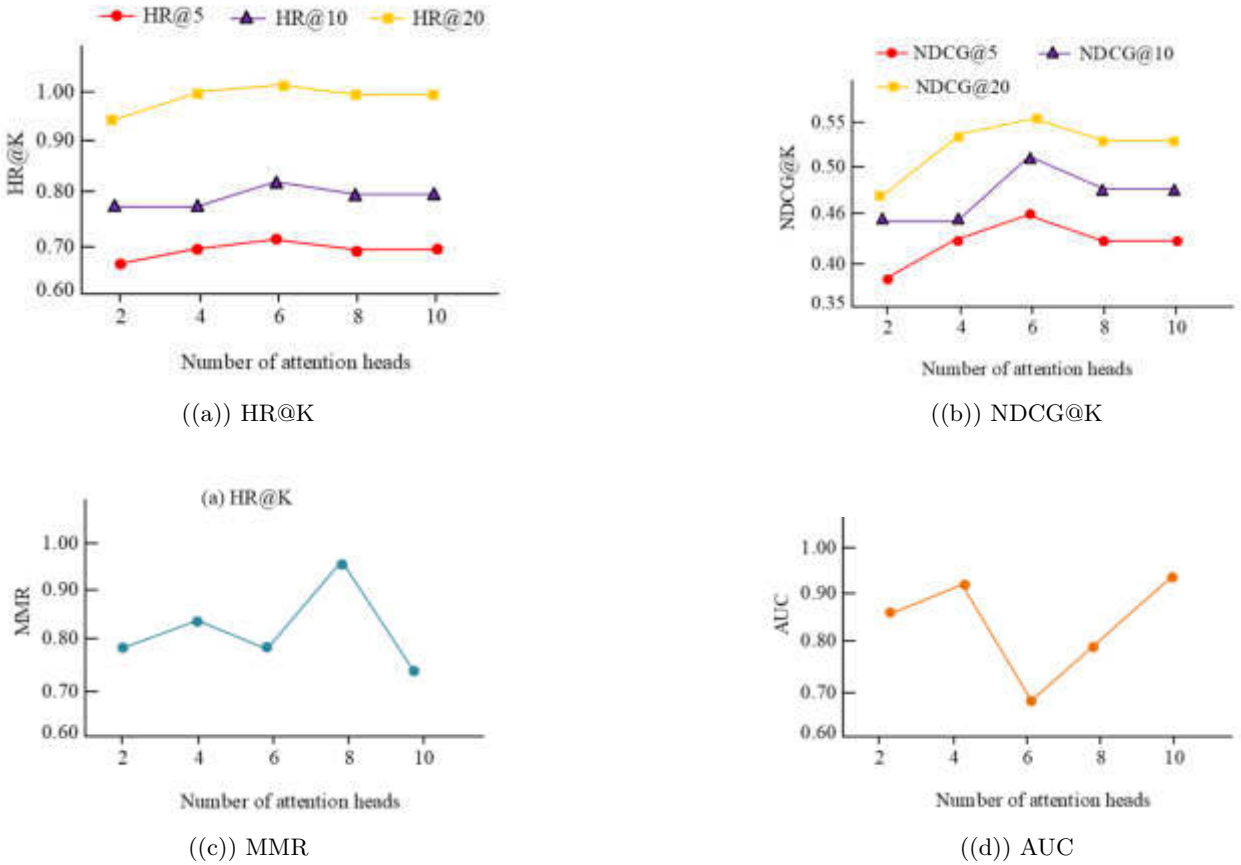
((b)) Monomer knowledge recommendation performance

Fig. 4.2: Performance Comparison between Single Knowledge Point and Complete Set

4. Application of physical education teaching mode integrating MOOC and flipped classroom.. Before the start of the formal class, the constructed recommendation model will be simulated and trained to ensure the effectiveness of the recommendation model and its practical value for tennis lessons. The data on tennis teaching on the MOOC platform from September 1, 2018, to September 1, 2022, will be used here as the data set, of which 80% will be used as the training set and 20% as the reference set. Firstly, the accuracy of the constructed model is tested, and the constructed HAN-RL model and the group learning attention network model HAN-SL are added to the overall recommendation and single knowledge point recommendation training respectively, and the evaluation metrics are divided into four categories, which are: hit ratio of rank (Hit Ratio, HR), Normalized Discounted Cumulative Gain (NDCG) Cumulative Gain (NDCG), Means Reciprocal Rank (MRR), and AUC (Area Under Curve, AUC). The specific test parameters 1-8 represent: HR@5, HR@10, HR@20, NDCG@5, NDCG@10, NDCG@20, MRR, and AUC, respectively. the results are shown in Figure 4.2.

As shown in the test results in Figure 4.2, the data set is more concentrated because the course is limited to tennis lessons. The set cut and hyperparameter verification show that both group learning and reinforcement learning performance can reach a certain level, and the parameters of both models have their own strengths in overall learning. However, the reinforcement learning constructed by the study is better in single knowledge point learning and fully satisfies the accuracy requirements of the recommended model. The optimal parameter settings of the model are explored here to better recommend instructional videos for students on MOOC platforms. The recommendation hit rate of the model is examined, where the number of heads in the attention mechanism affects the students' embedded expressions. Therefore, the number of heads is set separately to obtain the relationship with the parameter growth, and the above parameters are introduced into the performance evaluation. The results are shown in Figure 4.4.

From the results in Figure 4.4, the growth of the HR evaluation parameter does not change much under different numbers of attention heads, which indicates that the correlation between the hit rate of the rank and the number of attention heads is not significant. The change of NDCG is highly correlated with the number of attention heads, and the overall trend is characterized by increasing and then decreasing, with a peak at the number of attention heads of 6, which means that the cumulative gain will be the highest and the recommended effect of the model will be the highest under this condition. The trend of MMR is similar to that of NDCG, and the best effect is reached when the number of attention heads is 8. The AUC evaluation parameter is decreasing and then increasing, and the lowest parameter value of the model is reached when the number of heads is 6, and 10 is the peak. In summary, each parameter generally reaches the optimal value when the number of heads is 6. Since increasing the number of heads leads to a decrease in the convergence rate, the optimal value is



((a)) HR@K

((b)) NDCG@K

((c)) MMR

((d)) AUC

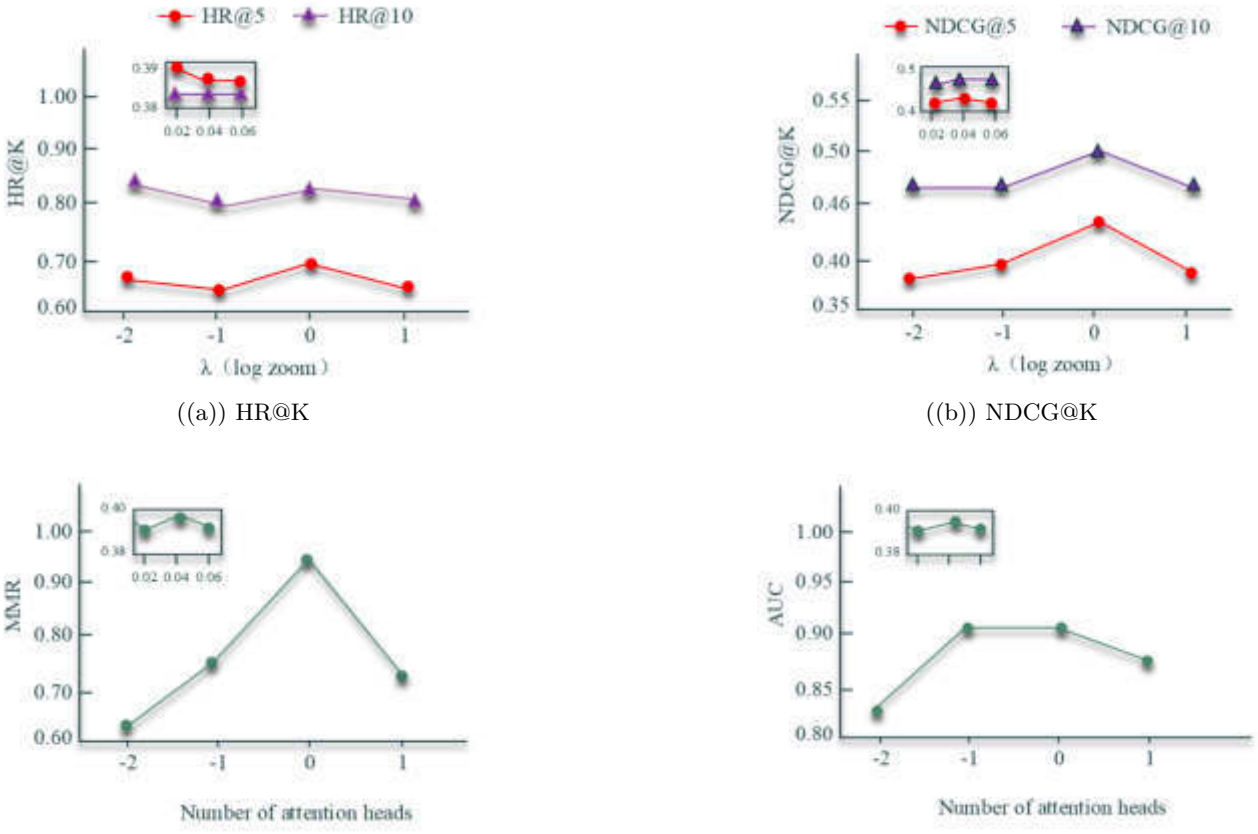
Fig. 4.4: Note the relationship between the number of headers and the parameters

chosen to be 6. The optimal value of the regularization coefficient will then be proposed for exploration, and the results are shown in Figure 4.6

From the results in Figure 4.6, although this coefficient has little effect on the weights, its variation is positively correlated with the performance metrics and is difficult to control, so this coefficient has some importance and therefore needs to be introduced in the recommendation model. After the MOOC-based recommendation model is prepared, the course experiment will be formally started. Two classes taught by Ms. M will be selected as the experimental group (n=51) and the control group (n=59), and two modes of instruction will be implemented, i.e., the experimental group will adopt the study-constructed instructional model and the control group will adopt the traditional instructional model for a semester-long course of 14 sessions, which will focus on tennis serve reception. With the exclusion of interference at the instructional level, the initial situation of the students will be understood to exclude the interference caused by their own situation. The results are shown in Table 4.1.

As per Table 4.1, it can be seen that the differences in students' scores on various tests before the start of the course were not significant ($p > 0.05$), which indicates that the interference of student factors can be eliminated. The course consisted of two parts, one of which was physical training and the other was skill practice. After half a semester, the physical fitness of the two classes changed, and the specific growth rates are shown in Table 4.2.

The comparison of the situations in Table 2 shows that there is little difference in the physical fitness of



((a)) HR@K

((b)) NDCG@K

((c)) MMR

((d)) AUC

Fig. 4.6: Relationship between λ coefficient and recommended performance

Table 4.1: Students' Mastery of Tennis Knowledge

Teaching Project	Experience Group	Control Group	t	P
Wave within 1 minute (p_1)	40.12 ± 3.21	39.33 ± 2.97	0.435	> 0.05
Shoot the ball in 1 minute (p_2)	15.46 ± 2.15	16.74 ± 1.63	1.321	> 0.05
Forehand stroke in 5 minutes (p_3)	19.21 ± 2.87	17.79 ± 2.73	0.637	> 0.05
Backhand within 5 minutes (p_4)	15.14 ± 2.51	14.23 ± 2.55	1.056	> 0.05
Serve within 3 minutes (p_5)	11.51 ± 1.85	13.16 ± 1.54	0.867	> 0.05
Rebound ball in 3 minutes (p_6)	18.25 ± 2.41	17.53 ± 2.07	0.513	> 0.05

the students before and after the course, which both exclude the interference caused by physical fitness factors to learning. It also shows that physical fitness training is used as a basic class for some time to prepare for the follow-up, and it is not an evaluation subject in itself. Subsequently, the two classes were tested separately at midterm, and the assessment results of each item were reflected as shown in Figure 4.7.

The midterm test scores are shown in Figure 4.7, and the scores of each item have improved compared to the beginning of the course. Comparing the performance of each item in both groups, it can be seen that the experimental group outperformed the control group in all items, and the difference was significant ($p < 0.05$). In

Table 4.2: Physical Fitness Changes of the Two Groups Before and After the Experiment

Teaching Project	Time of Test	Experience Group	Control Group	t	P
50-meter dash	Before the Course	7.49 ± 0.517	7.52 ± 0.539	-0.169	> 0.05
	Interim Test	7.33 ± 0.621	7.42 ± 0.544	-0.413	< 0.05
1000 meter run	Before the Course	253.51 ± 27.517	257.84 ± 29.972	-0.479	> 0.05
	Interim Test	242.33 ± 24.173	258.84 ± 30.251	0.431	> 0.05
Pull up	Before the Course	4.97 ± 3.703	4.58 ± 2.613	-0.482	> 0.05
	Interim Test	6.12 ± 3.241	5.36 ± 3.695	0.421	> 0.05
Turnaround run	Before the Course	18.72 ± 2.251	19.33 ± 2.873	-0.162	> 0.05
	Interim Test	18.07 ± 3.617	18.23 ± 2.34	-0.03	> 0.05

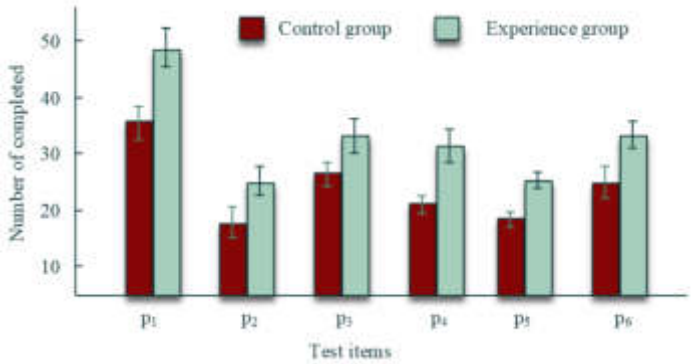


Fig. 4.7: Comparison of interim test results

tennis, the use of skills is closely related to the success of the project, and it is difficult to replace the instruction of skills with repetitive training alone, so the course recommendation based on the MOOC recommendation model will capture the characteristics of the students and make the effect twice as effective through the teaching of skills. After the midterm, the instructor will provide personalized instruction based on the problems that arise and determine the follow-up teaching plan, and the final results after intensive teaching are shown in Figure 4.8.

According to the results in Figure 4.8, it can be seen that the final scores of the experimental group were still better than those of the control group. Compared with the midterm, the scores of both groups have improved to a certain extent; the growth curve shows that the teaching model proposed in the study leads to consistently better growth in all test scores than the traditional teaching model, which indicates that the objective effect of MOOC integrated with flipped classroom meets the expectations. At this point, the students' subjective situation was used to understand the effect of the model, and their opinions were collected through the questionnaire method, and the results are shown in Figure 8, where the acceptance levels from 0-4 indicate: dislike, general attitude, basic acceptance, and very satisfied, respectively.

The results in Figure 4.9 show that the majority of students with acceptance levels of Basic Acceptance and Very Satisfied with the model, and the distribution and transformation of each grade level are similar, which indicates that the model has subjectively achieved the effect of satisfying students.

5. Conclusion. With the advancement of educational theory and technology reform, MOOC and flipped classrooms have long been richly explored experiences. The study uses these experiences in physical education to make the teaching model conform to the laws of the physical education curriculum. Since physical education courses require a lot of exercises to form muscle memory, proper skill instruction is the key to skill acquisition,

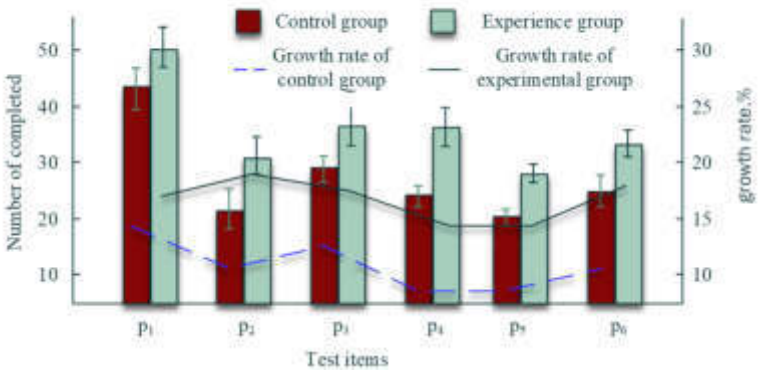


Fig. 4.8: Final result and change rate

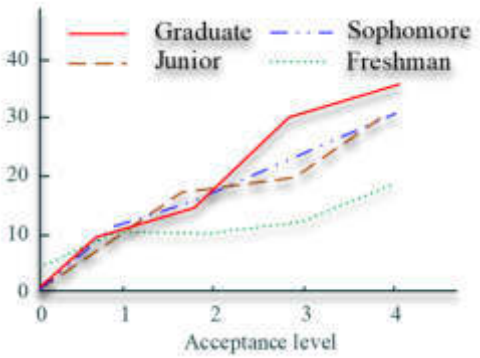


Fig. 4.9: Satisfaction of students of different grades with the new teaching model

and skill acquisition is closely related to individual students, so establishing a recommendation algorithm based on users and physical education courses is one of the focuses of the study. A recommendation method based on the sequence of knowledge points is formed through the attention mechanism and reinforcement learning, which can overcome the situation that students' interest points change due to the course progress, and the test proves that the parameters of this recommendation method can meet the automated recommendation of physical education courses. The new teaching model allowed students to improve their performance in a range of tennis training items, such as swing, by 20-30%, and each element showed a more significant improvement compared to the traditional teaching method ($p < 0.05$). Subjective findings showed that the majority of students in all grades in this elective physical education course had positive attitudes toward the instructional model. In the follow-up work, the study will advance teaching improvements in other skill-based sports courses to enrich and optimize the integration of MOOC and flipped classrooms.

REFERENCES

[1] Hewitt, M., Pill, S. & McDonald, R. Informing Game Sense Pedagogy with a Constraints-Led Perspective for Teaching Tennis in School. Ágora para la Educación Física y el Deporte. (2018)

[2] Zhao, D. & Kang, H. Practice exploration of the flipping classroom in a table tennis club in the informatization age. *Procedia Computer Science*. **166** pp. 175-179 (2020)

[3] Zhu, M., Sari, A. & Lee, M. A comprehensive systematic review of MOOC research: research techniques, topics, and trends

from 2009 to 2019. *Educational Technology Research And Development*. **68**, 1685-1710 (2020)
[4] Akçayır, G. & Classroom:, A. review of its advantages and challenges[J]. *Computers & Education*. **126** pp. 334-345 (2018)
[5] Fathi, J. & Rahimi, M. Examining the impact of flipped classroom on writing complexity, accuracy, and fluency: a case of EFL students. *Computer Assisted Language Learning*. **35**, 1668-1706 (2022)
[6] Deng, R., Benckendorff, P. & Gannaway, D. Progress and new directions for teaching and learning in MOOCs. *Computers & Education*. **129** pp. 48-60 (2019)
[7] Deng, R., Benckendorff, P. & Gannaway, D. Linking learner factors, teaching context, and engagement patterns with MOOC learning outcomes. *Journal Of Computer Assisted Learning*. **36**, 688-708 (2020)
[8] Zhang, N. Development and Application of an English Network Teaching System Based on MOOC. *International Journal Of Emerging Technologies In Learning*. **13**, 149-160 (2018)
[9] Mohammed, H. & Daham, H. Analytic hierarchy process for evaluating flipped classroom learning. *Computers, Materials & Continua*. **66**, 2229-2239 (2021)
[10] Wei, X., Cheng, I., Chen, N. & Others (2020) Effect of the flipped classroom on the mathematics performance of middle school students. *Educational Technology Research And Development*. **68**, 1461-1484 (0)
[11] Jdaitawi, M. The effect of flipped classroom strategy on students learning outcomes. *International Journal Of Instruction*. **12**, 665-680 (2019)
[12] Zhang, N., Han, Y., Crespo, R. & Others (2020) Physical education teaching for saving energy in basketball sports athletics using Hidden Markov and Motion Model. *Computational Intelligence*. **37**, 1125-1140 (0)
[13] De-kun, J. & Memon, F. Design of mobile intelligent evaluation algorithm in physical education teaching. *Mobile Networks And Applications*. **27**, 527-534 (2022)
[14] Forey, G. & Cheung, L. The benefits of explicit teaching of language for curriculum learning in the physical education classroom. *English For Specific Purposes*. **54** pp. 91-109 (2019)
[15] Ivashchenko, O., Iermakov, S. & Khudolii, O. Modeling: ratio between means of teaching and motor training in junior school physical education classe. *Pedagogy Of Physical Culture And Sports*. **25**, 194-201 (2021)
[16] Wang, Y., Sun, C. & Guo, Y. A multi-attribute fuzzy evaluation model for the teaching quality of physical education in colleges and its implementation strategies. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 159-172 (2021)
[17] Polat, H. & Karabatak, S. Effect of flipped classroom model on academic achievement, academic satisfaction and general belongingness. *Learning Environments Research*. **25**, 159-182 (2022)
[18] Chiu, Y., Clemente, F., Bezerra, P. & Others (2022) Day-to-day Variation of the Heart Rate, Heart Rate Variability, and Energy Expenditure during FIFA 11+ and Dynamic Warm-up Exercises. *Journal Of Human Kinetics*. **81**, 73-84 (0)
[19] Wang, Z., Liu, C. & Gombolay, M. Heterogeneous graph attention networks for scalable multi-robot scheduling with temporospatial constraints. *Autonomous Robots*. **46**, 249-268 (2022)
[20] Mei, G., Pan, L. & Liu, S. Heterogeneous graph embedding by aggregating meta-path and meta-structure through attention mechanism. *Neurocomputing*. **468** pp. 276-285 (2022)
[21] Brunke, L., Greeff, M., Hall, A. & Others (2022) Safe learning in robotics: from learning-based control to safe reinforcement learning. *Annual Review Of Control, Robotics, And Autonomous Systems*. **5** pp. 411-444 (0)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 17, 2023
Accepted: Nov 6, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 407–421, DOI 10.12694/scpe.v25i1.2306

RESEARCH ON THE EVALUATION MODEL OF STUDENTS' FOREIGN LANGUAGE LEARNING SITUATION BASED ON ORIENTED ONLINE TEACHING COLLABORATION PLATFORM

FEI HE*

Abstract. "Internet + Education" makes online teaching gradually penetrate the education industry, and makes the industry enter a great revolution based on information technology. The traditional student learning evaluation system cannot satisfy the actual demand of current learning evaluation. This paper constructs an evaluation model for the foreign language learning of online students. Firstly, the DBSCAN algorithm with distance optimization is used to conduct cluster analysis on the description indicators of student behavior, and the student groups with different behavior characteristics are obtained. Then the ANOVA F-test was used to extract the features of different student groups. Finally, a novel N-Adaboost algorithm based on multiple classifiers is proposed and a model is constructed to evaluate students' foreign language learning. The experimental results show that the accuracy of the evaluation model is 74.02% in the pass and fail groups and 73.74% in the excellent and non-excellent groups. Students' listening, speaking, and reading abilities are in a state of upward development overall through the online teaching collaboration platform, but their writing ability is obviously declining. There is a great improvement in foreign language vocabulary. This study provides a new perspective of thinking for the improvement of the quality of school teaching management, the analysis of students' behavior, and the evaluation of learning situations, and provides a new solution for the problem of students' learning situations in modern information teaching.

Key words: Online Education, Foreign Language Learning, Evaluation Model, N-Adaboost, DBSCAN

1. Introduction. With the popularization of Internet technology and the continuous development of cloud computing, big data, artificial intelligence and other technologies, Internet + education is gradually changing the traditional teaching mode. Now, students can study anytime and anywhere through the online learning platform, no longer limited by time and space [1, 2]. The online education platform not only breaks through the time and space restrictions, and provides students with rich learning resources, but also provides students with personalized learning recommendation and evaluation [3] through the intelligent learning system. At present, the online teaching model has attracted the attention of many educators at home and abroad, and many universities have deployed and implemented their own learning platforms. Some online platform courses and open courses have reached tens of millions of users. As an important subject of personal development and social needs, the teaching collaboration platform greatly promotes the efficiency and accessibility of students' foreign language learning. However, online foreign language learning also has the problem of high user dropout rate. Because learners have greater freedom and strong subjectivity, the lack of supervision and intervention when learning risk occurs [4]. Therefore, it is of great significance to establish a model that can evaluate students 'foreign language learning level to improve the teaching quality and students' learning effect. However, the traditional evaluation method often can not adapt to the characteristics of online teaching, so the research combines the density-based clustering algorithm (Density-Based Spatial Clustering of Applications with Noise DBSCAN) algorithm based on distance optimization with the multi-classifier based Adaboost algorithm (N-Adaboost) to build the foreign language learning evaluation model for students on the online teaching collaboration platform. It is hoped that through the establishment of such an evaluation model, students 'learning behavior and learning level can be better understood, provide teachers with more accurate teaching suggestions, and promote students' learning and development.

2. Related Work. Nowadays, the digital construction of colleges and universities continues to advance, the process of education reform continues to accelerate, and student learning evaluation is very important.

*School of Foreign Studies, Henan Polytechnic University, Jiaozuo, 454003, China (feihefhfh@outlook.com)

In the face of the problem of using digital technology to improve the efficiency of English learning and the effectiveness of teaching evaluation, researchers such as Susanty L use search engines to obtain data and use coding, interpretation, and other means to obtain results. Through data analysis and discussion, it is found that the use of digital technology in English teaching can significantly improve students' classroom participation; the introduction of this technology in teaching evaluation has effectively improved evaluation efficiency and accuracy [5].

Liu H's research team found that there are some problems in the current teaching evaluation system. To conduct a more comprehensive quality evaluation, an evaluation system and model that introduced the entropy weight method and gray clustering were proposed. Through example analysis, the results show that the quantitative indicators of the system are practical and innovative; the stability and accuracy of the constructed evaluation model are the best and have great practical application value [6].

Aiming at the problems of fuzzy evaluation index and imperfect system, scholars Li N constructed a fuzzy evaluation model of the analytic hierarchy process. Relevant experimental data show that the model integrates qualitative and quantitative indicators to conduct teaching evaluation from various aspects, which significantly improves the reliability of the evaluation results. At the same time, the model can effectively deal with fuzzy indicators, which greatly promotes the improvement of English teaching quality [7].

Zhang Y's researcher found that when teaching evaluation in online teaching, there are many fuzzy indicators, and the evaluation effect is greatly reduced. The key indicators that restrict the evaluation model are studied and analyzed, and a multi-attribute fuzzy evaluation model is constructed. The simulation experiment results show that the model can accurately evaluate the effect of online teaching, effectively quantify the quality of teaching in all aspects, and has a good development prospect [8].

Facing some of the problems existing in English teaching, Tran TQT and other teams expounded on the factors that affect the effect of English teaching from many aspects. Through questionnaire survey and data analysis, relevant data shows that students are not motivated enough in class, and interesting teaching strategies can improve their learning enthusiasm. At the same time, students' habits and interests will affect the teaching effect. Colleges and universities should formulate corresponding plans to improve students' classroom participation and improve teaching quality [9].

Student learning evaluation is an important means to improve teaching quality and reform teaching modes in colleges and universities. Wang and other scholars found that several factors such as teachers, students, teaching methods and teaching environment can significantly affect the teaching effect. The research starts with teaching methods, reforms the existing teaching strategies, and passes relevant experimental tests. The results show that after implementing the reformed teaching strategy, students are more motivated, the classroom atmosphere is more active, and their professional ability has been significantly improved. The results provide a reference for colleges and universities to formulate teaching effect evaluation indicators [10].

Li and other research teams used the attributes of evaluation indicators to construct an evaluation index system with different attribute dimensions. The system is used in the evaluation of English teaching ability, and the results show that the system has a simple structure, and can flexibly adjust the weight parameters to adapt to the evaluation needs of different grades, and objectively reflect the teaching level of colleges and universities from many aspects [11].

Researchers such as Sun have developed an English system that combines artificial intelligence and teaching assistance, which can mine potential connections between information and use decision tree technology to independently evaluate teaching effects. It can be seen from the actual application data that the system can accurately analyze the students' mastery of knowledge, help teachers provide the basis for improving teaching strategies, and improve the level of teacher education and the efficiency of students and students [12].

Facing the problems of slow speed and low accuracy of the existing English interpreting evaluation models, Lu et al. used principal component analysis to screen indicators, used a radial basis network evaluation model, and used genetic algorithms to optimize parameters. The simulation experiment data shows that the selected indicators are representative, the evaluation efficiency and accuracy of the constructed evaluation model are significantly improved, and it has high real-time performance [13].

Research scholars such as Fang C found that online evaluation can conduct an overall analysis of teaching activities and provide a basis for educational improvement, and proposed an evaluation model based on support

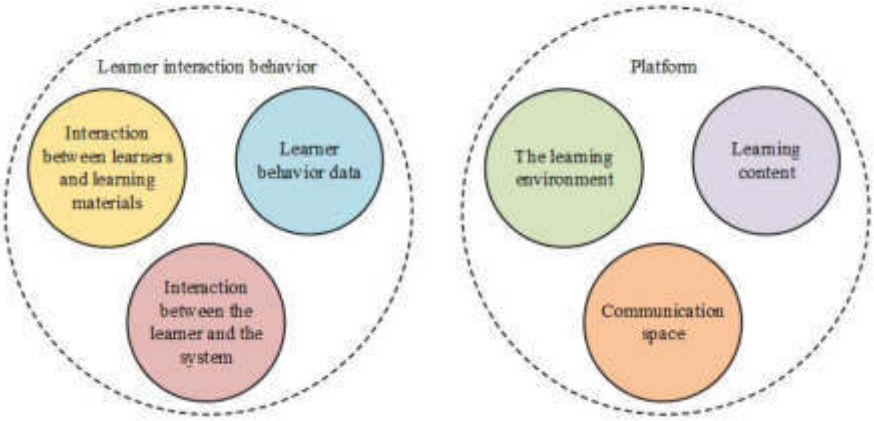


Fig. 3.1: Interaction between platform and learner in online teaching

vector machines. The experimental data show that the improved index dimensionality reduction method can improve the evaluation accuracy and reduce the interference factors. The evaluation effect of this model is better than that of the comparison method, and it has certain value in practice [14].

By expounding the achievements of domestic and foreign researchers, it is found that student learning evaluation plays an important role in education reform. Many scholars have put forward their own improvement plans for different problems, but almost all of them are based on shallow level analysis of simple data and models. Therefore, an improved N-Adaboost model based on multiple classiclassiers is proposed in this study. Through the use of modern intelligent technology to give a number of data evaluation of student learning.

3. Construction of an evaluation model for students' foreign language learning for an online teaching collaboration platform. With the rapid development of science and technology and the popularity of the Internet, more and more people choose to learn foreign languages through online platforms. However, compared with traditional face-to-face teaching, online learning is challenging. To address this problem, the research chose to construct an evaluation model specifically for online students' foreign language learning, providing students with personalized learning advice and feedback, as well as the role of providing guidance and supervision for teachers. Through the evaluation model, online students can better self-learning and improve learning results.

3.1. DBSCAN algorithm based on distance optimization. The rapid development of Internet technology and the innovation and transformation of educational concepts and methods have formed a new teaching method — online teaching [15]. The learners of online teaching break the predicament that traditional teaching is limited by time and space. And backed by big data, data mining and other network technologies, online learning has very rich learning resources, which can meet the diverse and personalized needs of many learners [16, 17]. Online learning is the sum of the teaching activities carried out by online learners and related learning groups through the interaction of online learning platforms in order to complete specific learning tasks, as shown in Figure 3.1.

From Figure 3.1, in online teaching, the platform can recommend a large number and various forms of learning content for learners. It also gives learners an independent space for self-directed learning and a collaborative environment for learning with others. At the same time, it can also meet the fast and efficient communication. At the same time, during online learning, the learner will leave a lot of behavior data, such as the interaction between the learner and the learning materials, the interaction between the learners, and the interaction between the learner and the system. The online learning platform uses cloud computing technology to process the massive data it owns, and then analyzes the characteristics of learners' learning behavior through data mining technology. Through the analysis, the learning rules are obtained, so as to carry out targeted push

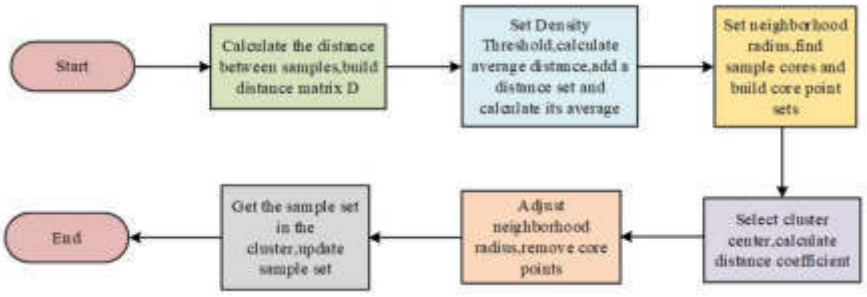


Fig. 3.2: Flow chart of DBSCAN algorithm based on distance optimization

services, and finally achieve the purpose of improving the learning efficiency and effect of learners. Faced with huge and complex online teaching students' academic behavior data, they need to be preprocessed. In 1996, Ester et al. proposed the DBSCAN algorithm. This algorithm differs from other principled clustering algorithms in that it determines the density of the dataset in space by neighborhood [18]. Another way to make the data set to achieve the clustering effect is to describe the density of the data point set in the space. The sample set is set as the neighborhood is for the hypersphere area with the center and the radius. In this area, all sample points constitute a sub-sample set, which satisfies equation 3.1.

$$N_\delta(x_j) = \{x_j \in A \,|\, \text{distance}(x_i, x_j) \le \delta\} \tag{3.1}$$

Improves some problems that may exist in the traditional DBSCAN algorithm, and proposes a distance-optimized DBSCAN algorithm. Set the sample set as $A = x_1, x_2, \ldots, x_n$, there is equation 3.2.

$$N(x_i) = \{x_j \in A \,|\, 0 < d(x_i, x_j) \ < \delta\} \tag{3.2}$$

For $x_i \in A, x_i$ the density $N(x_i)$ is the number of data points in x_i the δ neighborhood of. The distance coefficient of the core point is shown in equation 3.3.

$$\theta = \frac{N(x_j)}{N(x_i)} \tag{3.3}$$

In Equation 3.3, θ represents the distance coefficient. Therefore, the basic flow of the DBSCAN algorithm based on distance optimization is shown in Figure 3.2.

In Figure 3.2, the algorithm flow can be divided into nine steps. First, the distance between the sample points needs to be calculated to construct a matrix, as shown in equation 3.4.

$$D = \{D_{ij} \,|\, i, j \in R, i \ne j\} \tag{3.4}$$

In equation 3.4, D represents a matrix, i and j are two points, and R represents a sample data set. Then given the density threshold, calculate the average distance of the points with the nearest MinPts number, and put them into the distance set, and then find the overall average of the set. Then set the neighborhood radius to half the average distance. Find all core points among all sample points and put them in the core object set. Then select a core point for the cluster center to form a new cluster. Then calculate the distance coefficient between the sample point and the core point to adjust the neighborhood radius. Starting from the sample points in each neighborhood, recursively, using the same method, all the sample points that are density-reachable, and put them into clusters. Then the clusters that have been found are removed and the sample set is updated. Repeat the search and adjustment until all core points are traversed or removed. Finally, the result is obtained, as shown in equation 3.5.

$$C = C_1, C_2, \ldots, C_n \tag{3.5}$$

In equation 3.5, it C represents the whole cluster, C_1 , C_2, and C_n represent the sample points that meet the conditions. In 1987, the silhouette coefficient was proposed. It can evaluate different algorithms with the same original data. And for different parameter indicators of an algorithm, the clustering results can also be evaluated. Assuming that the clustering result has K clusters, and each cluster has I samples, then a sample x is shown in equation 3.6.

$$\begin{cases} a(x) = \frac{1}{i-1} \sum_i \text{distance}(x, x_i), & x \neq x_i \\ b(x) = \frac{1}{j} \sum_j \text{distance}(x, x_j), \end{cases} \tag{3.6}$$

In equation 3.6, within the same cluster, the average distance between the sample x and the rest of the samples x_i is represented by, and the average distance $a(x)$ from the nearest point is represented by. Then the silhouette coefficient of the sample x is shown in equation 3.7.

$$A(X) = \frac{b(x) - a(x)}{\max(a(x), b(x))} \tag{3.7}$$

In order to obtain the silhouette coefficient of the overall clustering results, $A(X)$ the average value, $A(X) \in [-1, 1]$.

3.2. Student behavior feature extraction and N-Adaboost model construction. This paper collects students 'school behavior data and network behavior data, and preprocesses the behavioral data, and builds a database of students' network behavior, network viscosity and life rules. The data in the database include students' network traffic data, network authentication data, canteen consumption data, library entry and exit data, dormitory access control data, the names of the courses, course attributes, credits, final examination scores, course semester and academic year of the course. At the same time, the study also desensistudent behavior logs, including electronic account desensitization, domain name desensitization, IP address desensitization, name desensitization, and student number desensitization. Desensitizing the five types of fields of student log information data can obtain an encrypted string with source data rules. This kind of string hides the user's privacy information, which ensures the data security, but also has certain recognition and readability. After desensitizing the school behavioral data, data preprocessing is required. Preprocessing includes data cleaning, data protocol, and data transformation. Study on the preprocessed data to construct a student "portraits" database. The study summarized the data into three aspects, namely, network behavior, network viscosity, and life regularity. Network behavior includes network behavior index, network viscosity includes network viscosity index, and life regularity includes canteen consumption data index and library learning index. Among them, the canteen consumption data and the library learning index jointly describe the offline behavior of students, and reflect the regularity of students' life in life. Therefore, this paper summarizes the two into the same item, namely the "regularity of life". According to the students "portrait" description index, the study using DBSCAN algorithm based on distance optimization of the student cluster, will be divided into different groups with different performance differences, to explore different academic performance, students daily network behavior and school behavior by distance optimization of DBSCAN algorithm after clustering students, can be intuitive analysis learning factors. From a macro perspective, the characteristic indicators with significant influence were found out, and the ANOVA F-test significance test was conducted on them. The F-test test value was used to characterize the influence degree of the influencing factors on the learning situation, and the characteristics of behavioral data were extracted to provide a data basis for the learning situation evaluation model. The influencing factors of students' learning are shown in Figure 3.3.

From Figure 3.3, the influencing factors of students' academic level can be mainly divided into three aspects: network behavior, network viscosity, and life regularity. Among them, network behaviors are divided into five types according to the types of access resources: video, knowledge, game, shopping, and social. In terms of network viscosity, after analysis, students' academic level is more affected by online time and online time. Considering the campus and social environment, the number of online days is not significantly different among different student groups, and the impact on the academic level is small. In terms of the regularity of life, it is believed that the consumption of breakfast can indicate whether students get up early, so it can show the regularity of students' life to a certain extent. ANOVA is a method to test the significance of the

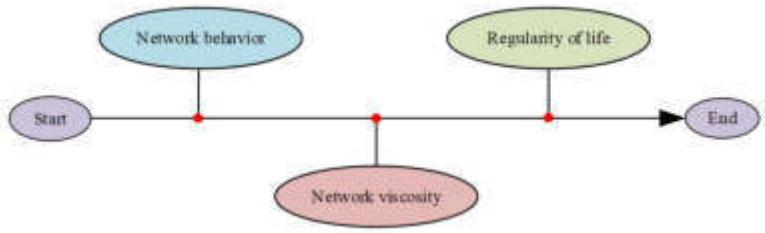


Fig. 3.3: Factors affecting students' academic level

difference between the means of two or more samples [19]. This method takes the F- distribution as the basis of the probability distribution and estimates the F value by the component and within-group mean square value calculated by the sum of squares and degrees of freedom. Then analyze the contribution of variation from different sources to the total variation, to determine the influence of different variables on the research results. Applied ANOVA F test to extract the characteristics of different student groups can understand the extent of different factors affect the evaluation of foreign language learning. In the F test, the study used the foreign language learning evaluation between different groups of students as the dependent variable, and the network behavior factors, network viscosity and life pattern as independent variables. Comparing the differences in variance between different student groups to determine which factors significantly influenced students' evaluation of foreign language learning. In ANOVA F-test, the sum of squares of the total deviation (Sum of Squares Total, SST) is the sum of squares of the error of the variable to the mean of the total sample, as shown in equation 3.8.

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2 \tag{3.8}$$

In equation 3.8, y_{ij} denotes the variable and $\overline{y}$ denotes the mean of the total sample. The Sum of Squares Regression (SSA) between groups represents the error sum of squares of the mean of each group to the total mean, as shown in equation 3.9.

$$SSA = \sum_{i=1}^{k} n_i \left(\bar{y}_l - \bar{y}\right)^2 \tag{3.9}$$

In equation 3.9, $\overline{y_i}$ represents the mean of each group. The Sum of Squares Error (SSE) within the group represents the sum of squares of the error between the sample data of each group and the mean of the group, as shown in equation 3.10.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_l)^2 \tag{3.10}$$

According to the content of the appeal, it can be known that the mean variance between groups MSA, the mean variance MSE within the group, and the F value, as shown in equation 3.11).

$$\begin{cases} MSA = \frac{SSA}{k-1} \\ MSE = \frac{SSE}{k(n-1)} \\ F = \frac{MSA}{MSE} \end{cases} \tag{3.11}$$

In 1996, Adaptive Boosting was proposed. The algorithm can automatically adjust the weight value according to the resulting feedback of the learner to adapt [20, 21]. It ensures that in the continuous iterative

process, the classifier can gradually focus on those samples that are difficult to classify, which improves the classification accuracy. However, because the classifier types are the same, there are certain limitations. Therefore, the research proposes a multi-classifier-based N-Adaboost model. Instead of using a single classifier as the base learner, the model integrates multiple classifier models to avoid the problem that homomorphic classifiers only perform well in one aspect, and underperformance in multiple problems. During the training process, the base learner is composed of multiple classifier models, each classifier model classifies the training samples, and the training results of the base learner are determined by the multiple classifier models. In each iteration, the training sample dataset successively passes through multiple classifier models to fit the model with the same weight. By integrating different classifier models, the model overcomes the classification limitations brought about by a single learner and makes the performance of the classifier complementary. The addition of N-Adaboost model can improve the universality of the online course evaluation model and solve the problem of insufficient performance in analyzing a large number of students and multiple online courses. Let the number of iterations be N, the training data set be D_i , and the weight distribution of the initialized training samples is shown in equation 3.12.

$$\begin{cases} W_1 = (w_{1,1}, w_{1,2}, \ldots, w_{1,i}) \\ w_{1,i} = \frac{1}{N}, \quad i = 1, 2, \ldots, N \end{cases} \tag{3.12}$$

Build a learning algorithm consisting of Γi classifiers $H(x)$, and then make classification evaluations on the training data. And count the classification results, and return the final classification results to the algorithm after processing. The learning algorithm is a model trained through the training data set, and input into the ensemble classifier model to obtain a weak classifier $G_n(x)$, as shown in equation 3.13.

$$G_n(x) = \Gamma(D_i, W_n, H(x)) \tag{3.13}$$

In equation 3.13, W_n represents the weight, where $n = 1, 2, , \ldots, n$. According to the classification error rate at this time, the weight in the strong classifier is calculated, as shown in equation 3.14.

$$\alpha_n = \frac{1}{2} \log \frac{1 - e_n}{e_n} \tag{3.14}$$

In equation 3.14, e_n represents the classification error rate. Then update the weight distribution of the training sample set, as shown in equation 3.15.

$$\begin{cases} w_{n+1,i} = \frac{w_{n,i}}{z_n} \exp\left(-\alpha_n y_i G_n\left(x_i\right)\right), \quad i = 1, 2, \ldots, N \\ z_n = \sum_{i=1}^{N} w_{n,i} \exp\left(-\alpha_n y_i G_n\left(x_i\right)\right) \end{cases} \tag{3.15}$$

In equation 3.15, Z_n represents the normalization factor that can make the probability distribution sum of the samples equal to 1. Repeat the previous steps N times to obtain the final classifier as shown in equation 3.16.

$$F(x) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_n G_n(x)\right) \tag{3.16}$$

When a model construction is completed, there needs to be a parameter to evaluate the relevant performance of the classification model, and the relevant parameters are called the model evaluation index. Different models have different tasks, and the corresponding ones can be evaluated with different indicators. The study predicted students' academic level mainly based on student behavior and belongs to a classification task. The performance evaluation indexes used in the study were accuracy, recall rate and F1 value. Accuracy refers to the percentage of the number of samples correctly classified by the classifier to the total number of samples, reflecting the situation that the classifier correctly identifies each sample. The calculation formula is shown in equation 3.17.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.17}$$

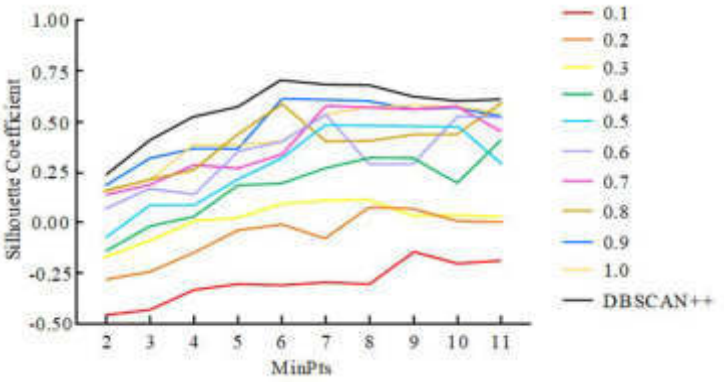


Fig. 4.1: Contour coefficient comparison diagram of traditional and optimized DBSCAN algorithm

In equation 3.17, TP represents true positive; TN represents true negative, FP and FN represent false positive and false negative. Accuracy refers to the number of true cases in the sample where the prediction result is positive, and the calculation formula is shown in equation 3.18.

$$Precision = \frac{TP}{TP + FP} \tag{3.18}$$

Recall refers to the percentage of positive predictions, with the formula shown in equation 3.19.

$$Recall = \frac{TP}{TP + FN} \tag{3.19}$$

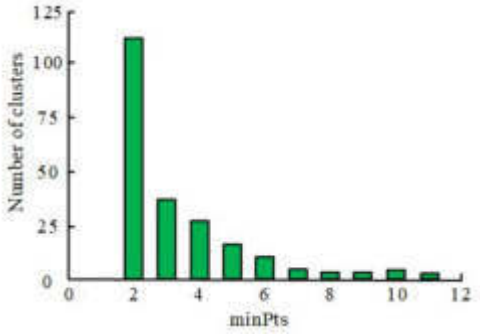
In order to better evaluate the performance of the classifier, the precision and the recall rate are called the measure, and the calculation formula is shown in equation 3.20.

$$F_\alpha = \frac{(1 + \alpha^2) \cdot \text{Precision} \cdot \text{Recall}}{\alpha^2 \cdot \text{Precision} \cdot \text{Recall}} \tag{3.20}$$

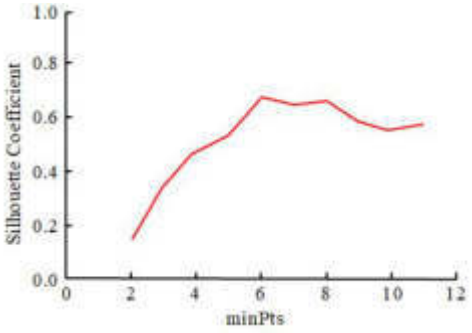
In equation 3.20, α is a non-negative real number. When α is 1, it is F1, which is the harmonic mean of precision and recall.

4. Analysis of the evaluation results of the evaluation model of students' foreign language learning. The data set of the study is the behavior data of students in a university in Henan province, which includes students 'network traffic data, network authentication data, campus "one-card" data and students' foreign language course test scores. To evaluate the clustering effect and determine the optimal number of clusters, the traditional DBSCAN algorithm and DBSCAN algorithm based on distance optimization were respectively used to cluster the data sets, and the corresponding contour coefficients of a different number of clusters were obtained, as shown in Figure 4.1.

According to the data in Figure 4.1, it can be seen that when δ is 0.1 and minPts is 2-11, the value of the silhouette coefficient is always below 0, indicating that the clustering effect is the worst at this time. When δ is 0.2 and 0.3, and minPts is 2-11, the value of the silhouette coefficient is basically below 0, but it can be around 0 when it is very few. When δ is in the range of 0.4-1.0, the value of the silhouette coefficient is constantly changing and can approach or even exceed 0.5 when the individual minPts is taken. Among them, when and when δ is 0.9 and minPts is 6, the contour coefficient exceeds 0.6, and the optimal parameter solution 0.661 is obtained at this time. In Figure 4.1, "DBSCAN + +" represents the distance optimized DBSCAN algorithm, the contour coefficient of the algorithm is always positive, the maximum value was 0.69, and the contour coefficient curve of the algorithm is always higher than that of the traditional DBSCAN algorithm. In conclusion, the results show that the clustering effect of the proposed modified DBSCAN algorithm is better.



((a)) The number of clusters under different values of minPts



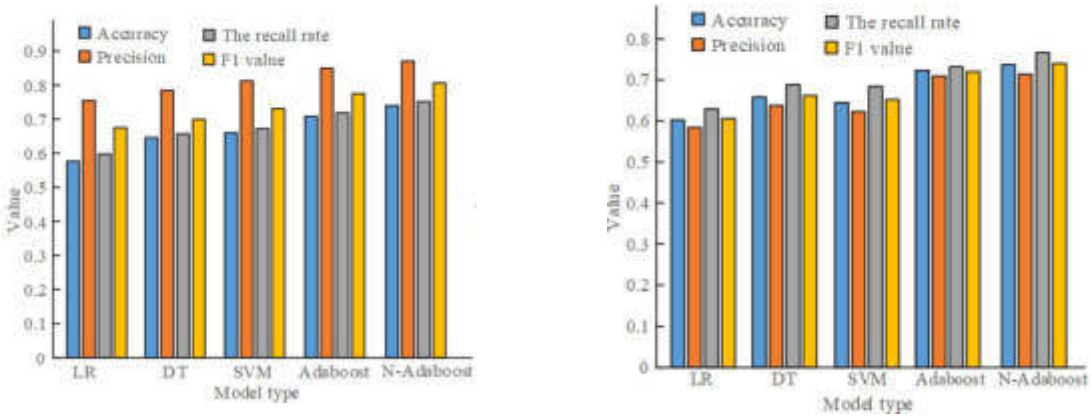
((b)) Contour coefficients at different values of minPts

Fig. 4.3: Value of DBSCAN algorithm after distance optimization under different minPts

Therefore, the distance-optimized DBSCAN algorithm is used to cluster the students' network behavior. The changes in the number of clusters under different minPts values are shown in Figure 4.3.

In Figure 4.3, the contour coefficient increases with minPts, and after reaching the highest value, it begins to decrease. At the minPts value of 6, the modified DBSCAN algorithm has the highest contour coefficient of 0.72, which is 0.06 higher than the highest round contour coefficient of the traditional DBSCAN algorithm, when the number of clusters is 4. In conclusion, the results show that the modified DBSCAN algorithm clusters the results best when the minPts value is 6 and the number of clusters is 4. In order to predict the risk of academic level, the study divided students into four groups: pass, fail, excellent and non-excellent. For each group of students, they use the decision tree model, SVM, Adaboost model, and the multi-classifier-based heteromorphic N-Adaboost model. SVM and decision tree model were used as a control benchmark model to verify the effectiveness of N-Adaboost model. Each model was trained and evaluated by ten-fold cross-validation. Both the training set and the test set of the model were randomly divided, and the prediction operation was repeated ten times. The average value of the accuracy and F1 measure were taken as the prediction accuracy of the final model and the F1 measure of the results. The classification measurement results of each model in the passing group, failing group, excellent group and nonexcellent group are shown in Figure 4.5.

Figure 4.4(a) shows the classification measurement results of each passing group and failure group of the comparison model. As shown in Figure 4.4(a), the accuracy of SVM and decision tree benchmark model is 57.63%, while the evaluation accuracy of the N-Adaboost model proposed by the study is 74.02%, which is higher than the benchmark model and also higher than other comparison models, and its accuracy performance is optimal. The accuracy rate of SVM and decision tree benchmark model is 75.46%, while the evaluation accuracy rate of N-Adaboost model is 87.05%, which is higher than other comparison models and has the optimal accuracy performance. Meanwhile, the recall rate and F1 value of the N-Adaboost model were 75.16% and 0.806, respectively, which are higher than the benchmark model and the comparison model, and the proposed N-Adaboost model has the best performance. Figure 4.4(b) shows the classification measurement results of each comparison model for the excellent group and the non-excellent group. As shown in Figure 4.4(b), the accuracy of SVM and decision tree benchmark model is 64.3%, while the evaluation accuracy of the N-Adaboost model proposed by the study is 73.74%, which is higher than the benchmark model and also higher than other comparison models, and its accuracy performance is optimal. The accuracy rate of SVM and decision tree benchmark model is 60.23%, while the accuracy rate of the proposed N-Adaboost model is 68.13%, which is higher than the best accuracy performance. Meanwhile, the recall rate and F1 value of N-Adaboost model were 74.32% and 0.716, respectively, which are higher than the benchmark model and the comparison model, and the proposed N-Adaboost model has the best performance. In conclusion, these results show that the proposed N-Adaboost model performs better than the other contrast models."N" represents the species of individual



((a)) Pass and fail

((b)) Merit and Non-Merit categories

Fig. 4.5: Evaluation of the learning outcomes of the pass and fail groups and the excellent and non-excellent groups under different models

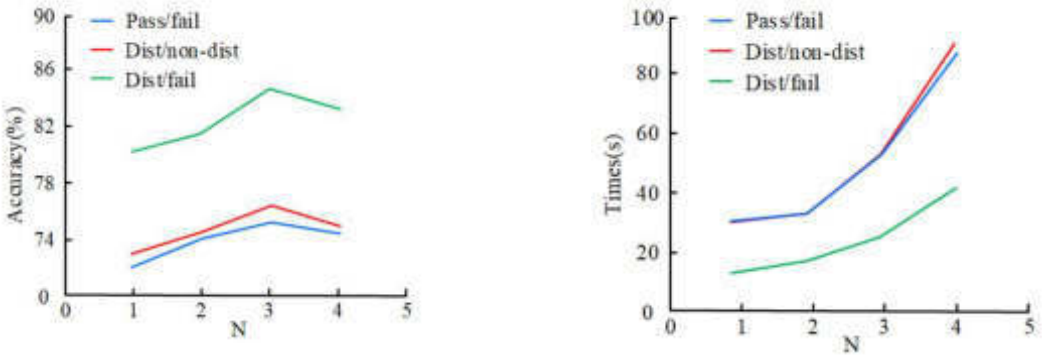
Table 4.1: Evaluation Results of the N-Adaboost Model for Different Values of N

	N=1	N=2	N=3	N=4
Accuracy	0.7138	0.7329	0.7441	0.7362
Precision	0.8535	0.8634	0.8759	0.8674
The recall rate	0.7231	0.7434	0.7474	0.7441
F1 value	0.783	0.799	0.807	0.801

learners in the modified multi-classifier-based N-Adaboost model. However, the variety of individual learners is not the more the better. According to the sample data of the " pass and fail group ", the research explores "N" in the N-Adaboost model and compares the model evaluation performance of N=1, N=2, N=3, and N=4. Among them, N=1 is the traditional Adaboost model, and its base classifier is composed of a decision tree model; N=2 is the N-Adaboost model used in the evaluation experiment, and its base classifier is composed of a decision tree model and an SVM model; N=3 Set the base classifier in N-Adaboost model to decision tree model, SVM model and Naive Bayes model (NB); N=4 set the base classifier in N-Adaboost model to decision tree model, SVM model, Naive Bayes model Yess model and logistic regression model. The results are shown in Table 4.1.

From Table 4.1, when N changes from 1 to 3, the evaluation accuracy is improved to a certain extent. The accuracy of N=2 is 1.91% higher than that of N=1, and the accuracy of N=3 It is 1.12% higher than N=2. However, when N=4, it is found that the accuracy rate has declined, and the accuracy is like that when N=2 and F1 is also reduced accordingly. Similarly, the same method is used to predict the samples of the "excellent and non-excellent group ""Pass and fail group"and "excellent and failing group ", and the results are shown in Figure 4.7.

Figure Figure 4.6(a) is the accuracy change curve of the three groups of samples under different N conditions. For excellent group and not excellent group, qualified group and unqualified group, excellent group and failing group, when N=3, the evaluation accuracy reached the highest, indicating that the more obvious the difference between the sample data, the higher the prediction ability of the model.(B) For the time-consuming change curve of three groups of samples under different N conditions. As the individual classifier types increase, the time consumption also gradually increases. When N=4, the highest time reached each group. Considering the balance



((a)) Accuracy of Prediction

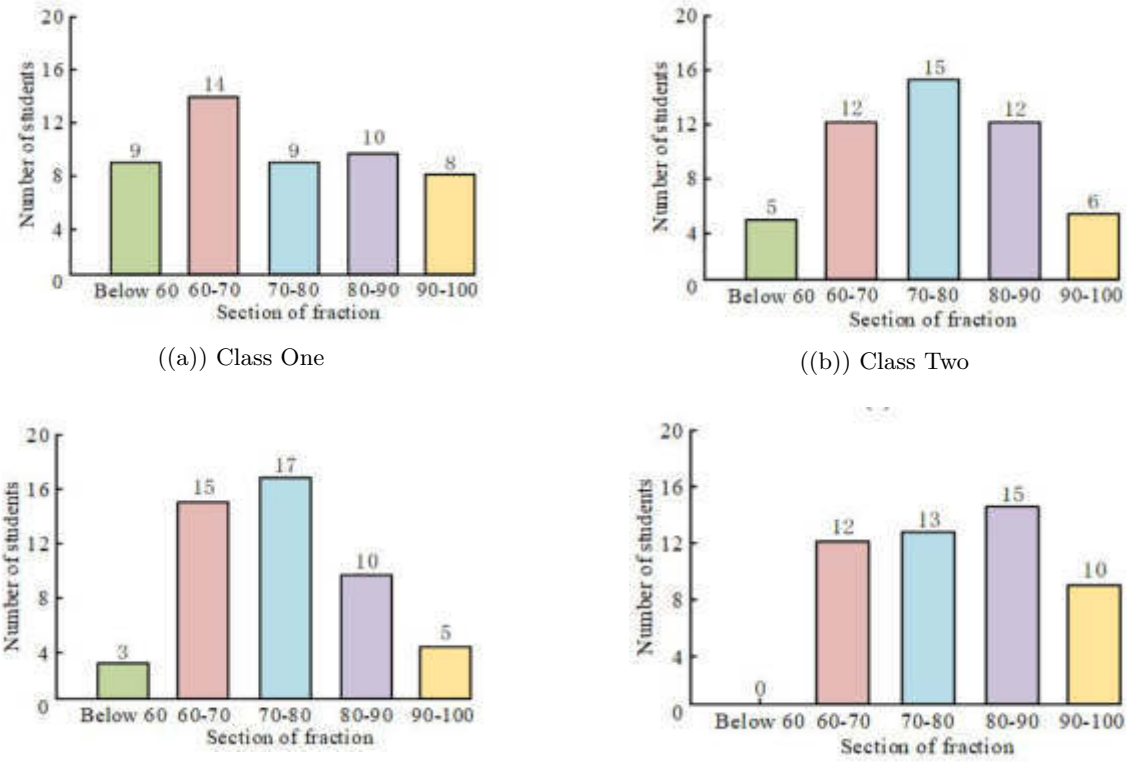
((b)) Elapsed Time

Fig. 4.7: Evaluation results of different models for excellent and non-excellent students

between the accuracy and the evaluation time, when N is 2 or 3, the operation time is also controlled within a reasonable range while ensuring the evaluation accuracy of the N-Adaboost model. Four classes of 50 students were selected to evaluate the online foreign language learning of a total of 200 students. Firstly, students' overall learning of the semester was evaluated based on their classroom performance, homework completion, and exam results, with score ranging from 0 to 100 points. The evaluation results consider the students 'understanding and logic, students' classroom performance, students' classroom understanding and summary ability, and the mastery of knowledge and skills. The comprehensive evaluation results of the students are shown in Figure 4.5.

Figure Figure 4.8(a) shows the study situation of Class 1. According to figure, 4.8(a), in Class 1, students with 60-70 account for 28%, students 90-100 account for 16%, and students below 60 account for 18%. Figure 4.8(b) shows the study situation of Class 2. From Figure 4.8(b), in Class 2, students with points of 70-80 account for 30%, and students with points below 60 and 90-100 account for 10% and 12%. Figure 4.8(c) shows the study situation of Class 2. From Figure 4.8(c), in Class 3, students with points of 60-80 account for 64% of the class, and students with points below 60 and 90-100 account for 6% and 10%. Figure 4.8(d) shows the study of Class 2, as shown from Figure 4.8(d). There are no students with scores below 60 points in class 4, and the other four grades are evenly distributed and have the best grades. To grasp students' learning status in more detail and adjust the teaching plan in time, the researchers conducted a detailed evaluation of students' listening, speaking, reading, and writing abilities through the model, as shown in Figure 4.11.

Figure 4.10(a) shows the listening learning situation of students in each class. From Figure 4.10(a), it can be seen that in terms of listening, the listening ability of students of the four classes is on the rise along with the learning progress, indicating that online teaching can improve students' foreign language listening level. Figure 4.10(b) shows the oral learning of students in each class. As can be seen from Figure 4.10(b), we can see that the oral ability of students in Class 1 is constantly improving, the oral ability of Class 2 and Class 4 remains unchanged, and the oral ability of Class 3 is declining. Figure 4.10(c) shows the learning situation of students in each class in reading. As can be seen from Figure 4.10(c), the reading ability of each class increases with the learning progress. Figure 4.10(d) shows the learning situation of students in each class, According to Figure 4.10(d), it can be seen that the writing ability of the four classes has decreased, and the study analyzes the causes of this phenomenon. It is found that the teaching plan has some deficiencies and the lack of relevant writing practice. When teachers assign classroom tasks, they are more inclined to cultivate the "listening", "speaking" and "reading" modules, and the writing tasks are assigned less. At the same time, the lack of students' vocabulary also leads to the decline of the writing ability of each class. In view of this problem, the study put forward corresponding solutions, such as students making personal learning plans, teachers regularly arrange thematic writing tasks and timely feedback and guidance, the school arranged writing competitions, learning communication activities, etc. Through the continuous practice, improve the



((a)) Class One

((b)) Class Two

((c)) Class Three

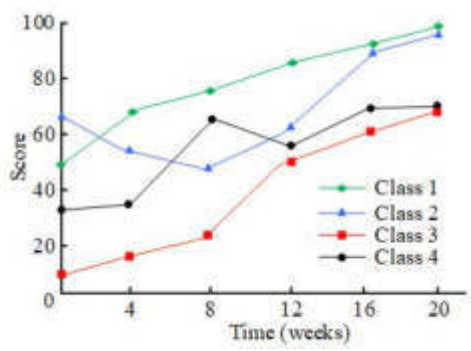
((d)) Class Four

Fig. 4.9: Student learning overall evaluation results

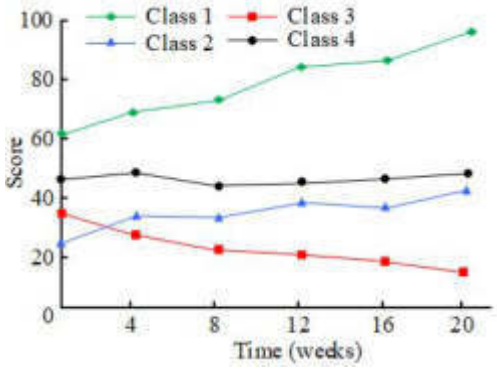
students' writing ability. Vocabulary is an important part of students' foreign language learning. Through the online teaching collaboration platform, teachers not only help students to accumulate vocabulary in class but also require students to independently accumulate at least 30 words a day through the online platform. After one semester, the evaluation model is used to evaluate group A and group B. The vocabulary of a total of 1000 students was investigated, as shown in Figure 4.7.

As can be seen from Figure 4.12, the foreign vocabulary learning of students in each group increased with the learning progress, and the vocabulary of students in Group A increased from less than 1000 to more than 6000. The vocabulary of students in group B increased from 500 to above 6000. In summary, the results show that the evaluation model of students' foreign language learning status for the online teaching collaborative platform can help students to improve their learning situation.

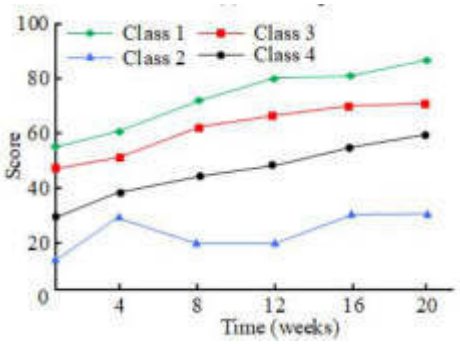
5. Conclusion. The rapid development of Internet technology has brought a new revolution to the traditional offline teaching mode. Aiming at the problem that the traditional teaching quality evaluation system is no longer suitable for modern information-based teaching evaluation, this paper constructs an evaluation model of students' foreign language learning based on the online teaching cooperation platform. DBSCAN algorithm based on distance optimization is used to cluster students' learning behaviors. Based on the traditional Adaboost model, the classification of the base classifier is improved and the N-Adaboost algorithm is proposed. Experiments show that the clustering effect and accuracy are improved by 8.7%, and the comprehensive performance is better than the traditional DBSCAN algorithm. The evaluation accuracy of the N-Adaboost model based on multiple classifiers in the two groups of the pass and fail, excellent and non-excellent is 74.02% and 73.74%, respectively, and the overall performance has been improved. In addition, when N is 2 or 3, the balance



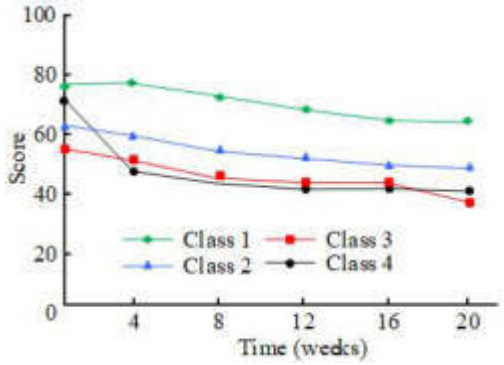
((a)) Listening



((b)) Speaking



((c)) Reading



((d)) Writing

Fig. 4.11: Student vocabulary changes over time

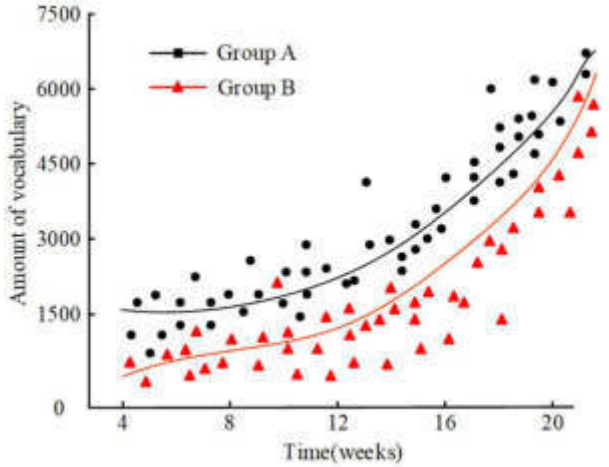


Fig. 4.12: Student vocabulary changes over time

between accuracy and evaluation time can be considered. Based on the performance of the class, the completion of homework, and the examination results, the learning situation of the four classes was evaluated. There were no failed students in the four classes, and the number of students with 90-100 scores was the largest. Class 1 had the worst overall situation, with 9 students below 60 points and 14 students between 60 and 70 points. The students' listening, speaking, and reading abilities showed an overall improvement, but their writing ability declined seriously, dropping to below 40 points at 20 weeks. In view of this problem, the study put forward corresponding solutions, such as students making personal learning plans, teachers regularly arrange thematic writing tasks and timely feedback and guidance, the school arranged writing competitions, learning communication activities, etc. Through the continuous practice, improve the students' writing ability. The proposed evaluation model of students' foreign language learning for online teaching collaboration platform has certain potential in practical application. This model can accurately evaluate students 'semester situation, and then provide guidance and feedback for teachers, promote students' independent learning, provide students with personalized learning guidance and support for students, optimize teaching design and resource allocation, promote educational reform and innovation, and then improve the quality of education. However, the application of the evaluation model still faces some limitations, the reality is often more complex and changeable, and the evaluation model needs to solve the problems of subjective evaluation factors and the difficulty of comprehensive evaluation. At the same time, the samples selected in this study also have some limitations, such as the under representation of the sample, which still needs to further explore the students who do not have the universal education information sharing platform. The future research direction is to establish a distributed cluster environment using high-performance platform and realize the parallelization of student behavior data processing and computing. The effectiveness verification of the proposed intervention based on the evaluation model proves that the evaluation model improves the teaching quality. At the same time, the research can also add more factors to the evaluation mode, and constantly optimize the online teaching mode, such as students 'participation in class, and evaluate the degree of students' participation and interaction in class, including ask questions, answer questions, discussion and cooperation. And students 'learning experience, such as assessing students' satisfaction and experience of online teaching mode, such as students' willingness to participate, learning motivation and learning interest. By comprehensively considering the factors of students 'classroom participation and students' learning experience, the research can more comprehensively evaluate the advantages and disadvantages of online teaching mode, and continuously optimize the teaching process to improve students' learning effect and learning experience.

Funding. The research is supported by the Henan Provincial Teaching Reform Research and Practice Project "Research and Practice on the Teaching Ecosystem of General Academic English in Science and Engineering Universities under the Double First Class Background" (No. 2019SJGLX060).

REFERENCES

[1] Wang, Y. On College English Teachers' Basic Skills of Classroom Teaching and Evaluation and the Promotion Strategies. *International Journal Of Social Science And Education Research*. **5**, 93-96 (2022)

[2] Gao, H. Application of Iwrite English Writing Teaching and Appraising System in College English Teaching. *International Journal Of Social Sciences In Universities*. **5**, 255-259 (2022)

[3] Jiao, F., Song, J., Zhao, X., Zhao, P. & Wang, R. A spoken English teaching system based on speech recognition and machine learning. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 68-82 (2021)

[4] Yang, Z. & Feng, B. Design of key data integration system for interactive English teaching based on internet of things. *International Journal Of Continuing Engineering Education And Life Long Learning*. **31**, 53-68 (2021)

[5] Susanty, L., Hartati, Z., Sholihin, R., Syahid, A. & Liriwati, F. Why English teaching truth on digital trends as an effort for effective learning and evaluation: opportunities and challenges: analysis of teaching English. *Linguistics And Culture Review*. **5** pp. 303-316 (2021)

[6] Liu, H., Chen, R., Cao, S. & Lv, H. Evaluation of college English teaching quality based on grey clustering analysis. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 173-187 (2021)

[7] Li, N. A fuzzy evaluation model of college English teaching quality based on analytic hierarchy process. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 17-30 (2021)

[8] Zhang, Y. The development of an evaluation model to assess the effect of online English teaching based on fuzzy mathematics. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 186-200 (2021)

[9] Tqt, T., Tmn, N., Luu, T. & Pham, T. An evaluation of English non-majored freshmen's attitude towards EFL learning at the Can Tho University of Technology. *International Journal Of TESOL & Education*. **1**, 72-98 (2021)

[10] Wang, H. & Cui, J. Evaluation of teaching effect in higher educational institutions and identification of its influencing factors. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 226-239 (2021)
[11] Li, Y. & Wang, X. The algorithm and implementation of college English teaching comprehensive ability evaluation system//EAI International Conference, BigIoT-EDU. (216-224,2021)
[12] Sun, Z., Anbarasan, M. & Praveen Kumar, D. Design of online intelligent English teaching platform based on artificial intelligence techniques. *Computational Intelligence*. **37**, 1166-1180 (2021)
[13] Lu, C., He, B. & Zhang, R. Evaluation of English interpretation teaching quality based on GA optimized RBF neural network. *Journal Of Intelligent & Fuzzy Systems*. **40**, 3185-3192 (2021)
[14] Fang, C. Intelligent online English teaching system based on SVM algorithm and complex network. *Journal Of Intelligent & Fuzzy Systems*. **40**, 2709-2719 (2021)
[15] Noor, A., Shahid, A., Ahmed, S. & Ahmad, M. An Evaluation of Communicative Language Teaching in Pakistan: A Study of Undergraduate English Learners of Pakistan. *Pakistan Journal Of Humanities And Social Sciences*. **9**, 259-264 (2021)
[16] Zhang, Y. & Yang, Y. The evaluation method for distance learning engagement of college English under the mixed teaching mode. *International Journal Of Continuing Engineering Education And Life Long Learning*. **32**, 159-175 (2022)
[17] Clayson, D. The student evaluation of teaching and likability: what the evaluations actually measure. *Assessment & Evaluation In Higher Education*. **47**, 313-326 (2022)
[18] Cook, C., Jones, J. & Al-Twal, A. Validity and fairness of utilising student evaluation of teaching (SET) as a primary performance measure. *Journal Of Further And Higher Education*. **46**, 172-184 (2022)
[19] Okoye, K. & Arrona-Palacios, A. Camacho-Zu niga C, Joaquín Alejandro Guerra. (2022) Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Education And Information Technologies*. **27**, 3891-3933 (0)
[20] Marshall, P. Contribution of open-ended questions in student evaluation of teaching. *Higher Education Research & Development*. **41**, 1992-2005 (2022)
[21] Marks, B. & Thomas, J. Adoption of virtual reality technology in higher education: An evaluation of five teaching semesters in a purpose-designed laboratory. *Education And Information Technologies*. **27**, 1287-1305 (2022)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 17, 2023
Accepted: Nov 1, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 423–434, DOI 10.12694/scpe.v25i1.2308

PRESCHOOL TEACHERS TEACHING QUALITY EVALUATION BASED ON NEURAL NETWORK ALGORITHMS

HONGXIA CAI*

Abstract. Based on the current teaching situation of preschool teachers, in order to comprehensively evaluate the effectiveness of early childhood teaching, research have constructed a teaching quality evaluation model using fuzzy synthesis method and expert method. This model can handle the fuzzy relationship between evaluation indicators and achieve the evaluation of teaching quality. Considering that the teaching evaluation is influenced by many factors, the Genetic Algorithm back propagation (GA-BP) neural network algorithm is chosen for the solution model construction of the preschool teacher teaching quality. The entropy method chosen for the data calculation is to complete the preschool teaching quality evaluation. In the mean square error test for solving the model, the improved GA-BP model converged after 40 iterations with the model convergence speed increased by 34.65%. In the evaluation and prediction of preschool teacher indicators, the improved GA-BP model accurately evaluated the teacher classroom teaching indicators. In sample 3, sample 6, and sample 9, the improved GA-BP model scored 91, 89, and 88 points, respectively, close to the true scoring results. The improved model's accuracy was high as 90% in teacher skills, personality charm, and academic research evaluation. The improved model also effected better in the application of the preschool teaching quality evaluation. The application effect of this model in teaching effectiveness evaluation and teacher quality evaluation is good, providing valuable reference for the establishment of early childhood education evaluation system.

Key words: Preschool education; GA-BP algorithm; Teaching quality evaluation; Fuzzy comprehensive method; Entropy method

1. Introduction. China has established a complete educational and teaching evaluation system in primary and secondary education to improve the current educational and teaching effects and standardize the educational quality evaluation mechanism, meeting the requirements of educational and teaching development [1]. However, there is no relatively complete evaluation system in the current stage for the preschool teaching quality, which results in significant differences in the effectiveness of preschool education in different regions [2]. Preschool education received widespread attention from more and more parents and social figures. Improving the preschool education is crucial to ensuring the comprehensive physical and mental development of children [3]. Modern education focuses on the scientific implementation of teaching, teaching content, and teaching methods. Teachers' personal abilities and teaching concepts have an important impact on teaching development. Therefore, the expert method studying the present preschool education and the fuzzy theorem constructing a preschool teaching evaluation system are used for the preschool teaching effect. The preschool teaching evaluation is affected by environmental factors, teaching content, and the overall quality of teachers, etc., which can affect the final quality evaluation effect, the Genetic Algorithm back propagation (GA-BP) neural network algorithm is used for the preschool teaching evaluation. By adjusting the initial BP model parameters, using the GA algorithm, and introducing the entropy method to optimize sample parameters, a solution model for preschool teachers' teaching quality is constructed to achieve an effective evaluation of preschool teachers' teaching quality. The research carries referential value to promote a standardized and scientific development of modern early childhood education.

2. Related Work. Constructing a teaching evaluation system is important for the modern education, and it is a very complex nonlinear research issue. Effective teaching evaluation would directly present the preschool teaching situation, thereby achieving optimization of the educational process. Domestic and foreign experts have studied this issue. Qianna et al. found that intelligent educational evaluation methods were the key to the development of modern education. A classroom evaluating model was constructed with the neural algorithm

*Preschool Education Department, Zhumadian Preschool Education College, Zhumadian, 463000, China (Hongxia_Cai2023@outlook.com)

technology for evaluating the teaching quality, and an empirical modal method is used to improve the evaluation model. Through data testing, it was concluded that the proposed method had excellent application effects in classroom quality assessment [4]. Hou et al. analyzed the existing online education evaluation model and found that the model faced the problem of insufficient data. Neural learning models were introduced to participate in training and applied to the evaluation environment for the online teaching evaluation. After testing, this method can effectively evaluate the teaching effectiveness of universities and improve the overall evaluation effect of online education in universities [5]. Liu et al. conducted research on current English education and found that existing English education faced the problem of inaccurate evaluation in teaching quality evaluation. The existing educational content was researched, and a classroom evaluation system was constructed for college English teaching through the entropy method and clustering method. Applying this evaluation method to a college English teaching environment can effectively evaluate college English classrooms and provide an effective reference for improving college English teaching [6]. Bao et al. conducted research on existing college physical education (PE) and found that existing college physical education faces problems in online and offline quality testing. A composite PE evaluation model was constructed. The first was to clarify the objectives, influencing factors, and relevant standards of physical education, thus building a indicator system for PE. The second was to divide the importance of physical education factor indicators, simplify physical education evaluation indicators through clustering and fuzzy evaluation, and obtain the weight of each indicator. Finally, it was applied to the specific teaching. The test showed that this method can evaluate the online and offline PE and meet the PE teaching requirements of physical education and teaching [7].

Neural network algorithm technology is widely used in education with a positive impact. PoL et al. conducted research on existing neural network technology and found that neural network technology had good application value in the field of quality assessment. Therefore, neural network technology is applied to construct a classroom education evaluation system in the existing educational environment for evaluating teachers' teaching quality. Applying this method to specific teaching data samples can accurately evaluate classroom teaching effectiveness and meet the development requirements of education [8]. Pag et al. conducted research on the existing biological field. Bioinformatics is a complex and systematic engineering, and the existing evaluation system cannot effectively evaluate the effectiveness of biological education. A common method was to combine multiple research results and construct different output features based on experts in different fields. A protein model was proposed for the problem in evaluating, and a convolutional neural network was applied for the optimization. Finally, the method was applied to a specific database, and the results showed a good evaluation effect and met the requirements of teaching development [9]. Siyan et al. conducted research on existing teacher quality assessment methods and found that teachers' abilities cannot be quantitatively analyzed, resulting in an imbalance in final classroom teaching. A method of teacher competency assessment was proposed to solve the problems faced by current classroom education. This method was based on advanced digital twin technology and neural network technology, and it constructed a data fusion model through the data of teacher professional information. At the same time, a decision tree algorithm can analyze teacher competency data and build a competency mining model. Applying this evaluation method to the current educational environment, the results showed the accurate evaluation of teachers' abilities [10].

According to their research, teaching evaluation can effectively reflect teaching characteristics and provide an important basis for teaching optimization. The development of neural networks and other technologies used in the education evaluation significantly improved its shortcomings and improved the teaching evaluation, which has important research significance for modern education development.

3. Preschool teaching evaluation model construction based on the neural network algorithm.

3.1. Preschool teaching evaluation model construction. Preschool teachers' teaching level directly affects the effectiveness of education implementation, and it is particularly important to effectively evaluate the preschool teaching. Therefore, a survey is conducted in multiple preschool education institutions in a city, strictly implementing the principles of scientific, objectivity, and pertinence. Through fuzzy principal component analysis and expert methods. Among them, the expert evaluation method is a commonly used evaluation method, which invites experts from relevant fields to evaluate the teaching level of teachers, in order to obtain more accurate and objective evaluation results. In this paper, expert evaluation method is applied to determine the evaluation indicators of Preschool teacher 'teaching quality. the evaluation indicators are

Table 3.1: Teacher teaching evaluation index system

Evaluation target	Level 1	Level 2
Preschool evaluation target	Classroom teaching U1	Scientific teaching U11
		Teaching artistry U12
		Teaching effect U13
		Clear teaching objectives U14
	Teacher Skills U2	Practical ability training U21
		Innovation cultivation U22
		Philosophy Education U23
		Thinking Inspiration Cultivation U24
	Teacher personality U3	High character U31
		Responsibility attitude U32
		Professional enthusiasm U33
	Academic Research U4	Discipline cognition U41
		Teaching Practice U42
		Theoretical depth U43

identified for preschool teachers' teaching quality. The first level of indicators includes classroom teaching, academic research, teacher personality, and teaching skills. The first level indicator layer is further divided by expert method to obtain the second level indicator factors. According to the expert evaluation method, the first level indicators of Preschool teacher 'teaching quality are determined, including classroom teaching, academic research, teachers' personality and teaching skills. These indicators are obtained by inviting experts from relevant fields to evaluate the teaching level of teachers. Then, the first level indicators were further divided using expert methods to obtain the second level indicator factors. This includes various aspects of classroom teaching, such as science teaching, teaching art, and teaching effectiveness; Various aspects of teacher skills, such as practical ability training and innovation cultivation; Various aspects of a teacher's personality, such as high personality, responsible attitude, and professional enthusiasm; And various aspects of academic research, such as subject cognition, teaching internships, and theoretical depth.Table 3.1 shows the preschool teaching evaluation.

The set of factors for the preschool teaching evaluation constructed represents the factors set that affect teaching as $(U = \{U_1, U_2, U_3, U_4\})$. Based on the set of factors constructed, a corresponding evaluation set is constructed, as shown in Equation 3.1.

$$V = \{V_1, V_2, V_3, V_4\}) \tag{3.1}$$

In Equation 3.1, V_1 represents excellent, V_2 represents relatively excellent, V_3 represents qualified, and V_4 represents unqualified. Based on the established system, further analysis is needed for the factor weight [11]. Considering that evaluation factors directly affect the teaching, the impact of different evaluation factors on teaching is vague, and it is necessary to consider the educational objectives and the students' needs. Therefore, the expert method scores the relationship between teaching quality and evaluation factors, reflecting the impact of different evaluation factors on teaching quality through scoring, and assigning corresponding weights to evaluation indicators based on the scoring. The corresponding fuzzy weight of the first level indicator pair is defined as A , as shown in Equation 3.2.

$$(A = (0.3, 0.4, 0.1, 0.2)) \tag{3.2}$$

The corresponding fuzzy weights of the secondary indicators are shown in Equation 3.3.

$$\begin{cases} A_1 = (0.2, 0.35, 0.2, 0.25) \\ A_2 = (0.28, 0.32, 0.25, 0.15) \\ A_3 = (0.3, 0.4, 0.3) \\ A_4 = (0.35, 0.4, 0.25) \end{cases} \tag{3.3}$$

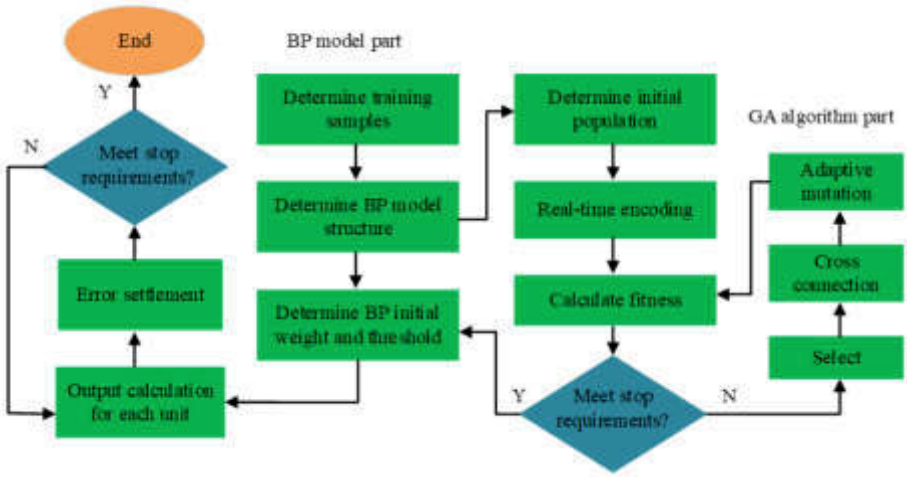


Fig. 3.1: Results of GA-BP Model Operation Process

After obtaining the fuzzy expression relationship of evaluation factors, it is also necessary to determine the fuzzy matrix relationship between evaluation factors and teaching quality. The relationship between evaluation set U and scoring set V is referred by R , the fuzzy matrix, and its relationship is shown below [12].

$$R = (r_{nm}) \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{3.4}$$

In Equation 3.4, r_{nm} is the evaluation vector, indicating the degree of the $n-$th evaluation factor subordination to the m grade. According to the fuzzy matrix relationship, the fuzzy evaluation matrix is calculated, and its calculation expression is shown in Equation 3.5.

$$B_i = A_i \times R_i \tag{3.5}$$

The expression relationship is calculated from the fuzzy matrix, and B is defined as the set of fuzzy relationship matrices for teacher teaching evaluation in Equation 3.6.

$$B = \{B_1, B_2, B_3, \ldots, B_n\}, \quad n > 1 \tag{3.6}$$

By calculating the fuzzy relational matrix of evaluation factors, a comprehensive evaluation of teachers' teaching quality can be achieved. Table 3.2 shows the final teacher index weights.

3.2. Teacher's teaching quality solving model construction based on GA-BP. In early childhood teaching, the early childhood teaching is influenced by many factors, and the traditional fuzzy evaluation method cannot effectively reflect teachers' comprehensive abilities and quality Considering that the teaching evaluation problem is non-linear, GA-BP is introduced to solve the problem. The GA algorithm optimizes the traditional BP's initialization parameters and improves its training accuracy [13]. The GA-BP model running process is shown in Figure 3.1.

Figure 3.1 is an optimization diagram of the GA-BP model, although the BP model has good solving ability in solving nonlinear problems. However, traditional BP models highly rely on initial thresholds and weights during initial training, and BP is prone to falling into local optimal solution problems during the

Table 3.2: The final teacher index weights

Evaluation target	Level 1	Level 2	Weight	Comprehensive weight
Evaluation target system of teaching quality for preschool teachers U	Classroom teaching U1	Scientific teaching U11	0.20	0.035
		Teaching artistry U12	0.35	0.045
		Teaching effect U13	0.20	0.035
		Clear teaching objectives U14	0.25	0.040
	Teacher Skills U2	Practical ability training U21	0.28	0.041
		Innovation cultivation U22	0.32	0.043
		Philosophy Education U23	0.25	0.040
		Thinking Inspiration Cultivation U24	0.15	0.025
	Teacher personality U3	High character U31	0.30	0.042
		Responsibility attitude U32	0.40	0.068
		Professional enthusiasm U33	0.30	0.042
	Academic Research U4	Discipline cognition U41	0.35	0.047
		Teaching Practice U42	0.40	0.067
		Theoretical depth U43	0.25	0.046

training process [14]. In practical optimization, the GA model is introduced to optimize the parameters of the BP model, obtaining the best fitness value through selection, crossover, and other behaviors, and using the optimization results as the training parameters of the BP model. Firstly, the rounding method determines the number of BP hidden layers, as shown in equation 3.7.

$$m = \sqrt{n + l} + a \tag{3.7}$$

In Equation 3.7, l is the total of output layer nodes, n is that of input layer nodes, and a is an arbitrary constant within 1 to 10. The number of BP model nodes has direct impact on the target variable dimensions and is represented by a hyperbolic function, as shown in Equation 3.8 [15].

$$f(x) = \frac{1}{1 + e^{x}} \tag{3.8}$$

In BP model training, the more sample data of teacher teaching quality indicators constructed, the more accurate the model training effect will be. However, there is a certain limit to the number of model training samples. If the size of the parameter setting is exceeded, the model training accuracy will decrease [16]. Therefore, in actual model training, it is necessary to reasonably select effective initial parameters and select a reasonable hidden layer model as the training model. Generally, the initial weight selection is based on the minimum initial weight, but the selected initial weight is not accurate. Therefore, genetic algorithms optimize BP's initial parameters and build a GA-BP solution model [18].

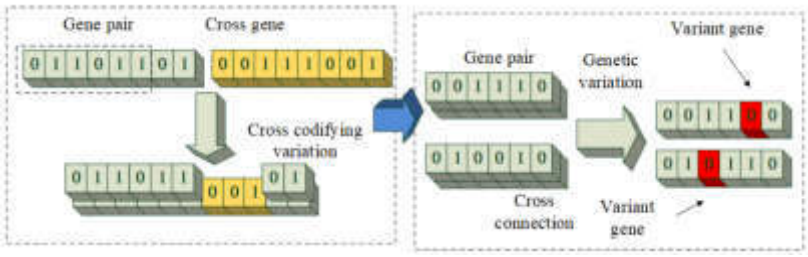


Fig. 3.2: Flow chart of gene crossover and gene mutation of GA algorithm

GA models do not directly search for feasible solutions when solving parametric problems, and require the first step of coding the feasible solutions. Effective coding will improve the global search ability of the GA model. Then, optimal gene selection is achieved through gene selection, gene crossover, and gene mutation. The crossover and mutation of genes are shown in Figure 3.2.

The training goal of GA models is to search for the initial parameter values of the minimum sum of errors in evolutionary iterations. GA models typically evolve in the direction of higher fitness. Choosing an appropriate fitness directly affects the search performance of GA models for initial parameters. Therefore, the individual error reciprocal is taken for the actual training as the fitness value, and the learning error is shown in Equation 3.9.

$$E = \frac{\sum_{k=1}^{P} \sum_{j=1}^{I} (y_j^k - o_j^k)^2}{2} \tag{3.9}$$

In Equation 3.9, $(y_j^k - o_j^k)$ represents the sample k's output error to the node j . k is P's sample number, and l is the total output nodes. According to the learning error, the fitness function of the GA model can be obtained, as shown in Equation 3.10.

$$\text{fitness} = \frac{1}{E} \tag{3.10}$$

In the research, roulette is selected as the method of selecting the GA model. The roulette method is selected based on the individual's fitness. The higher the fitness, the higher the probability of being selected, while the lower the fitness, the lower the probability of being selected. This selection method can maintain a proportional relationship of fitness, making individuals with high fitness more likely to be selected, thereby increasing the probability of retaining excellent individuals. First, the BP's initial parameters fitness value is calculated for individual genetic individuals, and the proportion of this value in the overall fitness value is calculated as the probability of individual selection. Then, the optimal individual value is obtained through crossover and mutation and chosen as BP's initial threshold and weight parameters [17].

When training sample data in the GA-BP model, the normalized data is processed through the entropy method for model's training effect. Equation 3.11 shows the standardized indicators of teaching evaluation.

$$x'_{ij} = \frac{x_{ij} - \bar{x}}{s_j} \tag{3.11}$$

In Equation 3.11, x_{ij} is the $i-$th evaluation indicator sample score on the $j-$th indicator factor, x'_{ij} represents the standardized value, $\bar{x}$ represents indicators' mean value, and s_j represents the standard deviation. Sample data also requires translation operations to meet training requirements on standardized data, as shown in Equation 3.12 [19].

$$Z_{ij} = x'_{ij} + A \tag{3.12}$$

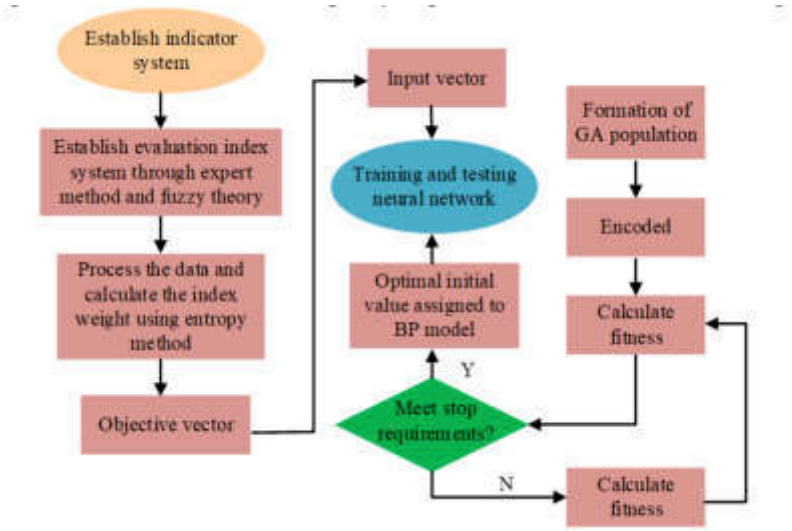


Fig. 3.3: Flow chart of preschool teaching quality evaluation model

In Equation 3.12, Z_{ij} represents the value after the translation operation, and A represents the translation length. Due to the differences among the preschool teaching evaluation indicators, it is also necessary to conduct a quantitative operation of indicator similarity, as shown in Equation 3.13 [19].

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{m} Z_{ij}} \tag{3.13}$$

In Equation 3.13, p_{ij} represents the proportion of the j index factor in the $i-$th evaluation index. The abnormal index coefficient of item j represents G_j , and if the abnormal index coefficient is normalized, the index weight of item j is as shown in Equation 3.14.

$$w_j = \frac{G_j}{\sum_{j=1}^{n} G_j} \tag{3.14}$$

According to the index weight in Equation 3.14, the $i-$th index sample can be calculated, as shown in Equation 3.15.

$$F_i = \sum_{j=1}^{n} w_j p_{ij} \tag{3.15}$$

The GA-BP model is used to construct a solution model for evaluating the preschool teaching. The teaching evaluation flowchart for the quality of preschool teachers is shown in Figure 3.3.

Due to the initial parameterization problem affecting the training accuracy faced by the BP model during training in the GA-BP model, the GA algorithm optimizes the BP's parameters for its training effect. Considering the interference of value differences on quality evaluation, the entropy method is adopted for the index deviation reduction, thereby improving the GA-BP's training.

4. Algorithm model simulation test. In the constructed preschool teaching evaluation, the number of primary indicators was 4, and that of secondary was 14. The indicators needed to be distributed to parents of young children and experts in the field of early childhood for scoring, with a total of 500 samples. The collected sample data needed to be standardized to obtain experimental training data, with a total of 1200 experimental sample data, The collected sample data needs to be standardized in order to obtain experimental training data, including 800 training sets and 400 testing sets. Table 4.1 shows some sample data information.

Table 4.1: Sample data after GA-BP standardization

Sample No	Classroom teaching	Teacher Skills	Teacher personality	Academic Research
1	0.86	0.78	0.86	0.82
2	0.86	0.65	0.56	0.75
3	0.84	0.87	0.68	0.68
4	0.71	0.45	0.78	0.82
5	0.68	0.74	0.68	0.71
6	0.68	0.82	0.75	0.68
7	0.85	0.74	0.82	0.71
8	0.65	0.67	0.75	0.82
9	0.72	0.68	0.82	0.83
10	0.75	0.68	0.72	0.72
11	0.85	0.75	0.62	0.48
12	0.71	0.72	0.81	0.75
13	0.82	0.80	0.75	0.82
14	0.75	0.82	0.82	0.75
15	0.72	0.68	0.72	0.81

Table 4.2: Parameters of improved GA-BP model training model

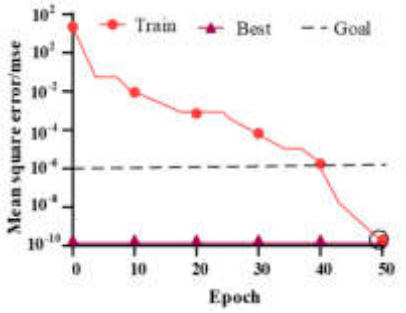
Model training parameters	Parameter value
BP model input layer node	15
Output layer node	1
Number of hidden layers	5
Maximum Iterations	100
GA model encoding length	80
GA model size	20
Cross probability	0.66

To verify the improved GA-BP's application in the evaluation of preschool teachers' teaching quality, the data in Table 4.1 were selected for model training. The test platform was Windows 10, with 64G of memory, an I7 64 core processor, and a graphics card NVIDIA RTX3080. Table 4.2 shows the training parameters.

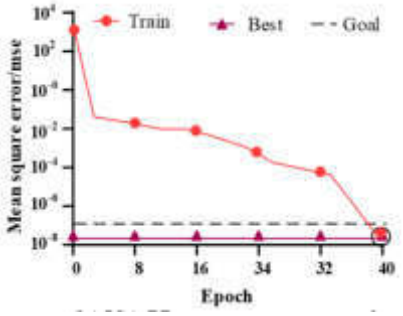
The simulation training of the preschool teaching evaluation was completed on the Matlab 2016 platform, and the mean square error training effects of the improved GA-BP model and the GA-BP model were compared. If the mean square error is smaller, the training accuracy will be higher. The training effect will be closer to the actual evaluation results of preschool teachers, as shown in Figure 4.2.

Figure 4.1(a) shows GA-BP's mean square error test results. From training curve changes, the GA-BP model converged after 50 iterations, and the model iteration speed was slower during the 10th to 40th iterations. After 40 iterations, the GA-BP model reached a training target position. The use of genetic algorithms to optimize BP's initial parameter performance significantly improved its accuracy. Thus, the GA-BP model accelerated its training speed after 40 iterations, converged after 50 iterations, and obtained the optimal training value. At this moment, the MSE value was 10-10. Figure 4.1(b) shows improved GA-BP's mean square error results. Using the entropy method to optimize training data during model training can improve the GA-BP model to converge faster and achieve higher accuracy values. The improved GA-BP model tended to converge after iteration 40. At this moment, the optimal MSE value was 10-8. The improved GA-BP model improved the convergence speed by 34.65%. Figure 4.4 shows the fitness training results of the two models.

Figure 4.3(a) shows the fitness value training results of the GA-BP model. According to the trend of training curve changes, the GA-BP model had a faster training speed during the first 40 iterations and gradually tended to a stable state after 40 iterations. After 80 iterations, the GA-BP optimal fitness curve tended to converge, and the average fitness curve coincided with the optimal fitness curve. At this moment, the optimal fitness

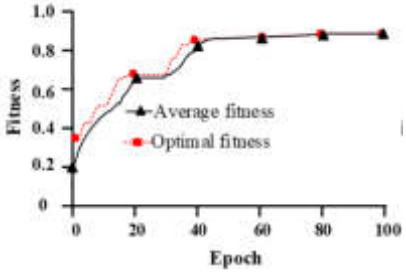


((a)) GA-BP mean square error result

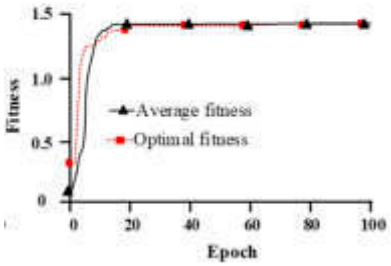


((b)) IGA-BP mean square error result

Fig. 4.2: Main Figure Caption



((a)) GA-BP model fitness



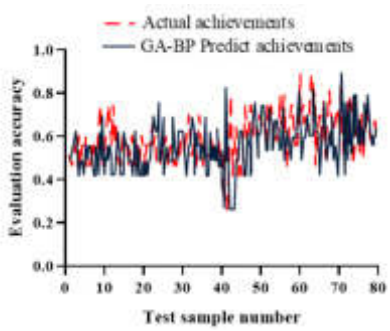
((b)) IGA-BPmodel fitness

Fig. 4.4: Fitness training results of two solving models

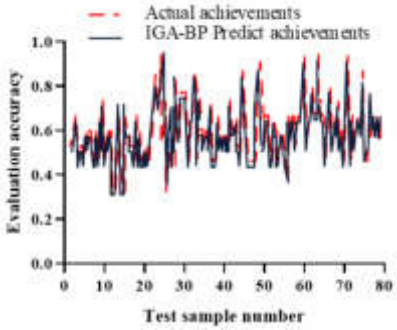
value was 0.856. Figure 4.3(b) shows the fitness value training results. The improved GA-BP model tended to converge after 20 iterations, and the optimal fitness value was 1.452 at this moment. The improved GA-BP model can obtain the optimal fitness value faster, while the fitness value was higher, which showed its better individual optimization ability. In Figure 4.6, the preschool teaching evaluation prediction are shown.

Figure 4.5(a) and Figure 4.5(b) show the evaluation and prediction results of both. Among them, the red dotted line is the actual result of sample evaluation, and the black solid line is the model training prediction result. In eighty training sets, the results of the GA-BP model had a significant deviation from the actual results. In samples 43 and 70, the prediction accuracy was less than 60%, and the average prediction accuracy was 73.85%. The improved GA-BP's average prediction accuracy in 80 groups of samples was 92.65%, while in sample 43 and sample 70, the prediction accuracy was 91.61% and 90.35%. The improved GA-BP's sample training accuracy increased by 39.65%. The prediction results of preschool teacher evaluation index scores are shown in Fig. 4.8.

Four types of primary indicator data for preschool teachers were selected for testing, and the number of model iterations was 100 to test the predictive effect of the two models on teacher performance scores. In Figure 4.8, the green line represents the actual grading results of preschool teachers, the red line represents the GA-BP's grading results of preschool teachers, and the blue line represents the improved GA-BP's grading results of preschool teachers predicted. Figure 4.7(a) shows the predicted scoring results of classroom teaching ability indicators. According to the curve changes in the figure, the actual teacher scores were 92, 90, and 89, respectively in sample 3, sample 6, and sample 9. There was a significant difference between the GA-BP model and the actual scoring results of preschool teachers. Its scores in samples 3, 6, and 9 were 70, 70,



((a)) GA-BP Prediction



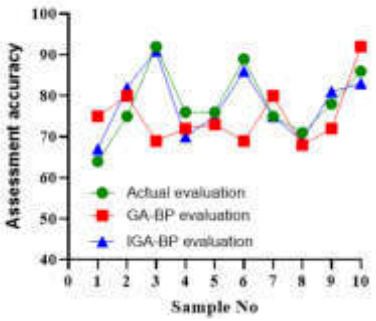
((b)) IGA-BPmodel Prediction

Fig. 4.6: Preschool teaching evaluation prediction

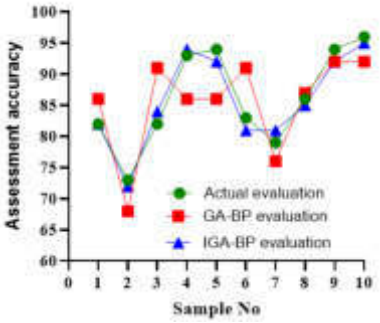
and 72, respectively. The improved GA-BP's scores were 91, 89, and 88, respectively, indicating that it was more accurate in scoring teacher teaching indicators. Figure 4.7(b) shows the predicted results of teacher skill index scores. According to the scoring results, the GA-BP model scored 91 points, 86 points, and 86 points, respectively in sample 3, sample 4, and sample 5, while the improved GA-BP model scored 72 points, 93 points, and 94 points, respectively. Compared with the actual scoring results, the GA-BP's accuracy rate is 82.65%, while the improved GA-BP's accuracy rate is 94.65%. The predicted results of teacher personality charm index scores are shown in Figure 4.7(c). The personality charm index is influenced by many factors, including teacher affinity, emotional state, and moral quality, which further tests the training and analysis ability. Based on the change results of the curve, the GA-BP model had a significant error in evaluating teachers' personality charm indicators, with a scoring accuracy of 76.65%, while the improved GA-BP model had a scoring accuracy of 90.35% for teachers' personality charm indicators. For example, there was a significant difference between the GA-BP model and the actual scoring results of teachers in Sample 2, 4, 5, 6, and 8 with the accuracy rate of the evaluation results being less than 70%, which cannot meet the quality evaluation requirements. The prediction results of teacher academic research index scores are shown in Figure 4.7(d). In sample 4, 5, and 7, the GA-BP model scored 75, 81, and 70 points, respectively, while the improved GA-BP model scored 85, 90, and 80 points, respectively. The improved GA-BP model training curve results were closer to the actual teacher scoring results, with a scoring accuracy of 94.65%, while the GA-BP model scoring accuracy rate was 82.65%. Table 4.3 shows the final scoring results of some preschool teachers' teaching quality.

In Table 4.3, the GA-BP model performed the worst in the preschool teaching evaluation. For example, the overall scoring accuracy of the GB-BP model was lower than 80% in sample 2, sample 7, and sample 11 of classroom teaching indicators, and the overall scoring results were poor, unable to meet the requirements of preschool teachers' teaching quality evaluation. However, the improved GA-BP model had an accuracy rate of more than 90% for evaluating teachers' classroom teaching indicators except for sample 9, meeting teaching requirements. At the same time, in the evaluation of teacher skills, teacher charm, and academic research indicators, the improved GA-BP model had a scoring accuracy of more than 89%, while the traditional GA-BP model had a scoring accuracy of less than 80% in individual sample tests, which cannot accurately evaluate the comprehensive personal abilities of teachers. The improved GA-BP model met the evaluation requirements for the preschool teaching and had better effects in evaluation.

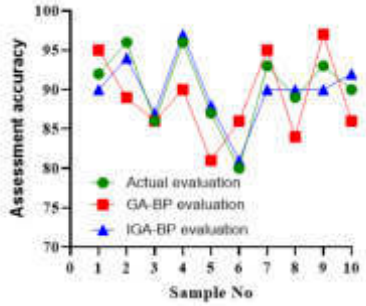
5. Conclusion. Preschool education is an important component of modern education, focusing on the cultivation of children's physical, mental, and creative abilities. The comprehensive quality level of teachers will directly affect the effectiveness of preschool education. Fuzzy theorem, analytic hierarchy, and expert method are used to analyze the current situation of preschool education, and constructs an evaluation system to improve the preschool teaching evaluation. Considering that the preschool teaching evaluation is a nonlinear problem that affects the effectiveness of quality evaluation, a genetic algorithm-optimized BP model is used to



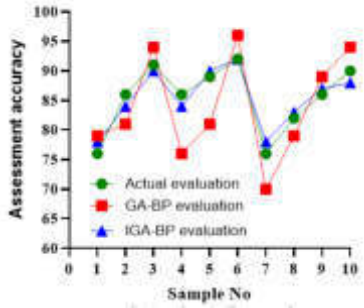
((a)) Classroom teaching



((b)) Teacher Skills



((c)) Teacher Personality



((d)) Academic research

Fig. 4.8: Predicted results of preschool teachers' index scores

Table 4.3: Final preschool teaching evaluation results

Sample No	Classroom teaching		Teacher Skills		Teacher personality		Academic Research	
	GA-BP	IGA-BP	GA-BP	IGA-BP	GA-BP	IGA-BP	GA-BP	IGA-BP
1	0.82	0.95	0.75	0.93	0.80	0.97	0.82	0.95
2	0.76	0.91	0.72	0.93	0.74	0.90	0.71	0.96
3	0.86	0.93	0.82	0.94	0.86	0.93	0.86	0.93
4	0.89	0.94	0.84	0.96	0.76	0.94	0.89	0.94
5	0.84	0.96	0.83	0.96	0.86	0.95	0.79	0.96
6	0.81	0.95	0.81	0.97	0.86	0.95	0.81	0.93
7	0.69	0.93	0.83	0.93	0.84	0.96	0.72	0.91
8	0.89	0.91	0.84	0.91	0.92	0.90	0.87	0.91
9	0.83	0.89	0.83	0.90	0.83	0.89	0.83	0.93
10	0.88	0.93	0.89	0.94	0.71	0.93	0.82	0.93
11	0.72	0.94	0.83	0.97	0.87	0.96	0.87	0.95
12	0.84	0.93	0.84	0.93	0.84	0.93	0.84	0.93
13	0.87	0.98	0.84	0.94	0.87	0.96	0.84	0.94
14	0.86	0.93	0.81	0.93	0.89	0.92	0.88	0.93
15	0.81	0.94	0.79	0.93	0.81	0.94	0.89	0.96

train indicator data and construct a GA-BP solution model. The entropy method is used for index data to optimize the training effect. When solving the fitness value, the improved GA-BP model tends to converge after 20 iterations, with an optimal fitness value of 1.452 and a fitness value of 0.856 for the traditional GA-BP model. Compared with the traditional GA-BP model, the improved GA-BP model has faster iteration efficiency and higher fitness values, proving that the improved GA-BP has better individual optimization ability. In preschool teaching evaluation, the overall prediction accuracy of traditional GA-BP is less than 80%, and the accuracy of teaching evaluation is relatively low. Except for sample 9, the improved GA-BP model has an accuracy rate of over 90% in all four primary indicators, which is superior to the traditional GA-BP model. The proposed model can accurately reflect the overall quality level of teachers. The proposed improvement GA-BP has good teaching evaluation results and meets the requirements of early childhood education development. However, there are also shortcomings in the research content. The effectiveness of preschool education is influenced by various factors, including teachers' teaching methods, students' family backgrounds, etc. Future research can comprehensively consider these factors and construct a more comprehensive evaluation system.

REFERENCES

[1] Zhou, W., Chen, Z. & Li, W. Dual-stream interactive networks for no-reference stereoscopic image qua lity assessment. *IEEE Transactions On Image Processing*. **28**, 3946-3958 (2019)

[2] Chen, X., Zou, D., Xie, H., Cheng, G. & Liu, C. Two decades of artificial intelligence in education. *Educational Technology & Society*. **25**, 28-47 (2022)

[3] Bacanin, N., Bezdan, T., Venkatachalam, K. & Turjman, F. Optimized convolutional neural network by firefly algorithm for magnetic resonance image classification of glioma brain tumor grade. *Journal Of Real-Time Image Processing*. **18**, 1085-1098 (2021)

[4] Qianna, S. Evaluation model of classroom teaching quality based on improved RVM algorithm and knowledge recommendation. *Journal Of Intelligent & Fuzzy Systems*. **40**, 2457-2467 (2021)

[5] Hou, J. Online teaching quality evaluation model based on support vector machine and decision tree. *Journal Of Intelligent & Fuzzy Systems*. **40**, 2193-2203 (2021)

[6] Liu, H., Chen, R., Cao, S. & Lv, H. Evaluation of college English teaching quality based on grey clustering analysis. *International Journal Of Emerging Technologies In Learning (iJET)*. **16**, 173-187 (2021)

[7] Bao, L. & Yu, P. Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing. *Mobile Networks And Applications*. **26**, 2188-2198 (2021)

[8] Po, L. & Liu, M. Yuen W Y F, Zhou C, Wong P. *A Novel Patch Variance Biased Convolutional Neural Network For No-reference Image Quality Assessment*. **29**, 1223-1229 (2019)

[9] Pagès, G., Charmettant, B. & Grudinin, S. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*. **35**, 3313-3319 (2019)

[10] Siyan, C., Tinghuai, W., Xiaomei, L., Zhu, L. & Wu, D. Research on the improvement of teachers' teaching ability based on machine learning and digital twin technology. *Journal Of Intelligent & Fuzzy Systems*. **40**, 7323-7334 (2021)

[11] Ouyang, F., Zheng, L. & Jiao, P. Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education And Information Technologies*. **27**, 7893-7925 (2022)

[12] Allugunti, V. machine learning model for skin disease classification using convolution neural network. *International Journal Of Computing, Programming And Database Management*. **3**, 141-147 (2022)

[13] Weis, C., Jutzeler, C. & Borgwardt, K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clinical Microbiology And Infection*. **26**, 1310-1317 (2020)

[14] Luan, H. & Tsai, C. review of using machine learning approaches for precision education. *Educational Technology & Society*. **24**, 250-266 (2021)

[15] Wenming, H. Simulation of English teaching quality evaluation model based on Gaussian process machine learning. *Journal Of Intelligent & Fuzzy Systems*. **40**, 2373-2383 (2021)

[16] Peng, X. & Dai, J. Research on the assessment of classroom teaching quality with q-rung orthopair fuzzy information based on multiparametric similarity measure and combinative distance-based assessment. *International Journal Of Intelligent Systems*. **34**, 1588-1630 (2019)

[17] Bao, L. & Yu, P. Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing. *Mobile Networks And Applications*. **26**, 2188-2198 (2021)

[18] Shukla, A., Pippal, S. & Chauhan, S. An empirical evaluation of teaching–learning-based optimization, genetic algorithm and particle swarm optimization. *International Journal Of Computers And Applications*. **45**, 36-50 (2023)

[19] Matosas-López, L., Aguado-Franco, J. & Gómez-Galán, J. Constructing an instrument with behavioral scales to assess teaching quality in blended learning modalities. *Journal Of New Approaches In Educational Research (NAER Journal)*. **8**, 142-165 (2019)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 18, 2023
Accepted: Nov 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 435–449, DOI 10.12694/scpe.v25i1.2310

BLENDED COLLEGE ENGLISH TEACHING MODEL AND EVALUATION BASED ON MOOC

JINGBO HAO* AND WULIAN WEI†

Abstract. MOOC teaching has been developing rapidly in the context of COVID-19, and its teaching quality has become the focus of social attention. Therefore, this study analyzes the blended college English teaching model based on MOOC by constructing a teaching evaluation model. The co-occurrence rate index is used to improve the K-modes algorithm, and then the important index in the teaching mode is extracted. The neural network is used to construct the prediction model of student learning effect, which reflects the advantages and disadvantages of the teaching mode. Through experimental analysis, the accuracy of the improved K-modes algorithm in the model reaches 0.985. The recall rate reached 0.982; The average error of the prediction model is less than 1 in the error analysis. Therefore, the model accurately reflects the problems existing in the teaching model, and has a high prediction accuracy, indicating that the teaching evaluation model has a good evaluation effect.

Key words: MOOC; Blended Teaching; Teaching Evaluation; Neural Network; Co-Occurrence Rate; K-Modes

1. Introduction. In recent years, due to the problem of the novel coronavirus pneumonia, most schools have been faced with the state of suspension of classes, Massive open online courses (MOOCs) have developed rapidly under the policy of continuous suspension of classes and become the focus of the society [1]. The main reasons for this are changes in educational needs, concerns about teaching quality, and support from technology and data. The epidemic has led to the suspension or restriction of traditional face-to-face teaching, and many students and educational institutions have turned to online learning and distance education. MOOC, as a large-scale online open course, has attracted widespread attention as it can meet the educational needs of a large number of students in a short period of time. With the rapid development of MOOC, people have begun to attach importance to the teaching quality of online education. Due to the characteristics of large-scale, remote, and heterogeneous student participation in MOOC courses, ensuring the effectiveness and quality of teaching has become an important issue. Ensuring the quality of MOOC teaching is not only related to students' learning outcomes, but also to educational equity and social development. In the context of the COVID-19 epidemic, the support of technology and data provides more possibilities for evaluating and improving the quality of MOOC teaching. Technologies such as data mining and machine learning can be used to analyze students' learning behavior and outcomes, thereby improving teaching methods and personalized learning. Therefore, the focus on the quality of MOOC teaching also involves the application of data mining and technology.

In college English teaching, the blended learning model based on MOOC has been widely explored and applied to address some specific challenges and problems. This includes a lack of motivation for students to learn; Insufficient interactivity and feedback; Technical application and facility conditions; Evaluation and certification. By designing motivational learning tasks, providing real-time interaction and personalized feedback, improving technical conditions, and diversifying evaluation methods, this model can improve teaching quality, promote students' active learning and participation. To ensure the quality of college students' English learning in MOOC, it is extremely important to evaluate the teaching of MOOC [2]. The reason is that MOOC based teaching evaluation can provide feedback and improvement opportunities, ensure learning quality and effectiveness, improve teaching design and resource allocation, and promote student participation and learning motivation. Through teaching evaluation, we can comprehensively focus on and improve the quality of learning, providing a better English learning experience and learning outcomes.

*School of Foreign Languages, North China Institute of Aerospace Engineering, Langfang, 065000, China (JingboOHao@outlook.com)
†School of Foreign Languages, North China Institute of Aerospace Engineering, Langfang, 065000, China (xybwwl@163.com)

Teaching evaluation can provide diagnostic functions for teaching models, including examining students' learning progress and mastery level, identifying problems and challenges in the teaching process, collecting student feedback and opinions. Through this diagnostic information, targeted improvements can be made to improve the effectiveness and quality of teaching models. Teaching evaluation has a feedback function, which can promote the further development of students and teachers and improve the quality of teaching by providing feedback on students' learning outcomes, teaching effectiveness, providing suggestions for teaching improvement, and promoting students' self-reflection. Teaching evaluation can stimulate teachers' enthusiasm and help them improve teaching quality by providing positive incentives, providing opportunities for improvement, promoting professional development, and utilizing student feedback. Through continuous self reflection and improvement, teachers can continuously enhance their teaching abilities and professional qualities, providing students with better educational experiences and learning outcomes. Teaching evaluation has the guiding function, establish the notarized teaching evaluation system, help teachers to find the direction of struggle, guide teachers to shift the focus of work to the teaching task; Teaching evaluation has management function and is an important basis for teacher promotion and evaluation [3, 4]. Establishing a comprehensive teaching evaluation system can provide accurate evaluation criteria, emphasize the importance of data and evidence, supervise the improvement of teaching quality, promote innovation and research in teaching methods, and promote the sharing and exchange of teaching experience. It can promote the scientification of teaching models, improve the quality and effectiveness of teaching. Therefore, the research carried out in-depth research on the teaching evaluation system, improved the teaching evaluation model by using improved K-modes algorithm and neural network, and realized student clustering analysis mainly through this method to identify student groups with similar characteristics and Learning styles. These characteristics may include students' interests, learning styles, learning tendencies, etc. By grouping students into different clusters, teachers can better understand their needs and differences, thereby providing personalized teaching support for each student. The improved K-modes algorithm can be used to mine and analyze student feedback data. By clustering students' feedback information, it can be discovered what similarities students in different clusters have in their evaluations and needs for teaching modes. Based on these common points, key points and directions for improving teaching models can be extracted, thereby further improving teaching quality.

2. Related Work. The MOOC teaching mode has developed rapidly in the environment of the COVID-19. Facing the situation of full offline suspension, the MOOC teaching mode has become the focus of many scholars due to its openness, flexibility, cross regional and cross-cultural, diverse learning methods and resources, social and Cooperative learning, data analysis and teaching improvement. Duan T scholars used ISM and MICMAC models to analyze the factors that affect ideological and political courses. Firstly, they selected 10 main factors that affect the teaching effectiveness of MOOC courses. Secondly, we establish a Adjacency matrix to clarify the basic Binary relation between these factors, find the reachability matrix through exponential operation, and obtain a 5-level interpretive structure model. Thirdly, based on MICMAC analysis, new ideas are provided for teaching optimization based on MOOC. Finally, through a detailed discussion of the survey results, we have proposed some suggestions for optimizing the effectiveness of ideological and political education MOOCs teaching [5]. Researchers such as Min Y D conducted a survey on the influencing factors of MOOC teachers' work engagement, and analyzed teachers' openness and teaching self-efficacy through online surveys. The results show that through learning and growth opportunities, innovation and teaching improvement, professional identity and social support, and interaction with students, Openness to experience can enhance teachers' self-confidence and professional ability, improve their commitment to teaching, and continuously improve teaching quality. Therefore, this characteristic has a direct impact on the self-efficacy and work engagement of MOOC teachers, while teaching self-efficacy has an indirect impact on work engagement [6].

Charo R team analyzed self-regulated learning strategies and MOOC related variables, understood students' self-regulated learning ability through questionnaires, and analyzed the data through Logistic regression model. The results showed that students who completed their studies were more capable of learning self-regulation than those who did not finish their studies. At the same time, it shows higher perception rate and participation in MOOC content [7]. The mixed teaching mode combines the advantages of traditional teaching and network teaching in the "Internet plus" era, and has become one of the important trends in the development of higher education teaching. In order to comply with this development trend, Yan R and other researchers proposed a

hybrid teaching model based on MOOC resources and digital experimental teaching platforms. This teaching method combines the advantages of online teaching with the advantages of offline teaching, so that both teachers and students have a positive attitude in the teaching classroom. It is suggested to continue to increase the construction of MOOC courses and create a high-level hybrid teaching model [8].

Yu Y et al. introduced the platform design of MOOC using a software based virtual experience, combining virtual experiments with MOOC to achieve the concept of open sharing, effectively integrating teaching resources, and visually displaying the teaching content of MOOC courses. The software virtual experience provides practical opportunities and advantages such as immersive learning, non-linear learning and personalized paths, collaboration and communication, feedback and evaluation, as well as cross regional and cross-cultural learning, thereby improving the teaching effectiveness of the MOOC platform and making the relationship between teaching theory and practice closer [9]. Cross cultural teaching of college English has shifted from offline to online. Driven by the MOOC teaching model, online cross-cultural teaching of college English has exposed problems such as insufficient intelligence and poor online teaching effectiveness. In order to improve the efficiency of cross-cultural teaching in college English, scholars such as Xie H have improved traditional algorithms based on the teaching needs of MOOC and established relevant functional modules, which have been verified through control experiments. The experimental results indicate that the improved new model is more attractive for students' online learning, effectively improving the efficiency of cross-cultural teaching of college English, and addressing the shortcomings of traditional online teaching [10].

With the development of society, various teaching methods emerge. To guarantee the effect of teaching methods, the research of teaching evaluation system has become the focus of social attention. Caldwell K and other researchers developed the Chief Resident Teaching Evaluation and Assessment System to rate physician skill and teaching standardization through classroom observers and to analyze the physician's teaching experience. The developed teaching evaluation system effectively evaluated the teaching performance of teaching assistants as reflected by the results of teaching assistant assignments [11]. To explore the impact of teaching leadership on effective teacher teaching practices and learning outcomes, Kazi M scholars analyzed the value of teaching evaluation and used structural equation models to analyze teaching data. The data shows that students' performance has been improved in the perfect teaching evaluation, and teaching practice has also been improved in the perfect teaching evaluation [12].

The Tarraga Menguez R team established a theoretical framework through literature review, analyzed teaching evaluation data, and evaluated teachers' teaching abilities. The results show that the problems existing in the current teaching of teachers are accurately reflected in the framework, including teaching methods and teaching ability [13]. Fans et al. believe that evaluation must consider the aspects of teaching, learning, and educational background that are missing from digital data. Therefore, they advocate for the participation of different types of data outside of teaching in the teaching evaluation system, avoiding the transformation of teaching evaluation systems constructed using digital data into narrow instrumental education methods. Through the evaluation and judgment of the quality of educational practice, the proposed teaching evaluation system can better reflect the educational technology and educational purpose [14].

Researchers such as Yz A utilized inter relationships, evaluation texts, and existing "user project" format rating matrices to form a multi-source and multimodal data structure, and proposed a hybrid recommendation model that integrates network structural features, graphical neural networks, user interaction activities, and tensor decomposition. Firstly, a teaching evaluation network based on a graph structure is proposed, which analyzes teaching scores and comments. The extracted personalized features are used as the third dimension of the rating tensor. Finally, Bayesian probability tensor decomposition is used to predict course evaluation. Through experiments with real teaching data, the results show that this method has smaller prediction errors [15]. Antoci team analyzed the influence of social factors on the teaching evaluation system, and found that teaching evaluation and teachers' performance in subsequent courses had a direct impact on teacher rating, which could easily lead to the polarization of social results [16].

To sum up, the MOOC teaching method has been vigorously developed in the current environment, which has a significant impact on student learning. However, few of the current MOOC teaching methods have perfect teaching evaluation results. Therefore, this study combines traditional teaching with MOOC classroom to achieve blended college English teaching, and constructs a teaching evaluation system through K-modes and

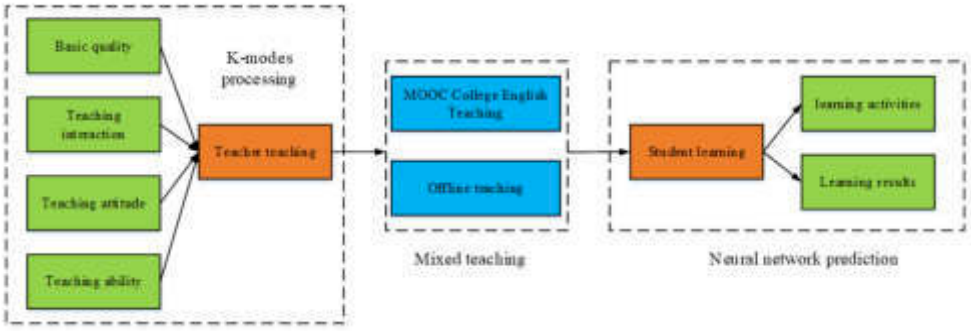


Fig. 3.1: MOOC based Mixed Teaching Evaluation Model for College English

neural networks to evaluate the proposed blended teaching, aiming to make the MOOC-based blended teaching model more perfect and scientific.

3. College English Blended teaching model and evaluation model construction based on MOOC.

3.1. Data selection of hybrid teaching model and evaluation model. At present, MOOC teaching methods lack a perfect teaching quality evaluation system, so it is impossible to make accurate and scientific judgment on the teaching quality of MOOC. Therefore, this study evaluates blended teaching by constructing a teaching evaluation model. The blended teaching model and teaching evaluation model are shown in Figure 3.1.

In Figure 3.1, the data generated by the hybrid teaching method are processed and analyzed by K-modes algorithm to obtain relevant evaluation indicators. The processed data will be sent to the neural network structure for learning and prediction, and the prediction effect will be compared with the actual situation to reflect the overall effect of the hybrid teaching method. In data selection and analysis, teaching behavior will produce more data information, and the information processing is difficult, so the model uses K-modes algorithm to process the data. The research adopts the MOOC teaching data of Y school for analysis, and sets four teaching evaluation indicators based on the evaluation of online teaching. The indicators include MOOC teachers' quality, MOOC teachers' teaching attitude, teachers' teaching methods and classroom interaction effects. Each evaluation index is divided into five grades, of which grade 0 is the lowest grade, and this grade indicates that the corresponding index fails; Level 4 is the highest level, which means that the corresponding indicators are excellent. The teaching data is mapped by setting evaluation indicators, but there are often abnormal data in the data. The elimination of abnormal data is an important link to ensure the accuracy of the evaluation model. The process of exception data elimination is shown in Figure 3.2.

In Figure 3.2, the classification of teaching evaluation is the first step of data cleaning. Its classification method can be classified according to student year, semester time, course selection number and teacher number. There are 1548 data samples of School Y used in the study. After the sample data is classified, the cosine distance similarity formula is used to calculate the samples to eliminate the wrong results caused by abnormal data. By indicator grade classification, the sample data has five dimensions, and the average dimension of the sample is . In the similar cosine distance formula, if the denominator is 0, it will lead to errors in the similarity calculation. To ensure the accuracy of the calculation, increase the value of 0.001 in the evaluation value to solve the problem of zero denominator. The dissimilarity formula after improving the denominator is shown in Equation 3.1.

$$Sim(X,Y) = \frac{\sum_{i=1}^{q}((x_i - p_x)\cdot(y_i - p_y))}{\sqrt{0.001^2}\cdot\sqrt{\sum_{i=1}^{q}(y_i - p_y)^2}} = \frac{X\cdot Y}{\|X\|\cdot\|Y\|} \tag{3.1}$$

In Equation 3.1, $Sim(X,Y)$ represents the difference between sample X and Y . The dimension of sample data is represented by q. The comparison of cosine distance similarity requires two samples, one of which is taken

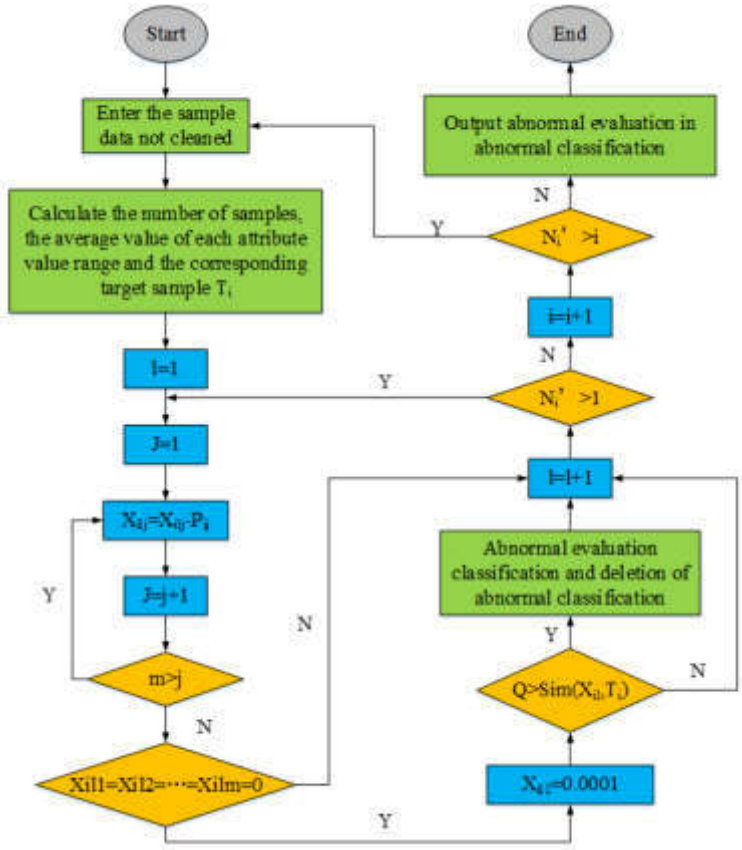


Fig. 3.2: Processing flow of abnormal data

as the target sample and the other is used as the sample of multidimensional data for comparison. According to the principle of center distance, the average value of the target sample in each dimension can be calculated. The calculation formula is shown in Equation 3.2.

$$T = \left(\frac{1}{N} \sum_{i=1}^{N} (x_{i1} - p_1), \frac{1}{N} \sum_{i=1}^{N} (x_{i2} - p_2), \ldots, \frac{1}{N} \sum_{i=1}^{N} (x_{iq} - p_q) \right) \tag{3.2}$$

In Equation 3.2, T represents the dimensional average of the target sample; N indicates the type of sample classification. The abnormal data can be eliminated by Equation 3.1 and Equation 3.2. After the data is processed abnormally, it needs to be standardized. Data standardization can be processed by min-max standardization method, which can shrink the data into the interval of [0,1]. The standardized data can be compared directly, so the min-max standardization is shown in Equation 3.3.

$$x_{ij} = (x_{ij} - min\ x_{ij} / max\ x_{ij} - min\ xij) \tag{3.3}$$

In Equation 3.3, $maxx_{ij}$ and $minx_{ij}$ represent extreme values; x_{ij} represents unstandardized raw data; x_{ij} represents the normalized data. After the data is standardized, the data in the same column need to be averaged and finally merged. The K-modes algorithm can be used to calculate the similarity of the merged data, as shown in Equation 3.4.

$$AVF(x_i) = \frac{1}{q} \sum_{j=1}^{q} f(x_{ij}) \tag{3.4}$$

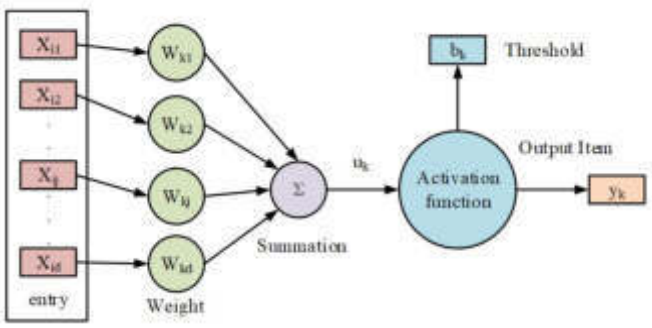


Fig. 3.3: Example of neuronal structure

In Equation 3.4, $AVF(x_i)$ represents the similarity based on frequency; $f(x_{ij}$ represents occurrences of the j attribute in the sample. When using K-modes algorithm to select the initial clustering center, problems such as unstable clustering results and influence of outlier data points on the availability of clustering results are likely to occur. Sum of Squared Error (SSE) can avoid such problems. The SSE expression is shown in Equation 3.5.

$$SSE = \sum_{l=1}^{k} \sum_{x \in L_l} \text{Dist}(x, Z_l)^2 \tag{3.5}$$

In Equation 3.5, SSE is used to determine the initial cluster center stably, where k represents cluster families. Z_l represents the clustering center of the item; $\text{Dist}(x, Z_l)$ represents the similarity between the data and the cluster center. In the same data, the value changes and value relationships of the same attributes and different attributes can affect the clustering effect, but this is not considered in the traditional K-modes algorithm. Therefore, the co-occurrence rate is used to improve the algorithm, and the distance measurement based on the co-occurrence rate is calculated as shown in Equation 3.6.

$$d(x, y) = \sum_{i=1}^{m} \sum_{j=1,\ldots,m, j \neq i} d_{ij}(x_{Ai}, y_{Ai}) \tag{3.6}$$

Co-occurrence rate is the probability that one thing will happen if another thing is certain to happen. In Equation 3.6, $d_{ij}(x_{Ai}, y_{Ai})$ represents the distance between an attribute in the sample data and another attribute. The smaller the distance, the greater the co-occurrence rate, indicating the higher the similarity of the samples. After the above analysis, the specific way to enhance K-modes through co-occurrence rate is as follows: firstly, the co-occurrence rate index is used to represent the degree of correlation between different discrete variables. Based on the characteristics and objectives of the data, an appropriate co-occurrence rate index is selected to measure the correlation between variables; Then, distance measurement is usually used to calculate sample similarity, thereby improving classification results.

3.2. Construction of student learning prediction model based on neural network. After the K-modes algorithm is used to process the teaching evaluation data and teaching data, the processed data is used to predict student learning [17, 18]. Firstly, the input data and output data of the model are determined. The second is to construct the network structure, including the network depth, the distribution form of neurons and the selection of excitation function. Finally, the training times of the model were adjusted [19, 20]. The output layer study takes K neurons as an example to construct a neural network, and the structural diagram of the neurons is shown in Figure 3.3.

The neuron structure in Figure 3.3 can be represented by mathematical expressions, as shown in Equa-

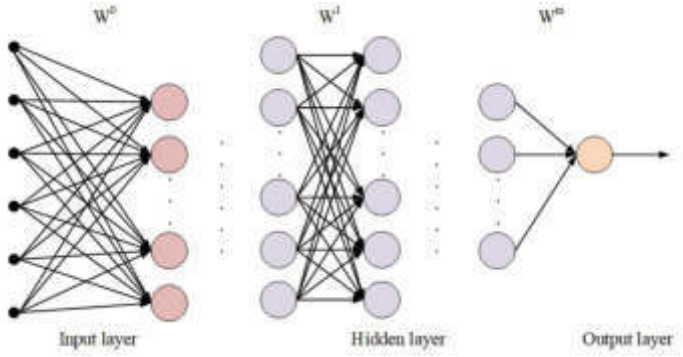


Fig. 3.4: Topological structure of neural network

tion 3.7.

$$\begin{cases} u_k = \sum_{i=1}^{n} w_{kj} a_{ij} \\ v_k = u_k + b_k \\ p_k = \varphi(v_k) \end{cases} \tag{3.7}$$

In Equation 3.7, a_{ij} represents the features extracted from the training sample, where i and j represent the number of the training sample and the sample features respectively. w_{kj} is the feature weight of the input data, k represents the input number; u_k indicates the weight relation between the input item and its corresponding weight. b_k represents the threshold in the structure; $\varphi(v_k)$ represents the excitation function of model selection; p_k is the output of the neuron. The topology structure of the constructed neural network is shown in Figure 3.4.

For feedforward neural network networks, there are only two types of neurons, one is the output unit, and the other is the calculation unit. For computing units, they can accept multiple different inputs, but due to the lack of feedback information, they can only have one output. However, this unique output can be coupled to any other unit as input, so for all layers other than the input room, the input is only related to the previous layer. The input and output layers are connected to peripherals, and the other layers in the middle are all hidden layers. When learning and training a network model, continuous testing can determine the appropriate number of layers, the number of neurons on each layer, the excitation function, and the number of training times. This model can then obtain the corresponding weights and ratings of each input item. In this way, we can input the learning behavior data of students' courses through this model to predict the students' Final examination scores, so as to predict their learning effects according to their daily learning behavior. Based on the above mathematical model, the network flow chart is shown in Figure 3.5.

In Figure 3.5, the input data is based on students' daily learning behavior and learning situation, such as class attendance, homework completion, online learning time and in-class test, etc. The input data is recorded as A_i . The final grade, which can best reflect the learning effect of students, is taken as the output data of the prediction model and denoted as B_i. The setting of network depth and the number of neurons has a direct influence on the model performance. The selection of excitation function can adjust the input weight of the model. The setting of training times can ensure that the weight value is adjusted to the best value and avoid over fitting. Assuming the network depth is m , the calculated values of each layer are shown in Equation 3.7.

If the output value of the first layer is H_k^1, the input value of the second layer is h_k^2, namely $h_k^2 = H_k^1$. The output calculation of the second layer is shown in Equation 3.8.

$$H_k^2 = \varphi(w_{kj}^2 h_k^2 + b_k^2) \tag{3.8}$$

It can be inferred that the input value of the m layer of the neural structure is the output value of the previous

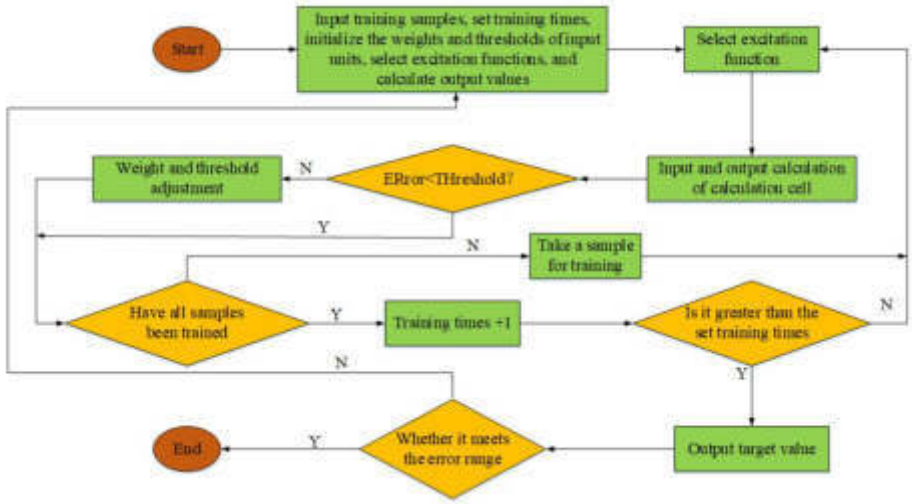


Fig. 3.5: The construction of neural network in student achievement prediction model

layer, namely $h_k^m 2 = H_k^m$, and the output expression of the m layer is shown in Equation 3.9.

$$H = \varphi\left(\sum_{j=1}^{N}(w_j^m h_j^m + b_j^m)\right) \tag{3.9}$$

During the training process, the prediction model can adjust the parameters of the network structure in real time to achieve the most appropriate value. Since the neural structure of the prediction model has only one output result and can accept multiple input information at the same time, the information is not the most useful feedback, so the model belongs to the feedforward neural network. As an important data in the teaching quality evaluation system, the smaller the error of the model in predicting students' performance, the better the system is. Now the excitation function is selected to optimize the constructed prediction model. Two kinds of functions are used as the excitation function of the model, in which the ReLu function is expressed as Equation 3.10.

$$Relu(x) = f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{3.10}$$

As a unilateral inhibition function, the ReLu function is 0 when its independent variable is not greater than 0. When the independent variable is greater than 0, the function value does not change at all. Such unilateral inhibition function makes the neural structure have sparse activation property in the model, so that the network can better mine the data features and ensure the fitting effect of the data. Another excitation function is Sigmoid function, whose expression is shown in Equation 3.11.

$$Sigmoid(x) = f(x) = \frac{1}{1 + e^{-x}} \tag{3.11}$$

The range of the function can be any value, but the range is in the interval (0,1). The image generated by the excitation function is continuous, smooth and derivable. In addition, when the domain value of the function is 0.5, the function is in central symmetry. The Sigmoid function has two properties that can describe the uncertainty of the decision with a smaller granularity. In the range close to the threshold, the convergence speed can be improved and the training times of the model can be reduced. The training times of the model

can be determined by superposition experiment. Therefore, the learning ability of the model can be reflected by Maximum Error (ME) and Cumulative Error (CE). The calculation of ME is shown in Equation 3.12.

$$ME = \begin{cases} \max(o_i' - o_i) \\ \min(o_i' - o_i) \end{cases} \tag{3.12}$$

The maximum errors of positive and negative values are respectively expressed in Equation 3.12. The calculation of CE is shown in Equation 3.13.

$$E = \begin{cases} \sum_{i=1}^{102}(o_i' - o_i) \\ \sum_{i=1}^{48}(o_i' - o_i) \end{cases} \tag{3.13}$$

Equation 3.13 respectively represents the cumulative errors of positive and negative values, and the corresponding average errors can be calculated through the cumulative errors. The error between the output value and the actual value is represented by . Predictability expressed by mean and variance. Then the average value is calculated as shown in Equation 3.14.

$$\text{Average error} = \begin{cases} \sum_{i=1}^{150} o_i'/150 \\ \sum_{i=1}^{150} o_i/150 \end{cases} \tag{3.14}$$

Formula 3.14 respectively shows the average error of the actual value and the output value. By comparing the difference between the output and the actual situation, the effect of the prediction model can be judged. The variance calculation of the model is shown in Equation 3.15.

$$\text{Variance} = \begin{cases} \frac{\sum_{i=1}^{150}\left(o_i' - \left(\frac{\sum_{i=1}^{150} o_i'}{150}\right)\right)^2}{150}/150 \\ \frac{\sum_{i=1}^{150}\left(o_i - \left(\frac{\sum_{i=1}^{150} o_i}{150}\right)\right)^2}{150}/150 \end{cases} \tag{3.15}$$

The variance of the actual value and the output value is respectively expressed in Formula (15). If there is a small error between the variance of the output value and the variance of the actual value, it indicates that the prediction ability of the model is stronger. However, the variance of the actual value is less than the variance of the output value, which indicates that the training samples of the model are insufficient or the training times are too few.

4. Performance analysis of blended College English teaching Model and evaluation model based on MOOC.

4.1. Performance analysis of K-modes algorithm in Blended College English Teaching evaluation Model. After the construction of the overall model is completed, the performance of the model is analyzed, and then the advantages and disadvantages of MOOC-based blended college English teaching are judged. The datasets used in the experiment were all from the actual teaching evaluation information of A school. Through organizing the data from A school, 2000 pieces of data were obtained as the experimental dataset, with the training and testing sets verified in a 9:1 ratio. Through the analysis of teachers' teaching behavior data, the model can judge the accuracy of students' learning effect improvement and finally reflect the feasibility of the teaching method. The research first analyzes the teaching evaluation data, reflecting the distribution of teaching quality through proportion, and then judges the classification performance of the model through three indicators: classification accuracy, recall, and error. Finally, the experiment adds different teaching groups and different prediction models for comparative analysis to verify the progressiveness of the proposed model. The accuracy, recall, and error calculations are shown in Equation 4.1.

$$\begin{cases} \text{Precision} = \frac{TP+TN}{TP+TN+FP+FN} \\ \text{Recall} = \frac{TP}{TP+FN} \\ \text{SSE} = \sum(y_i - \bar{y})^2 \end{cases} \tag{4.1}$$

Table 4.1: Statistics of teaching evaluation data of students in Y University in a semester

Evaluation grade	Basic quality	Teaching attitude	Teaching ability	Extracurricular links	Total	Proportion
Excellent	335	328	54	7	724	0.2233
Good	293	296	377	78	1044	0.3200
Secondary	182	182	328	500	1192	0.3677
Pass	1	16	43	191	251	0.0774
Fail	0	1	2	28	31	0.0096

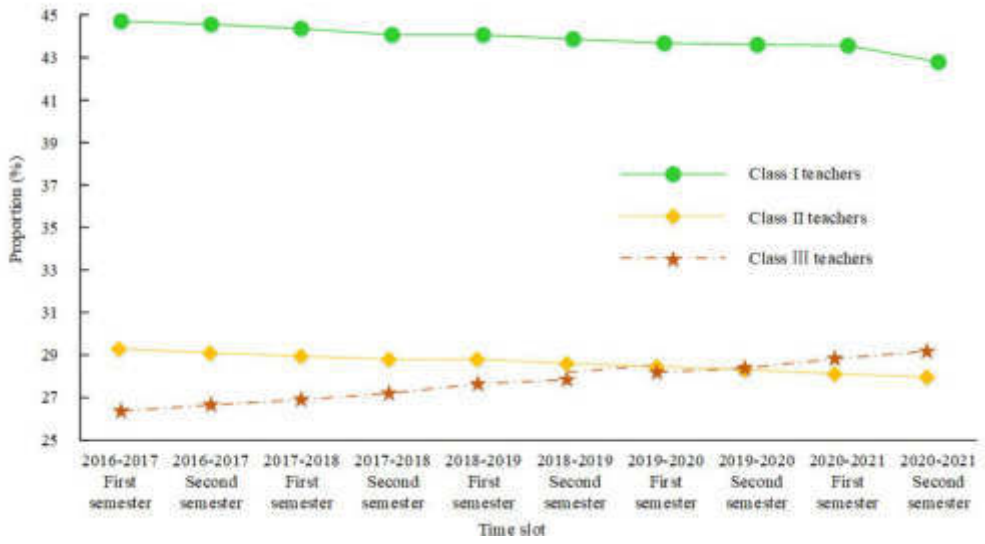
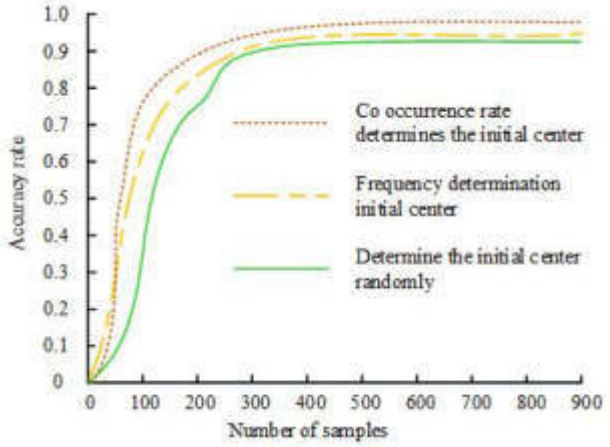


Fig. 4.1: Proportion of English teachers in different categories

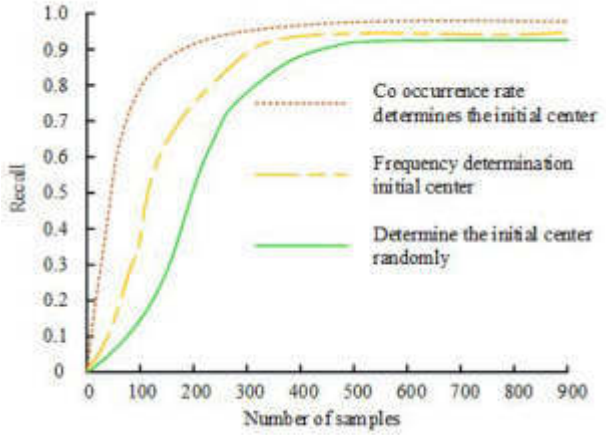
The study selected common teaching evaluation indicators in teaching quality evaluation through literature review, which are basic qualities, teaching attitude, teaching ability, and extracurricular activities [21]. At present, the data of teachers' teaching modes are collected and classified according to the indicators. The classified teachers are clustered by K-modes algorithm to analyze the existing problems in teaching. Finally, the performance of K-modes is compared.

In Table 4.1, teaching evaluation can be divided into four evaluation indicators. If three of the four evaluation indicators of teachers have good or above evaluation, they are regarded as first-class teachers, and the proportion of first-class English teachers is about 45%. If there is a medium evaluation, the teachers in this category are regarded as Class II teachers, and the proportion of Class II English teachers is about 28.5%. If the evaluation of the four indicators is below average, the teachers in this category are regarded as third-class teachers, and the third-class teachers account for about 26.5%.

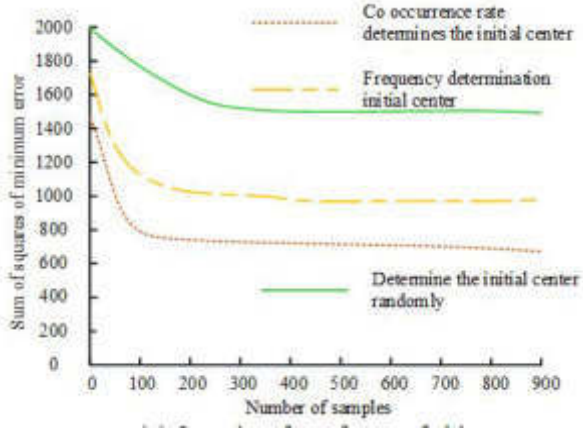
In recent years, the proportion of three types of English teachers in each semester through K-modes clustering algorithm is shown in Figure 4.1. Figure 4.1 and Table 4.1 are combined for analysis. Among the two evaluation indexes of teachers' basic quality and teaching attitude, about 76% are rated as excellent and good, and the remaining 24% are rated as medium, pass or fail. In the index of teachers' teaching ability, the number of teachers evaluated as excellent and good accounted for about 53%, and the total number of teachers evaluated as medium, pass and fail accounted for 47%. Among the indicators of teacher interaction, only about 10% are rated as excellent or good, 65% are rated as medium, and 25% are rated as pass or fail. Therefore, through the improved K-modes algorithm, shortcomings in blended teaching can be clearly found. Nearly half of the students feel inadequate about the teaching ability of MOOC teachers, and 90% of the students believe that MOOC teachers lack interactive links in teaching.



((a)) Comparison of the accuracy of three algorithms



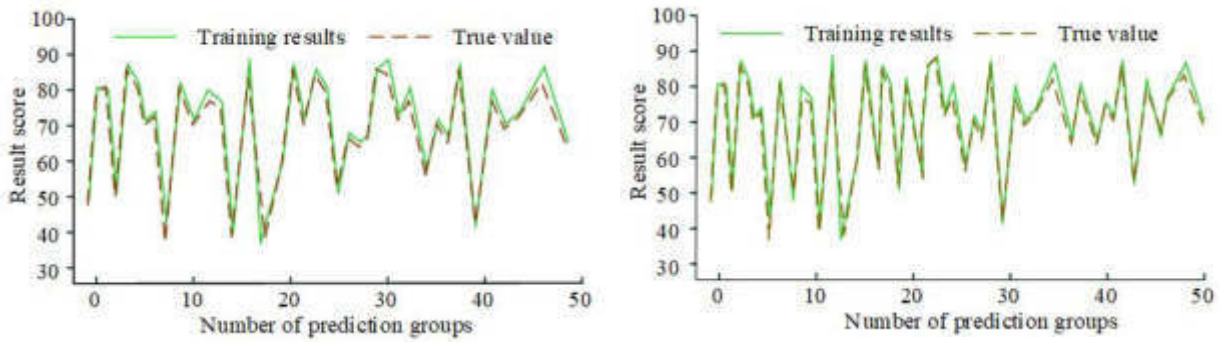
((b)) Comparison of recall rate of three algorithms



((c)) Comparison of sum of squares of minimum error of three algorithms

Fig. 4.3: Performance Results of Different Clustering Algorithms

In K-modes algorithm, three methods are used to test the initial clustering center. Figure 4.3 shows the comparison of the three methods. In Figure 4.2(a), with the same number of samples, the clustering center was determined by the co-occurrence rate, and the classification exactitude of the model reached 0.985. The prediction accuracy of the model is 0.936 when the clustering center is determined randomly. In Figure 4.2(b), the model of co-occurrence rate is adopted. When the number of samples is around 100, the maximum recall rate is reached, and the recall rate is 0.982. For the model whose clustering center is determined by frequency, its maximum recall rate is about 350 samples, and its recall rate is 0.925. The model of clustering center was determined by random method, and the maximum recall rate was reached only when the number of samples was 500. Its recall rate was 0.887. In Figure 4.2(c), the minimum error sum of squares (MESS) of the improved K-modes algorithm is 725. The MESS is 1022 for the model whose clustering center is determined by frequency. The MESS is 1526 for the model that randomly determines the cluster center. Therefore, the improved K-modes algorithm adopted in this study has strong learning ability and good performance.



((a)) Same Course and Same Teacher

((b)) Different Teachers for Different Courses

Fig. 4.5: Prediction effect of the model on two groups of different data

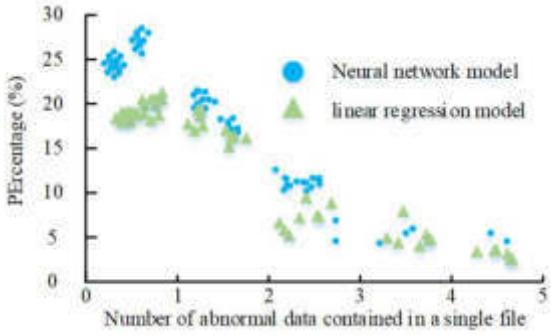


Fig. 4.6: Distribution of two groups of error prediction

4.2. Performance analysis of neural network in Blended College English teaching evaluation model. The improved K-modes algorithm has good data processing effect, which has been effectively verified in the experiment, and the predictive ability of the teaching evaluation model is analyzed. The effectiveness of the neural network model (NNM) was tested through two sets of different data, the performance of the NNM was reflected through the error distribution, and then the performance was compared with the linear regression model.

Two sets of data were used to analyze the performance of the NNM, and the specific results are shown in Figure 4.5. In Figure 4.4(a), this set of data is tested under the condition of the same course and the same teacher, and the predicted result overlaps highly with the real value, indicating that the model has a high prediction accuracy. The maximum error and cumulative error can be calculated by combining Equation3.12 and Equation3.13, and the positive value of the maximum error is 6.04. The maximum negative error is -5.82; The cumulative positive error is 1.87; The cumulative negative error is -2.26. In Figure 4.4(b), this set of data is tested without different courses and teachers, and its prediction effect is also highly overlapped with the real value. The maximum positive and negative errors obtained by the formula are 6.57 and -5.86. The cumulative error is 1.93 and -2.27.

The absolute error results of the two groups of data obtained by formula (14) are shown in Figure 4.6. In Figure 4.6, the absolute error of the two groups of data in the interval [0,2] accounts for the highest proportion, and the actual average of the first group is 72.15, the predicted average is 73.13, and its average error is 0.98. The actual mean of the second group was 73.06, the predicted mean was 73.77, and its mean error was 0.71.

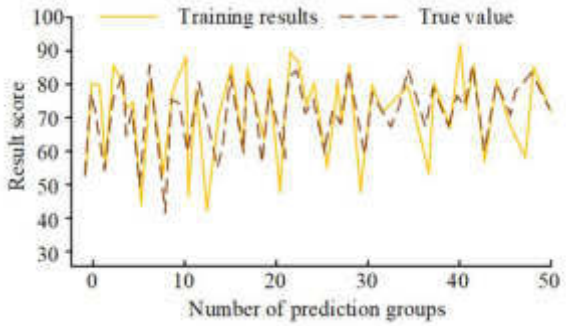


Fig. 4.7: Prediction results of linear regression model on data

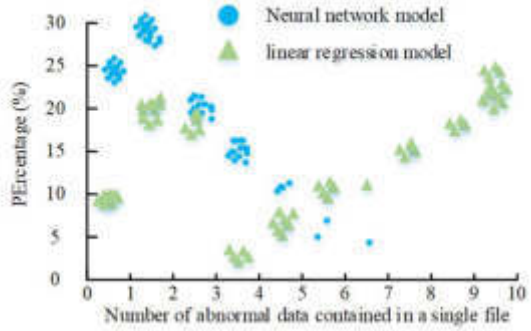


Fig. 4.8: Error distribution of linear regression model

There is a small difference in the calculated values of each index, and the prediction results of the model are more accurate under the teaching environment of different courses and different teachers. The results show that the prediction model has better prediction accuracy and stronger adaptability.

The traditional regression model is introduced to predict and analyze the learning effect of students, and the prediction effect is shown in Figure 4.7. In Figure 4.7, there is a large error between the predicted results of linear regression model and the real results, and the coincidence degree between the predicted linear and the real linear is low. Equation 3.12 and Equation 3.13 were also used to calculate the error values of the model, where the positive and negative error values were 22.46 and -13.19 respectively. The cumulative error is 6.84 and -3.54. Compared with the neural network prediction model, the prediction error of traditional linear regression model is larger.

The error distribution of the linear regression model is shown in Figure 4.8. In Figure 4.8, the linear regression model has a large error range, and the error value is relatively heavy in the interval [5,10]. The errors of the NNM are mainly distributed in the interval range of [0,3]. Therefore, neural network plays a huge role in the blended English teaching evaluation model based on MOOC classroom. The proposed evaluation model can accurately predict the learning effect of students. The experiment shows that the evaluation indexes selected by K-modes algorithm have an important influence on the teaching mode. Based on the above results, for teaching evaluation models, neural network models are usually able to more accurately predict students' learning outcomes and teaching quality. Compared to traditional linear regression models, neural network models can better capture nonlinear relationships and complex patterns, thereby improving the accuracy of prediction. Neural network models can perform end-to-end learning and representation learning on data, automatically extracting and learning features, thereby better reflecting the performance of evaluation models. By learning

hidden patterns and association relationships in data, neural network models can comprehensively consider the complex relationships between multiple evaluation indicators, thereby improving the accuracy and robustness of the evaluation model. The neural network model also has strong generalization ability and can handle data with noise and incomplete information. For evaluation models in the field of education, neural network models can model and predict students' learning processes and outcomes, thereby more comprehensively evaluating teaching quality.

5. Conclusion. In the context of COVID-19, MOOCs have seen rapid growth. To explore the effect of the blended college English teaching model based on MOOC, this paper uses data mining technology and deep learning to build teaching evaluation model. The improved K-modes algorithm is used to analyze the teaching methods of MOOC teachers, and the main evaluation indexes of the mixed teaching mode are obtained. The evaluation model is verified by experiments. In the experiment, the correct rate of the improved K-modes model in the experiment is 0.985. The recall rate was 0.982; The minimum error sum of squares is 725, and its performance is significant due to other algorithm models. The maximum positive error of the prediction model constructed by the neural network is 6.04. The maximum negative error is -5.80; The average error of positive and negative values is 1.87 and -2.26 respectively, and the prediction accuracy is that the actual results have high coincidence. Compared with the traditional linear regression model, the neural network prediction model has higher prediction accuracy and can better reflect the performance of the evaluation model. This paper establishes a teaching evaluation model, extracts the evaluation indicators of the MOOC hybrid college English teaching model, and obtains the accurate evaluation effect through the prediction model.

In the context of the COVID-19 and the rapid growth of MOOC, the results of this study have important significance and development contributions. First, it can adapt to the epidemic situation and future education development. The COVID-19 has had a great impact on the traditional face-to-face teaching method. The research results of the MOOC based blended teaching model indicate that this model can provide a way for the education field to respond to the epidemic and future education development. By combining online learning and face-to-face teaching, this model can continue to provide high-quality college English education and meet students' personalized learning needs. The second is to improve teaching quality and student engagement. The blended teaching model can provide more teaching resources, interactive opportunities, and learning support through online platforms. This helps to improve teaching quality and student engagement. Compared to traditional linear regression models, research results indicate that teaching quality evaluation and prediction models can more accurately reflect students' learning outcomes and teaching outcomes, thereby further improving teaching quality. The third is to promote the future development of online education, and the research on MOOC blended teaching mode has provided beneficial contributions to the future development of online education. By combining online education platforms with traditional teaching methods, the blended teaching model can balance its advantages, provide students with a more flexible and diverse learning experience, and meet personalized learning needs. In addition, the research results also provide a basis for evaluating and predicting the quality of teaching in online education, helping to develop strategies and measures to improve and optimize online education.

However, there are still shortcomings in the research. The teaching evaluation indicators used in the model are relatively broad and prone to incorrect evaluation methods. Therefore, more detailed classification of evaluation indicators can be carried out, which can be refined from multiple perspectives such as students' learning outcomes, coverage of course content, student evaluation and participation. More specific and quantifiable evaluation indicators can be determined through in-depth understanding of relevant research, disciplinary characteristics, and educational practices. Subsequent research can also further optimize the structure of the prediction model, considering the use of Transfer learning technology, through pre trained models in other fields or tasks, we can extract more abundant feature representations, and further improve the performance of the prediction model.

REFERENCES

[1] Lebedeva, M. Instructional Design of Skill-Balanced LMOOC: a Case of the Russian Language MOOC for Beginners. *Journal OF Universal Computer Science*. **27**, 485-497 (2021)

[2] Okoye, K., Arrona-Palacios, A., Camacho-Zuiga, C., Jag, A., Escamilla, J. & Hosseini, S. Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Education And Information Technologies*. **27**, 3891-3933 (2022)
[3] Sun, Q. Evaluation model of classroom teaching quality based on improved RVM algorithm and knowledge recommendation. *Journal Of Intelligent And Fuzzy Systems*. **40**, 2457-2467 (2021)
[4] Luo, Y., Zhao, X. & Qiu, Y. Evaluation model of art internal auxiliary teaching quality based on artificial intelligence under the influence of COVID-19. *Journal Of Intelligent And Fuzzy Systems*. **39**, 8713-8721 (2020)
[5] Duan, T. A new idea for the optimization of MOOC-based teaching. *Education And Information Technologies*. **27**, 3623-3650 (2022)
[6] Min, Y., Zhu, M., Bonk, C. & Tang, Y. The effects of openness, altruism and instructional self-efficacy on work engagement of MOOC instructors. *British Journal Of Educational Technology*. **51**, 743-760 (2020)
[7] Charo, R., Maite, A. & Guillermo, M. Self-regulation of learning and MOOC retention. *Computers In Human Behavior*. **6423**, 1-10642 (2020)
[8] Yan, R., Wang, L., Liu, L., Li, X. & Liu, H. A preliminary study on the mixed teaching of human parasitology based on MOOC resources and the experimental teaching digital platform. *Chinese Journal Of Schistosomiasis Control*. **33**, 74-78 (2021)
[9] Yu, Y., Li, F., Zhao, S. & Liu, H. Virtual experiment method for MOOC to solve teaching practice skills and difficult points. *Mechatronic Systems And Control (formerly Control And Intelligent Systems)*. **47**, 77-82 (2019)
[10] Xie, H. & Mai, Q. College English cross-cultural teaching based on cloud computing MOOC platform and artificial intelligence. *Journal Of Intelligent And Fuzzy Systems*. **40**, 1-11 (2020)
[11] Caldwell, K., Hess, A., Kramer, J., Wise, P., Awad, M. & Klingensmith, M. Evaluating chief resident readiness for the teaching assistant role: The Teaching Evaluation assessment of the chief resident (TEACh-R) instrument. *American Journal Of Surgery*. **222**, 1112-1119 (2021)
[12] Kazi, M. Instructional leadership: teaching evaluation as a key element for 6th grade student's achievement in mathematics. *International Journal Of Educational Management*. **35**, 1191-1204 (2021)
[13] Tarraga-Minguez, R., Suarez-Guerrero, C. & Sanz-Cervera, P. Digital Teaching Competence Evaluation of Pre-Service Teachers in Spain: A Review Study. *IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje: IEEE-RITA*. **16**, 70-76 (2021)
[14] Fawns, T., Aitken, G. & Jones, D. Ecological Teaching Evaluation vs the Datafication of Quality: Understanding Education with, and Around, Data. *Postdigital Science And Education*. **3**, 65-82 (2021)
[15] Yz, A., Hao, L., Ping, Q., A, K., B, J. & Znac, D. Heterogeneous teaching evaluation network based offline course recommendation with graph learning and tensor factorization - ScienceDirect. *Neurocomputing*. **415** pp. 84-95 (2020)
[16] Antoci, A., Brunetti, I., Sacco, P. & Sodini, M. Student evaluation of teaching, social influence dynamics, and teachers' choices: An evolutionary model. *Journal Of Evolutionary Economics*. **31**, 325-348 (2021)
[17] Suryanarayana, G., Lnc, P., Mahesh, P. & Bhaskar, T. Novel dynamic k-modes clustering of categorical and non categorical dataset with optimized genetic algorithm based feature selection. *Multimedia Tools And Applications*. **81**, 24399-24418 (2022)
[18] Yuan, F., Yang, Y. & Yuan, T. A dissimilarity measure for mixed nominal and ordinal attribute data in k-modes algorithm. *Applied Intelligence*. **50**, 1498-1509 (2020)
[19] Abdulkarim, S. & Engelbrecht, A. Time series forecasting with feedforward neural networks trained using particle swarm optimizers for dynamic environments. *Neural Computing And Applications*. **33**, 2667-2683 (2021)
[20] Chen, Z. & Cao, F. Construction of feedforward neural networks with simple architectures and approximation abilities. *Mathematical Methods In The Applied Sciences*. **44**, 1788-1795 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 18, 2023
Accepted: Oct 9, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 451–463, DOI 10.12694/scpe.v25i1.2333

RESEARCH ON THE APPLICATION OF INTELLIGENT GRADING METHOD BASED ON IMPROVED ML ALGORITHM IN SUSTAINABLE ENGLISH EDUCATION

LEI HUANG* AND LI MA†

Abstract. The current intelligent grading methods in English education have issues of feedback lag and incompatibility with manual grading, and manual grading is easily influenced by subjective consciousness. Based on this, this research selects Adaboost algorithm in ML algorithm to realize intelligent grading. The experiment improves the Adaboost algorithm according to the actual needs, constructs the Adaboost/CT algorithm, and verifies its effectiveness. The experimental results show that in the English intelligent scoring module, the adjacency accuracy of the Adaboost/CT algorithm for high- and low-quality English is 95.33%; 95.45% in the middle score; 94% in the breakdown. The comparison between Adaboost/CT model and DFA method shows that the accuracy and proximity accuracy of Adaboost/CT are 79.66% and 94% respectively, which are much higher than 55% and 92% of DFA. In addition, compared with Adaboost, the accuracy of Adaboost/CT is also significantly better than Adaboost. In practical application, the use of Adaboost/CT in the evaluation of English compositions can not only get more accurate scores, but also find out the shortcomings of each student, so as to improve them pertinently. Meanwhile, the accuracy, recall, and F1 values of the Adaboost/CT algorithm are 96%, 95%, and 95%, respectively, which are higher than those of the comparison algorithm. Overall, the improved Adaboost/CT algorithm has shown high effectiveness and practicality, and has high applicability and effectiveness in practical intelligent grading.

Key words: Machine learning; Intelligent marking; Sustainable English education; Adaboost/CT

1. Introduction. The intelligent development of computers has promoted the gradual prevalence of "machine marking". Machine grading can not only overcome the low efficiency of manual grading, but also effectively avoid the negative impact of teachers' subjective factors in the grading process [1]. In the context of sustainable English education, manual grading will not only add unnecessary burden to the grading teachers, but also bring some adverse effects. Therefore, intelligent grading is gradually widely used [2]. Zhou built an automatic composition scoring system under the mixed mode of English writing on the basis of machine learning (ML) for the related problems of manual grading [3]. Zhang et al. started with the automatic scoring of semi-open short answer questions, combined with general information and specific information in relevant fields, built an automatic scoring model for semi-open short answer questions [4]. Fang designed a scoring system for oral English intelligence based on dynamic time warping algorithm to solve the problem of low accuracy of oral English intelligence scoring [5]. In this context, this study selected Adaptive Boosting (Adaptive Boost) algorithm by analyzing natural language processing and ML, and introduced the concept of Centralized Trend (CT) to improve it, and constructed the Adaptive Boost/CT algorithm. The Adaboost/CT algorithm is mainly used to solve the problems of low efficiency and insufficient fairness in English manual grading, aiming to help teachers find effective ways to improve students' learning performance, and also provide reference for the development of sustainable English education.

2. Related Work. At present, the grading of objective questions in English has gradually matured, but there is no clear unified answer for the intelligent grading method of subjective questions [6]. At the same time, due to the low efficiency of manual grading, students' writing ability has not been improved, and the intelligent reading of English compositions has become a hot topic [7]. Aimed at the problems of high time consuming and low reliability of manual grading, Ramesh et al. made a comprehensive study on the automatic grading system and proposed a ML technology for intelligent grading. This technology provides help to judge student performance and improve student performance [8]. Liu proposed a hybrid scoring model for intelligent scoring

*General Education School, Chongqing Youth Vocational & Technical College, Chongqing, 400712, China (Corresponding Author: huangsweet@163.com)
†School of Artificial Intelligence, Chongqing Youth Vocational & Technical College, Chongqing, 400712, China

in English writing courses by combining real-time feedback from machines with manual evaluation. This model not only improves the accuracy of grading, but also improves student performance [9]. In order to evaluate learners' English level, Gaillat et al. built an intelligent grading system for English compositions based on the supervised learning method. The proposed system not only improves the accuracy of assessment, but also provides help to truly master learners' English level [10]. Liu et al. introduced convolutional neural network as an assistant in English automatic grading, aiming at the problem that the subjective consciousness of the rater will have a negative impact on the final grading. Finally, the experiment proposed a new automatic English composition scoring system, which provided help for the improvement of foreign students' English level [11]. In order to improve the spoken English ability of English learners, Wu et al. proposed a new oral English scoring method based on computer neural network. This method improves students' oral English ability and also promotes the standardization of students' oral English pronunciation [12].

In addition, Phophohuangpairoj et al. proposed a new method of English article grading and feedback by using automatic text analysis to solve the problem of poor writing performance of students when learning English. The proposed method can help teachers work effectively and also improve English learners' writing level [1]. Taskiran et al. conducted an in-depth study on students in the open education department of a university in Türkiye in order to verify the impact of intelligent marking on students' writing ability. The experiment verified the positive impact of intelligent grading, and also provided help for improving students' performance [14]. In order to solve the problem that the current intelligent scoring cannot achieve subjective scoring, Miao et al. proposed an intelligent scoring model for English subjective questions based on deep neural network and linear regression. This model can improve the accuracy of subjective scoring and also provide direction for the training system of English skills [15].

From the research of scholars at home and abroad, it can be seen that the current English intelligent scoring method is not very mature, and the research on natural language feedback is not deep enough. In the current intelligent scoring system built by intelligent scoring method, the connection between scoring module and feedback module is not smooth enough. Therefore, this research innovatively proposes the Adaboost/CT algorithm, which makes the English intelligent scoring model have a deeper connection between the two parts of scoring and feedback. In addition, another innovation of the study is the construction of the comment model, which can achieve the real meaning of promoting learning by evaluation.

3. Research on English intelligent grading method based on ML algorithm.

3.1. Research on natural language processing and intelligent grading technology. In order to improve the efficiency and fairness in the assessment of sustainable English education, Adaboost algorithm is selected in the ML algorithm. In the experiment, the concept of CT is introduced into the actual English intelligent grading to improve Adaboost, and the Adaboost/CT algorithm is obtained. As an important branch of artificial intelligence, natural language processing occupies a large proportion in computer science. Therefore, in order to understand machine learning, we need to fully understand natural language processing. Natural language processing is a discipline integrating linguistics, computer and mathematics. This discipline not only studies natural language, but also provides help for the development of computer systems capable of efficient natural language communication [16]. In the statistical language model of natural language processing, the relevant sequence model of words or sentences in the English corpus is essentially a probability model. The probability calculation expression of the word string in the complete sentence is shown in equation 3.1.

$$P(T) = P(r_1, r_2, r_3, , r_n) \tag{3.1}$$

In equation 3.1, $P(T)$ represents the probability of occurrence of English word strings in sentence T; r indicates the English words that make up the sentence; indicates the number of English words. In addition, you can also choose to use the chain rule inside the conditional probability to decompose the probability of the word string. Therefore, the probability calculation expression obtained by expanding equation 3.2 is shown in equation 3.2.

$$P(r_1, r_2, r_3, \ldots, r_n) = P(r_1) \cdot P(r_2 \,|\, r_1) \cdot P(r_3 \,|\, r_1, r_2) \cdot \ldots \cdot P(r_n \,|\, r_1, r_2, r_3, \ldots, r_{n-1}) \tag{3.2}$$

In equation 3.2, $P(r_1)$ represents the probability of the occurrence of the first word r_1 in the sentence; $P(r_1|r_2)$ indicates the probability of the occurrence of the second word on the basis that the first word is

known. Based on equation 3.1 and equation 3.2, the simplified calculation expression is shown in equation 3.3.

$$P(T) = P(r_1) \cdot P(r_2 \,|\, r_1) \cdot P(r_3 \,|\, r_2) \cdot \ldots \cdot P(r_n \,|\, r_{n-1}) \tag{3.3}$$

Equation 3.3 is essentially the relevant calculation expression of the binary grammar model. Suppose a is used to represent the beginning of a sentence. If you want to calculate the possibility of "Yesterday was a bad day", just multiply the probability of binary grammar of two adjacent words. Therefore, the model probability calculation expression of the sentence is shown in equation 3.4.

$$\begin{aligned}(Yesterday\,\text{was a bad day}) = {} & P(Yesterday \,|\, \langle a \rangle) \cdot P(was \,|\, Yesterday) \\ & \cdot P(a \,|\, was) \cdot P(bad \,|\, a) \cdot P(day \,|\, bad)\end{aligned} \tag{3.4}$$

On the basis of equation 3.4 the research first determines the corpus of English training, and according to the corpus, the number of occurrences of a binary grammar can be obtained. Normalize it on this basis. The calculation expression of conditional probability obtained is shown in equation 3.5.

$$P(r_n \,|\, r_{n-1}) = \frac{C(r_{n-1} r_n)}{\sqrt{C(r_{n-1} r)}} \tag{3.5}$$

In equation 3.5, C represents counting. According to equation 3.5, the model of multiple grammar can be deduced. For general multivariate grammar, the calculation expression of its parameter estimation is shown in equation 3.6.

$$P(r_n \,|\, r_{n-N+1}^{n-1}) = \frac{C(r_{n-N+1}^{n-1} r_n)}{C(r_{n-N+1}^{n-1})} \tag{3.6}$$

In equation 3.6, N represents the total number of words. Correspondingly, when the number of English words is more than two, the sentence probability calculation expression is shown in equation 3.7.

$$P(T) = \prod_{k=1}^{k+1} P(r_k \,|\, r_{k-n+1}^{k}) \tag{3.7}$$

In equation 3.7, k represents the serial number of the word, and the maximum is n. In addition, in the statistical analysis model of natural language processing, correlation analysis plays an important role in practical problems. According to the background of sustainable English education, the Pearson correlation coefficient is mainly summarized here. This coefficient mainly represents the linear correlation coefficient between variables, and its calculation expression is shown in equation 3.8.

$$c = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{3.8}$$

In equation 3.8, c represents correlation; x_i and y_i represent the values of the two under the sample; $\bar{x}$ And $\bar{x}$ represent the sample mean of both. In addition, in the manual grading of English, the composition grading standard is one of the main bases. The reviewers evaluate the quality of an English composition according to the general requirements of the grading criteria, and then determine an appropriate score through the grading criteria. The rules for grading CET-6 composition in the conventional sense are shown in Figure 3.1.

As can be seen from Figure 3.1, the grading rules for English compositions in the conventional sense mainly divide the grades of compositions into six levels. Each level is divided according to whether the composition is relevant to the topic, whether the thought is clearly expressed, whether the language is wrong, and whether the semantics are coherent. Therefore, it is necessary to convert the scoring elements in the scoring rules into characteristic indicators for intelligent English grading under sustainable education. The research proposes a three-level indicator system based on actual needs, and its contents are shown in Figure 3.2.

It can be seen from Figure 3.2 that the research has established a new evaluation index system on the basis of comprehensive consideration of some manual scoring standards, composition scoring rules at home and

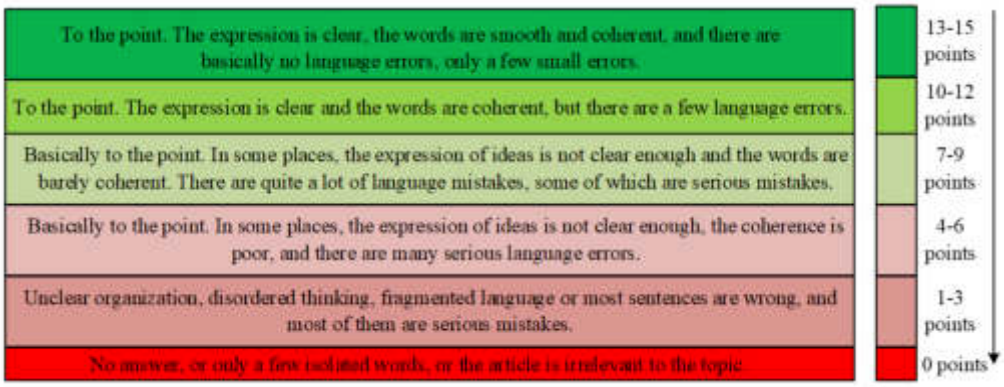


Fig. 3.1: Detailed Rules for Grading CET-6 English Composition

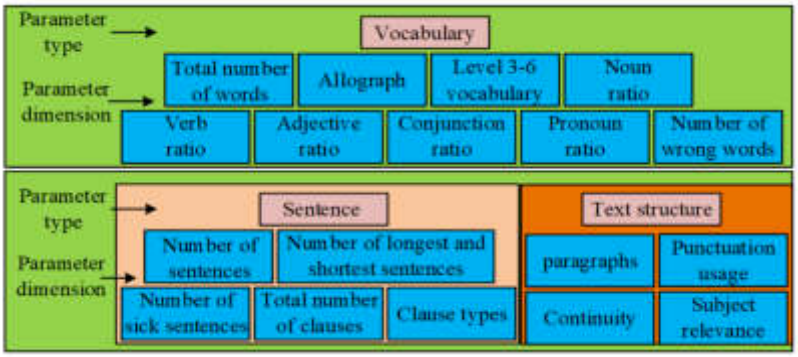


Fig. 3.2: Three-layer index system considering manual evaluation criteria

abroad and more mature scoring standards abroad. Among them, the characteristic index is calculated and classified from hard to soft, from large to small, and from shallow judgment to specific score. In addition, feature indicators also include lexical, syntactic and textual feature indicators. These indicators are based on the level of words, the total number of sentences, and the coherence of the text as the basic elements of the characteristic indicators.

3.2. Adaboost algorithm based on ML and its improved algorithm analysis. ML is interdisciplinary research, including probability theory, statistics, approximation theory, convex analysis, algorithm complexity, etc. It is specialized in studying how computers imitate or implement human learning behaviors to acquire new knowledge and skills. ML can reorganize the existing knowledge to continuously improve its own capabilities [17]. As an iterative algorithm in ML, the core idea of Adaboost algorithm is to train different classifiers (weak classifiers) in the same training set and combine them to form a stronger classifier (strong classifier) [18]. The main advantage of the Adaboost algorithm is that there are no requirements for the actual design of weak classifiers in the provided framework, so various methods can be used to construct weak classifiers without the need for prior knowledge of relevant experience. At the same time, its performance requirements for weak classifiers are not high, and the algorithm application is relatively simple. It does not need to be used for feature filtering, nor does it need to worry about overfitting. Therefore, it is studied as the basic algorithm for intelligent grading. However, when the traditional Adaboost algorithm is applied to English intelligent grading, the weak classifier is prone to repeat and error. Therefore, in order to remedy this

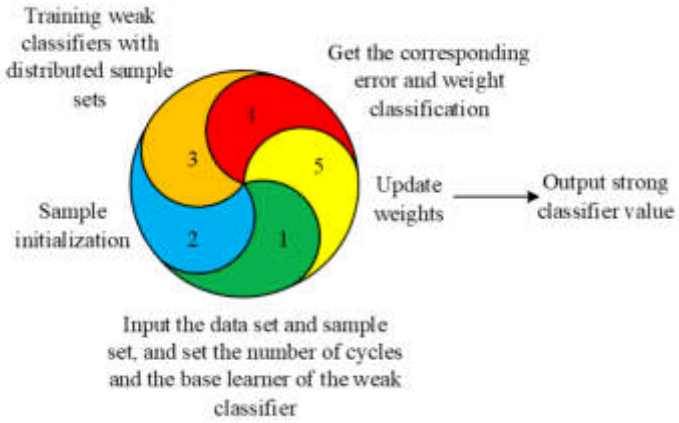


Fig. 3.3: Flow diagram of Adaboost/CT algorithm

defect, the experiment introduced the concept of CT and proposed an improved Adaboost algorithm, namely Adaboost/CT algorithm.

CT is a statistical concept, which represents a group of data close to a central value. At the same time, it also reflects the center of a data set [19]. The representative values of CT are numerical average and position average. In addition, the main factors reflecting CT include mean, central and median. It is worth noting that extreme data will break the balance of a group of data, so the average value does not represent the average level. The flow of the improved Adaboost/CT algorithm is shown in Figure 3.3.

From Figure 3.3, the Adaboost/CT algorithm process first inputs the corresponding dataset and sample set, and sets the number of cycles and the base learner of the weak classifier; The second step is to initialize the sample weights and use them to train weak classifiers using distributed sample sets, while using distributed sample sets to train base learners in the dataset. Then calculate the corresponding error and weight classification, and update the weights, that is, update the distribution; Finally, output the strong classifier value. Using Adaboost/CT algorithm can better avoid the over-fitting of traditional Adaboost, and can also effectively delete extreme data, thus solving the trap of stacking errors of weak classifiers. In the process, the equation expression of the dataset is shown in equation 3.9.

$$R = \{B_1, B_2, \ldots, B_K\} \tag{3.9}$$

In equation 3.9, R represents the input data set; B represents the elements in the data set, and the expression of their subscript related equations is shown in equation 3.10.

$$v = \{1, 2, \ldots, K\} \tag{3.10}$$

In equation 3.10, v represents the number of cycles, and its maximum value is K. In addition, the equation expression of the sample set is shown in equation 3.11.

$$B_v = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \tag{3.11}$$

In equation 3.11, x and y represent the characteristic value and result value of the data set samples respectively, and the maximum number of samples of the two is m . At the same time, in the Adaboost/CT algorithm flow, the calculation expression of error is shown in equation 3.12.

$$\lambda_v = \Pr_x \sim B_{(v,y)} I \tag{3.12}$$

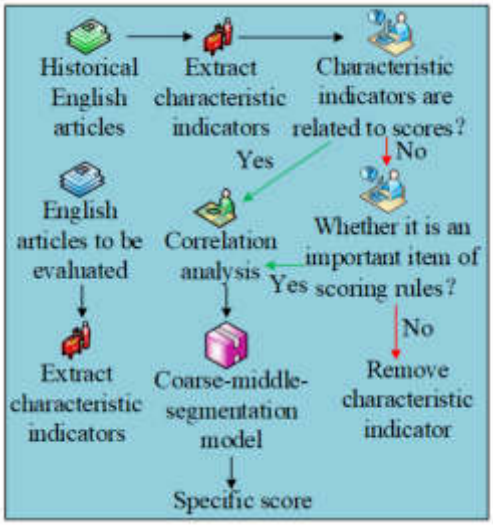


Fig. 3.4: Overall flow diagram of scoring model

In equation 3.12, λ represents error; I indicates the indicator function, which is 1 when it is correct and 0 when it is wrong. The calculation expression of weight classification is shown in equation 3.13.

$$\omega_v = \frac{1}{2}\ln\left(\frac{1-\lambda_v}{\lambda_v}\right) \tag{3.13}$$

In equation 3.13, ω represents the weight classification. Based on equation 3.13, the calculation expression of weight update is shown in equation 3.14.

$$B_{v+1}(i) = \frac{B_v(i)\exp\left(-\omega_v y_i h_v(x_i)\right)}{Z_v} \tag{3.14}$$

In equation 3.14, i also represents the maximum number of samples; h_v represents a weak classifier; Z_v represents a normalized constant. Finally, the output strong classifier numerical calculation expression is shown in equation 3.15.

$$U(x) = \text{sign}\left(\sum_{v=1}^{K}\omega_v h_v(x)\right) \tag{3.15}$$

In equation 3.15, $U(x)$ represents the strong classifier value. In the actual intelligent grading of English, the scoring model mainly adopts correlation analysis technology. This technology mainly removes irrelevant characteristic indicators and determines whether it is the key feature of the scoring standard [20]. The technology finally retains the characteristic indicators with high prediction ability, and establishes a scoring model based on the corresponding characteristic indicators. Finally, the model is integrated into the Adaboost/CT algorithm. The Adaboost/CT algorithm classifies the scores obtained by using the method of rough score, middle score and subdivision to get the final score. The overall process of the scoring model is shown in Figure 3.4.

It can be seen from Figure 3.4 that the scoring model will be based on the English article samples that have been reviewed and reviewed by experts. At the same time, the sample will be statistically analyzed and the characteristic index of the composition sample will be extracted. The experiment uses the correlation analysis method to retain the characteristic index with high correlation ($correlation\ coefficient \geq 0.6$). Then, according to the actual situation of sustainable English education, the model adds a feature index with high prediction ability to the mature automatic English article scoring system. According to the actual situation, the model

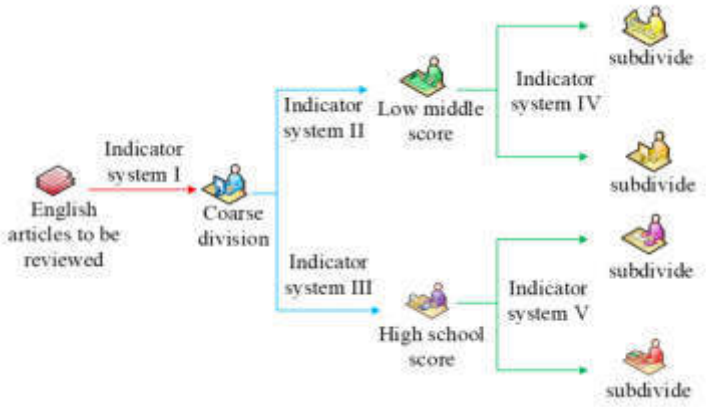


Fig. 3.5: Schematic diagram of course-middle-segmentation model

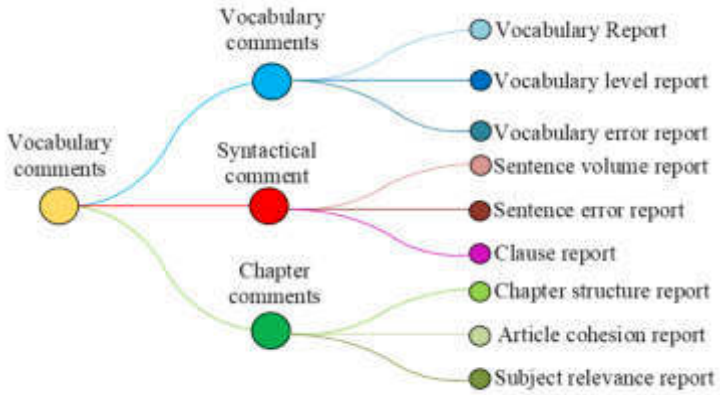
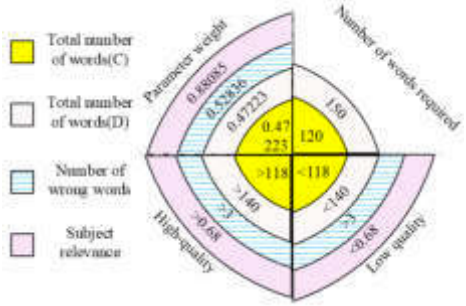


Fig. 3.6: Sketch map of comment generation model

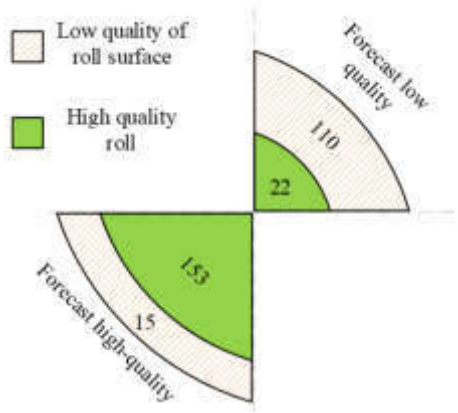
can include the required characteristic indexes into the corresponding index set and, finally, remove the remaining characteristic indexes. In this process, the final score will be obtained through the rough score-medium score-segmentation model. The coarse-middle-segment model built by the research is shown in Figure 3.5.

As can be seen from Figure 3.5, first input the English articles to be evaluated into the rough score-medium score-segment model. This process follows the scoring index system 1 composed of hard indicators such as the total number of rough scoring modules and the number of wrong words. Then, the article gives a low and middle score in index system 2, which is composed of the number of whole sentences and the number of grammatical errors. Then, the article carries out senior high school scores in the index system 3, which is composed of vocabulary cohesion and complex sentence patterns. Finally, the paper subdivides the index system 4 which is composed of sub-dispersion and punctuation frequency; The article is subdivided according to the index system 5 of lexical cohesion, complex sentence patterns and lexical relevance. In addition, real-time feedback in English articles can effectively help students find problems in time and improve their writing ability. Therefore, on the basis of intelligent English grading, the research also adds real-time natural language feedback to students. Among them, the generation of English article comments includes three aspects of comments: vocabulary, syntax and text. Based on this, the comment generation model built in the study is shown in Figure 3.6.

As can be seen from Figure 3.6, in the English article comment generation model, the vocabulary comment



((a)) Content of Rough Scoring Index System



((b)) High and low quality matching matrix content

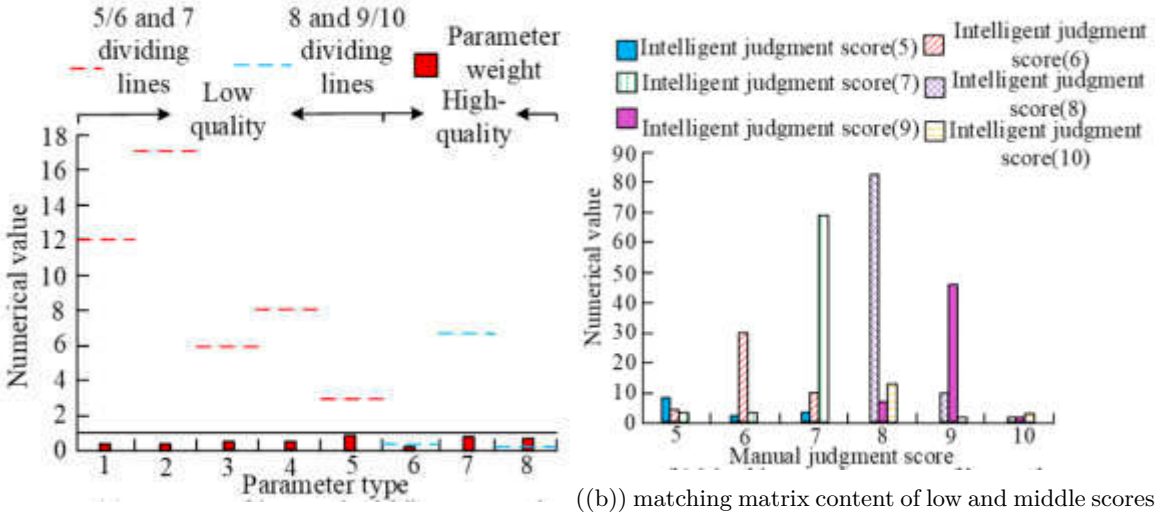
Fig. 4.2: Rough classification index system and high- and low-quality prediction classification content

includes the report of vocabulary size, vocabulary grade and vocabulary errors. Syntactic comments include the report of sentence size, syntactic errors and clauses. The text comments include reports on the partial structure, the cohesion of the article and the relevance of the topic. It is worth noting that the prerequisite for entering the comment generation model is to meet the relevant requirements, that is, whether the number of words in English articles meets the minimum requirements of the scoring model.

4. The practical application of Adaboost/CT algorithm in English intelligent grading. To verify the effectiveness of the Adaboost/CT algorithm in English intelligent scoring, the study validated the performance results of the algorithm in the scoring module and the commenting module, respectively. In the scoring module, the study used 312 English compositions from two grades of students in a certain university as test samples, and a total of 300 English compositions were included in the "coarse medium subdivision" model constructed in the study. 300 essays cover low, medium, and high levels, and are relatively comprehensive. At the same time, 300 English compositions that are higher than the mandatory indicators are considered high-quality compositions, while those that are lower than the mandatory indicators are considered low-quality compositions, denoted by A and B respectively. In the coarse grading - medium grading - subdivision model, the coarse grading index system and the content of the high- and low-quality matching matrix are shown in Figure 4.2.

In Figure 4.1(a), C represents 120 words; D means 150 words. It can be seen from Figure 7 that the coarse score indicator system focuses on topic relevance, and the parameter weight reaches 0.88085. At the same time, the high- and low-quality limits of 120 words and 150 words are 118 and 140; The high- and low-quality limits of the number of wrong words and topic relevance are 3 and 0.68. As can be seen from Figure 4.1(b), 110 samples of low quality and 15 samples of high quality are predicted under the low quality of the roll surface; There are 22 samples with low quality and 153 samples with high quality. Overall, 132 English compositions were divided into Group A and 168 into Group B. The accuracy rate of high- and low-quality English composition classification was 87.66%, and the adjacency accuracy was 95.33%. Therefore, Adaboost/CT algorithm shows good performance. When the coarse score - middle score - subdivision model is divided into stages, the low middle score and high middle score indicator system and the low middle score and high middle score matching matrix are shown in Figure 4.4.

In Figure 4.4, the horizontal axis 1-8 represents the parameter types of low, middle and high scores and high scores. Among them, 1-5 represents the integer sentence required by 120 words, the integer sentence required by 150 words, the number of words required by 120 words at level 3 to 4, the number of words required by 150 words at level 3 to 4, and the number of syntax errors. 6 8 means Latent Semantic Analysis (LSA), the



((a)) Content of low and middle score and high score index system

((b)) matching matrix content of low and middle scores and high scores

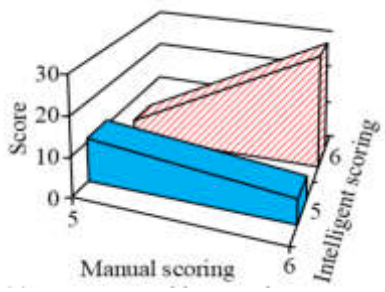
Fig. 4.4: The index system of low middle score and high high score and the content of the matching matrix of low middle score and high score

Table 4.1: Index system content of Group P and Group Q

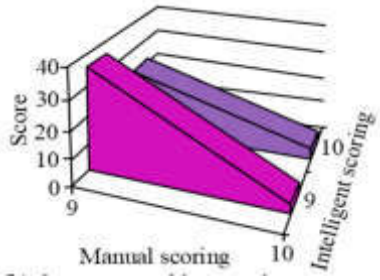
Group P indicator system			
	Sentence dispersion	Punctuation frequency	Part of speech
Parameter weight	0.549	0.805	0.908
5 points	>4.710	>0.670	Many pronouns
6 points	<4.710	<0.670	Many adjectives
Group Q indicator system			
Parameter weight	0.446	0.675	0.876
9 points	>0.763	<4.000	<0.550
10 points	<0.763	>4.000	>0.550

number of words at level 5 6 and grammatical cohesion of all sentences. In addition, there are high scores above the dotted line in the figure. It can be seen from Figure 8 that in the low and middle stages, 107 articles were classified using the Adaboost/CT algorithm, and the accuracy and adjacency accuracy were 81.06% and 95.45% respectively. Among them, 3 compositions with 5 points for manual review were divided into 7 points by algorithm; 10 manually-rated compositions with 6 scores were divided into 7 scores by algorithm; Three manually-reviewed compositions with 7 points were divided into 5 points by the algorithm; Two compositions with a manual score of 5 points were divided into 6 points by the algorithm. In the high score, the Adaboost/CT algorithm is used to classify 132 articles, and the accuracy rate and proximity accuracy rate are 78.57% and 91.07% respectively. In general, Adaboost/CT algorithm has high classification accuracy. In the subdivision stage, the study set the 5-6 score as group P, and the 9-10 score as group Q. The contents of the indicator system of Group P and Group Q are shown in Table 4.1.

It can be seen from Table 4.1 that the focus of the index system of Group P is on part of speech, and its parameter weight reaches 0.908. In the final prediction score, the pronouns under 5 are mostly, and the adjectives under 6 are mostly. The index system of group Q focuses on the verb related characteristics, and the

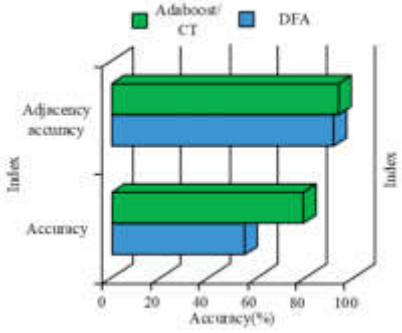


((a)) P group matching matrix content

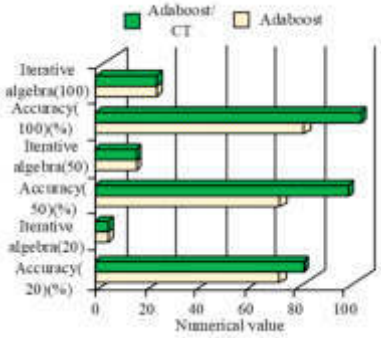


((b)) Q group matching matrix content

Fig. 4.6: Matching matrix content of group P and group Q



((a)) Accuracy comparison results of Adaboost/CT algorithm and DFT algorithm



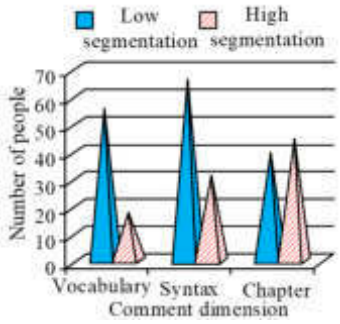
((b)) Comparison of classification results between Adaboost and Adaboost/CT

Fig. 4.8: Comparison of algorithm accuracy and classification results under different parameters

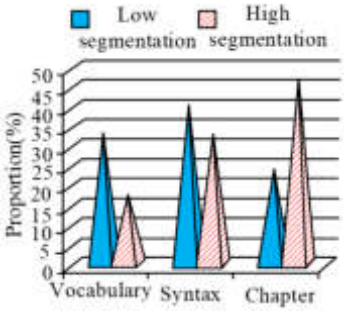
parameter weight reaches 0.87612. Under this indicator system, the contents of two sets of matching matrices obtained by using Adaboost/CT algorithm are shown in Figure 4.6.

According to Figure 4.6, 40 English compositions in Group P were correctly classified according to Adaboost/CT algorithm, with an accuracy rate of up to 80%. There are also 40 English compositions in Group Q that are correctly classified according to Adaboost/CT algorithm, with an accuracy rate of 61.54%. In addition, the predicted score of Group P is 5 or 6, while that of Group Q is 9. Therefore, there is no problem of adjacency accuracy, and there is no problem of adjacency accuracy of 100%. In general, the results of the evaluation of 300 English compositions by Adaboost/CT algorithm show that 239 of them are the same as those of manual evaluation. The prediction accuracy of Adaboost/CT algorithm reached 79.66%, and the adjacency accuracy reached 94%, showing high performance. In order to further verify the results, the Discriminant Function Analysis (DFA) algorithm was introduced and compared with the Adaboost/CT algorithm in accuracy [21]. The classification results of Adaboost and Adaboost/CT are shown in Figure 4.8.

It can be seen from Figure 4.8 that the accuracy and adjacency accuracy of Adaboost/CT algorithm are 79.66% and 94% respectively, which are higher than 55% and 92% of DFA. In addition, under different sample parameters, the classification accuracy of Adaboost/CT algorithm is higher than that of Adaboost algorithm under the same iteration. When the sample parameter is 100 and the number of iterations is 15, the accuracy rate of Adaboost/CT algorithm is up to 94%, far higher than 71.99% of Adaboost algorithm. In



((a)) Number of people to be improved under different dimensions



((b)) Proportion of people to be improved

Fig. 4.10: Comparison of algorithm accuracy and classification results under different parameters

Table 4.2: Comparison results of different indicators for different algorithms

Algorithm	Performance Indicators		
	Accuracy	Recall	F1
Adaboost/CT	96%	95%	95%
TA	85%	78%	81%
TF-IDF	93%	93%	81%

general, Adaboost/CT algorithm effectively solves the interference of singular value, and its performance is far higher than that of the comparison algorithm. In the evaluation module, we study the use of Adaboost/CT algorithm to evaluate 300 pre-processed English test papers here. On the basis of ensuring the reliability, the experiment invites experts to re-evaluate the scores with objections, so as to obtain reliable scores. Among them, the differences between the processed high and low score segments at different comment levels are shown in Figure 4.10.

It can be seen from Figure 4.10 that after the test paper is reviewed again using the Adaboost/CT algorithm, 163 people need to be improved in the low section, while 5 people need to be improved in the high section. In general, high-level students have a relatively stable grasp of vocabulary level, and only 10.2% of students still need to improve their vocabulary level. Among the students with low syntactic level, 53.6% of them need to make syntactic adjustments, which requires them to conduct deeper research at the grammatical level. In discourse, both junior and senior students must work hard on discourse in order to improve their writing ability. The experimental results show that the use of Adaboost/CT algorithm for English intelligent grading can not only obtain more accurate scores, but also give feedback. The result can help students improve their English level more pertinently. In addition, in the subsequent natural language module, the review results can be used to achieve positive language feedback for articles with elements up to standard; Articles that fail to meet the standards can also be targeted according to the review results. To further validate the performance of the Adaboost/CT algorithm, accuracy, recall, and F1 values were introduced to evaluate its classification performance. At the same time, the Triplet Algorithm (TA) and Term Frequency Inverse Document Frequency (TF-IDF) algorithms were introduced for comparison, and the results are shown in Table 4.2.

From Table 4.2, the classification accuracy of Adaboost/CT is 85%, which is higher than the TA algorithm's 85% and TF-IDF93%; And the recall rate is 95%, which is higher than 78% of TA algorithm and 93% of TF-IDF; The F1 value is 95%, which is much higher than 81% of the TA algorithm and 81% of the TF-IDF algorithm. Overall, the Adaboost/CT algorithm has high accuracy in essay classification. Based on this, the study applied it to actual English grading. A certain grade of students from a certain university covering low, medium, and

Table 4.3: Practical Application Results of Adaboost/CT Algorithm

Algorithm	Number of Tests	Correct Quantity	Accuracy	Average Accuracy	
				1	2
TA	1	400.00	190.00	48.71%	57.26%
	2	390.00	279.00	71.46%	
	3	385.00	166.00	43.00%	
	4	390.00	257.00	65.88%	
TF-IDF	1	400.00	246.00	61.53%	69.43%
	2	390.00	312.00	80.09%	
	3	385.00	231.00	59.88%	
	4	390.00	297.00	76.22%	
Adaboost/CT	1	400.00	308.00	76.91%	76.82%
	2	390.00	301.00	77.22%	
	3	385.00	275.00	71.49%	
	4	390.00	318.00	81.64%	

high levels of English test papers were selected and divided into four groups. The accuracy of the scoring results of the three algorithms on the four groups compared to manual grading was verified, and the results are shown in Table 4.3.

From Table 4.3, the accuracy of Adaboost/CT has decreased in practical applications, but it basically meets expectations. Its accuracy in the four groups is 76.91%, 77.22%, 71.49%, and 81.64%, respectively, most of which are higher than the comparison algorithm, and its average accuracy is also higher than the comparison algorithm.

5. Conclusion. In order to ensure the fairness of English grading and promote the development of sustainable English education, the research has improved and constructed the Adaboost/CT algorithm based on the actual English intelligent grading. In order to verify the proposed model, relevant experiments were carried out. The experimental results show that in the English intelligent grading module, the adjacency accuracy of Adaboost/CT in classifying high and low quality English compositions at the segmentation stage is 95.33%; The accuracy rate and adjacency accuracy rate of high and low quality English compositions in the middle stage are 81.06% and 95.45% respectively; In the subdivision stage, the Adaboost/CT algorithm evaluated 300 English compositions, 239 of which were the same as the manual evaluation, with the prediction accuracy of 79.66% and the adjacency accuracy of 94%. Compared with other DFA algorithms, the results show that the accuracy and adjacency accuracy of Adaboost/CT algorithm are 79.66% and 94% respectively, which are higher than 55% and 92% of DFA. Compared with Adaboost, the accuracy of Adaboost/CT algorithm also has advantages. In the English composition evaluation module, the use of Adaboost/CT algorithm can not only obtain more accurate scores, but also give feedback to each student, so as to improve it pertinently. Meanwhile, its average accuracy in practical applications is 76.82%, which is higher than the comparison algorithm. In general, the intelligent grading method using Adaboost/CT algorithm has shown high effectiveness and practicability in sustainable English education. It is worth noting that when using the Adaboost/CT algorithm in research, there is no syntactic or semantic knowledge involved in English discourse coherence issues. It is necessary to add these knowledge in the future to enhance discourse level coherence issues and further enhance the persuasiveness of intelligent grading.

Funding. The research is supported by: Science and Technology Research Program of Chongqing Municipal Education Commission, Design and Implementation of Cross-border E-commerce English Intelligent Learning Platform Based on Knowledge Map, (No., KJQN202204107); The research is supported by the 13th Five-Year Plan for Education Science of Chongqing Municipality in 2019, The International Development of Chongqing Higher Vocational Colleges under the Belt and Road Initiative: Current Status, Function and Path, (No., 2019-GX-047); The research is supported by Scientific Research and Innovation Team of Chongqing Youth Vocational & Technical College: A Research and Innovation Team for the Construction of the Ideological and

Political System of the English Curriculum, (CQY2021CXTDB01).

REFERENCES

[1] Phoophuangpairoj, R. & Pipattarasakul, P. Preliminary Indicators of EFL Essay Writing for Teachers' Feedback Using Automatic Text Analysis. *International Journal Of Educational Methodology*. **8**, 55-68 (2022)
[2] Dong, Y., Yu, X., Alharbi, A. & AI-and, A. and application of English multimode online reading using multi-criteria decision support system. *Soft Computing*. **26**, 10927-10937 (2022)
[3] Zhou, L. Construction of English Writing Hybrid Teaching Model Based on Machine Learning Automatic Composition Scoring System. *Procedia Computer Science*. **208** pp. 384-390 (2022)
[4] Zhang, L., Huang, Y., Yang, X., Yu, S. & Zhuang, F. An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*. **30**, 177-190 (2022)
[5] Fang, Y. Design of Oral English Intelligent Evaluation System Based on DTW Algorithm. *Mobile Networks And Applications*. **27**, 1378-1385 (2022)
[6] Canoy, D. & Loquias, A. Identifying Reading Miscues and Reading Performance in the Oral Reading Verification Test in English: Basis for an Intensive Reading Program. *International Journal Of English Language Studies*. **4**, 38-46 (2022)
[7] Štajner, S., Sheang, K. & Saggion, H. Sentence simplification capabilities of transfer-based models. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **36**, 12172-12180 (2022)
[8] Ramesh, D. & Sanampudi, S. An automated essay scoring system: a systematic literature review. *Artificial Intelligence Review*. **55**, 2495-2527 (2022)
[9] Liu, S. Research on the Blended Evaluation Mode in College English Writing Course. *Journal Of Language Teaching And Research*. **13**, 763-771 (2022)
[10] Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouye, M. & Zarrouk, M. Predicting CEFR levels in learners of English: the use of microsystem criterial features in a machine learning approach. *ReCALL*. **34**, 130-146 (2022)
[11] Liu, L. & Li, Y. An automatic composition scoring system for oversea Chinese students//International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP 2022). *SPIE*. **12456** pp. 479-484 (2022)
[12] Wu, X. & Zhang, Y. English Speech Scoring System Based on Computer Neural Network. *International Journal Of Education And Humanities*. **5**, 213-216 (2022)
[13] Phoophuangpairoj, R. & Pipattarasakul, P. Preliminary Indicators of EFL Essay Writing for Teachers' Feedback Using Automatic Text Analysis. *International Journal Of Educational Methodology*. **8**, 55-68 (2022)
[14] Taskiran, A. & And, G. and teacher feedback: Writing achievement in learning English as a foreign language at a distance. *Turkish Online Journal Of Distance Education*. **23**, 120-139 (2022)
[15] Miao, L. & Zhou, Q. Research on application of deep neural network model in college English skill training system. *International Journal Of Wireless And Mobile Computing*. **22**, 274-280 (2022)
[16] Regin, R., Rajest, S. & Shynu, T. An Automated Conversation System Using Natural Language Processing (NLP) Chatbot in Python. *Central Asian Journal Of Medical And Natural Science*. **3**, 314-336 (2022)
[17] Sheth, S., Giancardo, L., Colasurdo, M., Srinivasan, V., Niktabe, A. & And, K. and acute stroke imaging. *Journal Of Neurointerventional Surgery*. **15**, 195-199 (2023)
[18] Irfan, M., Ayub, N., Althobiani, F., Ali, Z., Idrees, M., Ullah, S. & Gas, P. Energy theft identification using AdaBoost Ensembler in the Smart Grids. *CMC-Computers, Materials & Continua*. **1**, 2141-2158 (2022)
[19] Walker, S., Solberg, S., Schneider, P. & Guerreiro, C. The AirGAM 2022r1 air quality trend and prediction model. *Geoscientific Model Development*. **16**, 573-595 (2023)
[20] Asif, M., Sheeraz, M. & Sacco, S. Evaluating the Impact of Technological Tools on the Academic Performance of English Language Learners at Tertiary Level: A Pilot Investigation. *Pegem Journal Of Education And Instruction*. **12**, 272-282 (2022)
[21] Vassallo, S., Davies, C. & Biehler-Gomez, L. Sex estimation using scapular measurements: discriminant function analysis in a modern Italian population. *Australian Journal Of Forensic Sciences*. **54**, 785-798 (2022)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Jun 7, 2023
Accepted: Oct 9, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 465–479, DOI 10.12694/scpe.v25i1.2338

RESEARCH ON THE PRACTICE EDUCATION PATTERN OF INNOVATIVE ENTREPRENEURSHIP IN COLLEGES IN THE INTERNET PLUS ERA

LI ZHUANG*AND LIN ZHU†

Abstract. Under the background of Internet plus, the innovative entrepreneurship practice education pattern (IEPEP) in colleges aims to enhance students' basic literacy and independent innovation ability of IE. The single form of PEP in IEPEP only focuses on theoretical education, lacking practical exercises. The teachers of IE courses are mainly part-time, lacking professional teaching team and sufficient teachers. Written evaluation is used as the IEPEP evaluation standard. This lack of practical exercises leads to an imperfect evaluation model. For this reason, the overall idea in this paper is constituted from raising questions to theoretical research to model construction and then to empirical analysis. By constructing three patterns to optimize the talent training mechanism of IE. The first is to construct IEPE curriculum pattern through literature, investigation and data analysis. The second is to construct IEPE activity pattern through expert guidance, in-depth exchanges, and case studies. The third is to construct IEPE evaluation pattern through questionnaire surveys, online assessments, counseling talks, and data analysis. Based on the systematic sampling method, 398 valid questionnaires are analyzed for data based on the questionnaires filled by the categorized 560 students and with the help of SPSS 22.0 and AMOS 26.0 software. A conclusion that IEPEP has a significant positive impact on entrepreneurial intention and ability is confirmed through empirical analysis. This effectively guarantees the quality of IE talents, and provides IE talents for society and enterprises.

Key words: Internet plus, innovative entrepreneurship (IE), practice, education system

1. Introduction. The Outline of the Medium and Long-Term Education Reform and Development Plan of the People's Republic of China (PRC) (2010-2020) clearly proposes to innovate the talent training pattern [1, 2, 3, 4]. The 20th National People's Congress of the Communist Party of China also makes important arrangements for the cultivation of innovative entrepreneurial (IE) talents, and puts forward clear and specific requirements for strengthening IE education. As a university, the cultivation of college students' innovative ability, entrepreneurial ability, and practical ability needs to be paid attention to. As China's higher education enters the stage of mass education, college students face unprecedented challenges in entrepreneurship and employment. How to train students to be capable of IE would become an urgent problem for colleges [5, 6, 7, 8].

Paper [4] examines the effectiveness of using a mobile learning platform for the improvement of students' entrepreneurial skills. The platform used allows users to post assignments, exchange ideas, generate photo and video content, etc. The average entrepreneurial competence assessment results of the experiment group are slightly higher than those of the control group, which are 4.5 points against 4.0 points, respectively. The results of the study are generalizable to educators and administrators of educational institutions or those involved in creating and conducting entrepreneurship training programs. On the basis of summarizing the innovation of practice and entrepreneurship integration system and pattern in colleges and universities, and combining the experience of IE education in Beijing Institute of Fashion, the proposal of constructing the practice and entrepreneurship integration system construction is put forward [5]. The practice and innovation fusion talent cultivation system is constructed from four aspects: cultivation goal, curriculum system, teaching system and guarantee mechanism. Paper [6] proposes a ranking refinement method based on optimal weighting model and determines the optimal weights using back propagation neural network. Results indicate that the number of employed people is always lower than the number of graduates, suggesting that it is difficult to find employment. In this case, many college graduates choose to start their own business, but the success rate of entrepreneurship is very low (only about 3%). This shows that the IE ability of university graduates is not high. Therefore, there

*School of Electronics and Computer Engineering, Southeast University Chengxian College, Nanjing 210088, China (Corresponding Author: Li_Zhuang2023@outlook.com)
†School of Electronics and Computer Engineering, Southeast University Chengxian College, Nanjing 210088, China

is a need to improve this ability and conduct self-assessment. An often-overlooked career pathway is medical device innovation, which can be mutually beneficial to both physicians and the industry. To this end, paper [8] explores a novel career path for early career physicians in the field of medical device IE, which provides a promising reference for expanding the field of IE.

There are three issues that need to be focused on.

1. For single form of IE practice education pattern (IEPEP) [9, 10]: IEPEP focuses on theoretical teaching and only offers courses related to IE. For example, courses such as career planning and employment guidance for college students and entrepreneurship education are offered. Since there is no practical teaching system corresponding to these courses, students are unable to apply the theory to practice in the learning process. It is found that theoretical education accounts for 80-90% of students' IE education. On the one hand, this method does not meet market demand. On the other hand, by combining the application-oriented undergraduate talent training program and integrating IEP into the classroom teaching process, this makes IEPE not diversified enough.
2. For insufficient teaching staff for IE talent cultivation [11, 12, 13, 14]: The education for IE talents in colleges only comes from school teachers (they make up about 90-96% of the faculty of the IE teacher team), but no experts are hired from society and enterprises to educate them. A professional team of IE teachers is also missing. This makes the teaching staff not sufficient enough, and has no experience in IE. Once educating students, teachers only have theoretical knowledge and no practical operation, which is not exemplary.
3. For imperfection of evaluation pattern for IE education [15, 16, 17, 18]: The evaluation pattern enables to evaluate the results of IE education. The existing evaluation patterns of IE education, such as course examinations and essay writing, are not scientific enough. About 50% of colleges and universities use course examinations, about 30% use essay writing, and the remaining 20% use a combination of the two. There is no reasonable evaluation index and system for students' IE ability so it is impossible to evaluate students accurately.

In general, in view of the obvious problems of single form, insufficient teachers, and imperfect evaluation model in IEPE pattern (IEPEP), this paper makes a qualitative and quantitative assessment of the basic elements contained in this pattern by analyzing and defining the relevant core concepts, based on the background of the Internet plus era. After IEPEP framework in colleges is initially constructed, the validity of the proposed pattern is demonstrated in depth by using the methods of design scales, questionnaire surveys, and data analysis. The form and composition of the proposed pattern are not found in existing work. Finally, a comprehensive analysis is made on demonstration results, and the opportunities and challenges in the future work are stated.

2. Significance of IEPEP in colleges in the Internet plus era.

2.1. A new pattern construction by taking the times as the background and changing education concepts. A new IEPEP aims to achieve high-quality employment for graduates and promote higher education to serve economic and social development. With IEPE as the theoretical guidance, and with the background of Internet plus, it finally builds an IEPE system [18, 19, 20, 21]. Through the combination of Internet plus, IEPE system is further constructed, and the construction of IE training base is further improved. Meanwhile, IE training program for college students is implemented, and the training plan for IE talents is formulated. In addition, IE training content is integrated into professional classrooms. A new IEPEP is finally constructed.

2.2. A scientific pattern construction by using practice as a guide and innovating the curriculum system. PEP significance is to apply theory to practice. For IEPEP, it mainly includes awareness training, ability improvement, environmental cognition and practice simulation. Aiming at these parts, IEPE system in colleges is constructed through methods such as literature, investigation and data analysis. Among them, the links of innovative thinking training, entrepreneurial ability training and entrepreneurial practice need to be strengthened, and the elective courses of IE training need to be increased. This aims to enlighten students' innovative consciousness and entrepreneurial spirit, analyze and cultivate students' critical thinking, insight, decision-making ability, organizational coordination ability and leadership and other innovative and entrepreneurial qualities. By guiding students to understand the current enterprise and industry environment,

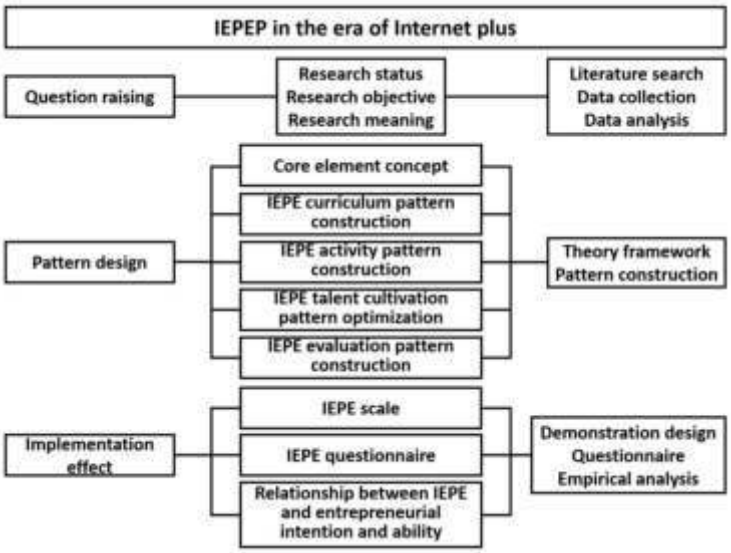


Fig. 3.1: A theoretical framework for IEPEP.

and encouraging students to experience all aspects of entrepreneurial preparation, their own IE capabilities are improved.

2.3. A education platform construction by taking cooperation as the basis and improving school-enterprise cooperation. The main bodies of the school-enterprise collaborative education platform are universities and enterprises. Through the combination of university leadership, enterprise participation and student training, the platform construction is perfected. IE platform is needed for social development, so it needs to be combined with the needs of social enterprises to highlight students' practical ability. The Outline of the Medium and Long-Term Education Reform and Development Plan of the People's Republic of China (PRC) (2010-2020) points out that it is necessary to implement the talent training mode of combining work and learning, school-enterprise cooperation, and internship. Through school-enterprise cooperation projects, the construction of entrepreneurship training projects is promoted. The information exchange between the school and the enterprise would hopefully realize the sharing of entrepreneurial information and create a good environment for IE. This helps the smooth project implementation, improve the quality of college students' personnel training, enhance the competitiveness of college students, and thus contribute to better serving the society.

2.3.1. A evaluation system construction by taking effects as guides and formulating evaluation indicators. The purpose of constructing a IEPEP in colleges based on the Internet plus era is to improve students' IE abilities, improve the quality of those with IE abilities, and provide innovative talents for social development. The evaluation system can be used to achieve timely feedback on educational effects, providing a scientific basis for further improving IE education pattern and cultivating IE talents.

3. Construction of IEPEP in colleges in the Internet plus era.

3.1. Theoretical framework. Figure 3.1 presents the overall theoretical framework of this paper, involving pattern design, theory framework, pattern construction, implementation effect, and so on.

3.2. Pattern construction.

3.2.1. Curriculum pattern construction in IE Education. Whether the curriculum pattern of IE education is constructed reasonably affects the implementation and basic guarantee effect of IE education [22, 23]. Through expert guidance, in-depth communication, case studies and other methods, we explore how

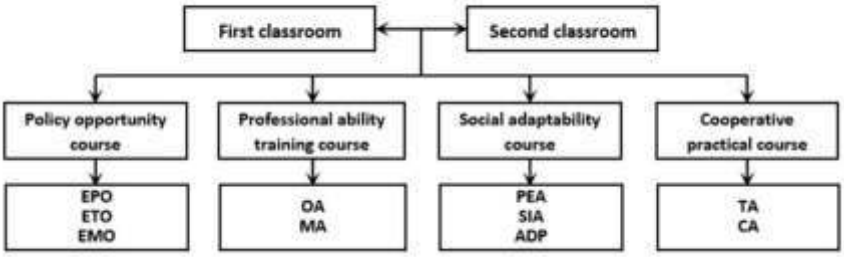


Fig. 3.2: A framework for IE curriculum pattern.

to create a IEPE system in colleges. The curriculum pattern of IE education is the basis for the realization of IE education. A large number of literatures are reviewed and the first classroom and the second classroom are combined to build this pattern. The pattern is validated by analyzing data from a questionnaire survey conducted with students. At the same time, the pattern is improved by modifying the existing problems. According to the talent training plan, the course system is divided into general education course, professional course, and entrepreneurship course platforms. On the basis of completing general courses and professional courses, the curriculum pattern of IE education is constructed (Figure 3.2). The teaching session is designed to connect the first and the second classrooms. The design concept is in line with the law of entrepreneurship education, so that students have sufficient time and space to absorb and digest the knowledge they have learned, and apply what they have learned through practical activities. A complete course model is divided into basic experiment, creative experiment, sociology practice, and cooperative practice teachings. On the basis of optimizing the training pattern of IE talents, colleges should use Internet plus thinking to establish a multi-integrated and dynamically optimized curriculum system. This not only reflects the professional characteristics, but also reflects the integration of professional and IE education. Meanwhile, this not only has a higher professional vision, but also improves IE quality.

In Figure 3.2, EPO, ETO, EMO, OA, MA, PEA, SIA, ADP, TA, and CA respectively represent entrepreneurial policy opportunity, entrepreneurial technology opportunity, entrepreneurial market opportunity, operation ability, management ability, psychological endurance ability, social interaction ability, ability to deal with people, teamwork ability and communication ability.

The framework of IE curriculum system is mainly divided into two parts. The first part is the integration of the first and the second classrooms. The second part is about the four types of courses involved. They are policy opportunity, professional ability training, social adaptability and cooperative practical courses. The policy opportunity course mainly helps students analyze the current policy, technology, and market opportunities for entrepreneurship, and provides relevant guidance so that students can understand the advantages brought about by entrepreneurship. The professional ability training course mainly helps students to cultivate the operation and management abilities in the entrepreneurial process, so that students can have the operation and management ability needed in the entrepreneurial process. The social adaptability course mainly cultivates students' psychological endurance, social communication and ability to deal with others. Cooperative practical courses are mainly to cultivate students' teamwork and communication abilities. Through the teaching guidance of social adaptability and cooperative practical courses, the basic emotional intelligence that students need to possess in the process of starting a business can be improved.

The first classroom is to equip students with the basic IE knowledge and ability, while the second classroom is to practice these abilities for students. Based on various subject competitions, the first and the second classrooms are integrated to achieve the implementation of students' IE education. For example, the Internet Plus Innovation and Entrepreneurship Competition integrates IE classroom system into it. By participating in such competitions, students acquire the practical operation ability, so as to flexibly use the basic IE knowledge and ability. This provides practical experience for future IE.

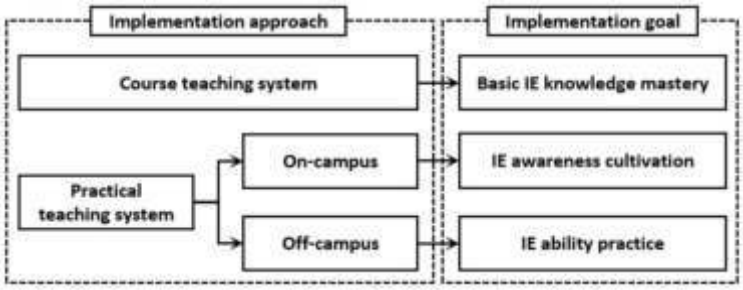


Fig. 3.3: A framework for IE education activity pattern

3.2.2. Activity pattern construction in IE Education. IE education activity pattern is an important platform for applying theory to practice. Using methods such as in-depth communication among school, enterprises and students, the objectives of IE education activity pattern are gradually realized through curriculum and practice. This pattern is validated using case study research. Eventually, the feasibility of the pattern is evaluated through case studies combined with expert guidance to develop the pattern.

From the perspective of curriculum setting, it is necessary to strengthen the cultivation of students' innovative spirit and entrepreneurial awareness, so that students have IE knowledge and ability. The curriculum system is mainly designed in accordance with the training direction of insight, thinking, understanding and innovation and entrepreneurship, and is combined with the actual situation of the enterprise. It involves the basic IE knowledge, such as Design and Management of Product Supply Chain. Such courses have been better used in IE process.

From the perspective of practical teaching, there are two parts that need to be paid attention to. One is the on-campus practice platform, whose main purpose is to cultivate students' IE awareness. The on-campus practice platform is to build a platform for students to practice IE, such as office hardware conditions. Various national policies are adopted to provide an environment for IE, or cooperate with enterprises to build a platform for students to innovate and start businesses. In addition, relevant subject competitions, such as college students' innovation and entrepreneurship competitions, college students' practical innovation training programs, etc., can be used to strengthen students' IE capabilities. In this case, a variety of ways is used to make the platform system of IE on campus an incubator for students' IE education activities. The second is on-campus and off-campus practice projects, which are mainly for the IE transformation and breakthrough. The off-campus practice project realizes school-enterprise cooperation by introducing enterprise projects. Through practical projects, students' IE ability can be further trained, and resources can be provided to enable students to participate in the entrepreneurial process independently, so as to realize practical IE teaching outside the school. The realization of IE education activity pattern is mainly composed of curriculum and practice. The realization goal is composed of the mastery of basic IE knowledge, the cultivation of initial IE awareness, and the practice of IE ability. After gradual improvement, the cultivation of IE talents is finally realized. Figure 3.3 illustrates an activity pattern in IE Education.

Through the development of various IE project activities, the first and second classrooms are integrated. Based on a variety of forms such as IE competitions, exchange meetings, and lectures, IE education courses are implemented to activate the campus IE atmosphere. Based on the school's IE competition projects, students are actively supported and encouraged to apply for national, provincial and school-level IE plan projects, and professional instructors are assigned to give them special guidance to form a diversified IE practice activity. At the same time, it is necessary to fully develop, integrate and make good use of internal and external resources, attract social resources to invest in IE talent training and social practice, and actively promote collaborative education and collaborative innovation inside and outside the school. According to the concept of Internet plus, a batch of convenient and open crowd-creation spaces combining IE, online and offline, incubation and investment are built through marketization.

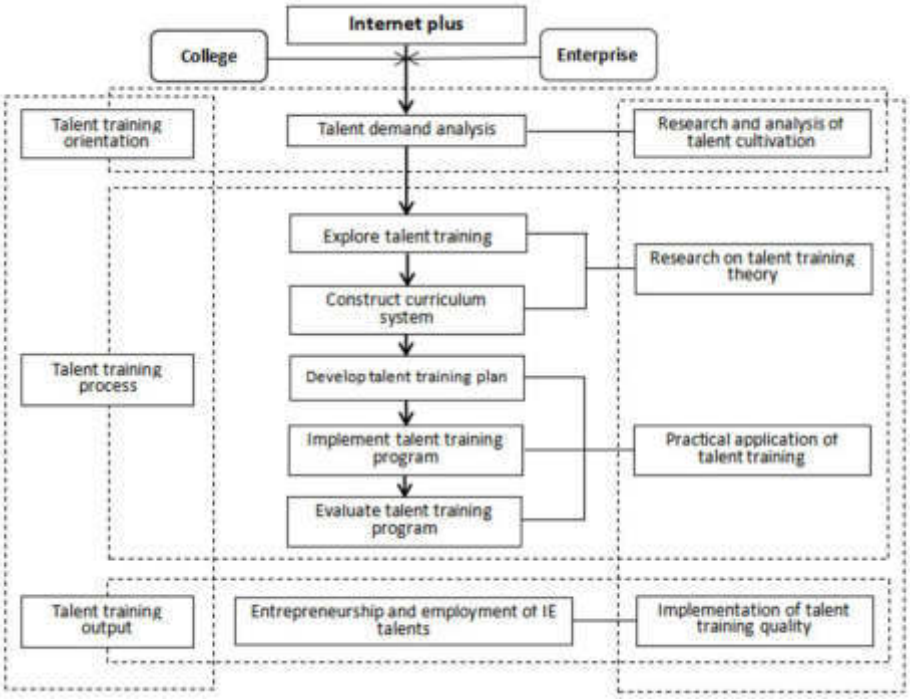


Fig. 3.4: A flowchart of talent cultivation under the Internet plus background

3.2.3. Training pattern optimization of IE talents. In view of the existing problems in the cultivation of IE talents, colleges need to use scientific methods to optimize the training pattern of IE talents [24, 25, 5] (Cai 2021; Yang & Luo 2020; Li & Li 2023). Under this pattern, a staged and progressive training method is adopted, and the roles of universities, students and enterprises are fully utilized to improve the talent training mechanism and ensure the quality of talent training. The optimized pattern is shown in Figure 3.4.

IE talent cultivation is divided into three parts: orientation, process and output. In terms of talent training orientation, through enterprise research and visits, we can understand the needs of society and enterprises for IE talents. Talent training is analyzed according to the analysis of talent needs, and the direction of IE talent training is provided. In terms of the talent training process, a scientific talent training plan is formulated to explore the talent training pattern, according to the talent training orientation generated by the talent demand analysis.

While social needs and enterprise needs are integrated into IE talent training pattern, IE education is integrated into the professional education teaching plan and curriculum system to build an IE talent training curriculum system. Enterprise experts are invited to guide and analyze the curriculum system, and how to combine the cultivation of students' IE ability with professional education is given priority consideration in order to formulate IE talent training programs.

According to the established IE talent training curriculum system, IE talent training program is formulated. In this process, different grades need to be considered to guide students to develop in different directions. Among them, first- and second-year students are mainly guided by ideas to let students understand what IE education is. Certain basic knowledge is mastered, and preliminary ideology is formed. This is also a kind of guidance for students' IE thinking in the future, so as to avoid the ideological deviation after failure due to insufficient IE ability in the later stage. Students in the second and third grades are mainly used in practical operations. Through the development of IE activities in the second classroom, such as IE lectures, IE competitions, etc., to cultivate students' practical experience in IE, so that students can understand how to use the professional knowledge in IE process. The third and fourth grade students are more focused on the guidance of the project.

By joining IE projects of enterprises, IE ability can be improved, students' interest in entrepreneurship can be stimulated, and finally college students can be encouraged to start their own businesses. Through the guidance of different grades, the cultivation of IE talents is realized.

According to the training plan for IE talents formulated, it will be implemented and applied among students. During this process, a guarantee mechanism needs to be formulated to ensure the effectiveness of the implementation process and the quality of IE talent training. The assessment methods for the cultivation of IE talents need to be changed. It is suggested that heuristic, discussion and participatory teaching methods be adopted to strengthen students' independent learning, cultivate students' critical and creative thinking, and stimulate innovation and entrepreneurship inspiration. At the same time, it is necessary to increase the proportion of experiments, practical training, and practical teaching, and pay attention to cultivating students' ability to solve practical problems. In addition, through the establishment of an online education platform, online education for minor majors, second majors and other course systems has been actively developed. Through the recognition of credits, teaching workload and teaching performance incentives, and the evaluation mechanism of undergraduate teaching work of departments, teachers and students are encouraged to actively participate in the teaching and learning of online courses.

After implementing the talent training plan, the feedback from all parties is summarized to understand the rationality and scientific of the development of IE talent training. It mainly evaluates and gives feedback through subject competition results, classroom teaching feedback, IE platform construction, and IE talent training assessment mechanism. The suitability of professional training programs, curriculum system settings and training objectives, the integration of IE education concepts in the training system, the cultivation of practical ability, the construction of teachers and teaching teams, the reform of teaching models and the improvement of learning effects, the quality of student training and IE needs to be evaluated with emphasis. In this way, the insufficient part of IE talent training program can be well adjusted.

3.2.4. Evaluation pattern construction of IE education. The evaluation pattern of IE education in the Internet plus era is mainly divided into the guarantee, construction evaluation and achievement evaluation patterns of IE education. A questionnaire is used to find out what kind of effect students expect IE education to achieve. Then, the online evaluation is used to understand the students' innovative consciousness and entrepreneurial ability after IE education. Based on the evaluation results, a sample of students are interviewed to refine and validate the results. Finally, all the data are analyzed and summarized to achieve scientific evaluation of the implementation effects of IE education.

IE education guarantee pattern includes the management institution, management system and funding.

1. Management institution: IE education is not only between schools and students, but also all aspects of schools, society, enterprises, and students. Therefore, it is necessary to establish an IE education management agency to coordinate various issues in the process of IE education and to manage IE education in an orderly manner. Schools should set up an IE education steering committee to guide the curriculum system, activity system, and talent training system in the process of IE education, and formulate relevant implementation policies and evaluation standards to ensure the cultivation of IE talents. As a link between society, enterprises and students, the school provides a communication platform for IE to ensure the sustainable development of IE education.
2. Management system: According to the spirit of the document issued by the Ministry of Education of PRC on Vigorously Promoting IE Education in Colleges and College Students' Independent Entrepreneurship Work (JGB [2010] No. 3) and the notice issued by the General Office of the Ministry of Education of PRC on the Issuance of Basic Requirements for Entrepreneurship Education Teaching in Ordinary Undergraduate Schools (for trial implementation) (JGB [2012] No. 4), the guiding ideology, objectives and implementation principles of IE education needs to be clarify, and the requirements and related countermeasures for IE education are put forward [27, 28]. Universities need to formulate IE education training programs, establish IE education curriculum systems, and formulate IE education quality evaluation systems based on their actual conditions. The quality of IE education system is guaranteed through the formulation of relevant management and the implementation of policies and regulations.
3. Fund setting: The school needs to set up special funds, such as college students' IE training programs,

subject competitions and awards, Internet plus college students' IE competitions, and entrepreneurship simulation training funds. According to the level and type of subject competition, credits will be awarded to students, and bonuses will be awarded to instructors and participating students. Through the investment of funds, the subject competitions, invention patents and scientific inventions of college students are effectively guaranteed.

The evaluation pattern of IE education construction includes soft and hard environment, teaching staff, curriculum system and scientific research construction evaluation [29, 30].

Evaluation of soft and hard environment. The evaluation of hardware environment is mainly aimed at the practice platform and related facilities (such as IE base for college students) required in the process of IE education practice. Colleges can provide students with funds for IE, and can also cooperate with enterprises to build industry-university-research cooperation bases as the environment for IE. The evaluation of soft environment is mainly aimed at the construction of campus IE, such as the launch of IE related forums, the holding of IE lectures, the formulation of various IE incentive policies, the publicity effect on IE campus, and the establishment of IE groups.

Evaluation of the teaching staff. It is done from the aspects of teachers' IE ability, IE teaching links, and IE teaching effects.

Curriculum system evaluation. The construction of IE talent training curriculum system needs to add IE courses, entrepreneurship education lectures, etc., and take relevant courses as compulsory courses and include them in the corresponding teaching plan. Meanwhile, IE education is integrated into the talent training program to realize the cultivation of IE talents. Based on the curriculum system of IE education, the relevant courses offered whether meet the objectives of cultivating IE talents, whether meet the talent needs of enterprises, and whether comprehensively improve students' IE capabilities are evaluated.

Evaluation of scientific research construction. It includes the evaluation of teachers' and students' scientific research construction. While it is necessary to evaluate the scientific research ability and academic level of teachers, it is also necessary to evaluate the ability of students to participate in scientific research projects. For example, in the practical innovation training project for college students every year, many students use the project to develop results and publish related papers. There are also students who have achieved core theses and patents. With the improvement of the scientific research ability of students and teachers, a number of scientific research achievements have been promoted, and the evaluation results of IE education are generally good.

The evaluation pattern of IE education achievements includes IE achievements and enterprise evaluation.

1. Evaluation of IE achievements: It mainly includes students' learning attitudes, assessment results of IE courses, participation in various IE activities, IE competition results, innovative works, participation in IE projects, etc.
2. Enterprise evaluation: According to the questionnaire filled out by the enterprise, the enterprise can track and manage the graduates and make a comprehensive evaluation while the feedback from the enterprise to the students is known. When the questionnaire is designed, it is necessary to reflect the application of students' IE abilities in enterprises, the degree to which units attach importance to students' IE abilities, and other evaluation information. This helps to understand the quality of IE talent training.

4. IEPEP implementation effects. Based on IEPE theory, this section dissects the connotation and divides the dimensions of key variables. By combining the actual situation of universities, rigorous logical reasoning is carried out, and the research model is further constructed with the proposed research hypothesis, based on the existing researches.

4.1. Argument design. IE education refers to a new type of education that takes into account both innovation and entrepreneurship education. By referring to the maturity scale and combining with the reality of this research, this article draws on the scale compiled by Frank & Luthje (2004) and Wang (2016) to measure IEPEP effects from the perspective of theoretical and practical education [31, 32]. Theoretical IE education includes classroom courses, lectures and competitions, while practical education includes entrepreneurial practice, communication, training or simulation. See Table 4.1 for the specific scale.

Table 4.1: An IEPE scale.

No.	Content
1	Classrooms and courses in IE education
2	Lectures and activities in IE education
3	Competition in IE education
4	Entrepreneurial practice
5	Entrepreneurial communication
6	Entrepreneurship training
7	Personal experience
8	Entrepreneurship intention
9	Entrepreneurship ability

4.2. Questionnaire. The questionnaire design needs to go through the following five steps. The first step is to preliminarily determine the problems to be solved in this questionnaire by reading relevant literature through databases such as China National Knowledge Infrastructure. The second step is to discuss the selection of variables and determine the feasible research plan. The third step is to process the survey questionnaire and process the sample data to ensure the reliability and validity of the scale. The fourth step is to interview the students and experts. After adjusting some of the questions, the final questionnaire is formed. The fifth step is to issue questionnaires on communication platforms such as WeChat and Tencent through Questionnaire Star.

The questionnaire in this paper includes research background introduction, personal information and main topics.

Background introduction includes research purpose, use, privacy protection commitment, author's acknowledgment and answering time. Its purpose is to emphasize the importance and privacy of this questionnaire, so that the students surveyed can answer the questionnaire more seriously and with confidence.

The personal information is about the demographic characteristics of the students surveyed. It includes gender, education, grade, major, academic performance ranking, student cadre experience and part-time jobs during school.

The main items of the questionnaire mainly involve 8 items of IEPE, 3 items of entrepreneurship intention, and 9 items of entrepreneurship ability.

This study adopts a systematic sampling method, and the surveyed subjects are undergraduate students from the school, including all grades and majors. These students are divided into different samples based on grade and major (for example, 20 freshmen majoring in computer science are taken as a sample). The 560 survey questionnaires distributed are divided into 28 groups of samples. A total of 503 questionnaires are recovered, of which 398 are valid, reaching an effective rate of 79.13%. Generally speaking, the sample size of this survey is relatively large and widely distributed, which is of reference value.

4.3. Empirical analysis. According to the results of the questionnaire, the collected sample data are processed and analyzed based on the principles of statistics, and with the help of SPSS 22.0 and AMOS 26.0 analysis software. First, the validity analysis is carried out to test whether the questionnaire design is reliable and effective. In the questionnaire survey, 398 questionnaires with effective answers are screened out according to the students' answers. Then, the difference analysis is carried out to examine whether there are differences among the variables. Based on the basic situation of the students in the questionnaire survey, the analysis is conducted. For example, in terms of gender, boys and girls account for 68% and 32% respectively. The answers to the questions of innovation and entrepreneurship in the questionnaire are similar, so there is no bias in the analysis of the questionnaire. Finally, the statistical results of the data are used to verify the research hypothesis. According to the results of student feedback for each question, the analysis and statistics are implemented to form the final data and graphs.

The results of the questionnaire survey show that IEPE can directly and significantly predict college students' entrepreneurial intentions, and the significance shows an increasing trend. There is a certain relationship between the strong entrepreneurial intention of students and the practice education of IE they receive, that

Table 4.2: Research Conclusions

No.	Research Content	Conclusion
1	IEPEP has a significant positive impact on entrepreneurial intention.	Support
2	IEPEP has a significant positive impact on entrepreneurial ability.	Support

is, when students receive more practical education on IE, their entrepreneurial intention will be stronger. The electives of IE classroom courses, and the participation of relevant lectures and competitions can enrich students' entrepreneurial knowledge and accumulate entrepreneurial experience. However, participation in entrepreneurial practice, exchanges and training will further strengthen students' entrepreneurial awareness and interest. This enables the theoretical IE study to be introduced into practice. The questionnaire survey also shows that IEPE can directly and significantly predict the entrepreneurial ability of college students, and the level of significance shows an increasing and steep trend. The entrepreneurial intention shown by students who received more practical education in IE is much higher than that of students with less education.

The above analysis shows that IEPE plays a significant moderating role in the relationship between entrepreneurial ability and intention. That is to say, when students receive more practical education on IE, their entrepreneurial ability is gradually improved, which makes it easier for students to generate entrepreneurial intentions.

4.4. Analysis conclusion. There is a significant relationship between IE theory education and college students' entrepreneurial intention, so it is necessary and significant to impart IE theory knowledge and experience. Meanwhile, the entrepreneurial practice education also plays a positive role in entrepreneurial intention to play a certain stimulating role. The empirical research illustrates that the participation of entrepreneurial competitions and other activities, the application of entrepreneurial practice platform and entrepreneurial theory can improve individual entrepreneurial skills, and to some extent stimulate students' entrepreneurial will. When receiving IEPE, college students can understand entrepreneurship at a deeper level by studying professional courses, signing up for relevant lectures and competitions, communicating with entrepreneurship patterns, visiting entrepreneurship bases, and conducting entrepreneurship simulation, which makes entrepreneurship intention enhanced.

Through IEPE, the entrepreneurial ability can be improved to enhance the individual's willingness to start a business. By making up for the gaps in the knowledge system, the entrepreneurial skills required for individual entrepreneurship are trained. This fills the vacancy of their own entrepreneurial practical experience, and thus has a complete entrepreneurial system. After combining IEPE with professional education, it helps educates improve their entrepreneurial ability and realize entrepreneurial behavior. This education also helps college students understand entrepreneurial laws and regulations and supporting policies more effectively and systematically, cultivate basic entrepreneurial skills such as planning, marketing, decision-making, and risk assessment that students need in practice, and encourage them to be willing, daring, and courageous to participate entrepreneurial activity refer Table 4.2.

Meanwhile, this study finds that the participation of entrepreneurial process and business management has the greatest impact on college students' entrepreneurial intentions, followed by entrepreneurial courses and lectures. Entrepreneurial practice helps to directly enhance college students' feelings, experiences and skills towards entrepreneurship, which significantly improves students' entrepreneurial intentions. However, entrepreneurship courses and lectures more indirectly teach entrepreneurship knowledge and skills, and can also improve students' entrepreneurial intentions. In contrast, some entrepreneurial competitions are only a formality, with heavy commercial publicity and utilitarianism, so the impact on college students is not obvious. This is a question that must be paid attention to when choosing a university and an entrepreneurial education method.

When carrying out the argument design, the entrepreneurial intention is mainly taken as the survey results, and no further research is done on entrepreneurial behavior. In the process of transforming entrepreneurial intention into final entrepreneurial behavior, there are many uncertain factors. As for how much can be converted into entrepreneurial behavior and how much entrepreneurial rate is, it needs further exploration. In

future research, it is necessary to conduct in-depth investigation and analysis on the entrepreneurial intention and entrepreneurial behavior of college students after receiving IE education. A more specific example is given here. Student San Zhang is a freshman computer science student. During his career planning education at school, he put forward the idea of starting his own business. However, he did not take any action at a later stage. If he had followed the proposed pattern, he would have had a clear entrepreneurial idea in the second half of his freshman year. In his second year, he started to contact his classmates around him and organized himself effectively according to his own conditions. By his third year, he had a preliminary business plan and was working effectively to implement it. By the time he reached his senior year, he was able to carry out simple entrepreneurial activities with a team of up to 10 people.

5. Discussion. On the theoretical level, this research helps to further improve the research on IEPE. IEPE course in colleges is taken as the research object of this research, instead of just presenting this course as a fragmented and fragmented part of the research on IE education. It mainly analyzes the relevant constituent elements of IEPE courses, and continuously expands the relevant theoretical research results. Based on the corresponding theoretical basis and elements, the optimization and reform of IE education curriculum system in colleges is promoted, thereby further enriching and deepening the research on IEPEP construction [33, 34]. This provides a valuable reference for future work.

On the practical level, this research is conducive to improving the quality of IEPE courses in colleges. With the goal of transforming college students' IE thinking consciousness, enriching professional IE knowledge, and enhancing IE ability, it systematically and step by step connects the practical education curriculum system with IE attributes in a deep and effective manner [34, 35]. In addition, the innovation in education and teaching methods has brought the effectiveness of talent cultivation into full play. Through the research and analysis of relevant practical education and its constituent elements, as well as the internal logical relationship between each element, the shortcomings in the actual construction of the university curriculum system are discovered. Based on the theoretical framework that has been summarized and established, suggestions for IEPEP improvement are proposed with a view to promoting certain improvements and assisting in the exploration of effective curriculum reform and construction paths [36, 37].

When drawing on the conclusions of this study, it is important to note that:

1. the conclusions are based on students' entrepreneurial intentions and abilities, so they do not fully guarantee the success of students' entrepreneurship. Students will encounter many problems in the process of entrepreneurship. It is necessary to analyze them according to specific problems;
2. entrepreneurial activities or practical education need to be carried out based on the characteristics of students in a particular school. Different schools have different training goals, so it is necessary to carry out activities or practical education on entrepreneurship according to the training goals of each school in order to improve the chances of students' entrepreneurial success;
3. the use of the evaluation model needs to pay attention to the combination of school and business evaluation. Schools focus on assessing students' innovation and entrepreneurial ability, while enterprises focus on assessing the effectiveness of students' entrepreneurial implementation. Only the full combination of the two can help students succeed in entrepreneurship more effectively.

The topic of this research focuses on the development of IE education in China, and it is necessary to strengthen the combination of practice and increase practical education courses [38, 39]. Therefore, while achieving the entrepreneurship education goal, it helps to achieve the quality education goal and solve the dilemma of insufficient entrepreneurship education in China. In addition, based on the literature collection and organization, the research results related to this research are obtained to construct the curriculum, activity, and evaluation patterns of IE education, through the analysis and understanding of IEPEP development and integration in colleges. On this account, IE talent cultivation pattern is also optimized. However, the applicability of the established pattern to different countries and regions needs to be further studied [40, 41, 42]. Different education and teaching methods in different countries and regions also put forward higher requirements for IEPEP. Meanwhile, using the results of this research can stimulate students' entrepreneurial intentions and enhance their entrepreneurial abilities, which provides students with the necessary conditions to start their own business. Students first need to have the idea of starting a business, and secondly, they need to have the ability to start a business in order to really implement the business plan. However, there are various risks in

the process of starting a business, such as national policies, market economy, industry development prospects, and marketing barriers. All of these will play an important role in the success of the business. Therefore, it is important for students to have sufficient abilities to better cope with various types of risks and to survive them. This study simulates the possible risks involved in entrepreneurship and gives examples of how to apply what they have learned to deal with them.

On the one hand, this research provides an effective reference case for the development of entrepreneurship courses in colleges. Compared with the existing work, this research takes college's IEPEP as the research object. Through the research on IEPE curriculum pattern in colleges, the detailed and real situation of the current college IEPE pattern is mastered and understood. On the other hand, this provides reliable data for the relevant research on IEPE courses and helps to build a more valuable curriculum system that deeply integrates IE and professional education.

6. Concluding remarks. The research on IEPEP construction in colleges under the Internet plus era can give full play to the optimization and integration of the Internet in the IEPE system, and can also realize the practical education activities based on cultivating students' innovative spirit and entrepreneurial ability. It is based on a sound IE education system, guaranteed by a sound IE education mechanism, and aims to enhance students' basic literacy of IE and their ability to innovate independently. By combining the country's needs for IE talents and the professional characteristics of students, it promotes the maximum development of students in society, and ultimately improves the chances of students' entrepreneurial success. The conclusion confirms that the IEPEP under Internet plus can enhance students' entrepreneurial intention and strengthen their entrepreneurial ability, providing guidance and guarantee for their future entrepreneurial success. Under the Internet plus era, IE education for college students is a long-term and complex project. Such an Internet plus action plan plays a positive role in promoting IE education in colleges. As the main body of college students' IE education, colleges need to actively carry out this education. This plays an important role in promoting college graduates' entrepreneurship, improving employment rate and employment quality, building a harmonious society, and boosting sustainable social and economic development.

Funding. This research was funded by the 2021 annual topic of the 14th Five-Year Plan of Jiangsu Education Science (Grant No. X-c/2021/42), and the 2021 Southeast University Party Building Research Project (Grant No. DJ202113).

Appendix: Questionnaire on the Impact of xx School IEPE on College Students' Entrepreneurship.

Dear Alumni:

Hello! Thank you for participating in this questionnaire survey. This survey focuses on the impact of IEPE on the entrepreneurship of college students in xx school, and its purpose is to promote the further improvement and development of IEPE in this school. We promise that this questionnaire is completely anonymous and confidential, and the results are only used for academic research, which will not affect you in any way. Please fill in the questionnaire according to your actual situation and feelings. We sincerely appreciate your support and cooperation!

Note that it takes about 3-5 minutes to complete this questionnaire!

1. What is your gender? [Single choice]
   - Male
   - Female
2. What is your education? [Single choice]
   - Undergraduate
   - Master
   - Ph.D.
   - Postdoctoral
3. What grade are you in? [Single choice]
   - 1st grade
   - 2nd grade
   - 3rd grade

- 4th grade
- 5th grade

4. What is your major? [Single choice]
   - Economics and management
   - Art
   - Science and engineering
   - Medicine
   - Others
5. How is your academic performance ranked in the grade? [Single choice]
   - Before 3
   - 3
   - 10
   - After 50
6. Have you ever had student cadre experience? [Single choice]
   - Yes
   - No
7. Have you ever had part-time job experience during school? [Single choice]
   - Yes
   - No
8. The status of receiving IE education during school: [Single choice]:

	1	2	3	4	5
Classes and courses in IE education	◯	◯	◯	◯	◯
Lectures and activities in IE education	◯	◯	◯	◯	◯
Competitions in IE Education	◯	◯	◯	◯	◯
Entrepreneurship practice	◯	◯	◯	◯	◯
Entrepreneurship communication	◯	◯	◯	◯	◯
Entrepreneurship training	◯	◯	◯	◯	◯

9. Regarding the entrepreneurship intention, have you ever had the following ideas? [Single choice]:

	1	2	3	4	5
1) If there is a suitable opportunity, I am willing to suspend my studies and start a business, and take the risk of postponing my graduation	◯	◯	◯	◯	◯
2) The possibility of starting a business while I am in school is very high	◯	◯	◯	◯	◯
3) Within three years after graduation, there is a great possibility that I can start a business	◯	◯	◯	◯	◯

10. Through IE education, how has your entrepreneurship ability been improved? [Single choice]:

	1	2	3	4	5
I am satisfied with the availability and sharing of resources needed to complete team tasks	◯	◯	◯	◯	◯
Available resources increase the overall effectiveness of my team	◯	◯	◯	◯	◯
I am satisfied with resource management across the organization	◯	◯	◯	◯	◯
I think I can easily recruit like-minded partners with similar value orientation	◯	◯	◯	◯	◯
I have the ability to make a scientific and reasonable design for the rights, responsibilities and equity of all stakeholders	◯	◯	◯	◯	◯
I can promptly discover conflicts and contradictions within the team and resolve them effectively	◯	◯	◯	◯	◯
I can mobilize my parents to solve my entrepreneurial funding problem	◯	◯	◯	◯	◯
I can use my social network to solve venture capital	◯	◯	◯	◯	◯
I can find suitable venture capitalists or well-funded partners through roadshows, conferences, etc.	◯	◯	◯	◯	◯

REFERENCES

[1] Chen, C. & Liu, Z. Evolution Multiple Logics, and Optimization Paths of Universities' Classified Management Policies-A Implementation Review of the Outline of the Medium and Long-Term Education Reform and Development Plan of the People's Republic of China (PRC) (2010-2020) for Ten Years. *Jiangsu Higher Education, No.* **10** pp. 35-45 (2020)

[2] Peng, C. Achievements Problems and Prospects of Cross-border School Running in my country-A Review of the Outline of the Medium and Long-Term Education Reform and Development Plan of the People's Republic of China (PRC) (2010-2020). *Journal Of Hebei Normal University (Educational Science).* **23**, 65-72 (2021)

[3] Gou, W. Ten Years of Higher Education Development in Western China: Achievements, Challenges and Prospects-A Implementation Review of the Outline of the Medium and Long-Term Education Reform and Development Plan of the People's Republic of China (PRC) (2010-2020). *Research On Higher Education Of Nationalities.* **10**, 20-31 (2022)

[4] Tretyakova, N., Lyzhin, A., Chubarkova, E. & Others (2021). *Mobile-Learning Platform For The Development Of Entrepreneurial Competences Of The Students.* **15**, 118-135 (0)

[5] Li, C. & Li, W. Research on the Construction of "Practice and Innovation" Integration System and Model Innovation in Higher Education Institutions-A Case Study of Beijing Institute of Fashion Technology. *Creative Education.* **14**, 2023 (2023)

[6] Guo, W., Liu, L., Yao, Y. & Others (2022). Refinement Method of Evaluation and Ranking of Innovation and Entrepreneurship Ability of Colleges and Universities Based on Optimal Weight Model. (Discrete Dynamics in Nature,0)

[7] Xing, R. Research and Practice on Innovation and Entrepreneurship Model of Higher Vocational College Students. *Education Science, No.* **1** pp. 2022 (2022)

[8] Janardhan, V. & Janardhan, V. Engaging Early-Career Physicians in Medical Device Innovation and Entrepreneurship. *Stroke: A Journal Of Cerebral Circulation, No.* **5** pp. 2022 (2022)

[9] Liu, Y., Tang, Y., Bi, X. & Others (2018). *Cultivation And Practice Of College Students' Innovative Entrepreneurship Based On Second ClassroomInternational Conference On Entrepreneurial Economics-Science Topic.* pp. 113-116 (2018)

[10] Liu, L. & Wang, Y. Innovation and Entrepreneurship Practice Education Mode of Animation Digital Media Major Based on Intelligent Information Collection. *Mobile Information Systems, No.* **11** pp. 1-11 (2021)

[11] Akhmetshin, E., Mueller, J., Chikunov, S. & Others (2019). *Innovative Technologies In Entrepreneurship Education: The Case Of European And Asian Countries.* **22** pp. 1 (0)

[12] Ashar, M., Kamdi, W. & Kurniawan, D. Professional Skills Development Through the Network Learning Community Using an Online Learning Platform. *International Journal Of Interactive Mobile Technologies (iJIM).* **15**, 202-210 (2021)

[13] Xia, J. & Li, Y. On the Construction of Teaching Staff in Higher Vocational Colleges under the Background of Innovation and Entrepreneurship. (DEStech Transactions on Engineering,2017)

[14] Shi, J. Research on Innovative Entrepreneurship Teaching in New Undergraduate Universities. (The Guide of Science & Education,2018)

[15] Clercq, D., Dimov, D. & Belausteguigoitia, I. Perceptions of Adverse Work Conditions and Innovative Behavior: The Buffering Roles of Relational Resources. *Entrepreneurship: Theory And Practice.* **40** pp. 3 (2016)

[16] Servoss, J., Chang, C., Olson, D. & Others (2017). (An Experiential Learning Program for Surgery Faculty to Ideate,0)

[17] Zhu, M., Kim, I. & An, Z. Optimizing the Construction of Multidimensional System of Entrepreneurship Education from the Perspective of the Second Classroom. (Scientific programming,2021)

[18] Li, L. & Yang, W. Research on Time and Space Features of Hot Sports of College Students' "Internet Plus" Innovation & Entrepreneurship and Its Enlightenment to High Education. (Future,2019)

[19] Han, X. & Polytechnic, H. Research on Innovation and Entrepreneurship Platform of Higher Vocational Colleges under the Background of "Internet Plus Double Creation". (Heilongjiang Science,2019)

[20] Fang, Z., Razzaq, A., Mohsin, M. & Others (2022). *Spatial Spillovers And Threshold Effects Of Internet Development And Entrepreneurship On Green Innovation Efficiency In China.* **68** (0)

[21] Gomes, S. & Lopes, J. ICT Access and Entrepreneurship in the Open Innovation Dynamic Context: Evidence from OECD Countries. *JOItmC.* **8** pp. 2022 (2022)

[22] Zhao, L. Research on the Path of Higher Vocational Innovation and Entrepreneurship Education from the Perspective of Curriculum Ideology and Politics. *Education Science, No.* **2** (2022)

[23] Sheng, D., Wang, Y. & Mehmood, T. Design of Innovation and Entrepreneurship Education Ecosystem in Universities Based on User Experience. (Mathematical Problems in Engineering,2022)

[24] Cai, J. Current Situation of Entrepreneurship Education and Talent Training Mode of College Students in China. *CIPAE 2021: 2021 2nd International Conference On Computers.* (2021)

[25] Yang, X. & Luo, M. Research on the Talent Training Mode of Application- oriented Undergraduate Cross-border E-commerce Innovation and Entrepreneurship Education. 2020 International Conference on Big Data and Informatization Education (ICBDIE). (IEEE,2020)

[26] Li, C. & Li, W. Research on the Construction of "Practice and Innovation" Integration System and Model Innovation in Higher Education Institutions—A Case Study of Beijing Institute of Fashion Technology. *Creative Education.* **14** pp. 1 (2023)

[27] Liu, T., Sun, H., Fung, W. & Others (2021). (An Artifact-based Simulation Method for Teaching Intellectual Property Management in an Innovation,0)

[28] Chen, L., Zhang, X., Huang, L. & Others (2018). (Career Development Education for College Students. Chongqing University Press,0)

[29] Wang, Z. & Hao, W. Innovation and Entrepreneurship Education in New Liberal Arts: Connotation, Model and Evaluation. *Journal Of Ningbo University (Educational Science Edition).* **45**, 102-110 (2023)

[30] Zhang, Y. Research on the Innovation and Entrepreneurship Course Construction of Social Sports Major Basketball Under the Background of Internet Plus. (Journal of Nanchang Hangkong University (Social Sciences,2019)
[31] Collins, C., Hanges, P. & Locke, E. The Relationship of Achievement Motivation to Entrepreneurial Behavior: A Meta-Analysis. *Human Performance*. **17**, 95-117 (2004)
[32] Wang, X., Bo, F. & Lei, J. Research on the Influence of Entrepreneurship Education on College Students' Entrepreneurship Intention-Comparison of Undergraduates and Vocational Students. *Tsinghua Journal Of Education*. **37**, 116-124 (2016)
[33] Li, X. & Huang, J. A Survey of the Undergraduates' Satisfaction with Innovation and Entrepreneurship Education—Taking as an Example Public Management Specialty in Fujian Agriculture and Forestry University. (Journal of Jimei University (Education,2017)
[34] Xu, X. The Construction of Innovative Practical Curriculum System for Business English Major Under the Background of Cross-Border E-commerce. *Advances In Higher Education*. **3** pp. 4 (2019)
[35] Shen, Y. Research on Practical Education of Speciality Teachers in Vocational Colleges to Promote Curriculum-based Ideological and Political Education. *Education Reserch Frontier*. **12** pp. 2 (2022)
[36] Fang, F., Zhai, X., Chen, H. & Others (2015). Teaching Reform of Traditional Chinese Medicine Practical Technology Curriculum System Based on Military Medical Position Demand. (Chinese Medicine Modern Distance Education of China,0)
[37] Pratikto, H., Hanafiya, R., Ashar, M. & Others (2021). Entrepreneurship Game Apps to Enhancement Student Skill Thinking Analytic in Class Online. *International Journal Of Interactive Mobile Technologies (iJIM)*. **15**, 155-162 (0)
[38] Liang, G., Walls, R. & Lu, C. Standards and Practice for Physical Education in China. *Journal Of Physical Education Recreation & Dance*. **76**, 15-19 (2005)
[39] Li, D. Training doctors for primary care in China: Transformation of general practice education. *Journal Of Family Medicine And Primary Care*. **5** pp. 1 (2016)
[40] Holmes, A., Tissot, S., Mcleod, K. & Others (2022). *Optimizing Surgical Training In The Time Of COVID-*. **92**, 336-340 (0)
[41] Adriana, D. Education and Italian regional development. *Economics Of Education Review*. **27**, 94-107 (2008)
[42] Liu, T., Zhao, J. & Li, S. Research on Regional Basic Education Quality Assessment Based on Deep Convolutional Neural Network. *Journal Of Circuits, Systems And Computers*. **32** pp. 4 (2023)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Jun 9, 2023
Accepted: Oct 7, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 481–494, DOI 10.12694/scpe.v25i1.2399

IMPROVE ENGLISH LEARNING THROUGH ARTIFICIAL INTELLIGENCE FOR ONLINE AND OFFLINE MIXED TEACHING PATH

LI HUANG*

Abstract. With the development of higher education in China, English has been listed as a compulsory course in basic education, and people's demand for English teaching level is also increasing. However, with the advent of the information age, there have been significant changes in the dissemination of knowledge, teaching methods, and teaching environment. Traditional English teaching models can no longer meet the requirements of modern society for versatile talents. Therefore, exploring the mixed online and offline English teaching path has become particularly important. Artificial intelligence (AI) is an advanced technology today. It can evaluate students' pronunciation and provide corresponding guidance, thereby increasing students' interest in English and improving English pronunciation accuracy. It can also provide personalized learning content and learning paths according to students' learning needs and levels. By analyzing students' learning situation and feedback, it automatically adjusts and optimizes learning resources so that students can better master English knowledge and skills. The traditional English teaching model is often limited by time and space, while the AI-based hybrid teaching model can integrate a variety of educational resources around the world, including high-quality course content, learning resources and online tutoring services, so that students can get the most suitable educational resources anytime and anywhere, which can improve learning efficiency.

Key words: Artificial Intelligence, Speech Recognition Technology, Mixed Teaching, English Teaching

1. Introduction. In the context of global economic integration, students must proficiently master English skills in order to achieve higher levels of development. In English teaching, teachers should recognize the current development trends and teaching changes, and reflect on the shortcomings of traditional teaching models. By establishing a blended teaching model of "online+offline" in English, English teaching can exhibit a completely different new style from the past and achieve greater development in the new era. People should extend the application of AI to the field of education, and language is the proudest of mankind. Whether it is natural language processing, machine translation, speech recognition, or other AI technologies, language learning is indispensable. Applying AI to English teaching would be an important breakthrough in the field of online and offline hybrid teaching of English. In practice, "flipped classroom model" is a common and widely practiced blended learning style. Flipped classroom mode refers to the transfer of traditional classroom teaching content to the online learning platform, and the practice, exploration and discussion and other activities are carried out in the classroom. In this mode, students can independently learn relevant knowledge and concepts through online learning resources (such as videos, textbook readings, etc.) before class. Then, in class, students interact with peers and teachers to engage in practical activities such as problem solving, case studies, and experimental operations to deepen understanding and apply what they have learned.

The rapid development of information technology has had a profound impact on human life and work, and has brought unprecedented changes to English teaching. However, simple classroom teaching can no longer meet students' learning needs, so it is imperative to carry out blended English teaching. Albiladi Waheeb proposed that blended learning is a relatively new field, which combines traditional distance learning and online learning methods. Blended learning can be effectively used to develop language skills and improve the English learning environment, thus promoting students' language learning [1]. Millorpe Naomi introduced a cooperative teaching project in English courses, called "Mixed English", which involves the use of online and mobile technology to develop, implement and evaluate learning and teaching activities for students' English units. The hybrid approach he described has the ability to enhance subject learning and increase enrollment opportunities for students in remote areas, thereby promoting deeper academic research [2]. Huang found that in recent years, blended learning has become a popular teaching model at all levels of education and

*Nanchang University College of Science and Technology, jiujiang 332020, Jiangxi, China; Li_Huang332@outlook.com

in different disciplines. He believed that from education to the role of teachers, there are paradigm shifts in different areas of blended learning. He studied how students view their teachers' role in blended English learning. The results showed that in the eyes of students, teachers have more influence in blended learning than in online learning [3]. Ginaya Gede described a structured attempt to investigate the impact of blended learning and traditional teaching on students' oral proficiency. By analyzing the results before and after the test through planning, action, and observation, the survey results showed that students who participated in blended learning significantly improved their English-speaking ability and enhanced their learning motivation and interest [4]. Sari aimed to reveal the advantages of blended learning and students' motivation for learning English. Blended learning combines the positive aspects of the traditional model with improved technology to maintain, improve and attract students' enthusiasm and participation. Blended learning improves access to materials and learning activities, which can support and strengthen the role of teachers, students' experience and social environment [5]. The immersive teaching model of the flipped classroom is indispensable in blended and online language learning. Using a quasi-experimental design, Yulian R conducted pre-test and post-test paired T-test for critical reading. The participants were 37 second-semester students in an academic English class. The results showed that flipped classroom teaching model improved students' critical thinking from the pre-test average (12.4865) to the post-test average (18.3243). In terms of self-directed learning, students have a positive view of the implementation of the model [6]. Based on the cloud computing artificial intelligence model, Liang X made an in-depth summary and analysis of the interactive English teaching mode, explored the characteristics of intelligent classrooms, and practiced the reform of the interactive teaching mode. According to the investigation results, the construction of artificial intelligence course teaching model is optimized to make it more perfect. Based on cloud computing technology, the system architecture and functional module division of online open course platform are designed according to the overall demand, and the development and implementation are carried out on this basis [7].

The above scholars believed that the application of blended learning mode in practical teaching can improve the quality of English teaching.

Introducing AI technology into English blended learning is a practical approach. AI is a discipline that studies how to use computers to simulate human intelligence. With the development of modern computer technology and the rapid advancement of social informatization, people have increasingly focused their attention on the application of AI in real life. There is no doubt that the application field of AI should be expanded from the perspective of education. How to effectively teach English has become a consensus among many linguists and a problem faced by many teachers and students. In summary, the application scope of AI urgently needs to be extended to the construction of English teaching paths, and the construction of blended English teaching cannot be separated from the promotion of AI.

2. Construction of English Blended Teaching Based on AI. In the new era, students' learning status and enthusiasm have undergone significant changes, and their dependence on smart devices such as mobile phones has become increasingly strong, bringing new challenges to traditional English teaching.

In blended learning, various types of AI tools can be used to assist the teaching and learning process. The virtual experiment platform enables students to operate and observe experiments and gain practical experience by simulating experiment environment and virtual experiment equipment. AI based blended learning refers to the combination of online learning and offline traditional classroom teaching. Before class, teachers would formulate the content they want to learn into preview tasks, and then send them to students on online platforms [6, 7]. Students can use online platforms to engage in online learning. When they encounter problems, they can communicate and discuss with teachers and classmates at any time, and can solve these problems in a timely manner. In the offline teaching stage, teachers can provide centralized answers to common problems that students encounter in online learning and engage in deeper discussions on these issues.

Blended teaching can effectively mobilize students' subjective initiative and expand limited in class knowledge to unlimited extracurricular knowledge, thereby achieving the goal of improving teaching quality [8, 9]. This article constructed blended English teaching from three perspectives: pre class collaborative preview, in class teacher-student interaction, and post class collaborative expansion, which integrated online and offline teaching and promotes each other. The English blended teaching based on AI is shown in Figure 2.1.

As can be seen from Figure 2.1, with the rapid development of artificial intelligence technology, traditional

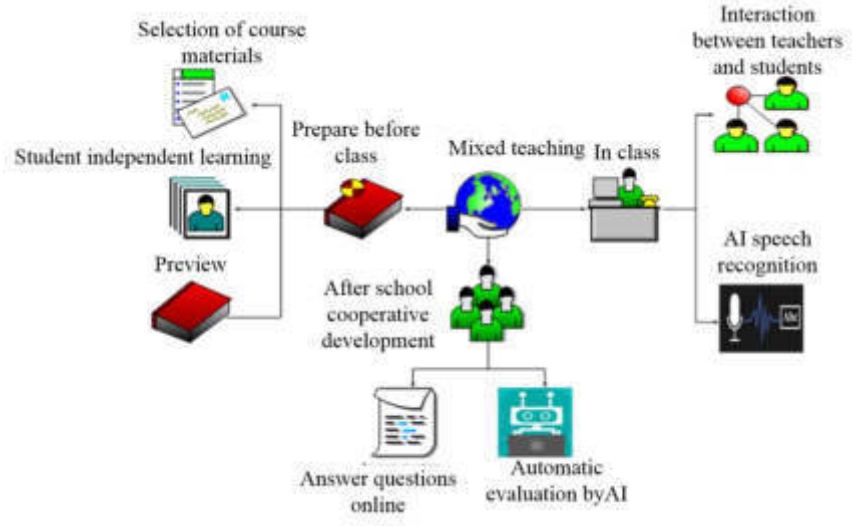


Fig. 2.1: Schematic diagram of AI based blended English teaching

English teaching is gradually shifting to a hybrid teaching method combining online and offline. Teachers can easily use the network platform to carry out teaching activities, so as to complete the comprehensive integration and storage of digital media resources, as well as teaching activities such as answering questions and correcting homework after class.In blended learning, students can easily preview, study in class, review after class, interactive discussion, online question and answer and other processes [10, 11]. The blended teaching based on online and offline teaching not only gives full play to the comprehensive supervision and guidance role of teachers, but also fully considers the needs of students' independent learning [12].

In Figure 2.1, with the rapid development of AI technology today, traditional English teaching is gradually shifting towards a hybrid approach that combines online and offline teaching. Teachers can conveniently use online platforms to carry out teaching activities, thereby completing the comprehensive integration and storage of digital media resources, as well as teaching activities such as answering questions and correcting homework after class. In blended learning, students can easily engage in preview, classroom learning, after-school review, interactive discussions, online Q&A, and other processes [10, 11]. The blended teaching based on online and offline teaching not only fully utilizes the comprehensive supervision and guidance role of teachers, but also fully considers the needs of students' autonomous learning [12].

2.1. Preparation before Class.

2.1.1. Selection of course materials. In terms of selecting and developing course materials, educators should select and develop corresponding course materials based on the current teaching objectives and students' basic conditions, and the difficulty should not be too high or too low. Moreover, when setting the teaching time, it should not be too short or too long. Within the specified time limit, students can ensure that their attention is fully utilized. If the time is too short, students may have finished before they can enter the learning state. The arrangement of extracurricular homework should be in line with the teaching objectives, with a focus on laying a solid foundation rather than promoting it. Too much or too difficult content can cause students to develop resistance towards English, which cannot effectively improve their English proficiency and can actually dampen their enthusiasm for learning. The selection of course materials is a key link in combining offline and online teaching, which can be considered from several perspectives such as students' requirements, teaching objectives, and learning outcomes.

2.1.2. Students' autonomous learning. Using artificial intelligence technology to analyze students' learning behavior: by monitoring the activity record, browsing history, learning time and other data of stu-

dents' online learning platform, it can understand students' learning habits, learning progress and learning effect. In traditional face-to-face teaching, if the teaching time is set at noon, it would cause some students to doze off during this period, affecting the efficiency of classroom teaching. In blended teaching, students can choose their own time to learn. The blended teaching mode refers to a new type of experiential teaching that integrates modern teaching methods and traditional teaching concepts by utilizing multiple experiences, combining multiple contents, and integrating multiple methods. Since the outbreak of the epidemic, the call of the Ministry of Education to "leave school without leaving teaching, suspend classes without stopping" has further accelerated the deep integration of online and offline teaching. Blended teaching not only fully stimulates students' subjective consciousness, but also allows them to perceive the charm of learning through creative activities. Especially among English majors in colleges and universities, the effective application of blended teaching can, to a large extent, enable students to use fragmentation time for English conversation practice and situational learning. This not only effectively improves students' enthusiasm for active learning English, but also cultivates students' intercultural communication ability.

2.1.3. Preview. In blended English teaching, preview is the primary link. Due to the constraints of the number of students, class hours and other factors, the traditional classroom teaching cannot be fully discussed, so there are few opportunities for students to practice in the classroom. In contrast to the pre class preview in the network environment, teachers can set learning goals and topics according to the teaching content, so that students can form a group to carry out cooperative learning [13, 14]. At the same time, teachers can also send relevant audio, video, and courseware to students. When students encounter problems during the preview process, students can communicate with their group members and solve the problems as soon as possible. Teachers can divide students into groups based on their interests, professional background, or other factors. It can choose random groups or choose group members according to the students' willingness, and ensure that the number of members in each group is moderate, so that the effect of cooperative learning can be fully utilized.

When encountering controversial issues, students can seek help from teachers. Teachers can use the online platform where students preview to monitor and guide their pre class learning in real-time, preparing and laying the groundwork for classroom explanations. During the process of preview and discussion, students can deeply reflect on the content of the topic, and find their own sense of value through expressing their own opinions and listening to others, thereby improving their subjectivity. In addition, they can also improve their sense of belonging and relieve pressure in group cooperative learning. Finally, each group summarizes the results of the discussion and communicates with teachers and other students in class.

2.2. Teacher-student Interaction in Class. In blended teaching, the teaching method of teachers has changed from simple teaching to student-centered, in order to stimulate students' autonomy and initiative [8]. Students can use online platforms to interact and communicate with teachers and classmates, which can truly achieve teacher-student interaction, student-student communication, and joint learning, thereby achieving understanding and application of knowledge and skills.

Teachers can also showcase their excellent teaching achievements online, providing students with learning references. At the same time, in order to enable students to consolidate and expand their knowledge, teachers should also analyze the learning situation and abilities of each student, and set up expansion content for students who have spare time, encouraging them to dare to challenge themselves. For students with poor learning abilities, teachers should encourage them more and guide them in understanding and learning knowledge. AI plays an increasingly important role in English teaching. Therefore, it is necessary to integrate AI technology into blended learning. Speech recognition technology can enhance listening and speaking abilities, or achieve automated evaluation of students' reading and writing skills through natural language understanding.

2.2.1. Expansion of After-school Cooperation. Extracurricular expansion refers to extending the learning time in the classroom to extracurricular activities. In the past, in English classes, everyone had to be wary of missing important knowledge, so the atmosphere in the classroom was quite oppressive. In blended English teaching, teachers can use various software to post classroom videos and courseware online, allowing students to consolidate and review after class. At the same time, they can also ensure that there are no omissions through their own self-learning and inquiry [16, 17].

Teachers can also assign some after-school exercises and follow-up discussion topics to students, allowing

them to engage in group discussions and exchanges. Students can express their opinions and exchange learning experiences with classmates and teachers at any time. Teachers can monitor students' learning situation and promptly solve difficulties encountered, which makes up for the time limitations of traditional classroom teaching. Google offers a cloud-based speech recognition API that converts speech to text. Developers can do voice recognition by calling the API provided by Google.

2.2.2. AI Based Speech Recognition. AI can analyze students' learning records to gain a better understanding of their learning tendencies and needs, and provide corresponding content and functions to meet their requirements, timely correcting their grammar, pronunciation, vocabulary use, and language application. AI has also received increasing attention in English assisted teaching, and the application of AI in teaching research has gradually matured. The availability of AI in terms of interactivity, flexibility, and more space for choosing answers would be of great interest to language teachers. In AI, intelligent speech recognition technology can achieve automatic evaluation, defect localization, and problem analysis of the speaker's speech level [18, 19].

Speech recognition technology is to convert the vocabulary in human voice into computer readable input, such as keyboard, binary code or string. In the process of speech signal processing, it is often necessary to segment it and use window function to truncate it, so that it has less distortion.

By multiplying a certain number of window function $W(n)$ by $\overline{S(n)}$, the windowed voice signal S_w can be obtained:

$$S_W = \overline{S}(n) * W(n) \tag{2.1}$$

In speech processing, the window function is also the basic data to be used. There are generally two types: one is the rectangular window, and the other is the Hamming window. The rectangular window is:

$$W(n) = \begin{cases} 1, & \text{if } 0 \le n \le N-1 \\ 0, & n = \text{else} \end{cases} \tag{2.2}$$

Hanming Window $W(n)'$ is:

$$W'(n) = \begin{cases} 2\pi n, & \text{if } 0 \le n \le N-1 \\ 0, & n = \text{else} \end{cases} \tag{2.3}$$

The larger the value of the window function, the better the effect of the filter and the smoother the signal. On the contrary, if the window function is narrow, the filtering effect cannot be achieved. For extracted speech, its slope should gradually decrease to 0 to avoid truncation effects.

When analyzing speech signals in time domain, it is needed to extract their features. The easiest to get is the short-term average Zero-crossing rate, that is, the signal in each frame, whose frequency passing through the zero point is called Zero-crossing rate. The Zero-crossing rate can truly reflect the spectrum properties:

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} |\text{sgn}[x_n(m)] - \text{sgn}[x_n(m-1)]| \tag{2.4}$$

The Zero-crossing rate of voiceless $x_n(m)$ is generally high, because its energy is concentrated in the high-frequency signal segment; the Zero-crossing rate of voiced $x_n(m-1)$ is generally low because its energy is concentrated at a lower frequency. What lies between these two is usually noise.

Before training and recognizing speech, it is necessary to normalize the data to avoid saturation during output. The normalization formula is:

$$X_{mid} = \frac{X_{max} + X_{min}}{2} \tag{2.5}$$

X_{min} and X_{max} represent the minimum and maximum values of a row of data, respectively.

Due to the fast fluctuation speed and large fluctuation of speech signals in the time domain, they are usually transformed into frequency domain detection, where the fluctuation of their spectrum is relatively

Table 3.1: Basic Information

	Category	Group A	Group B	Total
Gender	Male	23	23	46
	Female	23	23	46
Age (years old)	18-20	25	22	47
	More than 20	21	24	45

gentle. Therefore, it is generally necessary to perform Fourier transform on windowed frames to obtain the spectral coefficients c_m of all frames:

$$c_m = \sum_{K=1}^{N} E_K \cos(m(k - \frac{1}{2})\frac{\pi}{2} \tag{2.6}$$

Intelligent speech recognition technology has a voice library. After the teacher or student reads through the microphone, the software would automatically record the user's voice and compare it with the correct pronunciation in the built-in voice library. After that, it would provide a score and mark the error points in the pronunciation. Intelligent speech recognition technology can provide timely feedback on the accuracy of learners' pronunciation, allowing teachers or students to have a better understanding of their own pronunciation problems, and then carry out targeted correction exercises. In English exams, teachers can use intelligent speech recognition technology to evaluate students' learning situation and timely correct students' incorrect pronunciation based on feedback information provided by the system. In addition, in the classroom teaching process, students can also use intelligent pinyin technology to learn and consolidate vocabulary, sentence structures, etc.

3. Evaluation of Blended Teaching Effectiveness Based on AI.

3.1. Impact of Intelligent Speech Recognition Technology on Hybrid Teaching. According to the hierarchical English teaching background of the school and the statistics of the final English scores of the last two semesters, the two classes with the smallest difference in the level of students within the class were selected.This article selected 92 English major students from a certain university for experimental analysis, and divided them into two groups: Group A and Group B, with 46 students in each group. Group A used traditional English teaching methods for teaching, while Group B used AI based English blended teaching for a period of 24 weeks, followed by comparative experiments. A survey was also conducted on the teaching effectiveness of 50 English teachers. At the end of the experiment, evaluations were conducted and the experimental data was statistically analyzed. The basic situation of students is shown in Table 3.1.

In Table 3.1, it can be seen that the number of males and females in Group A and Group B was the same. This was because in order to make the experiment comparable, the variables were deliberately controlled to ensure the same gender ratio, making the experiment more rigorous.

The reading function in AI based speech recognition technology can help students master English pronunciation correctly, which effectively solves some non-standard pronunciation problems that students may encounter during the learning process. In teaching, teachers can use intelligent speech recognition technology to avoid the negative impact of their incorrect pronunciation on students' English learning, so that they can learn correct pronunciation and correct their pronunciation in a timely manner. During the teaching process, teachers can also improve students' learning effectiveness by controlling their speaking speed and volume according to their own requirements. AI can provide more materials for English teaching and provide more convenience for teachers to carry out various English teaching activities. The application time of intelligent speech recognition technology in the classroom is shown in Table 3.2.

In Table 3.2, the cumulative time for classroom preview was 12.5 minutes, while the cumulative time for using intelligent speech recognition technology was 8.8 minutes. The usage time ratio of intelligent speech recognition technology was 70.4%, with the highest usage time ratio of intelligent speech recognition technology in pronunciation evaluation, which was 92.6%.

Table 3.2: Application Time of Intelligent Speech Recognition Technology in the Classroom

Types	Cumulative Time (minutes)	Use Time with Intelligent Speech Recognition (minutes)	Intelligent Speech Recognition Usage Time Ratio (%)
Class Preview	12.5	8.8	70.4
Import	9.8	6.5	66.3
Word Teaching	21.6	16.4	75.9
Sentence Pattern Teaching	19.3	13.3	68.9
Oral Communication Teaching	28.2	22.5	79.8
Review and Consolidate	37.1	27.7	74.7
Pronunciation Test	14.9	13.8	92.6

Intelligent speech recognition technology can not only enable teachers to achieve teaching objectives, but also utilize existing or self-made personalized teaching resources to continuously learn English for students. In addition, personalized reading pen cards made by teachers using intelligent speech recognition technology tools can also create a diverse learning environment, thereby stimulating students' learning enthusiasm. At the same time, through the speech evaluation function, students' voices can be tested and corrected.

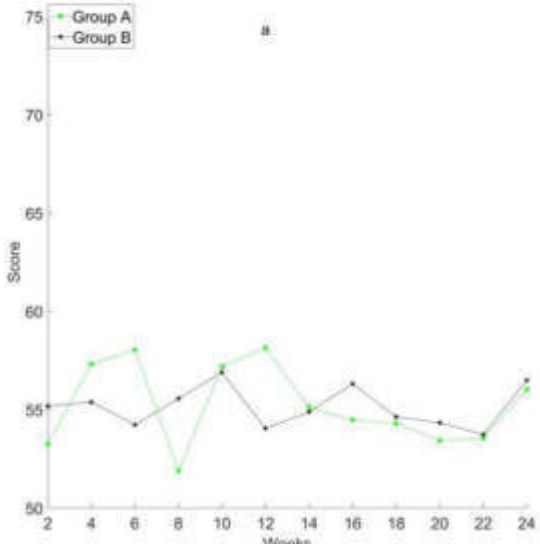
3.2. AI Based Blended English Teaching Effectiveness. This article aimed to compare the effectiveness of AI based blended English teaching with traditional teaching, and further explore whether AI can improve English teaching level. AI can provide students with an interactive English learning environment in English teaching and change the current situation of "deaf mute English", which allows for a high degree of freedom in teaching time, students' initiative in learning, personalized teaching, and improving students' English grades.

AI can provide an interactive English teaching environment and change the current situation of "deaf and mute English".

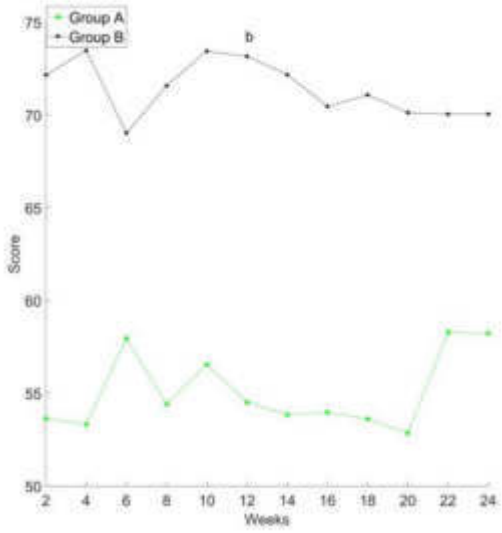
The traditional teaching method in English classrooms is relatively rigid, usually with teachers holding English textbooks to explain theoretical things to students, while students passively listen. This situation is no different from English learning in junior high school and high school before. If the English class is still the monotonous and tedious teaching method it used to be, it cannot stimulate students' enthusiasm for learning. The blended teaching of "online+offline" is the transfer of theoretical knowledge learning to online platforms. Teachers can guide students to self-study before class and organize various practical activities in the classroom, which enables students to proficiently apply English knowledge. In this way, traditional teaching methods can be effectively changed and a brand-new English classroom can be established. The interactive evaluation of teaching methods in Group A and Group B before and after the experiment is shown in Figure 3.2. The score is on a percentage scale.

In Figure 3.1(a), two weeks before the experiment, Group A and Group B scored 53.25 and 55.18 points respectively on the interactivity of their teaching methods; 24 weeks before the experiment, Group A and Group B scored 56.04 and 56.50 respectively on the interactivity of their teaching methods. Figure 3.1(b) shows that two weeks after the experiment, Group A and Group B scored 53.65 and 72.17 respectively on the interactivity of their teaching methods. At this point, compared to the two weeks before the experiment, the score change in Group A was relatively small, while the score change in Group B was significant. After 24 weeks of the experiment, Group A and Group B scored 58.25 and 70.06 respectively on the interactivity of their teaching methods. After 24 weeks of the experiment, both Group A and Group B improved their scores on interactivity. However, Group A's rating for interactivity improved very little, while Group B's rating for interactivity increased significantly, indicating that Group B's teaching methods have strong interactivity.

The English classroom requires an interactive teaching environment, and universities can use various methods such as English corners, foreign teachers, exchange students, and internships with foreign companies to create an interactive English teaching atmosphere. However, in traditional classrooms, students find it difficult to understand English, let alone communicate fluently in English, which leads to the problem of "deaf mute



((a)) Interactive evaluation of teaching methods in Group A and Group B before the experiment



((b)) Interactive evaluation of teaching methods in Group A and Group B after the experiment

Fig. 3.2: Interactive evaluation of teaching methods by Group A and Group B before and after the experiment

Table 3.3: Degree of Freedom in Traditional Offline Teaching

Degree of Freedom	Teachers	Percentage
Very Free	4	8%
Relatively Free	2	4%
In General	5	10%
Unfree	12	24%
Very Unfree	27	54%

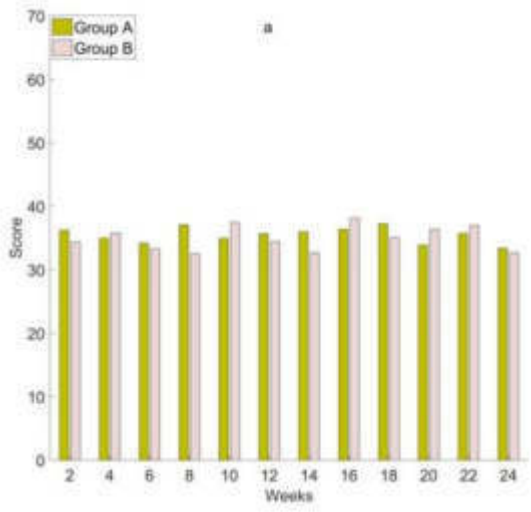
English" among students. However, AI can provide technical support for the development of English teaching environments. AI can be used to comprehensively process various media information such as text, images, sound, etc., and logically associate and integrate them to form a complete set of representations, thus providing multiple interactive modes for English teaching.

3.2.1. High degree of freedom in teaching time. Adopting a blended teaching approach can effectively expand the space of English teaching and improve the freedom of teaching time. Traditional English classroom teaching is limited by both time and space, which brings great pressure to teachers' teaching. In order to complete teaching tasks, English teachers often adopt a "one talk" teaching model, while students only mechanically memorize some words, grammar, and problem-solving skills, which makes English classrooms dull and unable to achieve individualized teaching. 50 teachers experienced a session of traditional offline instruction as well as a session of blended instruction, and then analyzed the freedom they perceived between the two types of instruction. The degree of freedom that 50 teachers believe in traditional offline teaching is shown in Table 3.3.

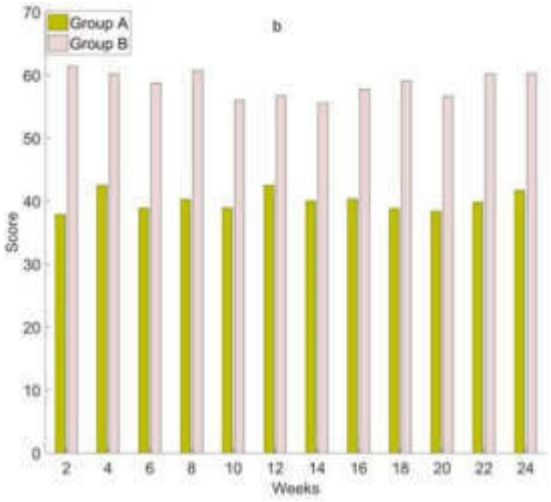
In Table 3.3, only 4 teachers believed that traditional offline teaching is very free, accounting for only 8% of the total; only 2 teachers believed that traditional offline teaching is relatively free, accounting for only 4% of the total; 27 teachers believed that traditional offline teaching is very unfriendly, accounting for 54% of the

Table 3.4: Degree of Freedom in Blended Teaching

Degree of Freedom	Teachers	Percentage
Very Free	42	84%
Relatively Free	5	10%
In General	3	6%
Unfree	0	0%
Very Unfree	0	0%



((a)) Proactive rating of students in groups A and B before the experiment



((b)) Proactive rating of students in Group A and Group B after the experiment

Fig. 3.4: Proactivity scores of students in Group A and Group B before and after the experiment

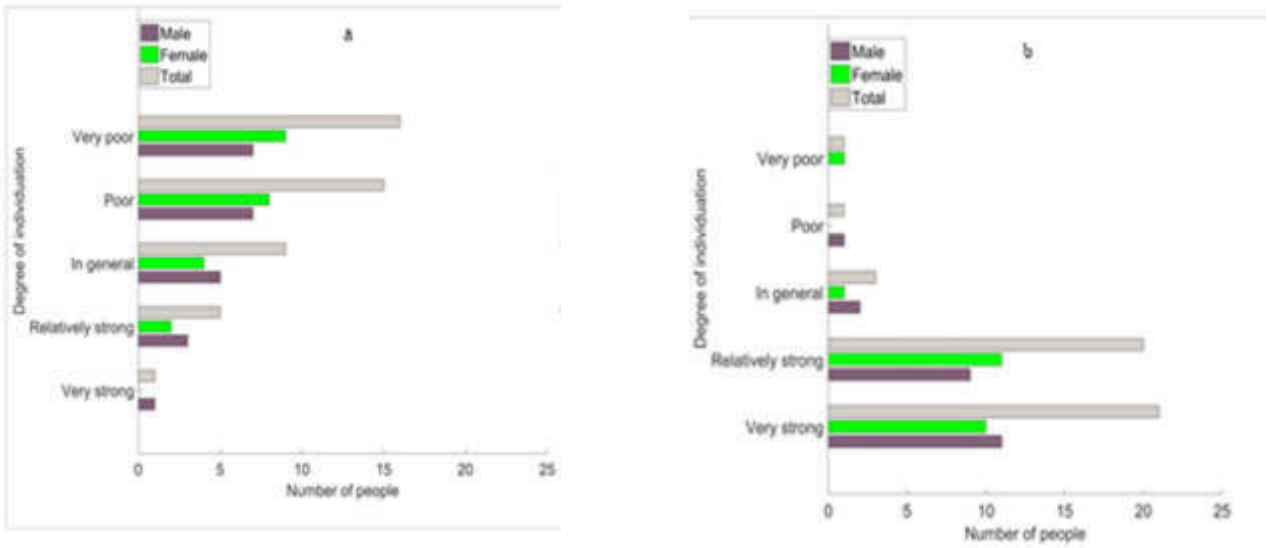
total. The degree of freedom perceived by 50 teachers in blended learning is shown in Table 3.4.

In Table 3.4, there were 42 teachers who indicated that blended learning is very free, accounting for a high percentage of 84%; there were 5 teachers who indicated that blended learning is relatively free, accounting for a percentage of 10%; the proportion of teachers who indicated that blended learning is very free and relatively free added up to 94%. It can be seen that blended learning has a very high degree of freedom.

The most prominent feature of blended English teaching under AI is the high degree of freedom in teaching time. With the support of AI, teachers no longer need to teach at fixed times and locations, but can assign learning tasks at any time and location before and after class. This allows students to learn fragments according to their actual situation, enhancing teaching flexibility.

3.2.2. Strengthening students' initiative. Mixed teaching focuses on the cultivation of students' active learning ability, thus forming a good atmosphere for students' active learning. Blended teaching can effectively solve the lack of students' initiative cultivation in traditional teaching models, and students would not actively raise their own problems during the learning process. The initiative scores of students in Group A and Group B before and after the experiment are shown in Figure 3.4.

In Figure 3.3(a), two weeks before the experiment, students in Group A and Group B rated their learning initiative as 36.21 and 34.42, respectively; 24 weeks before the experiment, students in Group A and Group B rated their learning initiative as 33.39 and 32.69, respectively. Before the experiment, students in both groups had low learning initiative; during the 24 weeks before the experiment, students in Group B rated their learning



((a)) Comparison of personalization levels in Group A after the experiment

((b)) Comparison of personalization levels in Group B after the experiment

Fig. 3.6: Comparison of personalization levels between Group A and Group B after the experiment

initiative even lower than those in Group A.

Figure 3.3(b) shows that two weeks after the experiment, students in Group A and Group B rated their learning initiative as 37.84 and 61.40, respectively; after 24 weeks of the experiment, the scores of students in Group A and Group B on their learning initiative were 41.69 and 60.23, respectively. After the experiment, the learning initiative of Group B students made significant progress, while the change in learning initiative of Group A students was not significant.

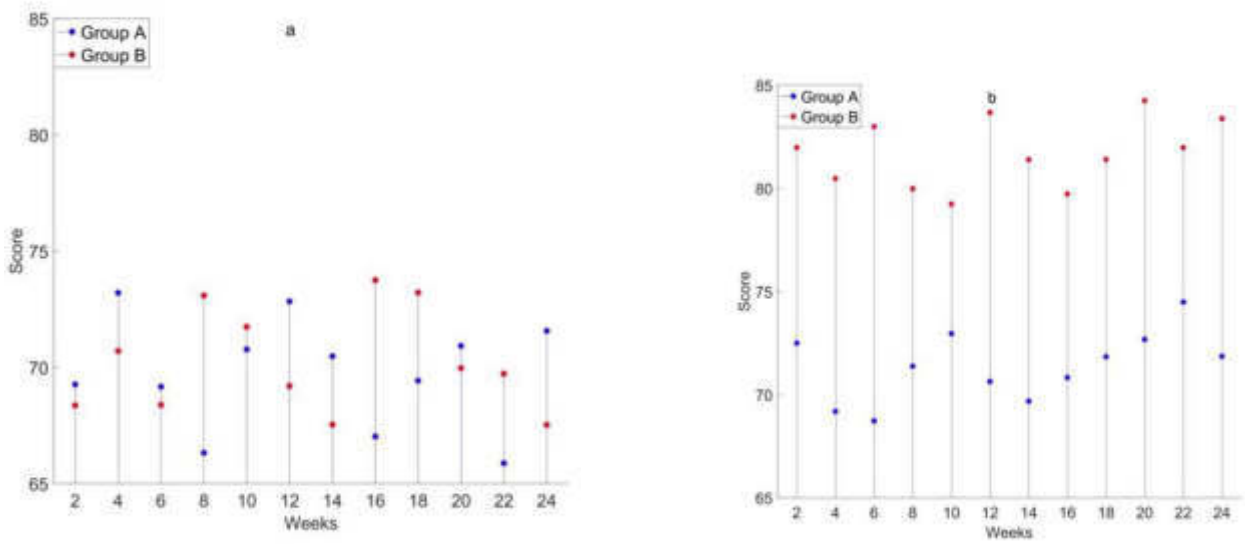
Blended teaching combines technological means with teaching methods, and follows the rules of students' learning and teachers' teaching. It does not emphasize online teaching, nor does it ignore teachers' words and deeds in offline classroom teaching. Teachers play a guiding role in teaching, and they can guide students from mastering language knowledge to acquiring language skills by designing various forms of teaching activities and tasks.

3.3. Providing personalized teaching. In this information age dominated by technology, computers can provide personalized services for different users. Nowadays, teachers are able to use smart computers to provide personalized education for different students, which is the characteristic of AI in English blended teaching.

The biggest advantage of AI technology in English teaching lies not only in its ability to simulate real English environments, but also in its ability to provide students with an efficient and personalized teaching platform. For example, AI can help students better train English vocabulary, speaking, and writing, while also helping students correct errors in pronunciation, grammar, and grammar. Expanding extracurricular knowledge that meets user learning needs and interests beyond grammar and vocabulary, AI can provide personalized teaching methods that can optimize teaching effectiveness. The comparison of personalization levels between Group A and Group B after the experiment is shown in Figure 3.6.

In Figure 3.5(a), it was observed that only 1 boy and 0 girl in Group A felt that the teaching personalization in this group was very strong, but 7 boys and 9 girls felt that the teaching personalization in this group was very poor, with a total of 16 students feeling that the teaching personalization in this group was very poor.

Figure 3.5(b) shows that there were 11 boys and 10 girls in Group B who felt that the teaching personaliza-



((a)) Comparison of average English scores between Group A and Group B students before the experiment

((b)) Comparison of average English scores between Group A and Group B students after the experiment

Fig. 3.8: Comparison of average English scores between Group A and Group B students before and after the experiment

tion of this group was very strong, while there were 0 boys and 1 girl who felt that the teaching personalization of this group was very poor. It can be seen that most students believed that the teaching personalization of Group B was very strong.

Teachers can use AI to bring vitality to English classrooms and provide students with more extracurricular English learning resources. In addition, online English teaching resources are rich and diverse. It is very important to ensure the efficient use of English resources by both teachers and students in English teaching. AI's intelligent retrieval technology can provide English teachers with more information, allowing them to choose suitable English resources based on their different levels, personalities, and learning stages, thereby formulating and adjusting personalized learning strategies.

3.4. Comparison of Grades. In traditional English classrooms, the teaching methods of teachers are all one-on-one. For students with good grades, a single class does not allow them to learn more, while students with poor grades would be unable to keep up due to the teacher's teaching speed. Over time, they would lose interest in learning. However, under the English blended teaching with AI, there have been significant changes in English learning methods. Teachers can design a set of targeted, hierarchical, and targeted courses for each student based on their learning abilities, and allow students to learn according to their own abilities before and after class. The comparison of average English scores between Group A and Group B students before and after the experiment is shown in Figure 3.8.

Figure 3.7(a) shows that the average English scores of students in Group A and Group B in the first two weeks of the experiment were 69.27 and 68.37, respectively; the average English scores of Group A and Group B students in the 24 weeks prior to the experiment were 71.57 and 67.53, respectively.

Figure 3.7(b) shows that the average English scores of students in Group A and Group B after 2 weeks of the experiment were 72.50 and 81.98, respectively; after 24 weeks of the experiment, the average English scores of students in Group A and Group B were 71.86 and 83.39, respectively.

One student said, "Blended learning allows me to learn at my own pace. I can study the course content at

home or anywhere through the online learning platform, so I can plan my study according to my own pace and time, which is more flexible and convenient."The data showed that the average English score of Group A students after the experiment slightly improved compared to the average English score of Group A students before the experiment. However, the average English score of Group B students after the experiment significantly improved compared to the average English score of Group B students before the experiment, indicating that the teaching method of Group B is more conducive to improving students' English grades.

AI based blended English teaching is essentially an innovation of the traditional English teaching model. It emphasizes the combination of "teaching" and "learning" and complements each other, thus overcoming the shortcomings of "cramming" teaching. Through collaborative previewing before class, teacher-student interaction in class, and collaborative expansion after class, innovation in English courses has been achieved. This makes online and offline classroom teaching complement each other, effectively improving teaching quality.

4. Strategies for Constructing a Blended English Teaching Path.

4.1. Hybrid Construction of Online Learning Platform and Traditional Classroom. Based on the current situation of English teaching, there is generally an online learning platform, but in most cases, teachers do not personally participate and instead allow students to log in to the online learning platform on their own computers or laptops after class. Due to the lack of teacher participation, some students do not engage in serious learning on online learning platforms, making it difficult to achieve good learning outcomes. Therefore, English teachers must pay attention to this and combine online learning platforms with traditional theoretical courses to establish a teaching model that combines online and offline. Specifically, in theoretical classroom teaching, by logging into the online learning platform through the teaching computer, learning materials are introduced into the classroom and collective learning is carried out in the classroom. In addition, theoretical teaching can also be transferred to a computer room, allowing students to log in to an online learning platform. Teachers can explain theoretical knowledge while allowing students to practice through the online learning platform, achieving an effective combination of teaching and learning. This can not only improve students' theoretical knowledge accumulation but also enhance their practical level.

While fully utilizing existing online learning platforms, teachers should also attach importance to the development of new media and combine it with English classroom teaching to create a new and diverse teaching method. Teachers can organize and summarize the content to be taught, and design it into a learning micro lesson. After that, it would be released through WeChat official account. This allows students to receive WeChat official account pushed learning lessons on their mobile phones, and use their free time before classes to learn by themselves. During class, practical activities can be set up to allow students to participate in free discussion activities for a certain period of time based on self-study before class, while teachers provide guidance and observation on the side. Teachers can also open a second classroom online and allow students to interact, learn, and exchange English with teachers on devices such as mobile phones and computers.

4.2. Optimizing Teaching Content. Before English teaching, teachers need to publish the teaching plan of the entire course and the teaching arrangements for each stage on the online teaching platform, so that students can make choices and prepare in advance. During teaching, the content of each class would be posted online for students to preview and review. When publishing lecture content, the application and extension of related knowledge can also be published together to stimulate students' interest in learning and pay attention to the knowledge points. This makes communication and interaction in classroom teaching more convenient, and can also promote the comprehensive development of students. In order to explain the key and difficult points of English knowledge, teachers can first use some pre prepared micro lessons or MOOCs in the classroom, including videos, audio, animations, text materials, etc. These can provide visual and auditory stimuli to students, enabling them to better understand and remember this information. The "teacher student" interaction in online teaching can be utilized to emphasize students' initiative. In the classroom discussion segment, attention should be paid to cultivating students' awareness of multi angle and comprehensive thinking, and teachers should provide correct guidance for students' interaction. So, in the classroom teaching process, teachers should try to encourage students to speak up and respect their opinions. Positive emotions between teachers and students can stimulate students' interest in learning.

By publishing the content of after-school exercises on online teaching platforms, students can consolidate

the learning content and deepen their understanding of knowledge through online after-school exercises and reviews after classroom teaching, thereby effectively improving the quality of teaching. The content of the post-class learning test can be a re-presentation of classroom learning content, or resources such as test questions and papers used to evaluate the effectiveness of teaching and learning, as well as some references and web site materials used to extend the learning content. In addition to the function of teaching content, the online teaching platform also provides an after-school exam module published by teachers to conduct online testing of the teaching content in the classroom or in front of it. The platform can automatically provide test results for objective questions and provide real-time feedback to teachers and students, thereby evaluating their learning effectiveness. Students can also use the teaching platform to evaluate the teaching effectiveness of teachers. Teachers can adjust the teaching plan for this lesson or the next step in a timely manner based on students' mastery and learning needs, striving to achieve the best teaching effect.

5. Conclusions. With the rapid development of AI technology, its application in learning and daily life is increasing. Applying it to blended English teaching would have a direct impact on the development of AI technology in modern information education. AI based blended English teaching can fully leverage the enormous resource advantages of the internet. This enables students to acquire more knowledge that is suitable for themselves and provides a new approach and method for English teaching. This article introduced AI based blended online and offline teaching. It was found that AI based speech recognition technology has been used for a long time in English classrooms and can improve students' English learning performance to a certain extent. In order to demonstrate that blended learning based on AI can promote the development of English teaching in the experiment, a comparative analysis was conducted between traditional English teaching and blended learning. After experiments, it was found that blended teaching based on AI can change the current situation of "deaf mute English" and provide teachers with a high degree of freedom in their teaching time, thereby improving students' learning initiative and providing personalized teaching. Moreover, it has also improved students' English grades. In summary, blended learning based on AI has its unique advantages, which can assist students in English learning. Schools should combine online and offline courses organically based on the actual situation of students, in order to maximize their advantages and improve their English learning abilities.

REFERENCES

[1] Albiladi Waheeb, S. & Alshareef, K. Blended learning in English teaching and learning: A review of the current literature. *Journal Of Language Teaching And Research*. **10**, 232-238 (2019)

[2] Naomi, M. Robert Clarke. *Lisa Fletcher,Robbie Moore,Hannah Stark*. **17**, 345-365 (2018)

[3] Qiang, H. Comparing teacher's roles of F2f learning and online learning in a blended English course. *Computer Assisted Language Learning*. **32**, 190-209 (2019)

[4] Gede, G., Rejeki, I. & Astuti, N. The effects of blended learning to students' speaking ability: A study of utilizing technology to strengthen the conventional instruction. *International Journal Of Linguistics, Literature And Culture*. **4**, 1-14 (2018)

[5] Sari Ima Frafika, A. & Apriliandari, D. Blended Learning: Improving Student's Motivation in English Teaching Learning Process. *International Journal Of Languages' Education And Teaching*. **6**, 163-170 (2018)

[6] Yulian, R. The flipped classroom: Improving critical thinking for critical reading of EFL learners in higher education. *Studies In English Language And Education*. **8** pp. 2-508 (2021)

[7] Liang, X. & Haiping, L. Liu J ,et al. *Eform Of English Interactive Teaching Mode Based On Cloud Computing Artificial Intelligence – A Practice Analysis[J].Journal Of Intelligent And Fuzzy Systems*. **40**, 1-13 (2020)

[8] Zibin, A. & Altakhaineh, A. The effect of blended learning on the development of clause combining as an aspect of the acquisition of written discourse by Jordanian learners of English as a foreign language. *Journal Of Computer Assisted Learning*. **35**, 256-267 (2019)

[9] Wang, N. Juanwen Che. *Mankin Tai,Jingyuan Zhang. "Blended Learning For Chinese University EFL Learners: Learning Environment And Learner Perceptions*. **34**, 297-323 (2021)

[10] Altay, I. & Altay, A. A Review of Studies on Blended Learning in EFL Environment. *International Journal Of Curriculum And Instruction*. **11**, 125-140 (2019)

[11] Tupas, F. & Linas-Laguda, M. Blended Learning–An Approach in Philippine Basic Education Curriculum in New Normal: A Review of. *Universal Journal Of Educational Research*. **8**, 5505-5512 (2020)

[12] Simbolon, N. EFL students' perceptions of blended learning in English language course: learning experience and engagement. *Journal On English As A Foreign Language*. **11**, 152-174 (2021)

[13] Dahmash, N. I couldn't join the session': Benefits and challenges of blended learning amid Covid-19 from EFL students. *International Journal Of English Linguistics*. **10**, 221-230 (2020)

[14] Rianto, A. Blended Learning Application in Higher Education: EFL Learners' Perceptions, Problems, and Suggestions. *Indonesian Journal Of English Language Teaching And Applied Linguistics*. **5**, 55-68 (2020)

[15] Zhang, R. Exploring blended learning experiences through the community of inquiry framework. *Language Learning & Technology*. **24**, 38-53 (2020)
[16] Gunes, S. What are the perceptions of the students about asynchronous distance learning and blended learning?. *World Journal On Educational Technology: Current Issues*. **11**, 230-237 (2019)
[17] Wichadee, S. Significant predictors for effectiveness of blended learning in a language course. *Jalt Call Journal*. **14**, 25-42 (2018)
[18] Sriwichai, C. Students' Readiness and Problems in Learning English through Blended Learning Environment. *Asian Journal Of Education And Training*. **6**, 23-34 (2020)
[19] Syakur, A., Fanani, Z. & Ahmadi, R. The Effectiveness of Reading English Learning Process Based on Blended Learning through. *Absyak" Website Media In Higher Education*. **3**, 763-772 (2020)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Jul 7, 2023
Accepted: Oct 7, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 495–509, DOI 10.12694/scpe.v25i1.2424

ASSESSING DIGITAL TEACHING COMPETENCE: AN APPROACH FOR INTERNATIONAL CHINESE TEACHERS BASED ON DEEP LEARNING ALGORITHMS

LIQING YANG[*], QICHENG WANG[†], BORUI ZHENG[‡], XUAN LI[§], XITONG MA[¶], AND TIANYU WANG[‖]

Abstract. Digital Teaching Competency (DTC) is an important skill in the professional development of international Chinese language teachers. This study developed a new deep learning-based assessment model of DTC for international Chinese language teachers. To build this model, the researchers first collected data on DTC from 221 international Chinese language teachers at different levels in 26 countries to ensure that these sample data are representative; secondly, clustering and feature dimensionality reduction techniques were used to preprocess the data and constructed the Siamese architectural model; and finally, the researchers confirmed through experimental validation and expert evaluations that the model has a high accuracy rate of 96.33%. The innovation of this model is to use the traditional three-level network as an improved constructed digital twin network, so as to extract some features that are more accurate and to characterize those features that are most predictive. The improved network is able to extract all the inputs globally and also locally that are of most interest to the user/researcher, the final prediction results are weighted, and those weighted results are used as the final prediction output of the model. This model not only provides systematic and adaptive support for improving teachers' DTC, but through the comprehensive result output, it can provide targeted improvement strategies for teachers to improve their DTC.

Key words: International Chinese teachers, professional development, digital teaching competence, assessment models, deep learning

1. Introduction. Due to advancements in information technology and the penetration of artificial intelligence research, online teaching has become one of the most important pedagogical methods for international Chinese language education [51, 19]. In the context of language education, digital teaching competence (DTC) has emerged as a crucial competency for teachers to acquire, emphasizing the need for effective methods to develop digital competence globally [37]. Assessment, as the primary method of evaluating teachers' competency, helps Chinese as foreign language (CFL) teachers improve their DTC and teaching effectiveness [61]. There has been little research on the digital competence of international Chinese teachers, and the current situation necessitates the development of more convenient and efficient methods of assessing DTC. However, existing research has highlighted the challenges that Chinese language teachers face in achieving high levels of proficiency in DTC and the lack of standard assessment methods available [54]. Teacher assessment does not exist in isolation from other aspects of the education system and must be considered alongside student development goals, curriculum objectives, and professional development [1]. Traditional evaluation methods rely on weights determined by experts to calculate competence scores, but these weights are neither flexible nor dynamic [7]. Furthermore, in teaching practice, the weights of dimensions such as student difference and teaching environment are supposed to be adjusted, which can be expensive when each adjustment necessitates a single expert decision.

It is undeniable that the education system faces challenges in the development and assessment of DTC, and more attention should be paid to these areas [12]. Artificial intelligence has the potential to achieve complex goals [46]. Intelligent algorithms are useful in teacher education because they reconfigure teaching practice and

[*]Department of International Chinese Education, Yunnan Normal University, 650000 Kunming, China;
[†]Nanjing Research Institute of Electronics Technology, 210000 Nanjing, China;
[‡]College of International Education, Shanghai University, 200444 Shanghai, China;
[§]Department of Foreign Language, Baotou Teachers' College, Inner Mongolia University of Science and Technology, 014030, Baotou, China;
[¶]Beijing Bai Zhi Xiang Technology Co., 100000 Beijing, China;
[‖]Department of International Chinese Education, Yunnan Normal University, 650000 Kunming, China (Corresponding Author: wangtianyu18@ynnu.edu.cn)

teachers' ideological construction [41]. Based on this, intelligent algorithms can be applied to teacher evaluations and personalize guidance during ongoing evaluations [22]. Deep learning algorithms are particularly effective in constructing machine learning models and improving classification or prediction accuracy. In this context, our research team has made a meaningful attempt to apply the algorithm technology to the field of international Chinese teachers [53]. By leveraging deep learning algorithms and training data, intelligent systems can learn valuable features, resulting in higher prediction accuracy of teachers' abilities [35]. Notably, these predictions only require the appropriate parameter settings, as the machine autonomously learns to assign weight to each dimension based on input [21]. However, limited studies have investigated the relationship between technology and DTC assessmentv [13]. To achieve sustainable development for international Chinese teachers, continuous innovations and algorithm design are necessary. From the perspective of technology, the question of how to assist teachers in effectively mastering digital teaching skills and accessing their digital competence in Chinese teaching has emerged as a primary concern [42].

Based on these issues, the authors of this paper conducted research on DTC for international Chinese teachers using deep learning algorithms. The first part provides a brief introduction to the evaluation of DTC for international Chinese teachers and outlines the research design. The second part presents an overview of domestic and international research on DTC and its assessment, summarising current research gaps. The third part designs a model for assessing DTC in international Chinese teachers, utilizing deep learning algorithms to address the shortcomings of traditional assessment methods. The fourth part analyses the constructed DTC model for international Chinese teachers in this study and validates its accuracy through experiments. The novelty of this paper lies in the use of algorithmic construction for a new evaluation approach. Considering the complexity of the classroom teaching environment, the author proposes a data-driven evaluation method based on deep learning.

2. Related Works.

2.1. Digital Teaching Competence. Digital teaching competence (DTC) is derived from digital competence (DC), which focuses on the acquisition of digital competence in the context of tools and technologies rather than a pedagogical model [14, 43]. DTC refers to the skills, competencies, and knowledge that teachers should master and develop in order to improve their teaching quality and efficiency in the classroom. In general, DTC can be defined as a set of knowledge, skills, or strategies used by teachers to address educational issues and challenges posed by society in the information age [39]. Krumsvik [23] defines DTC as "the teacher proficiency in using information and communications technology (ICT) in a professional context with good pedagogic-didactic judgement and awareness of its implications for learning strategies and the digital Bildung of pupils and research is primarily concerned with investigating the factors that influence DTC in order to improve it for teachers." Age, gender, teaching experience, and level of education are among these factors [4, 15, 38]. Other studies have focused on influencing factors such as teachers' perspectives on technology and ethical safety [11]. Furthermore, teachers' individual characteristics, particularly motivation and self-efficacy, have received a lot of attention [16].

However, most current research is disconnected from the study's context and focuses on a specific teacher competency while ignoring the impact of the larger environment [36]. Changes in education caused by AI have resulted in the growth of online distance learning, particularly in foreign language teaching [62]. It is therefore essential to investigate teachers' DTC. As technology advances, concepts around teaching competencies continually evolve, making it imperative to prioritize the development of technologies and solutions that fully support the education industry and thus lead to widespread improvements in education [24]. Research is increasingly exploring the relationship between technology and teachers, highlighting the need for more real-time support from AI mentors to determine when students require human assistance, assess the impact of their own help, and manage student motivation [17]. Tondeur [47] emphasizes the importance of preparing the next generation of teachers for the integration of technology into education. Therefore, the evaluation of DTC should consider a variety of technological aspects. However, how to apply new technologies, particularly AI, in education remains a grey area [25], requiring teachers to be prepared to introduce advanced technology into education. Hence, teacher education plays a critical role in preparing teachers for the future [32].

2.2. Assessment of DTC. Teachers' self-assessment provides the majority of feedback on teacher competence [10], and current DTC assessment is based on a number of frameworks. The European Framework for Digital Competence of Educators is the most commonly used assessment method, with a large number of empirical studies demonstrating its reliability [6]. Currently, the majority of DTC research in language teaching focuses on quantitative studies, employing instruments such as questionnaires for self-assessment, or expert judgments [65, 64]. However, self-assessment is frequently regarded as less accurate [16], and expert-based rubrics can be time-consuming. Although various frameworks have been used to assess teachers' DTC, their use in language teaching research has been limited [29]. According to Chinese scholars' research, there was no research framework on DTC for international Chinese teachers, and most scholars focused solely on theoretical aspects. Xu [20], for example, previously discussed the meaning, evaluation system, and cultivation of digital competence for teachers in a CFL context, which primarily focused on theoretical aspects with little attention paid to DTC in teaching practice. Lin [52]developed a DTC assessment system for international Chinese teachers at various levels of competence based on the Belt and Road initiative. Furthermore, Liu [54] proposed an assessment index for international Chinese teachers' DTC based on a survey of 205 working Chinese teachers. Furthermore, the proliferation of online teaching in Chinese language education has necessitated stricter DTC requirements [52]. The framework has made theoretical contributions to the evaluation of international Chinese teachers' DTC, but further empirical research support or promotion is needed.

In previous studies, algorithms have been utilized to assess teacher competence [44]. Fuzzy clustering algorithms have emerged as a more scientific, reasonable, and straightforward teaching evaluation method. However, the accuracy of these methods depends on the sensitivity of the initial data, and the clusters produced may not always be meaningful or accurate [53]. Bayesian algorithms, on the other hand, are faster, but have lower accuracy and are unable to effectively apply their findings to new datasets [44]. Decision trees are advantageous in that they provide interpretability and can effectively handle high-dimensional data. However, they are susceptible to overfitting and may not generalize well to new data [44]. Random forests, by contrast, can avoid overfitting, process quickly, and handle high-dimensional data, but have poor interpretability and may not always yield the most accurate predictions [34]. These studies have created a new perspective, linking DTC assessment with AI, but they also reflect that the application of AI technology in the field is not mature enough, and algorithms need to be improved in follow-up research.

At present, sentiment recognition and prediction models based on deep learning have achieved performance beyond existing algorithms, as reported by Liu [56]. Zhou [59] proposed a deep learning-based approach for analyzing interactive classrooms and assessing teaching effectiveness, which offers a faster and more accurate recognition of teacher behaviors and assessment of teaching outcomes, resulting in improved efficiency. This highlights the significant potential of deep learning in online teaching assessment and personalized recommendations. Additionally, Ning [2] conducted research on the relationship between behavior and cognition, revealing that deep learning algorithms consistently outperformed other methods, yielding an enhancement in classification accuracy ranging from 2% to 7%. The study by Hussan Munir et al. [58] demonstrated the wide applicability of deep learning algorithms in performance prediction, adaptive learning, and automation, showcasing their exceptional performance. Notably, deep learning can leverage deep neural networks to construct highly accurate prediction models using large volumes of unlabelled and unstructured data, surpassing the limitations of traditional approaches and effectively improving feature learning capabilities [57].

In summary, the majority of studies in this field have centered on assessing the DTC of EFL instructors, with little attention paid to CFL teachers. Moreover, it is worth mentioning that only a small number of articles have employed artificial intelligence algorithms to evaluate teacher DTC, and existing research has certain limitations. Many scholars have highlighted the difficulties that teachers encounter when attempting to enhance DTC. One major obstacle is the disconnect between technology and teaching methods [44]. This clearly underscores the necessity for further exploration and investigation in this particular area. As a result, it is crucial to conduct more in-depth investigations into DTC evaluations for international Chinese teachers.

3. Materials and Research Methods. The current study is comprised of two distinct parts: The first part focuses on the teaching status of international Chinese teachers and utilizes a questionnaire method adapted from DigCompEdu. The second part involves the creation of an algorithmic assessment model using deep learning algorithms, SQL database, and t-SNE technology. The model was fitted using validated data

Table 3.1: Sampling Characteristics

Age	20-25 (25.8%) 25-30 (41.6%) 31-40 (30.3%) > 40 (2.2%)
Country	China (73.8%) Other country (26.2%)
Gender	Female (84.16%) Male (15.84%)
Education Background	Doctor (19.46%) Master (65.61%) Bachelor (14.48%) other (0.45%)
Profession	MTCSOL (73.8%) Pedagogy (15.8%) Literature (5.9%) Language (3.6%) other (0.9%)
Years of Teaching	1-3 (57.01%) 3-5 (14.48%) 5-10 (17.19%) 10-20 (9.95%) > 20 (1.36%)

Fig. 3.1: European Framework for the Digital Competence of Educators (DigCompEdu)

obtained from the first part of the questionnaire survey. After multiple iterations, the fit of the model was finalized. Experts were then invited to conduct a second assessment of DTC in order to further improve the model's experimental effect. The primary objectives of this study are as follows:

1. To describe the current state of DTC for international Chinese teachers.
2. To create an algorithmic assessment model for the DTC of international Chinese teachers.
3. To design an experimental study that tests the feasibility of the model.

3.1. Participants. Participants in this study come from 26 countries, with a total of 221, including 163 (73.8%) international Chinese teachers from mainland China and 58 (26.2%) international Chinese teachers from other countries. There are 35 (15.84%) males and 186 (84.16%) females. One (0.45%) is a specialist, and the remaining 32 (14.48%) are undergraduates. 206 (93.2%) have a bachelor's degree in Chinese literature, Chinese language, or TCSL; 10 (4.5%) have a bachelor's degree in English or another foreign language; and 5 (2.2%) have a bachelor's degree.

3.2. Instruments. Since there is no DTC framework specified for international Chinese teachers, this study turned to DigCompEdu as the basis for its questionnaire design. Numerous studies have demonstrated DigCompEdu's practicality and dependability [6]. Many Chinese scholars have also used DigCompEdu to investigate the DTC of international Chinese teachers [20]. DigCompEdu contains 22 entries across three domains as shown in Figure 3.1. We designed a questionnaire with 33 questions, 1-10 for basic information shown in Table 3.1 and 12-33 for scale questions Table 4.2 using a 6-point Likert scale (1=strongly disagree, 6=strongly agree). Five TCSL experts were consulted about the quality of the questionnaire in this study. The questionnaire was revised and tested for reliability after receiving expert feedback.

3.3. Data Collection and Analysis. The International Chinese Teachers' DTC Survey utilized a snowball method to collect data. This involved identifying and recruiting potential participants who were international Chinese teachers currently teaching (in-service) and who had utilized digital teaching tools in their teaching process. An electronic questionnaire was developed with the help of experts, based on the DigCompEdu, which has been shown to have excellent reliability and validity $Cronbachs\alpha = 0.97, KMO = 0.94$. A total of 232 teachers who agreed to participate in the study were selected and sent electronic questionnaires via Questionnaire Star, a web-based user survey tool used to collect data. To comply with GDPR guidelines, the questionnaire was anonymous [8]. After careful screening by experts and consideration of factors, such as the data collection environment and participants' levels of enthusiasm, the authors selected 221 groups of high-quality data. The questionnaire-reclaiming efficiency was 95%, which ensured the effectiveness of the study. The collected questionnaires were entered into SPSS26.0 for analysis of the original data.

3.4. Methods. DL (Deep Learning) [50] is a sub-field of Artificial Intelligence (AI), which uses neural networks to learn from data and make predictions. Deep learning has unique advantages over traditional methods, such as intelligence and robustness. It can automatically screen data without the need for manual extraction, making it convenient and fast. Additionally, it can extract potential features of data, providing valuable insights. It has been applied to a wide range of fields, including natural language processing, computer vision, and speech recognition, and has demonstrated satisfactory performance in a wide range of tasks. Deep learning has recently been used in educational research, including studies of teacher performance. Therefore, we use deep learning to extract the competency distribution of teachers, establish an assessment model, and predict their DTC. The prediction process in deep learning includes four steps:

1. Data preprocessing: During this stage, the raw data is transformed into a format that can be understood by the neural network. This may involve techniques such as normalization and scaling to prepare the data for input into the model,
2. Model building: The neural network models are built based on specific problems at hand. This includes choosing the appropriate architecture and configuring the parameters of the model to optimize its performance,
3. Training: The optimization algorithm is used to train the model on the dataset to adjust the parameters of the model and improve its accuracy. During the training process, the model learns to recognize patterns in the data and make predictions based on those patterns.
4. Prediction: Once the model is trained, it can be used to make predictions about new data. The input data is fed into the model, which produces an output that represents its prediction. According to the characteristics of deep learning, there are six steps in the development of DTC assessment model in this study.

3.4.1. Step 1: Database → Name TDSet. The primary data was collected through a questionnaire survey. A highly advanced and maintainable database utilizing SQL has been constructed to ensure scalability and facilitate future research on the assessment of digital competency among teachers. This database, named TDSet, serves as a benchmark for subsequent studies due to the validity of our data collection methods and the reliability of our data sources. The collected data has been stored in the SQL database (Figure 3.2) for convenient expansion.

Figure 3.2 depicts the detailed data recorded in the SQL, where each row corresponds to the scores of specific DTC items, and the distribution of data in each row is visualized on the right. The distribution characteristics of the data are used to inform both the teacher competence classification criteria and the data processing methods.

3.4.2. Step 2: DTC level classification criteria. The study included a 22-question survey designed to assess DTC comprehensively, yielding 221 high-quality and meaningful data points. In order to more accurately assess and predict teachers' proficiency, it is proposed that, instead of using the total score as the sole criterion for determining DTC, each teacher be labeled into one of five levels based on their individual scores (see Table 3.2). The problem was converted from a regression task to a classification task using this method. Experiments revealed that, using techniques such as clustering and feature dimensionality reduction, the new grading criteria could clearly visualize the distribution characteristics of data within different categories. Secondly, the one-hot

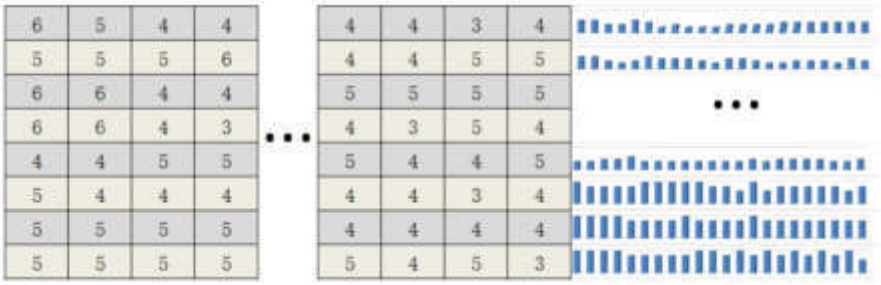


Fig. 3.2: A sample data image

Table 3.2: Digital Teaching Competence Level Classification Criteria

Score	Level	One-hot Encoding
(114,132]	A	[1,0,0,0,0]
(96,114]	B	[0,1,0,0,0]
(78,96]	C	[0,0,1,0,0]
(60,78]	D	[0,0,0,1,0]
(42,60]	E	[0,0,0,0,1]

encoding technique was used, which can handle discontinuous numerical features and expands the feature space to make data preparation for model training easier.

3.4.3. Step 3: Data analysis and downscaling. In supervised learning, the quality of the data plays a crucial role in determining the effectiveness of the model. Assessing the quality of the data is a valuable task that requires advanced techniques for data classification and analysis. To achieve a comprehensive understanding of the collected data, t-SNE (t-Distributed Stochastic Neighbor Embedding) was employed for visualizing the data. This technique is particularly useful in identifying the variations between different clusters of data and assessing the quality of the data collection process. This technique excels at detecting differences between data clusters and assessing the quality of the data collection process. The algorithm was demonstrated in the main process of t-SNE by using the Algorithm 1.

Algorithm 1: t-SNE

Input: The collected data
Output: Sample data of different categories
while *low-dimensional representation has not converged* **do** Set the number of categories;
Initialize the low-dimensional representation;
Compute the pairwise similarities between all data points in the high-dimensional space;
Minimize the divergence;
end while

3.4.4. Step 4: Data transformation and pre-processing. In order to predict and evaluate DTC using neural networks in this study, the data must be translated into a format that makes network training easier. The platform Pytorch [55] was used to create deep learning models. By converting the data to decimal values in the (0, 1) range, normalization aims to make the data processing easier. The pseudocode is shown in Algorithm 2.

3.4.5. Step 5: Data Modeling. This study demonstrates that the extraction of local features is a more effective and meaningful approach. To achieve this, a Siamese architecture [26] was employed for model

Algorithm 2: Normalize the Data

Input: The collected data
Output: The normalized data
for *each data point in the feature vector* **do** Get the maximum and minimum values of the vector;
Normalize the data point: $x_i = \frac{x_i - \text{Min}}{\text{Max} - \text{Min}}$;
end for

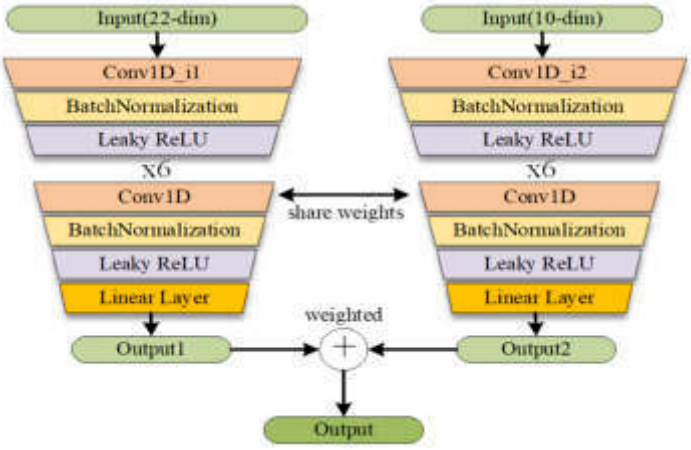


Fig. 3.3: The main structure of the Digital Competence Forecasting Network

construction, which is capable of enhancing both local and global feature extraction. In this approach, the input was divided into two parts - one comprising all the data and the other including more significant indicators and test items, which were identified through expert guidance or interviews. To construct the DTC prediction network (as shown in Figure 3.3), a two-input model was used. The first input consisted of the scores of 22 test items, and the second input of manually filtered scores for 10 essential questions related to Assessment, Empowering Learners, Teaching and Learning. This method helped to extract rich feature data from both local and global sources. For feature extraction, input was first passed through a 1D convolutional layer, followed by a batch layer, and a non-linear activation function. 6-layers of convolutional layers with shared weights were then applied, and a classifier was used to classify the extracted features into their respective types. Finally, the two inputs were weighted for output, with weights defined as 0.8 and 0.2 respectively to maintain integrity for the global features.

The formula used for classification is as follows. The model applies the following equation.

$$out(N_i, C_{out_j}) = bias(C_{out_j}) + \sum_{k=0}^{C_{in}-1} weigth(C_{out_j}, k) a^* \ input(N_i, k) \tag{3.1}$$

Inside C_{in} and C_{out} are the dimensions of input channel and output channel in a Conv1D layer. Bias is utilized to adjust neuron sensitivity and primarily serves to characterize the various neighborhoods within input features. It aids in training the response to input signals and contributes significantly to subsequent classification operations. Batch Normalization is an optimization technique for neural networks, ensuring that input feature variation remains within an appropriate range. This allows input values to smoothly traverse the sensitive part of the excitation function. The activation function is of utmost importance in neural networks as it enables effective utilization of multiple layers. Non-linear activation functions are essential because, without them, multiple linear layers would behave similarly to a single layer, thereby limiting the network's ability to learn from data and handle complex tasks. Additionally, the activation function helps regulate network output by suppressing

irrelevant values and amplifying crucial ones. In this study, the LeakyReLU activation function [5] is adopted as the network's activation function. It assists in eliminating unnecessary values, promoting network convergence, and preserving feature integrity. To achieve this, a negative slope parameter of 0.05 is employed. The following represents the corresponding form.

$$LeakyReLU(x) = \begin{cases} x, & if x \geq 0 \\ negative\_slope.x, & \text{otherwise} \end{cases} \tag{3.2}$$

3.4.6. Step 6: Training Details. To accelerate network training, a $GTX1660TI6GB$ with CUDA core was used to train the deep network model. Adam optimizer was additionally used to make network optimization more efficient. The estimation steps for the Adam optimizer are as follows.

$$g_t = \bigtriangledown \widehat{L}(\Theta t) \tag{3.3}$$

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1+\beta) g_t \\ v_t = \beta_2 v_{t-1} + (1+\beta_2) g_t^2 \end{cases} \tag{3.4}$$

$$\begin{cases} \widehat{m_t} = \frac{m_t}{1-B_1^t} \\ \widehat{v_t} = \frac{v_t}{1-B_2^t} \end{cases} \tag{3.5}$$

$$\theta_{t+1} = \theta_t + \frac{\eta}{\sqrt{\widehat{v}} + \epsilon} \widehat{m^t} \tag{3.6}$$

The first term gradient g_t is the loss function biased against v_t. The second term is the first-order moment estimate and the second-order moment estimate of the gradient at moment t in the momentum situation. The third term is the first-order moment estimate and second-order moment estimate after using bias correction. The last item is the weight update step. To better quantify the distance metric between the predicted values and the labels, a cross-entropy loss function was used in the training process [31], as follows.

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) log(q(x_i)) \tag{3.7}$$

The cross-entropy loss function measures the variability between two probability distributions. In the above equation, $P(x)$ represents the true distribution of the sample and $q(x)$ represents the distribution predicted by the model.

4. Results.

4.1. The Status Quo of DTC of international Chinese teachers.

4.1.1. Descriptive analysis. A descriptive analysis method was used to gain a comprehensive understanding of the current status of DTC for international Chinese teachers. First, descriptive statistics on six dimensions of digital competence (table 4.1) were conducted, followed by descriptive analysis on 22 specific items. The specific findings of the analysis are shown in Tables 4.1 and 4.2.

The highest and lowest scores, mean scores, standard deviation (SD), and median of the five dimensions of digital competence, as well as 22 specific competence scores, are shown in tables 4.1 & 4.2. The scores for all six dimensions ranged from 4.4 to 4.9, indicating that teachers scored in the middle of the scale and that international Chinese teachers had a high level of overall DTC. The standard deviation for each dimension, however, is quite high, around 0.9, indicating that there are significant differences in scores among teachers.

With a mean of 4.817 and a standard deviation of 0.854, the Professional Engagement dimension received the highest score. Teaching and Learning and Digital Resources also scored highly, with means of 4.632 and 4.668, respectively, indicating that international Chinese teachers are adept at acquiring teaching resources to

Table 4.1: International Chinese Teachers' Digital Competence in Each Dimension

Variables	Max	Min	Mean	SD	Md
Facilitating Learners' Digital Competence	6	1	4.502	0.895	4.6
Empowering Learners	6	1.333	4.407	0.946	4.333
Assessment	6	1	4.478	0.935	4.667
Teaching and Learning	6	2.25	4.632	0.851	4.75
Digital Resources	6	2	4.668	0.904	4.667
Professional Engagement	6	2.25	4.817	0.854	5

Table 4.2: Questionnaire Information

Professional Engagement	
Organisational communication	5.03
Professional collaboration	4.8
Reflective practice	4.63
Digital CPD	4.8
Digital Resources	
Selecting	4.71
Creating and modifying	4.61
Managing and protecting and sharing	4.69
Teaching and Learning	
Teaching	4.73
Guidance	4.7
Collaborative learning	4.59
Self-regulated learning	4.51
Assessment	
Assessment strategies	4.4
Analysing evidence	4.46
Feedback and planning	4.58
Empowering Learners	
Accessibility and inclusion	4.32
Differentiation and personalisation	4.48
Actively engaging learners	4.42
Facilitating Learners' Digital Competence	
Information and media literacy	4.53
Communication	4.63
Content creation	4.22
Responsible use	4.6
Problem solving	4.52

support teaching students' digital competence in accordance with corresponding teaching needs. Empowering Learners, Assessment, and Facilitating Learners' Digital Competence all scored lower, at 4.407, 4.478, and 4.502, respectively, indicating that international Chinese teachers generally focus on self-development in their teaching practice while ignoring differentiated and personalized teaching, as well as awareness of digital competence assessment.

Among the 22 items, organizational communication received the highest score (5.03), while content creation received the lowest (4.22), followed by accessibility and inclusion (4.32), indicating that teachers' instruction is teacher-centered rather than taking into account learners' individual factors. Overall, teachers are well-equipped in terms of Professional Engagement and Digital Resources, but Assessment, Teaching and Learning may need to be improved.

Table 4.3: Correlation analysis of international Chinese teachers' digital competence in each dimension

	Professional Engagement	Teaching and Learning	Digital Resources	Assessment	Empowering Learners	Facilitating Learners' Digital Competence
Professional Engagement	1	.656**	.701**	.713**	.658**	.640**
Teaching & Learning	.656**	1	.770**	.809**	.740**	.753**
Digital Resources	.701**	.770**	1	.717**	.656**	.673**
Assessment	.713**	.809**	.717**	1	.835**	.804**
Empowering Learners	.658**	.740**	.656**	.835**	1	.871**
Facilitating Learners' Digital Competence	.640**	.753**	.673**	.804**	.871**	1

** At level 0.01 (2-tailed), the correlation was significant.

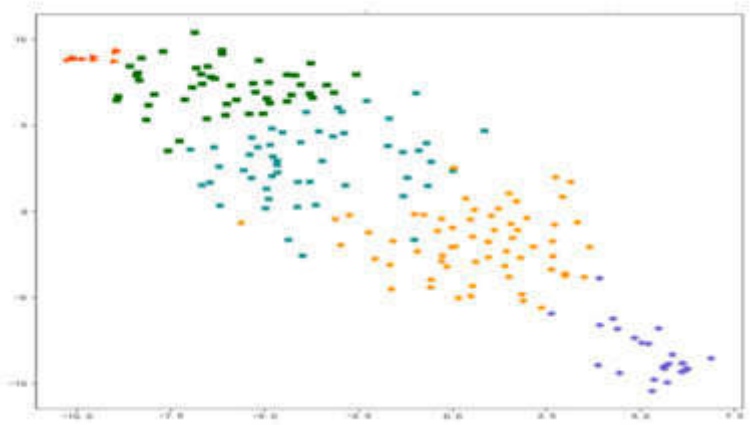


Fig. 4.1: Data dimensionality reduction feature map

4.1.2. Correlation analysis. This study used Pearson correlation analysis to validate the relationship between the six dimensions of international Chinese teachers' DTC, with higher numbers indicating stronger correlations. According to the data analysis, all six dimensions were positively correlated. The six dimensions had a significant impact on international Chinese teachers' DTC. Based on this, the author investigated the relationship between the assessment dimensions and the other dimensions further. As shown in Table 4.3, the correlation coefficients were: Empowering Learners (0.835) > Teaching and Learning (0.809) > Facilitating Learners' Digital Competence (0.804) > Digital Resources (0.717) > Professional Engagement (0.713), and the results indicated that the assessment of DTC was most correlated with empowering learners followed by teaching and learning. This showed that digital competence assessment and students' perspective had a certain impact on teaching effectiveness.

4.2. Experiment on the DTC Assessment Model.

4.2.1. Data quality analysis. Figure 4.1 illustrates the outcomes of feature distribution for several data categories following feature reduction scaling and t-SNE operation. Internally, it can be seen that different colors stood in for the five pre-established categories, distances between points denoted variations in features, and different regions showed the distribution of several categories. They had little overlap and were each rather small in their own divisions. It demonstrated the accuracy and value of the data collection process used in the study.

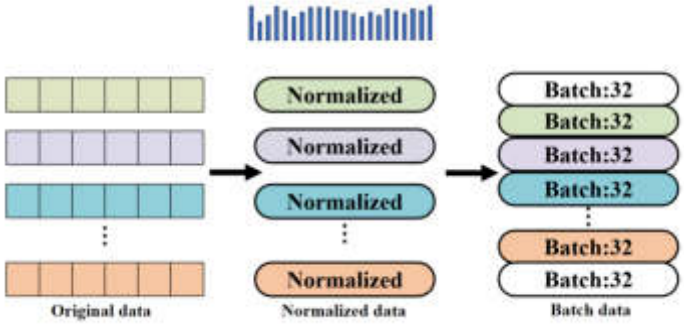


Fig. 4.2: DTC data conversion

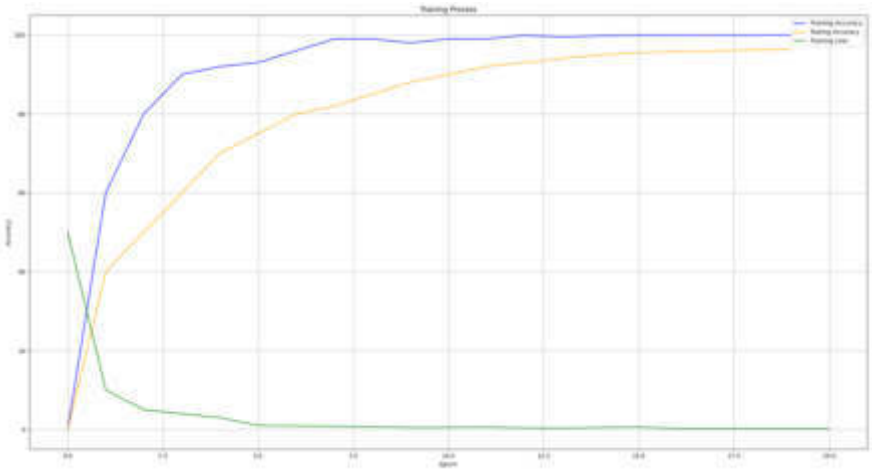


Fig. 4.3: New TDSet model predictive value

4.2.2. Data processing analysis. The experimental findings demonstrated that data conversion sped up the deep learning model's gradient descent procedure, allowing it to reach the best solution more quickly and increasing the classifier's prediction accuracy. As seen in Figure 4.2, the outcomes of the data transformation procedure could change the DTC data into neural network training data that could be applied to batch iterative neural network optimization.

4.2.3. Analysis of model prediction results. In this work, a 32-person batch size was used to train the model over 10,000 iterations. Also, in order to aid in the network's process optimization, we employed an approach to lower the learning rate. It was established that deep learning could be used as an algorithm for teachers' digital competence evaluation model because the predictive value of our research results reached 96.33% (Figure 4.3), which was significantly higher than that of the Bias Classifier, Decision Tree, Fuzzy Clustering, and Random Forest algorithms (Table 4.4). The high score also demonstrated the objectivity and accuracy of the deep learning algorithm for evaluating the DTC of Chinese teachers, and that deep learning could successfully gauge teachers' overall proficiency through a large number of experiments.

Expert evaluation was also used in the study to guarantee the model's accuracy. To evaluate the DTC of Chinese language teachers, five foreign experts in the field of Chinese language teaching were invited. The outcomes revealed agreement between the model's predictions and the expert assessment, proving the model's

Table 4.4: Model experiment results

Methods	Accuracy (%)
Bayes Classifier	93.67
Decision Tree	91.74
Fuzzy Clustering	90.86
Random Forest	91.22
New TDSet	96.33

Table 4.5: Chinese teachers' digital competence assessed by experts

	Professional Engagement	Teaching and Learning	Digital Resources	Assessment	Empowering Learners	Facilitating Learners
Expert1	4.80	4.70	5.30	4.00	4.00	4.20
Expert2	4.70	5.00	5.50	4.50	4.50	4.90
Expert3	4.90	4.80	5.30	4.30	4.60	4.80
Expert4	4.50	5.00	4.00	3.00	3.00	4.00
Expert5	5.00	4.80	5.00	4.50	4.60	4.80

viability from a different angle.

5. Discussion. The research findings indicate that international Chinese teachers demonstrate a relatively high level of DTC, with an average score surpassing 4.4. Notably, the dimensions of Assessment (4.478) and Empowering Learners (4.407) received lower scores, underscoring their significance and necessity of future development in these areas. Furthermore, the teachers' DTC exhibits variations and continuous development that aligns with their years of experience (standard deviation greater than 0.8). These results support the notion that foreign language teachers generally possess lower digital competence [30]. These observations could be attributed to evolving standards of proficiency and insufficient training and self-assessment [54, 52]. As technology advances, teaching methods have become more diverse and complex, transitioning from traditional face-to-face instruction to blended learning. This rapid transformation presents a significant challenge for teachers, as they are required to enhance their DTC within a limited time frame [19]. Previous studies mainly focused on specific factors such as age and teaching experience when assessing teacher's DTC, neglecting the complexity of teaching practices and the dynamic nature of teacher abilities.

The findings also indicate that insufficient attention has been paid to DTC, reflecting issues with international Chinese teachers' training methods, thinking, and development, as well as a demand for training sessions to address DTC deficits and improve the overall quality of teacher training [54]. The teachers' DTC assessment can also assist them in enhancing their teaching skills and identifying gaps in their practice and training, ultimately promoting their professional development. Teachers are expected to engage in practical reflection, focusing on the needs of their students, differentiating their instruction appropriately, and using digital technology responsibly to improve learner-centered teaching effectiveness. Furthermore, through focus group discussions and interviews with ten participants from the initial questionnaire survey, the effectiveness of the proposed model in assessing digital teaching competence and providing strategies for improvement was acknowledged. It facilitated self-enhancement and professional development. The study thoroughly examined the current status of teachers' DTC and proposed pathways for improvement.

This study presents compelling evidence regarding the efficacy of deep learning models and their potential to enhance teacher research, reinforcing the promising role of artificial intelligence in education [53, 35]. The proposed model has the capability to gather a broader range of classroom data, including fixed achievement tests, self-reports, and multimodal data from interactive classroom sessions, encompassing teacher behaviors and expressions. Currently, the assessment methods for assessing teachers' digital teaching competence often rely heavily on self-assessment or framework-based assessment, which are susceptible to subjective factors and yield limited and inconsistent outcomes. Furthermore, in the teaching and learning process, there is now greater

emphasis on general pedagogical skills rather than digital technologies [13]. This study adopts a pedagogical standpoint to create an algorithmic DTC assessment model that takes self-reporting and other important measures into account, ensuring convenience and validity [12, 47, 9]. The future implementation of the model will have the potential to be applied and promoted to teachers from other backgrounds.

The construction of this deep learning-based model represents a novel and intelligent approach to personalized and digitized evaluation. It is characterized by automatic feature extraction, data-driven analysis, adaptive learning, and high accuracy. By automatically extracting meaningful features from data without the need for complex feature engineering, deep learning algorithms simplify the assessment process of teachers' digital competence, thereby reducing research complexity. This approach facilitates a comprehensive understanding and capture of teachers' behaviors and skills in real-world teaching scenarios. Moreover, deep learning algorithms possess the ability to adaptively adjust their parameters to accommodate dynamic data changes, ensuring the effectiveness and accuracy of the evaluation method when confronted with new teaching contexts and technologies. The approach promotes accurate self-reflection among teachers, enhances teaching effectiveness, and fosters their professional development capabilities. Aligned with the imperative future trend of empowering teachers with technology, this study contributes by providing a novel and effective assessment model. In summary, this model surpasses prior studies in terms of data extension and maintenance, noteworthy visualization features, and high prediction rates, thereby broadening the scope of previous research and aligning more closely with the definition of competencies.

6. Conclusion. With the advent of the digital era, the education community has increasingly focused on the assessment of Digital Teaching Competence (DTC). However, existing assessment methods primarily rely on self-reports and traditional frameworks, lacking relevance, accuracy, and in-depth research on international Chinese teachers. This study delves into the current state of DTC among international Chinese teachers and proposes a new assessment method based on an intelligent algorithmic model, thereby improving the relevance and accuracy of the assessment. The findings of this study will help place the assessment of working teachers' DTC in an interrelated and integrated perspective from the standpoint of professional development for Chinese international teachers [3]. The results indicate that the deep learning algorithm performs effectively on the given data set, leading to an improvement in classification accuracy by 2%-5%. The model evaluation, established through the deep learning algorithm, allows for the identification of teacher ability characteristics, aiding in proper self-evaluation and preparation for promotion. Furthermore, the classification criteria established provide clear visualization of the distribution characteristics of teacher DTC in different categories. The use of t-SNE technology during data processing significantly reduces calculation cost.

However, this is just the beginning of a long journey. Future studies should continue to innovate the algorithm and incorporate the latest technology into the evaluation model. The model will thus be improved along the following two aspects. On the one hand, the researchers will continue to use new types of data screening to realize data visualization and data analysis methods, in order to improve the quality and effectiveness of data collection. Second, in future algorithm research, the researchers will make the network multi-functional while adjusting for current computing speed. For example, while predicting teacher's DTC, it should also be possible to generate new and specific improvement strategies. In the future, the researchers will also attempt to deploy the network on mobile devices to enhance network reusability. Beyond this, empirical research can be carried out to promote the implementation and application of the model.

7. Limitations. It must be acknowledged that there are some limitations in the considerations of model construction. Deep learning algorithms rely on a large amount of data, and in this study, we only collected data from teachers in 26 countries, which may not be representative of all international Chinese teachers. Therefore, ensuring the quality and quantity of data remains an issue that needs to be addressed in future research. In addition, although deep learning can automatically recognize data characteristics, setting hyperparameters is still dependent on both technology and experience. Therefore, it is essential to set reasonable parameters to ensure the reliability and validity of the assessment model.

Funding. This work is supported by the Scientific Research Fund project of the Education Department of Yunnan Province "Research on the Digital Competence Assessment of International Chinese Teachers' Online Teaching Based on Intelligent Algorithms". Education Department of Yunnan Province (No.2023Y0536).

REFERENCES

[1] National Education Association Teacher Assessment and Evaluation: The National Education Association's Framework for Transforming Education Systems to Support Effective Teaching and Improve Student Learning. (Available online,2010), https://eric.ed.gov/?id=ED583104, Accessed: March 5, 2022

[2] Ning, X. Behavior Recognition of College Students Based on Improved Deep Learning Algorithm. *International Journal Of Web-Based Learning And Teaching Technologies*. **10**, 1-16 (2023)

[3] Bingham, E. et al. (Deep Universal Probabilistic Programming,2018)

[4] Cabero-Almenara, J. et al., Digital competence of higher education professor according to DigCompEdu. Statistical research methods with ANOVA between fields of knowledge in different age ranges. *Educ Inf Technol (Dordr)*. **26**, 4691-4708 (2021)

[5] Chicco, D. Siamese Neural Networks: An Overview. (2020)

[6] Caena, F. & Redecker, C. Aligning teacher competence frameworks to 21st century challenges: The case for the European Digital Competence Framework for Educators (Digcompedu). *European Journal Of Education*. **54**, 356-369 (2019)

[7] Dong, L., Ping, L. & Ping, Y. Research on the Digital Competence Model Construction of International Chinese Teachers. *Journal Of Research On Education For Ethnic Minorities*. **33**, 153-160 (2022)

[8] Dong, Z., Fei, M. & Yan, Y. European Framework for Digital Competence of Educators: A New Guide to Technological Innovation for Teacher Development. *E-Education Research*. **42**, 121-128 (2021)

[9] Engen, B. & Others Digital Natives: Digitally Competent?. (2014)

[10] Fernández-Batanero, J. & Others Digital competences for teacher professional development. Systematic review.. *European Journal Of Teacher Education*. **45**, 513-531 (2020)

[11] García-Vandewalle García, J. & Others Analysis of digital competence of educators (DigCompEdu) in teacher trainees: the context of Melilla, Spain. *Technology, Knowledge And Learning*. (2021)

[12] Garzón Artacho, E. & Others Teacher Training in Lifelong Learning—The Importance of Digital Competence in the Encouragement of Teaching Innovation. *Sustaincompetence*. **12** (2020)

[13] Garzon-Artacho, E. & Others Teachers' perceptions of digital competence at the lifelong learning stage. *Heliyon*. **7** (2021)

[14] Gimeno, A. & Almenara, J. Las tecnologías de la Información y Comunicación y la formación inicial de los docentes. Modelos y competencias digitales. Profesorado. *Revista De Curriculum Y Formación Del Profesorado*. **23** pp. 247-268 (2019)

[15] Guillén-Gámez, F. & And, M. and F. *J. Álvarez-García, A Study On The Actual Use Of Digital Competence In The Practicum Of Education Degree*. **25**, 667-684 (2018)

[16] Hatlevik, O. Examining the Relationship between Teachers' Self-Efficacy, their Digital Competence, Strategies to Evaluate Information, and use of ICT at School. *Scandinavian Journal Of Educational Research*. **61**, 555-567 (2016)

[17] Holstein, K., Aleven, V. & N. Rummel, A. . *Conceptual Framework For Human–AI Hybrid Adaptivity In Education*. pp. 240-254 (2020)

[18] Hsu, L. Examining EFL teachers' technological pedagogical content knowledge and the adoption of mobile-assisted language learning: a partial least square approach. *Computer Assisted Language Learning*. **29**, 1287-1297 (2017)

[19] Hui, W. International Chinese teaching under the influence of the COVID-19:Problems and strategies. *Language Teaching And Linguistic Studies*. **4**, 11-22 (2021)

[20] Juan, X. and S. *J. Hua, The Connotation, Evaluation System, And Cultivation Of Information Literacy Of International Chinese Teacher*. **1**, 26-31 (2006)

[21] Jun, W. Teaching Evaluation of Based on Fuzzy Clustering Analysis. *Journal Of Gansu Normal Colleges*. **27**, 34-36 (2022)

[22] Koedinger, K. & Corbett, A. . (Technology Bringing Learning Science to the Classroom,2006)

[23] Krumsvik, R. Digital competence in the Norwegian teacher education and school. *H"ogre Utbildning*. **1** pp. 39-51 (2011)

[24] Laura, R. & McPherson, A. The technologisation of education: Philosophical reflections on being too plugged in. *International Journal Of Childrens Spirituality*. **14** pp. 289-298 (2009)

[25] Gray, L. & S., A. Towards a democratic future. *London Review Of Education*. **18** (2020)

[26] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature*. **521**, 436-44 (2015)

[27] Lei, W. Research on development trend of digital Chinese language teaching for foreigners. *International Journal Of Chinese Studies*. **1**, 109-113 (2010)

[28] Lin, H. A meta-synthesis of empirical research on the effectiveness of computer-mediated communication (CMC) in SLA. *Language Learning And Technology*. **19** pp. 85-117 (2015)

[29] Lucas, M. et al., The relation between in-service teachers' digital competence and personal and contextual factors: What matters most?. *Computers & Education*. **160** (2021)

[30] Maderick, J. et al., Preservice Teachers and Self-Assessing Digital Competence. *Journal Of Educational Computing Research*. **54** (2015)

[31] Masetti, G. & Giandomenico, F. Analyzing Forward Robustness of Feedforward Deep Neural Networks with LeakyReLU Activation Function Through Symbolic Propagation. in ECML PKDD 2020 Workshops. 2020. (Springer International Publishing,0)

[32] Mayer, D. & Oancea, A. Teacher education research, policy and practice: finding future research directions. *Oxford Review Of Education*. **47** pp. 1-7 (2021)

[33] Mei, W. et al., Conception Continuum, System Elements and Integrated Model: From Digital Literacy to Digital Competence. *Journal Of Distance Education*. **31**, 24-29 (2013)

[34] Miao, W. & Yue, Z. The Transition of Teachers'Roles in the Context of MOOC Education Based on a Narrative Research. *Curriculum,Teaching Material And Method*. **3**, 26-31 (2017)

[35] Minghua, Y., Xiang, F. & Zhiting, Z. The Educational Applications and Innovative Explorations of Machine Learning in the

View of Artificial Intelligence. *JOURNAL OF DISTANCE EDUCATION*. **3**, 11-21 (2017)
[36] Pettersson, F. On the issues of digital competence in educational contexts – a review of literature. *Education And Information Technologies*. **23**, 1005-1021 (2017)
[37] Santos, P., A., C. & Cortés-Pe na, O. . (An Evaluative Perspective of the Digital Teaching Competence in the Context of a Pandemic,2021)
[38] Sánchez, P. S., et al., Análisis correlacional de los factores incidentes en el nivel de competencia digital del profesorado. *Revista Electrónica Interuniversitaria De Formación Del Profesorado*. **23** pp. 1 (2020)
[39] Prendes, P. . (Indicadores y propuestas para la definición de buenas prácticas,2009)
[40] Ramírez-Montoya, M., Mena, J. & Rodríguez-Arroyo, J. In-service teachers' self-perceptions of digital competence and OER use as determined by a xMOOC training course. *Computers In Human Behavior*. **77** pp. 356-364 (2017)
[41] Redecker, C. . (DigCompEdu,2017)
[42] Sánchez-Cruzado, C., Campión, R. & Sánchez-Compa na, M. Teacher Digital Literacy: The Indisputable Challenge after COVID-19. *Sustaincompetence*. **13** pp. 4 (2021)
[43] Semerci, A. & Aydın, M. Examining High School Teachers' Attitudes towards ICT Use in Education. *International Journal Of Progressive Education*. **14** pp. 93-105 (2018)
[44] Siyan, C. et al., Research on the improvement of teachers' teaching competence based on machine learning and digital twin technology. *Journal Of Intelligent & Fuzzy Systems*. **40**, 7323-7334 (2021)
[45] Taghizadeh, M. & Yourdshahi, Z. Integrating technology into young learners' classes: language teachers' perceptions. *Computer Assisted Language Learning*. **33**, 982-1006 (2019)
[46] Tegmark, M. Life 3.0: Being human in the age of artificial intelligence. (Knopf,2017)
[47] Tondeur, J. et al., Developing a validated instrument to measure preservice teachers' ICT competencies: Meeting the demands of the 21st century. *Br*. **48** pp. 462-472 (2017)
[48] Tondeur, J. et al., Teacher educators as gatekeepers: Preparing the next generation of teachers for technology integration in education. *British Journal Of Educational Technology*. **50** (2019)
[49] Tseng, J. et al., A critical review of research on technological pedagogical and content knowledge (TPACK) in language teaching. *Computer Assisted Language Learning*. **35** (2020)
[50] Wolford, B. . *The EU's New Data Protection Law?*. **2018** (2018), https://gdpr.eu/what-is-gdpr/
[51] Xi, Z. International Chinese language teacher preparation in the post-epidemic era. *Language Teaching And Linguistic Studies*. **5** pp. 1 (2020)
[52] Yan, L. & Yu, Z. Cultivation of Information Literacy of International Chinese Teachers under the" Belt and Road " Initiative. *Information Science*. **38**, 108-115 (2020)
[53] Yang, L. & Others Study on Evaluation Model of International Chinese Teachers' Digital Competence in Online Teaching Based on K-Means Clustering Algorithm. in Proceedings of the 4th IEEE Eurasia Conference on IoT. *Communication And Engineering*. **2022** (2022)
[54] Yu-ping, L., Xiao-dong, L. & Jia-xin, H. Research on Status and Influencing Factors of International Chinese Teachers'Digital competence. *Journal Of esearch On Education For Ethnic Minorities*. **32**, 139-146 (2021)
[55] Zhang, Z. & Sabuncu, M. . (Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels,2018)
[56] Fan, Y., Lu, X., Li, D. & Liu, Y. . *Video-Based Emotion Recognition Using Cnn-Rnn And C*. **3** (2016)
[57] Fenghua, X. and Yu le. "Pedagogic coresearcher deep learning: Origin. *Connotation And Prospect*. **58** pp. 147-56 (2022)
[58] Munir, H., Vogel, B. & Jacobsson, A. Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision. *Information (Switzerland)*. **17** pp. 203 (2022), https://doi.org/10.3390/info13040203
[59] Nan, Z. & Zhou, J. The Students' Learning Behavior Analysis and Teaching Effect Evaluation Based on Deep Learning. *Modern Educational Technology*. pp. 102-111 (2021)
[60] Redecker, C. . (Digcompedu,2017)
[61] Jianling, L. Framework of evaluation literacy training for Chinese teachers [J]. *International Chinese Education (English And Chinese)*. **6** pp. 2 (2021)
[62] Huang, W., Hew, K. & Fryer, L. Chatbots for language learning – are they really useful?. *A Systematic Review Of Chatbot-supported Language Learning*. **38**, 237-257 (2022)
[63] Gisbert-Cervera, M., Usart, M. & Lázaro-Cantabrana, J. . (2022)
[64] Gisbert-Cervera, M., Usart, M. & Lázaro-Cantabrana, J. Training pre-service teachers to enhance digital education. *European Journal Of Teacher Education*. **45**, 532-547 (2022)
[65] Wong, K. & Moorhouse, B. Digital competence and online language teaching: Hong Kong language teacher practices in primary and secondary classrooms. *System*. (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Jul 18, 2023
Accepted: Sep 5, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 511–518, DOI 10.12694/scpe.v25i1.2431

DESIGN OF ENGLISH INFORMATIONIZATION TEACHING SYSTEM BASED ON POSITIVE PSYCHOLOGY

XIAO CHANG*

Abstract. Early computer-aided teaching systems fully played the leading role of teachers, but ignored the embodiment of students' subjectivity. In English teaching practice, there is still a current situation of emphasizing cognition and neglecting psychology, seriously neglecting students' psychological needs such as learning interest, self-confidence, and happiness. Therefore, in order to explore the promoting role of positive psychology in English teaching, this study, guided by subjective well-being theory and self-determination theory, adopted JSP+Struts+SQL Server technology to design an English information-based teaching system. The system was based on English knowledge, with students' personalized reasoning as the premise, and teachers' teaching strategies as the key. Personalized assistance was provided to English teachers for information-based teaching. Finally, through a controlled experiment, it was found that after half a semester of English information technology teaching, the excellent final grade rate of the intervention group X students was 56.67%; the pass rate reached 90%, and only 3 students failed. There was a significant improvement compared to the performance of the control group Y students. This experiment can prove the effectiveness of the English information teaching system designed in this article. The English information teaching system is a supplement to the traditional classroom teaching mode, which can enhance students' learning interest and English oral expression ability, as well as improve the quality of teachers' education.

Key words: English Teaching; Information Based Teaching System; Positive Psychology; Fuzzy Comprehensive Evaluation

1. Introduction.

In the era of Internet plus, information globalization has become a trend. English, which is commonly used all over the world, is playing a leading role. In terms of English curriculum standards, teachers are also required to pay attention to, stimulate, and cultivate students' emotions, stimulate their interest in English, and cultivate their confidence in English. Positive psychology has a great promoting effect on improving the learning enthusiasm and innovative thinking of college students [1, 2]. As an auxiliary tool in English teaching, computers also play an important role in cultivating students' autonomous learning ability. Therefore, in English teaching, teachers need to improve their thinking, guided by the theory of positive psychology, and cultivate students' enthusiasm and creativity. Improvements should also be made in teaching methods, and computer-aided language teaching is used to organically integrate teachers' teaching and learning processes in a complete, continuous, interactive, and personalized training format [3, 4]. In this way, teachers can be prompted to carry out teaching reforms, resulting in a fundamental change in English teaching, transforming students from passive learning mode to active mode, and forming a new combination of teachers, students, textbooks, and teaching modes [5].

2. Literature Review.

Application of Positive Psychology in Language Teaching. There are many studies on the application of positive psychology in language teaching. Eva Gajdosova believed that the social and emotional health of students and teachers is one of the most important standards of school quality. He also investigated the social and emotional health of students and teachers in an inclusive primary school in Slovakia, and introduced the main organizational principles of an inclusive primary school in Bratislava, Slovakia. The research results indicated that teachers and students reported a high level of social and emotional health related to the school's core organizational principles guided by a positive educational framework [6]. Li examined the complex relationship between emotional intelligence, foreign language enjoyment, and English as a foreign language learning performance of 1307 Chinese high school students based on the theories and assumptions of positive psychology. The research results indicated that there was a small to moderate correlation between students' emotional

*Huaiyin Institute of Technology, Huaian 223003, China (Xiao_Chang2023@outlook.com)

intelligence, foreign language enjoyment, self-perceived English scores, and actual English scores; academic performance indirectly mediated the partial impact of learning ability on perceived and actual achievements [7].

Computer-Aided English Teaching. Computer-Aided Instruction (CAI) is a pioneering topic, and its development and popularization have already reached a large scale. There are also many studies on computer-aided teaching. Kaye developed a model to explore which factors must be in place to ensure that CAI helps improve learning outcomes. This model outlined key trends that promote or hinder the deployment of CAI tools in low - and middle-income countries. Finally, he discovered that key factors to consider when designing CAI interventions include the operating environment, stakeholder participation, infrastructure, trust in technology, CAI tool design, content management creation, student participation, classroom integration, teacher's ability, student's ability, and data collection and use [8]. Rosali explored the impact of implementing computer-aided teaching on the academic performance of high school physics students in his paper.

A quasi experimental pre test posttest control group design was adopted, involving 157 10th grade students from Philippine private schools. Finally, the study found that both CAI and traditional teaching methods can significantly improve students' physics grades. However, when comparing the effectiveness of the two methods, there was no significant difference in their impact on academic performance. Therefore, CAI can serve as an alternative teaching method [9].

Jing D introduced the characteristics of the currently very practical Windows application programming tool Visual Basic 6.0 in his paper, and combined some program examples to illustrate the practicality of Visual Basic in the field of foreign language teaching. Moreover, he found that using VB to develop computer-aided teaching courseware can improve classroom efficiency and courseware production efficiency. After being applied in the teaching process, students' learning enthusiasm has significantly improved, providing a guarantee for the smooth implementation of professional English course design. The courseware produced is flexible and has a beautiful interface [10].

Although the research of the aforementioned scholars plays an important role in improving classroom efficiency and student academic performance, CAI lacks guidance for user learning, and cannot fully leverage the teacher's dominant position and guiding role in teaching. It also cannot automatically adjust learning strategies based on learners' existing knowledge system. The current CAI systems in the world also generally have the following shortcomings: (1) they do not support networks; (2) there is no intelligence; (3) supervisors are unable to effectively supervise, resulting in a weak student-centered role and low self-discipline in learning during the teaching process. Based on this, this article took positive psychology as the theoretical guidance, adopted JSP+SQL Server technology, fuzzy comprehensive evaluation technology, and designed a personalized English information assisted teaching system. This system can provide hierarchical teaching to students based on their cognitive level, and has more advantages compared to traditional CAI teaching systems in English information assisted teaching.

3. Overall Architecture of Information Technology Teaching System.

3.1. Introduction to Relevant Technologies.

JSP technology. JSP (Java Server Pages) is a technology based on the Java language and closely integrated with HTML. JSP programs can run on different platforms [11, 12]. Currently, Web/Server/Application Server systems that support Servle/JSP can be seen on most platforms. Software developers can develop, deploy, and expand in any environment. JSP has a rich and powerful development tool. It has the characteristics of separating content generation and display, emphasizing reusable components, simplifying page development with identification, and being widely applicable to various platforms [13, 14].

Struts technology. Struts technology is an MVC (Model View Controller) framework based on SunJ2EE, implemented through technologies such as Servlet and JSP, and has been widely used [15, 16]. Struts integrates Servlets, JSPs, and other things into one framework, allowing developers to implement a complete MVC pattern without the need to write code, thus saving a lot of time.

SQL Server 2008 database. SQL Server 2008 databases can directly extract data from structured, semi structured, and unstructured files. It can also perform operations such as querying, searching, synchronizing, reporting, and data analysis [17, 18]. Data can be stored on different devices, from large computers to desktop computers to mobile devices, and can be controlled by them no matter where they are stored.

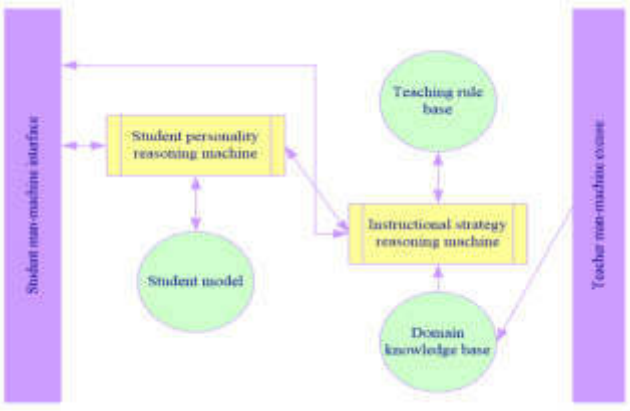


Fig. 3.1: System model of information-based teaching

Overall System Design Plan.

1. *English knowledge base.* The English knowledge base is an information system that stores English learning resources and knowledge point attributes in a knowledge expression manner. The organization and expression of knowledge are key factors that affect system efficiency.
2. *Student model.* The student model is a prerequisite for an information-based teaching system, which includes two parts: a student personality inference machine and a student model library. The function of the student personality inference machine is to evaluate students' English learning outcomes, diagnose problems they encounter in English learning, and evaluate their cognitive abilities during their English learning process. In addition to students' basic information, the student model library also stores information related to their English learning status, English learning ability, etc., which is the basis for formulating teaching strategies.
3. *Teacher model.* The teacher model is the key to information-based teaching, which includes a teaching strategy inference engine and a teaching rule library. The role of the teaching strategy inference engine is to use relevant inference algorithms to select suitable textbooks and teaching methods for students based on the personalized information provided by the student model. When inferring teaching strategies, it is necessary to select them based on relevant laws and store the laws in the teaching rules database to make them computer recognizable.

4. System Detail Design.

4.1. Teacher Model Design. In the theory of positive psychology, it is necessary to help students discover their potential and stimulate their interest in learning by leveraging their outstanding qualities and positive power. English teaching design needs to fully leverage students' initiative. Therefore, when designing a teacher model, it is necessary to divide and prune the knowledge tree based on student types to obtain the knowledge points that students should learn. The knowledge point tree is used to represent the hierarchical relationship between knowledge points, as shown in Figure 4.1.

The fuzzy support relationship between each knowledge point is utilized to determine the optimal order between each knowledge point. It is assumed that knowledge point TD_p is the direct precursor of TD_q. $W(TD\_p, TD\_q)$ is used to represent the direct support level of vertices $TD_p$and TD_q, and let $W(TD_p, TD_q) = \lambda(X_pq)$; By using WumW to represent the sum of the weights of the learning paths provided by a teaching sequence, it can be obtained:

$$WumW = \sum_{p=1}^{n} W(TD_p, TD_q) \tag{4.1}$$

Among them, the larger the value of WumW, the higher the correlation between the knowledge points on the

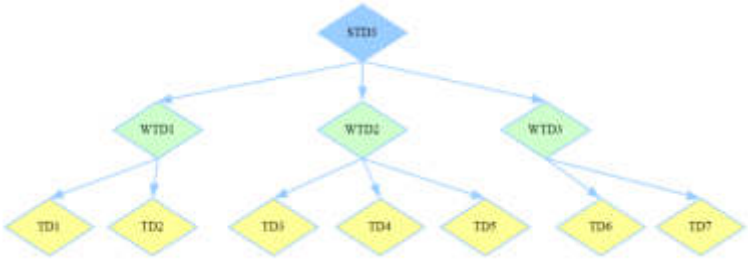


Fig. 4.1: The knowledge tree is used to represent the hierarchical relationship between knowledge points

corresponding learning path, and the better the learning effect of this teaching sequence. Therefore, the best teaching order is the largest topological order in WumW.

4.2. Student Model Design. Due to the existence of many highly subjective and fuzzy concepts in the teaching system, which cannot be accurately described using traditional mathematical methods, this article introduces fuzzy comprehensive evaluation to quantitatively and mathematically handle fuzzy phenomena and concepts when evaluating students' English learning outcomes, diagnosing problems they encounter in English learning, and evaluating their cognitive abilities.

1. Designing a module for evaluating students' English learning level In evaluation set A, the evaluation indicators are determined: English cognitive ability is A1, and learning interest is A2. It can be obtained: evaluation set A=A1, A2. Among them, cognitive abilities include: a11 memory ability, a12 understanding ability, a13 application ability, a14 analysis ability, and a15 comprehensive ability [19, 20]. Therefore, the corresponding elements can be represented as:

$$A1 = a11, a12, a13, a14, a15 \tag{4.2}$$

In comment set B, B=Excellent, Good, Medium, Pass, Relatively Poor, Poor, and the corresponding elements can be represented as:

$$B = b1, b2, b3, b4, b5, b6 \tag{4.3}$$

The indicator set A2 for evaluating students' learning interest is defined as A2=interest in learning English, and the corresponding elements can be represented as A2=a21. Comment set B is defined as B=very interested, interested, average interested, not interested, corresponding to element B=b1, b2, b3, b4.

2. Constructing membership functions The first level factor set is A=A1, A2; The second level factor set is: A1=a11,a12,a13,a14,a15; A2=a21, so the secondary evaluation matrix is:

$$A1 = \begin{bmatrix} a11 \\ a12 \\ a13 \\ a14 \end{bmatrix}, A2 = \begin{bmatrix} a21 \end{bmatrix} \tag{4.4}$$

3. Weight determination: For each subset of factors that have already been constructed, a comprehensive decision evaluation should be conducted. B is the decision set, and factor set A can be recorded as $A = \{A1, A2\}$. The weight coefficient of A is denoted as $M = \{M1, M2\}$, and the weight coefficient of A1 is denoted as $M1 = m11, m12, m13, m14, m15$ [21, 22].
Therefore, the first level evaluation can determine the student's English cognitive ability R1, and the weight of cognitive ability factors is determined as: $Ai \circ R1$. After normalization, $F1 = \{f1, f2, f3, f4, f5, f6\}$. Learning ability can be achieved by using fuzzy matrix synthesis operations to obtain a second level evaluation vector of $F = M \circ R = \{f1, f2, f3, f4\}$.

Table 5.1: English Scores of Students in Group X and Group Y at the End of the Semester Before the Experiment

Group	Number	Final English average	Pass rate (%)	Rate of excellence (%)
Group X	30 (30)	88.03 (88.03)	76.6 (76.6%)	36.6 (36.6%)
Group Y	30 (30)	88.2 (88.2)	76.6 (76.6%)	40.0 (40.0%)

* Note: A score of 0-72 on the final English test paper is considered a failure; a score of 72-96 is considered good; and a score of 96-120 is considered excellent.

5. Implementation and Testing Experiment of English Informationization Teaching System.

System Architecture. According to the requirements of the information based teaching system, this English teaching system provides users with the following functions for operation: administrator function - login, teacher management, department management, and other functions; student functions - login registration, query course related information, searching for true question explanations, answering questions, online quizzes and data statistics, submitting assignments, etc; teacher functions - student and class management, course related information module management, test question type management, real question explanation module management, and other functions.

This article uses a Browser/Server structure, which is mainly divided into three layers: Web server, Apache server, and database management.

System Development Environment. The operating system adopts Windows XP Professional; The database system adopts SQL Server 2008; The webpage creation tool uses JSP.

5.1. Experiments.

Experimental Content and Objects. In order to test the effectiveness of the English teaching information system, this article conducted an experiment in a middle school in Nanchang City. The experiment was divided into an intervention group and a control group, with the intervention group consisting of Class 1, Grade 7 (hereinafter referred to as Group X) and the control group consisting of Class 2, Grade 7 (hereinafter referred to as Group Y). Before the experiment began, there was no significant difference in academic performance between Group X and Group Y students, and the teaching conditions were consistent. In the experiment, Group X adopted a combination of English information-based teaching and traditional lectures. Firstly, Group X students were preliminarily evaluated and graded based on their cognitive ability and interest values. Then, corresponding learning content was selected for them based on their level, and after they have completed a knowledge point, they were given a test. The test results were combined with their interest values to obtain their comprehensive score on this knowledge point, which was then used to adjust the students' learning ability and level. Only traditional teaching methods were used for Group Y. Teachers use information technology to create language contexts for students, and students follow the teacher's progress in learning. Before the experiment, the English scores of students from Group X and Group Y at the end of the semester are shown in Table 5.1.

Experimental Results. After half a semester of experimentation, two groups of students were given written English exams to explore the effectiveness of English information technology teaching in students' learning of English subjects. The learning effectiveness was demonstrated by the passing rate and excellent rate of students' written English test scores.

Figure 5.1 shows the descriptive statistical results of the final written English test scores for Group X and Group Y, respectively. From the graph, it can be seen that the performance of Group X was better than that of Group Y. In terms of passing rate, Group X achieved 90%, with only 3 students failing. In terms of excellent rate, Group X achieved 56.67%, and more than half of the students achieved excellent English written test results. From it, it can be seen that using English information technology teaching can enhance students' enthusiasm for English learning, help them correct their learning habits, and improve their academic performance.

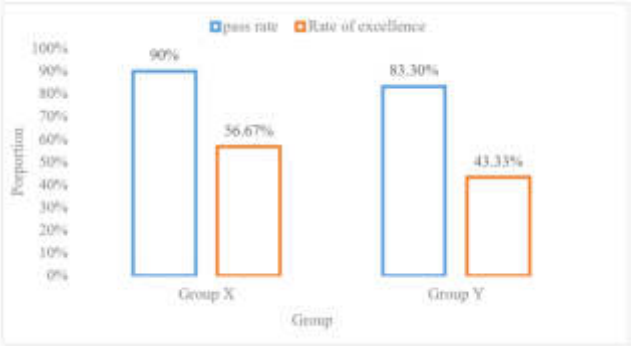


Fig. 5.1: Comparison of the passing rate and excellent rate of the written English test between Group X and Group Y after the experiment

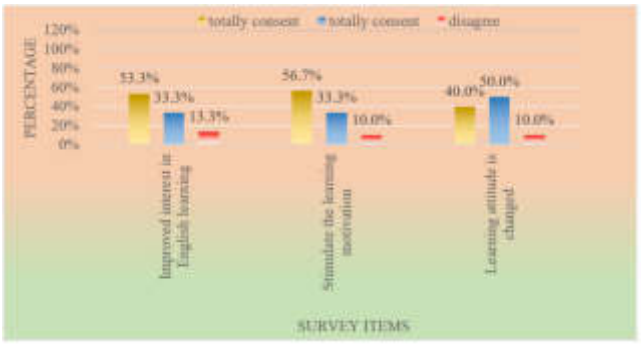


Fig. 5.2: The attitude of X group students towards English learning before the experiment

Experimental Discussion. Before and after the experiment, discussions and exchanges were held with students in Group X, and they were asked to fill out a survey questionnaire to explore their attitudes towards English learning.

From Figures 5.2 and Figures 5.3, it can be seen that the attitudes of Group X students towards English learning before and after the implementation of English information technology teaching showed significant changes in the three dimensions of students' interest, motivation, and attitude towards English learning. Before the experiment, 53.3%, 56.7%, and 40% of students held positive attitudes towards learning interest, motivation, and attitude, respectively. Only about half of the students had a positive attitude. After conducting English information technology teaching, 96.7%, 90%, and 83.3% of students held positive attitudes towards learning interest, motivation, and attitude, respectively. It can be seen that after English information technology teaching, most students have improved their confidence in learning English, maintained a positive attitude, and showed a strong interest in English learning.

5.2. Comparison between English Informationization Teaching System Based on Positive Psychology and Traditional Informationization Teaching. In order to further understand the innovation of the English information-based teaching system constructed in this article, this section compares the English information-based teaching system constructed in this article with traditional information-based teaching systems from five aspects: functional integrity, richness of teaching resources, personalized learning, evaluation of teaching effectiveness, and satisfaction of teachers and students. The evaluators are X groups of teachers and students, with an evaluation level of 1-5 points. The higher the evaluation level, the better the system evaluation effect. The comparison results are shown in Figure 5.4.

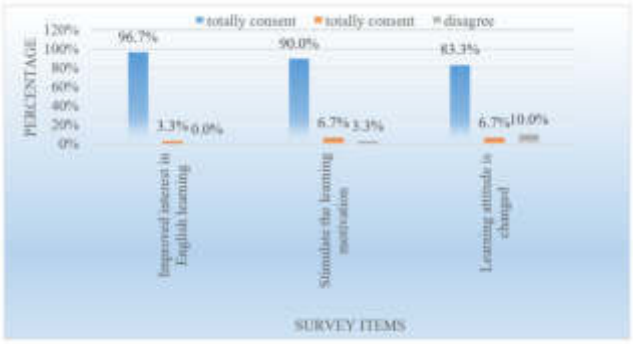


Fig. 5.3: X group's attitude towards English learning after the experiment

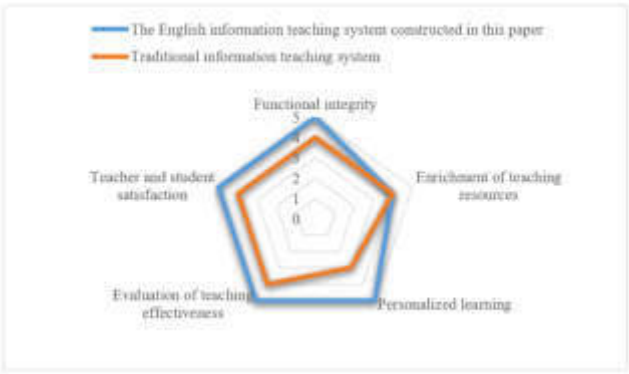


Fig. 5.4: Comparison results

From Figure Figure 5.4, it can be seen that the English teaching system constructed in this article performs well in five aspects: functional integrity, rich teaching resources, personalized learning, teaching effectiveness evaluation, and teacher and student satisfaction.

6. Conclusions. How can traditional English teaching be combined with computer-aided English teaching to achieve better teaching by teachers? What is the purpose of students learning better English? Through experimental research and analysis and discussion of the experimental results, a conclusion of "yes" was reached. Under the guidance of positive psychology theory, this article used JSP+SQL Server technology and fuzzy comprehensive evaluation technology to design an information-based teaching system that assists English teaching. Recently, a middle school student in Nanchang City was selected as the experimental object, and students from two classes were divided into a control group and an intervention group for a controlled experiment. The results of the controlled experiment showed that the intervention group had a significant improvement in the excellent rate and pass rate, reaching 56.67% and 90% respectively. This indicates that combining English teaching with information technology can fully mobilize and enhance students' learning enthusiasm and initiative, and can also improve the quality and efficiency of English teaching activities. However, the English information-based teaching system established in this article still has some shortcomings: (1) So far, the system does not have autonomous learning function; (2) The system has not yet implemented automatic knowledge extraction; (3) The system has weak error detection and correction functions.

Acknowledgment. 1) This paper is a part of achievement of the project 2022SJYB 1932 entitled "A Study of College English Vocabulary Teaching from the Perspective of ECL" that guides Universities' Philosophy and Social Science Researches in Jiangsu Province. 2) This paper is part of achievement of Higher Education

Research of Huaiyin Institute of Technology, No.: 2022GJ07 entitled “Research on the Promotion Path of School City Integration from the Perspective of Symbiosis”. 3) This paper is sponsored by Shanghai Foreign Language Education Press for 2023 China Tao Xingzhi Research Association’s Reading and Teacher Development Research No.2023JS0028 “Research on the Connotative Development of Foreign Language Teachers in Universities under the Blended Learning Environment in the Information Age”. 4) This paper is a part of achievement of the project ”Research on Teaching Chinese Sentence Structure as a Foreign Language from the Perspective of Dynamic Syntax ” No. 220900816285121 of the 2022 Ministry of Education’s Industry and University Collaborative Education Program.

REFERENCES

[1] Noori, S. & Narafshan, M. Enhancing Self-Esteem in Classroom Language Learning: The Potential of Implementing a Strength-Based Positive Psychology Intervention at Higher Education[J]. *International Journal Of Language Teaching And Education*. **2**, 334-345 (2018)
[2] Review, B. Positive Psychology Perspectives on Foreign Language Learning and Teaching. *Edited Collection By Danuta Gabry-Barker And Dagmara Galajda[J]*.. **1**, 2019 (0)
[3] And, L. and rules of college English education based on cognitive process simulation[J]. *Cognitive Systems Research*. **57** pp. 11-19 (2019)
[4] Leidy, J. Using Computer Assisted Instruction in an ESL Language Program[J]. *IALLT Journal Of Language Learning Technologies*. **15**, 13-24 (2019)
[5] Miller, N., Wyatt, J., Casey, L. & Others Using computer-assisted instruction to increase the eye gaze of children with autism[J]. *Behavioral Interventions*. **33**, 3-12 (2018)
[6] Gajdosova, E. Veronika Bisaki,Silvia Majercakova Albertova. *N Application Of Positive Psychology In An Inclusive Primary School In Slovakia[J].Psychology Research*. **10**, 2020 (0)
[7] Li, C. positive psychology perspective on Chinese EFL students’ trait emotional intelligence, foreign language enjoyment and EFL learning achievement[J]. *Journal Of Multilingual And Multicultural Development*. **41**, 246-263 (2020)
[8] Kaye, T. & Tools:, A. model to guide use in low-and middle-income countries[J]. *The International Journal Of Education And Development Using Information And Communication Technology*. **17**, 82-99 (2021)
[9] Rosali, L. Effect of Computer-Assisted Instruction (CAI) on the Academic Achievement in Secondary Physics[J]. *Open Access Library Journal 0*. **7**, 1-11 (2020)
[10] Jing, D. & Jiang, X. Optimization of Computer-Aided English Teaching System Realized by VB Software[J]. *Computer-Aided Design And Applications*. **19** pp. 139-150 (2021)
[11] Rafamantanantsoa, F., Analysis, R. & Modeling, S. of the Performance of Dynamic Web Server Using JSP and PHP[J]. *Communications And Networks*. **10**, 196-210 (2018)
[12] Wu, Z. & Zheng, X. Dynamic Web Page Development Technology based on JSP Technology [J]. *Information And Computer*. **2018** pp. 13-15 (0)
[13] And, Z. and implementation of book management system based on JSP technology[J]. *Heilongjiang Science*. **9**, 11-13 (2018)
[14] Huang, J., Chen, S., Song, H. & Others A VR resource Site Design based on JSP[J]. *Digital Technology & Application*. **36** pp. 07 (2018)
[15] Ahmad, S., Rana, T. & Maqbool, A. Model-Driven Framework for the Development of MVC-Based (Web) Application[J]. *Arabian Journal For Science And Engineering*. **47**, 1733-1747 (2022)
[16] Domenico, D. & Ricciardi, G. Shear strength of RC beams with stirrups using an improved Eurocode 2 truss model with two variable-inclination compression struts[J]. *Engineering Structures*. **9359**, 1-10935 (2019)
[17] Malik, A., Burney, A. & Ahmed, F. Comparative Study of Unstructured Data with SQL and NO-SQL Database Management Systems[J]. *Journal Of Computer And Communications 0*. **8**, 59-71 (2020)
[18] Gao, H., Jiang, G., Gao, X. & Others An equine disease diagnosis expert system based on improved reasoning of evidence credibility[J]. *Agricultural Information Processing00*. **6** pp. 003 (2019)
[19] Zhang, W., Lai, T. & Li, Y. Risk Assessment of Water Supply Network Operation Based on ANP-Fuzzy Comprehensive Evaluation Method. *Journal Of Pipeline Systems Engineering And Practice*. **13**, 1-40210 (2022)
[20] Wang, W., Jing, Z. & Others Assessing effect of grassland resources policies using AHP and fuzzy comprehensive evaluation:A case study of Ningxia Hui Autonomous Region,China[J]. *Ecological Economy V*.. **16** pp. 03 (2020)
[21] Chen, Z., Shi, M. & Zou, J. Application of improved fuzzy comprehensive evaluation method in eutrophication assessment for tributary bays in the Three Gorges Reservoir, China[J]. *Water Environment Research*. **96**, 808-816 (2020)
[22] Li, L., Lin, H., Wan, J. & Others MF-TCPV: A Machine Learning and Fuzzy Comprehensive Evaluation-Based Framework for Traffic Congestion Prediction and Visualization[J]. *IEEE Access PP(*. **99** pp. 1-1 (2020)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Jul 21, 2023
Accepted: Oct 7, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 519–534, DOI 10.12694/scpe.v25i1.2316

CROP FIELD BOUNDARY DETECTION AND CLASSIFICATION USING MACHINE LEARNING

D.BHAVANA* AND MYLAPALLI JAYARAJU†

Abstract. Crop classification and detection of crop field boundaries empower farmers which helps the agricultural businesses to estimate crop field dimensions and yields accurately. Our research focuses on estimating crucial agricultural inputs such as seeds, pesticides, insecticides, and fertilizers to enhance overall production. The conventional method of manually identifying field boundaries is both time-consuming with labour-intensive. In contrast, our study harnesses data from diverse satellites such as Sentinel, Landsat, and MODIS, encompassing valuable land usage information. By integrating this data with machine learning algorithms, we achieve real-time monitoring of crop fields through effective classification and boundary identification. For the classification of crop fields within our study area, we recommend employing the Classification and Regression Tree (CART) algorithm. Additionally, we leverage normalized difference indices, such as the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI), as features for classification. We compare these features with Support Vector Machine (SVM) and Random Forest (RF) algorithms. Subsequently, we utilize the Canny edge detection technique to identify boundaries within the classified crop areas. Notably, our approach utilizes the Google Earth Engine (GEE) as a primary platform for extracting features, conducting data training, and visualizing information. The proposed algorithm yields impressive results with a high level of accuracy. Notably, the CART algorithm achieves a remarkable accuracy rate of 96.1%. Furthermore, we incorporate NDWI-based Canny edge detection into our methodology. The outcomes convincingly underscore the practicality and applicability of our research in real-world scenarios.

Key words: Crop field boundary detection,Normalized difference indices, Canny edge detection

1. Introduction. To satisfy the expected rise in food demand, agricultural production must be raised while reducing its negative environmental effects. There are many powerful tools and technologies for optimizing agriculture. Machine learning is one such technology. Utilizing satellite data and applying machine learning algorithms provides farmers with a healthy environment for farming, thus after detection of the boundaries right amount of water, fertilizers and pesticides can be supplied to the farmers and this helps them reach optimal yield while using less amount of renewable resources.

Land-use details are very important in modern farming, hence field boundary detection has been opted in order to give accurate and up-to-date results [1]. Boundary detection helps the suppliers, government, and policymakers know about the areas under crops and their yield. Usually, existing administrative maps are considered and field boundaries are detected manually based on the surveyed data. But it involves a huge amount of manual labor as many number of maps are to be updated [2]. Due to this, the yield predictions produced will not be accurate. Thus, Automated Detection is way better and more helpful it easily identifies the boundaries of different crop fields across the country with minimal human involvement. This will be beneficial in the countries like India, where digital records are not available excessively. The exact detection of boundaries helps in obtaining more precise information about the crop yield. This is where Earth Observational satellite data comes into play.

Lately, Satellite imagery is abundantly available which is cost-effective and frequently updated. Boundary detection is typically seen as a mid-level method for determining the borders of (and between) objects in scenes, with tight linkages to both grouping/segmentation and object form.However, the satellite imagery have their limitations. They are usually available at low image resolution [3]. The properties of the images also change depending upon the land area covered. Hence, there is little research interest on field boundary detection

*Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur District, AP 522502, India (Corresponding author: bhavanaece@kluniversity.in)
†PG student, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur District, AP

when compared to other applications of satellite imagery. Therefore, the robust solutions for automatic field boundary detection are rare.

Many recent studies on field boundary detection, make use of several machine learning algorithms. There are several other studies like, Turker and Kok used Gestalt laws along with perceptual grouping in order to detect the boundaries in known fields. Rahman et al used an approach which makes use of statistical data on crop rotation patterns. Tiwari et al. used fuzzy logic rules along with color and texture information of the images and finally after identifying the boundaries they refined them using snakes. Yan, L.; Roy, D.P used web enabled Landsat data (WELD) along with watershed algorithm in identifying the automated crop field boundaries [4]. Recently, Watkins and van Niekerk compared various edge detection kernels along with watershed, multi-threshold, and multi-resolution segmentations to identify their potential in field boundary detection. Their results have shown that Canny edge detection and watershed have produced best results when compared to other algorithms [4].

In this study we propose a method which is NDWI based canny edge detection which helps in automatic detection of crop field boundaries. Normally, traditional edge detection techniques make use of spatial information of an image but this approach makes use of both spatial and spectral information [5-7]. In this study the identification of crop field boundaries is performed in unknown fields with minimal amount of prior information. A normal human cannot know the crop areas in a particular region; they must perform the study manually in order to identify the crop areas in a location. Hence, in order to reduce this effort the classification is proposed in this research. Crop classification helps in finding out the areas where the crops are present in a particular location. Our approach use Sentinel-2 and Landsat-8 datasets from Google earth engine where a classification algorithm CART is applied and proposed in order to classify and identify the crop area and furthermore NDWI based Canny Edge algorithm is applied to the detected agricultural area in order to obtain the boundaries for crop fields. This data is used as training data and further after classification, all the algorithms were analysed and compared in order to propose a better algorithm through this research.

2. Literature Survey. In the Corresponding approach 3 different locations around Southern India are selected as the region of interests. The Vijayawada region situated in Krishna basin in Andhra Pradesh state is assumed as ROI-1, Mydukur area which is situated in Rayalaseema region of Andhra Pradesh state was assumed as ROI-2 and the Alappuzha region of the Kerala state was considered as the ROI-3 [5]. The 3 regions were selected in order to visualize and identify the performance of the algorithm for any region when it is applied. The goal of this research is to identify crop lands by using classification and detect the crop field boundaries at the identified crop lands.

3. Data Acquisition. Satellite imagery and Ground truth data are explained in in detail within this section.

Satellite Data. Satellites are referred as the 'eye of the sky'. Satellite imagery also known as Satellite Data is a collection of group of images which consists information about the Earth which is gathered by man-made satellites. Satellite data is generated by using remote sensing technologies. Satellite Data has authentic information about Earth surface, weather and others. This data helps us understand long term changes and act accordingly. Benefits of using satellite data are: We can monitor large areas at a time. We can demonstrate movements on large areas especially sea. Satellites can deliver the data irrespective of the conditions on Earth atmosphere like light and weather conditions. The satellite images show the Earth surface conditions extremely well.

Sentinel – 2. Sentinel-2 is mission for observing Earth by Copernicus programme which is a European union's earth observation programme [6]. The Sentinel 2 is maintained and operated by the European space agency. Sentinel captures optical imagery at high spatial resolution around about 10m to 60m over the land and coastal regions.

This Sentinel is a constellation with 2 satellites Sentinel-2A which was launched on 23 June 2015 and Sentinel-2B which was launched on 7 March 2017; and also a third satellite Sentinel-2C is under testing and it is preparing to launch in 2024. This mission specializes in broad range applications like agriculture monitoring, land cover classification and water use and it's quality. These satellites were manufactured by consortium led by Airbus defence and space [7].

Sentinel-2 has multispectral data with 13 bands with a swath width of 290 Km. It provides visible to NIR spectral bands and SWIR spectral bands. Sentinel 2 revisits the same place with same viewing angles after every 10 days. It captures the images at spatial resolution 10m/pixel, 20m/pixel and 60m/pixel.

Landsat - 8. Landsat 8 is developed collaboration between NASA and U.S. Geological Survey (USGS). It was called as the Landsat Data Continuity Mission. Landsat 8 was launched on February 11 in the year 2013 from Vandenberg Air Force Base. The design, Construction, Launch and on-orbit revolution was led and taken care by NASA. Landsat 8 captures images at high resolution which is from 15m to 100m per pixel. Landsat 8 consists of multispectral data with 11 bands. Landsat 8 consists of two sensors – the Operational Land Imager (OLI) and Thermal Infrared Sensor (TRIS). These two sensors provide the images at spatial resolution via bands covering Visible, NIR, SWIR, Thermal and Panchromatic [8-10].

OLI collects the data for visible region, NIR, SWIR and Panchromatic bands. TIRS collects the data for two more bands in Thermal Region. Landsat 8 is regularly providing with around 725 Scenes per day to USGS data archive. Landsat 8 takes approx. 99 minutes per revolution and it completes 14.5 orbits per day. It repeats the coverage which means revisits the same geographical location for every 16 days. The swath width of Landsat 8 is 185 Km. In this Study, the Landsat 8 collection 1 Tier 1 TOA reflectance raster collection from GEE was considered assuming that it is best suited data for the application [11-13].

The Ground Truth Data. The Ground data truth was obtained from Google Earth.This is a user-friendly resource which is helpful for beginner and intermediary learners who are interested in learning more about GIS and wants to perform some analysis and operations in GIS [8]. The data obtained was mapped to the three region of interests as shown in the given below Fig. 4.2.

This research selected the region of interests and mapped the data in sentinel dataset. The image collection were selected in such a way that they possess least cloud cover during the corresponding dates i.e., 01-01-2020 to 01-02-2021. Furthermore, the classification and boundary detection was applied using the ground truth data obtained through Google Earth [9].

4. Methodology. As discussed earlier in this research we have identified agricultural fields in the study site using CART algorithm and then applied NDWI based canny edge detection in order to identify the boundaries of the agricultural fields. The workflow and methodology is explained in the following steps in detail.

The procedure has been done in five steps: 1) feature definition, 2) feature extraction, 3) dataset, 4) classification 5) boundary detection. For any machine learning approach the data cleaning and preprocessing are considered key steps which shown in detail in the below sections.

4.1. Feature definition. To remotely detect and sense any kind of land cover, the basic mechanism to be carried out is to acquire the electromagnetic wave reflectance information from sensors onboard the satellite. This information is then processed and analyzed for variations that could help detecting the targeted land cover type. The reflectance of spectra from every area differs with regard to the vegetation, water that area possess [10]. The information collected through the spectra of the visible, near-infrared, and mid-infrared, as well as the ultraviolet, is what remote sensing of vegetation is principally dependent on. However, as was previously said, the satellite data that we are employing in this study only offers high resolution in the blue, green, red, and near-infrared bands. Therefore, this work could only use certain spectral areas. Further, indices like NDVI and NDWI were computed with the formulae which include coefficients calibrated specifically for satellites like MODIS, Landsat-8 and Sentinel-2 As a result, we employed appropriate indicators in the form of normalised differences relevant to the detection of vegetation, as well as standardised equations that had been carefully considered and calculated [11]. NDVI and NDWI are used for feature extraction, where these two metrics are able to extract vegetation and water content in the region of interest shown in Fig. 4.1.

NDWI. While McFeeters NDWI index is frequently used to define water bodies, it can also be used to track changes in the water content of plant leaves. Given the variety in moisture levels among the crops used in this study, this is particularly helpful. As indicated in equation 2, it may be calculated [12].

$$NDWI = \frac{G - N}{G + N} \tag{4.1}$$

where N is the near-infrared band's surface reflectance and G is the green band's surface reflectance. The NDWI exhibits good results for the boundary identification. As it measures the water content of the plants it

Fig. 4.1: Flow chart of Methodology

was chosen so it can reduce the number of infield edges and helps in finding the edges perfectly.

The fundamental contrast between the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI) is rooted in the spectral bands employed during calculation and the intended focus of analysis. NDVI is derived from the near-infrared (NIR) and red light bands, primarily indicating the presence and vitality of vegetation. In contrast, NDWI utilizes the green and NIR bands to detect and evaluate the moisture content within various features, such as water bodies and moisture-laden vegetation. While NDVI centers on vegetation, NDWI is tailored for investigations related to water content.

The Normalized Difference Vegetation Index (NDVI) assumes a pivotal role in crop classification due to its capacity to quantify and assess vegetation density and health. By examining NDVI values derived from remote sensing data, distinctions among diverse crop types can be established, enabling the monitoring of growth stages, health conditions, and spatial distributions. This wealth of information contributes to agricultural management, precision farming methodologies, and informed decision-making encompassing aspects like crop yield estimation, irrigation scheduling, and the detection of pests and diseases.

4.2. Feature Extraction. In all classification and border detection algorithms, feature extraction is regarded as the important step. Ground Truth data mapping and identification of the features is considered

as the preliminary steps for the Feature extraction. As stated earlier the Ground truth data was obtained from Google Earth. The images used in this research consists of 10m/pixel and 30m/pixel resolution which are provided by Sentinel level 1c dataset and Landsat8 TOA Reflectance dataset. We mapped the ground truth data on to the sentinel level 1c dataset hosted by Google Earth Engine.

4.3. Dataset. The Sentinel-2 Level-1C and Landsat 8 collection 1 Tier 1 TOA reflectance raster collection were used to perform the land cover classification along with the boundary detection of the selected region [13].

4.4. Classification. Using the features extracted above we have performed the proposed machine learning algorithm CART, and also 2 other machine learning algorithms Random Forest and Support vector machine taking the indices and raw bands as the input features to the algorithms.

Classification and Regression Tree. Classification and regression trees are a way of understanding decision tree techniques which are used for classification and regression learning tasks. CART was developed for regression tasks in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. It is also a prediction model that aids in the discovery of a variable reinforced by other labelled variables. To be more specific, the tree topologies anticipate the outcome by asking a series of if-else arguments.

Classification and regression trees (CART), a basic yet effective prediction method. CART, unlike logistic and linear regression, does not create a prediction equation. Instead, data is partitioned along predictor axes into subsets with homogenous dependent variable values—a procedure illustrated by a decision tree that would be used to create predictions from fresh observations. The CART method is a classification technique that is used to construct a decision tree based on Gini's impurity index. It is a simple machine learning method with a wide range of applications. Leo Breiman, a statistician, created the concept to characterise Decision Tree methods that are often used for classification or regression predictive modelling applications.

A decision Tree is a predictive analytic approach used in statistics, data mining, and machine learning. The decision tree is used as the predictive model in this case, and it is used to progress from observations about an item, which are depicted by branches, to the product's predicted values, which is represented by leaves. Decision trees are among the most popular machine learning approaches due to their accessibility and flexibility.

The CART algorithm does this by utilising the Gini Index criteria to find the optimal homogenization for the subnodes. A decision tree's structure is made up of three major components: root nodes, internal nodes, and leaf nodes. The root node is used as the validation set, and it is categorised into two halves based on the best attribute and threshold value. Furthermore, the subsets are divided using the same rationale. This process is repeated until the tree's last pure sub-set is discovered or the maximum number of leaves feasible in that developing tree is reached. This is sometimes referred to as tree pruning.

Gini's Impurity Index is given by the equ.4.2 shown below.

$$Gini = \sum_{i=1}^{c}((pi))^2 \tag{4.2}$$

Gini's impurity index is a measure used in decision tree algorithms to quantify the impurity or disorder of a set of data points. It ranges from 0 to 1, where a value of 0 indicates a completely pure or homogeneous set, and a value of 1 represents maximum impurity or heterogeneity. The index is calculated by summing the squared probabilities of each class within the data set and subtracting the sum from 1.

Fig. 4.2 combines both training and testing data in order to obtain the better results. The CART algorithm works best for the classification of the land cover type.

4.5. The Random Forest. RF classifier is a collection of decision trees in which randomly sampled rows and attributes are supplied to replacement decision trees. The classification is carried out based on the majority vote that is gathered from decision trees, and the prediction is carried out on regression data using the mean of all the decision tree outputs. Simple trees typically have high variation and low bias. The variance is decreased by increasing the number of trees, making it the ideal model to fit the data. The hyperparameters that must be selected in order to train the model are the number of trees and tree depth. It is the best classifier for all kinds of data because it almost never over fits the model and is immune to the curse of dimensionality. Random forest classifier suits well for land cover classification and classifies better than many existing statistical methods.

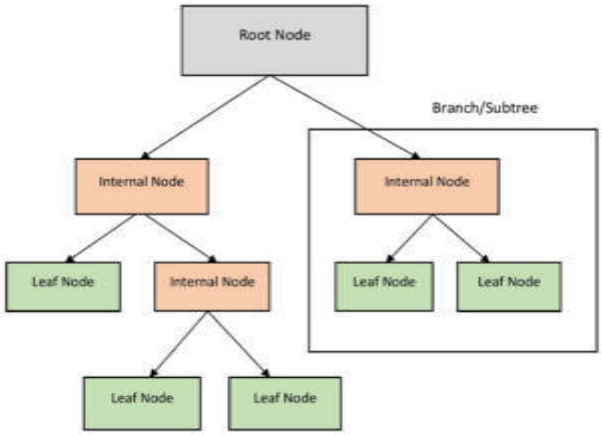


Fig. 4.2: Architecture of CART

4.6. The Support Vector Machine. SVM is considered as one of the best classification models. It has been introduced in the 1960's and later improvised in 1990's. Hyper planes are margins that help in classification of data points. Different classes are constructed by hyper planes that is caused by data points falling on either side of it. The number of features define the dimension of the hyper plane i.e., n features require n-1D hyper plane. Kernel function is a term used for hyper planes. In the current study, a radial-based kernel that classifies in infinite dimensions is applied. Kernel functions assume that the points in space exist in higher dimensions without actually transforming them before calculating the relationship between each pair of points. SVM is used on this data because it works well when the data does not have dimensionality issues and trains more effectively with less samples [14-15].

4.7. Boundary Detection. After the classification is performed we have implemented boundary detection for various crop fields in the different region of interests using canny edge detection algorithm.

4.8. Canny Edge. Canny edge detector is an edge detection algorithm that works in multiple levels in order to identify the edges in given input images. Among all the edge detection algorithms the canny edge detector provides good and reliable detection. This algorithm is adaptable to various environments as it helps in the detection of the boundaries even if the images consists of different characteristics within it. Hence this algorithm was choosen for our crop field boundary detection shown in Fig.4.3.

The Canny Edge Detector was developed by John F Canny in 1986. This technique only extracts the required information from the input and by this it reduces the amount data to be processed. It is being widely applied in computer vision systems. Due to the simplicity of the processing and implementation the canny edge detection is widely used. The canny edge detection algorithm works in 5 different steps those are 1) Applying Gaussian filter in order to remove the noise and smoothen the image, 2) identifying the different intensity gradients in the input image, 3) Applying the gradient magnitude thresholding to get rid of the spurious response to edge detection, 4) Applying double threshold to determine potential edges, 5) Track the edges by hysteresis.

As the first step A Gaussian filter is applied to the input image in order to filter out and reduce the noise in the image. If there is any noise present, it may lead to the false detection. Hence the Gaussian filter is convolved with the input image to smooth it.

The process of utilizing the Gaussian equation to eliminate noise from an input image revolves around convolving the image with a Gaussian kernel. This kernel represents a two-dimensional distribution adhering to the Gaussian probability density function. Its characteristics are defined by both its size and standard deviation (sigma), which dictates the distribution's extent [14-15]. The application of the Gaussian filter involves

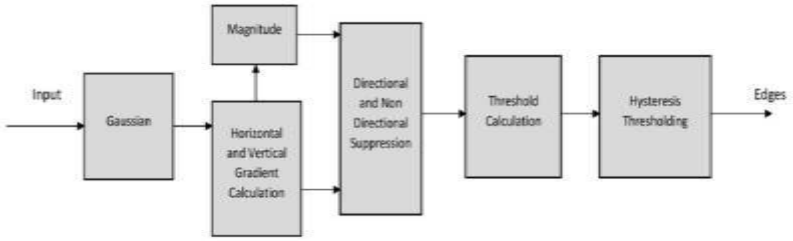


Fig. 4.3: Working of the canny edge detection algorithm

convolving each individual pixel within the image with the Gaussian kernel. This action entails computing a weighted mean of the pixel values found in the vicinity of each pixel. The weights are determined by the values assigned by the Gaussian distribution at those specific positions. Subsequently, the computed weighted average value takes the place of the original pixel value, leading to a notable reduction in noise and the attenuation of high-frequency intricacies present within the image. This iterative procedure is conducted for every single pixel throughout the image, culminating in a filtered image that exhibits significantly diminished noise levels.

In the given input image the edge maybe present in different directions, so in order to overcome this canny edge detector uses four filters to detect vertical, horizontal and diagonal edges. In order to thin out the edges the non-maximum suppression is performed. After performing this step the edges become thinner. But, it can be observed that there is a difference between the intensity of different pixels.

In the process of applying the Gaussian filter to an input image, every pixel within the image undergoes convolution with the Gaussian kernel [14]. This operation entails computing a weighted mean of pixel values surrounding each individual pixel, where the weighting factors are governed by the Gaussian distribution. The computed weighted average is then utilized to substitute the initial pixel value. The outcome is a perceptible reduction in noise and a decrease in high-frequency intricacies present in the image. This iterative procedure is carried out for each pixel within the image, ultimately yielding a filtered image characterized by diminished noise levels.

In the further steps threshold is calculated by identifying the high intensity and low intensity pixels. After getting the threshold results the hysteresis thersholding is performed. This is the final and important step which transforms the weak pixels into strong ones. In this way the Canny edge detection algorithm works. The working flow diagram of the above discussed five steps is shown in Fig. 4.3.

The compactness of the vertical gradient and horizontal gradient depends on the specific characteristics of the proposed work and the image being processed shown in Fig.4.3.

Algorithm:

1. Identification of agricultural fields
2. Partition the datasets into 80% training data and 20% testing data
3. CART classification algorithm is applied on datasets
4. The NDWI based canny edge detection algorithm is applied on the classified data for boundary detection.
5. Qualitative assessment

5. Results and Discussion. As discussed earlier this research is done in two major steps those are classification and boundary detection

5.1. Classification:. The land cover classification for the different region of interests is performed using ML algorithm CART, and two other machine learning algorithms RF and SVM are also performed in order to

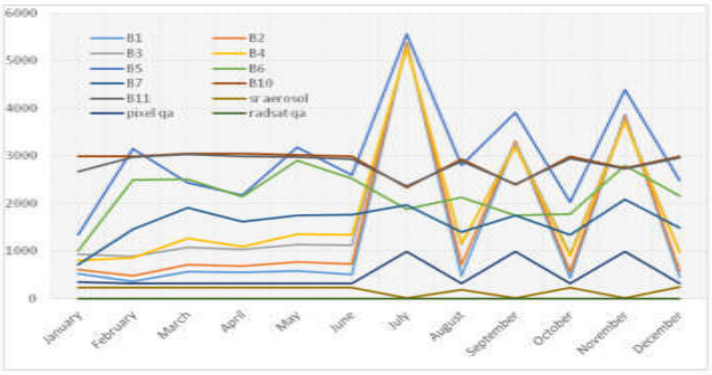


Fig. 5.1: Reflection differences of the all the bands of Landsat 8 in the year of 2021

Fig. 5.2: Visualization of NDVI for the year 2021 in the region of interest 1

compare the performance of the proposed algorithm. Our data is trained with vegetation indices NDVI and NDWI and also all the bands in the corresponding year of 2021. 20% of the dataset was used as testing data. The following represents the visualized outputs obtained by using the classifiers.

As observed in the figures of Figs. 5.1-5.4 gives the visualized output differentiates between the land cover type and the colors red, green and blue represents the Urban, Croplands, River respectively. The classified output was very satisfactory with the good amount of accuracy shown in Fig. 5.6.

5.2. Evaluation Matrices. As discussed earlier the data was extracted and cleaning, pre-processing is done. After these steps the data is fed to several machine learning algorithms like CART, RF, SVM and Canny. The effectiveness of a model can be determined by performing evaluation using some standard metrics. The classification models in this study were evaluated using confusion matrix.

Confusion matrix is actually a table which is used to test the performance of a classification model where the true values a known. The table is a combination of Actual values vs Predicted values.

The matrix consists of 2 classes 'Yes' and 'No'. Using this confusion matrix, Accuracy, Recall and Precision are calculated. Let us know some basic terms used in a confusion matrix. These basic terms are used in calculating the Overall accuracy and Kappa Coefficient. Those basic terms are given below:

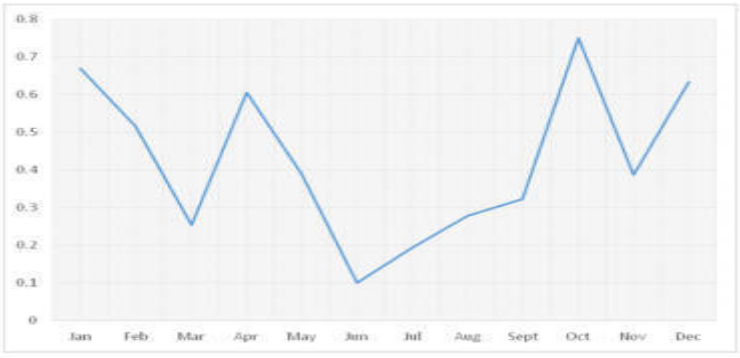


Fig. 5.3: Visualization of NDVI for the year 2021 in the region of interest 2

Fig. 5.4: Visualization of NDVIfor the year 2021 in the region of interest 3

True positive (TP). This is a case where the predicted class is Yes and the actual class is also a Yes.

True negative (TN). This is a case where the predicted class is No and the actual class is also a No.

False positive (FP.) This is a case where the predicted class is a Yes and the actual class is No. This is also known as "Type I error".

False negative (FP). This is a case where the predicted class is a No and the actual class is Yes. This is also known as "Type II error". In a confusion matrix the rows correspond to actual values and the columns correspond to the predicted values. Using this confusion matrix, Recall, Precision and Accuracy can be calculated.

Recall. Out of all affirmative classes, how many are successfully predicted is calculated. This value must be high for good classifier. Recall is given by following equation.

$$Recall = \frac{TP}{TP + FN} \tag{5.1}$$

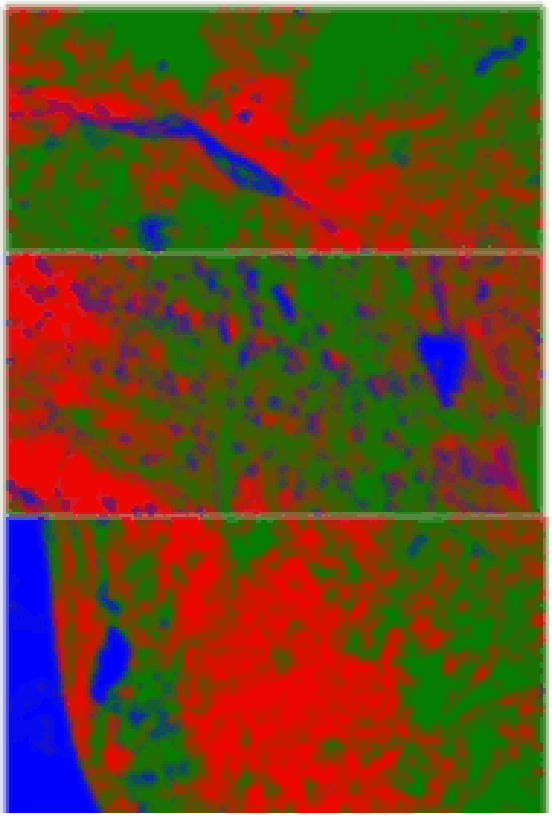


Fig. 5.5: Classified output of CART Classifier for 3 ROIs

Precision. It determines how accurate the model is out of predicted positive, number of actual positive classes. It must be high for a good classification result. It is formulated as shown in equation.

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

Accuracy. It determines the accuracy of the model based on True positive and true negative values. It must be high for good classifier. Which is also termed as overall accuracy. It is given by following equation

$$Accuracy = \frac{TP + TN}{Total} \tag{5.3}$$

Cohen's Kappa. Cohen's Kappa coefficient (k) is a measure of how closely the classified data by a machine learning classifier matches the actual data or the ground truth data. Generally a classifier with kappa statistic value between 0.4 to 0.75 is considered as a good classifier and above 0.75 (>0.75) is considered to be as excellent. It is also calculated by using a confusion matrix. The Cohen's Kappa formula can be given as:

$$Cohen'sKappa = \frac{2 * (TP * TN - FN * FP))}{((TP + FP) + (FP + TN) * (FN + TN))} \tag{5.4}$$

To calculate Cohen's Kappa in the proposed work, the agreement between two raters is measured using the observed agreement and expected agreement. The observed agreement is the proportion of cases where the raters agree, and the expected agreement is the agreement expected by chance. Cohen's Kappa is then

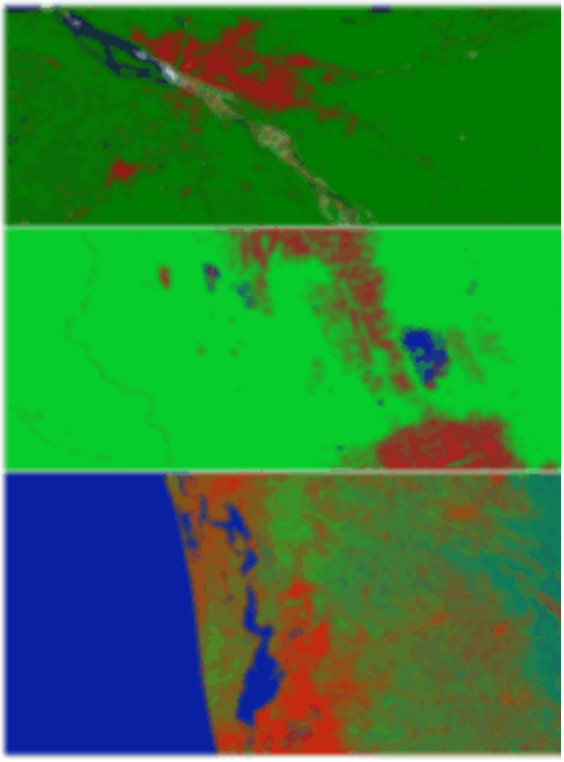


Fig. 5.6: Classified output of the SVM for 3 ROIs

Table 5.1: Confusion Matrix of CART Classifier

+/-	0	1	2
0	66	8	4
1	0	128	30
2	0	5	981

calculated by subtracting the expected agreement from the observed agreement and normalizing it by dividing by 1 minus the expected agreement [16-17].

The overall accuracy in a confusion matrix are calculated by adding all the correctly classified values and further by dividing it with the total number of values in the matrix. To show the performance of classifier models the overall accuracy (OA) is evaluated using the confusion matrix provided by the Google Earth Engine along with Cohen's kappa coefficient for each classifier using the confusion matrix.

It can be observed that some classes are missing in Table 5.1. As these classes don't have enough samples, the model excluded them while being trained and hence, the matrix does not reflect them. The confusion matrix shown above is calculated in Google earth engine and further the evaluation matrices were calculated according to the formulae. The overall accuracy (OA) was obtained as 96.1% and the kappa coefficient (k) is 0.87 for CART which is very good. Furthermore, the classified output was visualised as shown below in Figure 5.7.

It is interesting that other classifiers also have given good results with good amount of accuracies. The outputs were visualized in Google Earth Engine platform for the Random forest classifier also the visualized outputs are shown in below figure 10. The Random Forest Classifier have shown less accuracy when compared to CART classifier because it found difficulty in differentiating between the features given to it as the input.

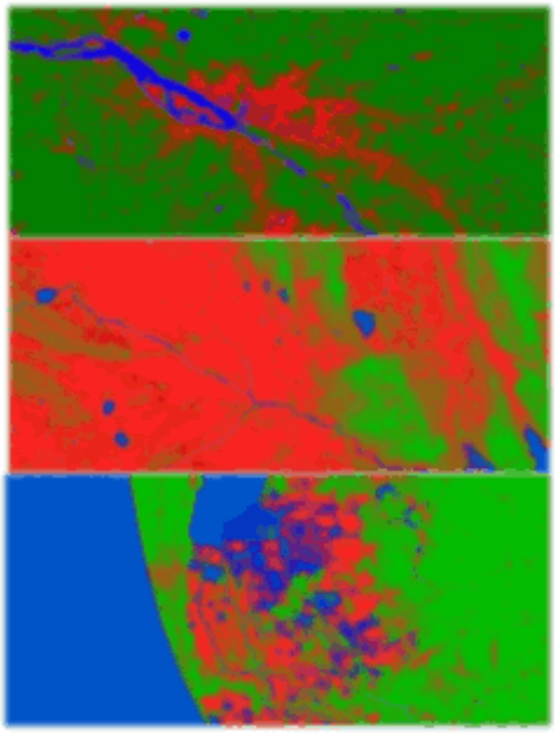


Fig. 5.7: a) Overall accuracy of classifiers b) Kappa coefficient of each classifier

The Random Forest classifier model hosted by Google Earth Engine was applied to the same training data and the results obtained were overall accuracy (OA) is 83.6% and the kappa coefficient (k) is 0.71 which means it was a good statistic. The random forest found difficulties in differentiating among the wetlands, rivers and croplands. The confusion matrix was used to calculate the accuracy and kappa statistic [18]. The support vector machine classifier model hosted by Google Earth Engine was applied to the same training data and the results obtained were overall accuracy (OA) is 85.7% and the kappa coefficient (k) is 0.73 which means it was a good statistic. The confusion matrix was used to calculate the accuracy and kappa statistic. It was applied in order to find an efficient algorithm for the classification of the land around the selected region of interest. But the SVM algorithm was showing accuracy better than RF but lesser than CART shown in Fig. 5.7.

6. Crop field Boundary Detection. In this study canny edge detection was used for detecting the boundaries of the crop fields identified by the CART classifier. NDWI index is calculated for the time period of Kharif season (The months of June – October in southern India) and Rabi Season (The months of October – February in southern India). NDWI is given as an input to the Canny. Number of images are added during the period of each season this summation of the images improve the results and obtain better output data.The NDWI Index was calculated for each month and fed as the input to the canny edge detection algorithm. Several images added upon each image in order to obtain a better output. The limiting factor was the resolution which is 10m/pixel yet still the data provided good results. As shown in Table. 6.1 the bands with high resolution were used to calculate the NDWI those bands are B3 and B8 these bands are used to calculate the NDWI, usage of the bands with low resolution such as B6, B7 need to be avoided in order to obtain good results. The spectral response was influenced by the dry crop growth and total yield decline in 2020, as illustrated in Fig. 6.1.

In this study we have calculated the NDWI for 2 different seasons and 3 different region of interest in order test the efficiency of the algorithm. The reflectance values and NDWI values are dependent on the weather conditions and the growth progress of the crops in the selected region of interest. Hence, The NDWI is shown in

Table 6.1: Comparison of different algorithms with the proposed algorithm

S.No	Algorithm	Kappa coefficient	Accuracy
1	SVM	0.73	85.6%
2	RF	0.71	83.1%
3	CART	0.87	96.1%

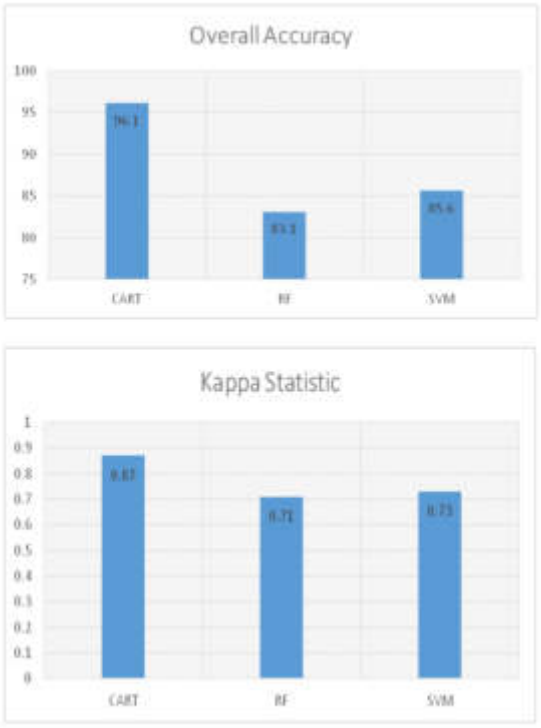


Fig. 6.1: a) Band Reflectance values for ROI-1 in Kharif season b) Rabi season

next figures for each region of interest. The Vijayawada region situated in Krishna basin is assumed as ROI-1, Mydukur area which is situated in rayalaseema region of Andhra Pradesh state was assumed as ROI-2 and the Alappuzha region of the Kerala state was considered as the ROI-3. The visualized output of the canny edge detector which is applied for the Region of interests is shown in Fig. 6.2 and Fig. 6.3.

The trained model, which was previously given, was used for a site close to Alappuzha and for the period of May 2020 to February 2021. This location is not part of the training data set, so we can evaluate how well our model performs in a real-world situation. As seen in the Fig. 6.3 field boundaries are similar in both the Kharif and Rabi seasons by this we can say that our model is efficient and the weather conditions doesn't affect the algorithm.

The results obtained through classification and the crop field boundary detection are satisfactory and the input data show the influence on the performance of the algorithms. Data preprocessing and data cleaning are very important steps that are to be carried before every data science application. As expected, the NDWI-based canny edge detection produced excellent results. The NDVI index and raw bands were used as the features, and the results were excellent and satisfactory. The results are significantly impacted by the use of the raw bands B6, B7, B8, B11, or B12.

Data pre-processing, cleansing, and cloud filtering was another key aspect. The presence of the clouds does

Fig. 6.2: NDWITime series of our ROI-1 during May 2020 – Feb 2021

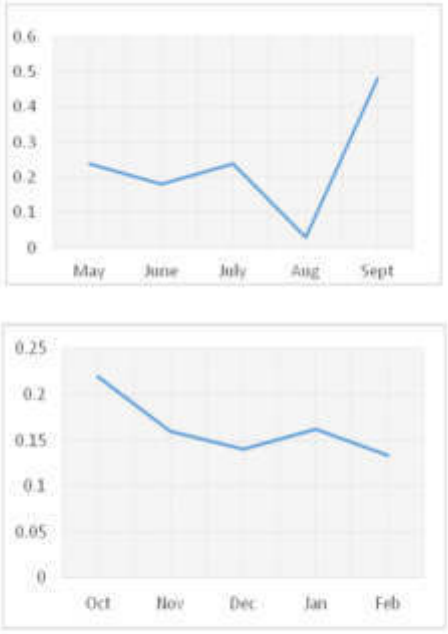


Fig. 6.3: NWDI based Canny output for ROI-2 in Kharif (left) and in Rabi (right) seasons

effect the boundary detection but does not affect much to the classification as compared to the field boundary detection.

The results produced very good classification accuracies and the boundary detection was also very good

when compared to the ground truth data. The limiting factor was the resolution which is 10m/pixel yet still the model provided good results. As a result, we contend that our model's performance is excellent and that it is ideally suited for use in real-time scenarios in everyday life. Sentinel 2 and Landsat 8 were preferred as the datasets can be obtained on the regular basis.

7. Conclusion and Future Scope. This study comprehensively tackled aspects of land cover classification and the detection of crop field boundaries, while striving to propose an efficient machine learning algorithm. These applications are particularly vital for analysing crop yields in cases where field information is scarce. The study employed a variety of algorithms, including CART, RF, SVM, and Canny Edge detection. Additionally, indices like NDVI and NDWI were computed, with the incorporation of raw bands as inputs leading to the discovery of supplementary information and increased result accuracy.

Key stages of data pre-processing and cloud filtering were integral to both classification and crop field boundary detection processes. The algorithmic evaluations spanned the time frame of 2020 to 2021. The application of Canny edge detection in conjunction with NDWI yielded notably favourable outcomes. To ensure temporal diversity, this approach was executed on different days during the Kharif and Rabi seasons across 2020 and 2021. Comparative analysis revealed the CART model's superior performance among classifiers. Trained on the specified data, CART exhibited an impressive 96.1% Overall Accuracy and a 0.87 Kappa Coefficient, signifying its efficacy as a classifier. Similarly, RF demonstrated an Overall Accuracy of 83.1% with a Kappa Coefficient of 0.71, while SVM achieved an Overall Accuracy of 85.6% and a Kappa Coefficient of 0.73.

Based on these quantitative metrics, the CART model was advocated for land cover classification due to its robust performance. For precise crop field boundary detection, NDWI-based Canny edge detection was applied during the Kharif and Rabi seasons, producing satisfactory results aligned with ground truth data. Thus, the NDWI-based Canny edge detection method was proposed as a feasible solution. The study's potential extends to predicting crop yields and optimizing pesticide and insecticide usage based on detected boundaries. By leveraging insights from this research, it is possible to make informed decisions that enhance agricultural practices and resource allocation.

REFERENCES

[1] Achanta R and Süsstrunk S. (2017) Superpixels and Polygons Using Simple Non-iterative Clustering. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4895-4904, doi: 10.1109/CVPR.2017.520.
[2] Breunig, Fábio Marcelo, et al. (2020) Delineation of management zones in agricultural fields using cover–crop biomass estimates from PlanetScope data. International Journal of Applied Earth Observation and Geoinformation 85: 102004
[3] Chang L, Peng-Sen S, Shi-Rong L, (2016) A review of plant spectral reflectance response to water physiological changes. Chinese Journal of Plant Ecology, vol. 40, no. 1, pp. 80–91.
[4] Dixon, Dan J, et al. (2021) "Satellite prediction of forest flowering phenology." Remote Sensing of Environment 255): 112197.
[5] NagaJyothi, G., and SriDevi, S. (2017, March). Distributed arithmetic architectures for fir filters-a comparative review. In 2017 International conference on wireless communications, signal processing and networking (WiSPNET) (pp. 2684-2690). IEEE.
[6] Nagajyothi, G., Kumar, G. P., Kumar, B. S., Kumar, B. D., and Damodaram, A. K. (2023, March). High-Speed Low Area 2D FIR Filter Using Vedic Multiplier. In Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022 (pp. 569-578). Singapore: Springer Nature Singapore.
[7] Jyothi, G. N., Anusha, G., Kumar, N. D., and Kundu, D. (2019). Design of FINFET based DRAM cell for low power applications. Computer-Aided Developments: Electronics and Communication, 35-43.
[8] Gamanya, R., P.D. Maeyer, and M.D. Dapper. (2007) An automated satellite image classification design using object- oriented segmentation algorithms: A move towards standardization, Expert Systems with Applications, 32:616– 624.
[9] Kokhan S, Vostokov A. (2020) Application of nanosatellites PlanetScope data to monitor crop growth. In Proceedings of the E3S Web of Conferences, Lublin, Poland, 17–20 September 2020; Volume 171, p. 2014
[10] Kumar, L.,and Mutanga, O. (2018). Google Earth Engine applications since inception: Usage, trends, and potential. Remote Sensing, 10(10), 1509.
[11] Pandey, A., and Jain, K. (2022). An intelligent system for crop identification and classification from UAV images using conjugated dense convolutional neural network. Computers and Electronics in Agriculture, 192, 106543.
[12] Roopashree, S., Anitha, J., Mahesh, T. R., Kumar, V. V., Viriyasitavat, W., & Kaur, A. (2022). An IoT based authentication system for therapeutic herbs measured by local descriptors using machine learning approach. Measurement, 200, 111484.
[13] Kumar, V. V., Raghunath, K. K., Rajesh, N., Venkatesan, M., Joseph, R. B., & Thillaiarasu, N. (2022). Paddy plant disease recognition, risk analysis, and classification using deep convolution neuro-fuzzy network. Journal of Mobile Multimedia, 325-348.

[14] Vinoth Kumar, V., Wang, L., Chen, J. I. Z., Sikdar, B., & Nones, M. (2022). Guest editorial: Trends, perspectives and prospects of sensor technologies in hydrological sciences. Acta Geophysica, 70(6), 2837-2839.
[15] Piao, Shilong, et al. (2019) Plant phenology and global climate change: Current progresses and challenges. Global change biology 25.6: 1922-1940.
[16] Qayyum, Nida, et al. (2020) Glacial lakes mapping using multi satellite PlanetScope imagery and deep learning. ISPRS International Journal of Geo-Information9.10: 560.
[17] Bhavana, D., Likhita, N., Madhumitha, G. V., & Ratnam, D. V. (2023). Machine learning based object-level crop classification of PlanetScope data at South India Basin. Earth Science Informatics, 16(1), 91-104.
[18] Chang L, Peng-Sen S, Shi-Rong L, (2016) A review of plant spectral reflectance response to water physiological changes. Chinese Journal of Plant Ecology, vol. 40, no. 1, pp. 80–91.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for future generation IoT systems
Received: May 26, 2023
Accepted: Sep 25, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 535–545, DOI 10.12694/scpe.v25i1.2320

HYBRID ARCHITECTURE STRATEGIES FOR THE PREDICTION OF ACUTE PULMONARY EMBOLISM FROM COMPUTED TOMOGRAPHY IMAGES

PRIYANKA YADLAPALLI* AND D. BHAVANA†

Abstract. The timely identification of pulmonary embolism is of utmost importance, as the condition has the potential to be life-threatening if not promptly addressed. The assessment of the severity of a pulmonary embolism (PE) frequently necessitates a time-consuming and potentially life-threatening estimation by a medical practitioner. The primary objective of the study was to investigate the potential utility of an artificial neural network in assisting physicians with the identification and prediction of pulmonary embolism risk in patients. Deep learning algorithms are frequently employed in medical imaging to enhance image analysis due to their ability to automatically learn representations from large datasets, as opposed to relying on pre-programmed instructions . The implementation of automated systems has the potential to decrease the level of physical effort needed and enhance the efficiency of diagnostic procedures for medical professionals. Efficient training and calculation processes are crucial for the proper execution of the implementation. The Tensor Processing Units (TPUs) developed by Google are employed to expedite the process of training., with the computational tasks being executed through Google Colab, a platform offered by Google Cloud TPUs. In order to achieve outcomes comparable to human judgment, deep learning algorithms engage in reasonable assessments of data based on a predetermined logical framework. Diagnosing pulmonary embolism (PE), a potentially lethal yet curable condition, poses challenges in early detection. A distinctive convolutional neural network (CNN) model was developed and examined for the purpose of distinguishing between pulmonary embolism (PE) and computed tomography (CT) pictures.

The proposed study yields a precision rate of 91.2%, showcasing an enhancement compared to current convolutional neural network (CNN) architectures that include limited trainable parameters. Furthermore, our model provides interpretability by the utilization of computed tomography (CT) images, specifically in the inferno and bone models. Our proposed deep learning model has the potential to predict the presence of PE and other associated features in current cases.

Key words: Deep Learning, Pulmonary Embolism, Dense Net, CT Scan, Artificial Intelligence

1. Introduction. Machine learning is a constituent part of both artificial intelligence (AI) and machine learning itself, and conversely, AI and machine learning encompass machine learning as a subset. Artificial intelligence (AI)[1], a broad concept, pertains to computer programs that exhibit human-like behavior. The advent of machine learning, which encompasses a variety of algorithms[2,3], has facilitated the realization of these achievements. Machine learning is a subfield that falls under the domains of both artificial intelligence (AI) and machine learning itself, and conversely, AI and machine learning encompass machine learning as a subset. Artificial intelligence (AI), a broad concept, pertains to computer programs that exhibit human-like behavior. The advent of machine learning, which encompasses a range of algorithms, has facilitated the achievement of these outcomes.

Deep Learning, on the other hand, is a subset of machine learning that is inspired by how the human brain is structured [4,5]. In order to analyse data and come to conclusions that are comparable to those of humans, deep learning algorithms employ a predetermined logical structure. Deep learning with neural networks is used to achieve this [6,7]. The basic diagram of deep learning is as shown in Fig.1.1. and the neural network is shown in Fig.1.2.

Deep learning algorithms are constructed dynamically to work across many layers of neural networks and are nothing more than a collection of previously trained decision-making networks. Then, each of these is put through a series of simple layered representations before moving on to the next layer [8].

*Research Scholar, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur, India; Department of Electronics and Communications, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

†Associate Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, AP, 522502, India (Corresponding Author: bhavanaece@kluniversity.in)

Fig. 1.1: Deep Learning ref [2]

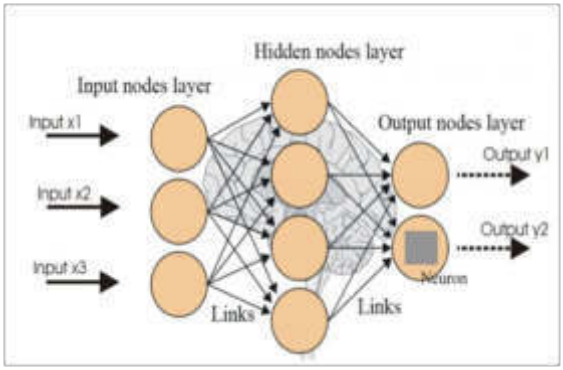


Fig. 1.2: Neural Network of ref [4]

The principal benefit of deep convolutional neural networks (DCNNs) illustrated in Fig.1.3 lies in its hierarchical architecture, which is distinguished by several layers. A three-dimensional neural network design, known as a deep convolutional neural network (DCNN), is utilized to concurrently process the Red, Green, and Blue components of an image [9,10]. The application of this methodology results in a substantial reduction in the number of artificial neurons required for image processing, as compared to traditional feed forward neural networks. Deep convolutional neural networks (CNNs) are a machine learning model specifically developed for the purpose of image processing and analysis [11,12]. These neural networks accept photos as input and employ them in the training process of a classifier. The network employs a unique mathematical operation called a "convolution" instead of doing matrix multiplication.

The accuracy of CNN classification in medical image analysis surpasses that of human visual perception, enabling the detection of anomalies in X-ray or CT scan pictures. These systems possess the capability to examine sequences of images, such as those obtained over an extended duration, and detect small variations that may elude human analysts. Additionally, this enables the execution of predictive analyses. The training of classification models for medical pictures is conducted using extensive public health datasets. The resultant models possess the capability to be applied to patient test outcomes[13], enabling the identification of medical disorders and the automated generation of a prognosis.

PE refers to a group of conditions that cause scarring in the lungs. Diffuse parenchymal lung diseases is another term for it. Scarring causes the lungs to stiffen, making breathing and absorbing oxygen more difficult. PEs cause permanent lung damage that worsens with time and shown in Fig.1.4.

1.1. Pulmonary Embolism Types:.

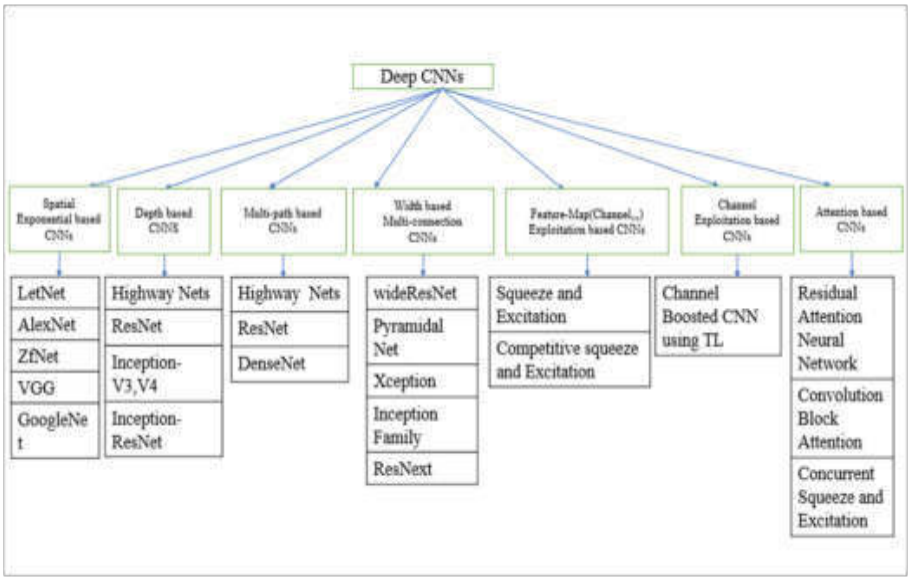


Fig. 1.3: Types of Deep Learning Algorithms

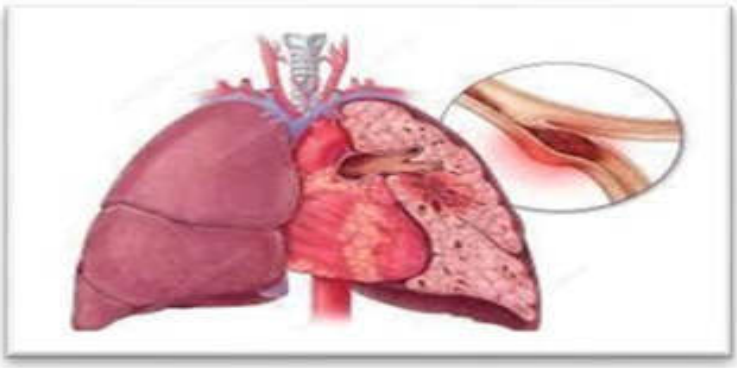


Fig. 1.4: Lung effected with PE of ref [8]

1. Acute pulmonary embolism: Acute pulmonary embolism is a condition where blood clots develop in the pulmonary arteries and block them, making breathing extremely difficult and causing chest pain.
2. Chronic pulmonary embolism: Multiple blood clots and repeated pulmonary embolism caused by 'deep vein thrombosis' do not dissolve and continue to obstruct the blood supply to the lungs, resulting in chronic pulmonary embolism.

2. Literature Survey. In the study conducted by the author, a number of machine learning models were developed and employed, including neural networks, gradient boosted trees, and logistic regression [14]. The author provided training to the participants using a dataset consisting of health information from 63,798 inpatients who received medical and surgical care at a prominent medical center in the United States. The XGBoost model had superior performance in predicting pulmonary embolisms. The XGBoost model achieved an AUROC of 0.85, demonstrating a sensitivity of 81% and a specificity of 70%.

The neural network model proposed by the author in reference [15] incorporates the InceptionResNet and CNN architectures, together with a long-short-term memory network, to effectively process whole stacks

of CTPA images in the form of sequential slices. The analysis of the model's efficiency was conducted by considering multiple parameters. The author's conclusion is that, at the stack and slice levels, both models demonstrated strong performance, with specificity and sensitivity rates of 93.5% and 86.6% respectively.

In the aforementioned scholarly article [16], the researcher conducted a comparative analysis of two variants of Artificial Neural Networks (ANNs). The study was conducted utilizing 294 patients from three hospitals, employing feed-forward and Elman backpropagation models. The researchers utilized an enhanced artificial neural network in conjunction with a perfusion scan diagnostic technique. The study achieved accuracy rates of 93.23% and 86.61% for the relevant tasks. This study has the potential to enhance the accuracy of risk assessment for patients and contribute to reducing mortality rates, benefiting physicians, medical assistants, and healthcare professionals. The author utilized four pre-existing convolutional neural network (CNN) architectures, namely Inception, VGG-16, ResNet50, and Mobilenet, in order to classify cases of pulmonary embolism, as stated in reference [17]. The experimental results indicate that the Inception-based CNN model exhibits superior performance compared to other CNN architectures.

The study described in Reference [18] proposes a systematic approach for the identification of pulmonary embolisms through the utilization of computed tomography (CT) scans. The U-net network is employed to identify potential embolism candidates and afterwards conduct classification. The approach achieved a Dice coefficient of 0.81 and an Intersection over Union (IoU) score of 0.79. The author [19] compared pretrained design performance. This comparison uses loss, accuracy, and AUC. The study found that slimmer architectures including MobileNet, VGG, ResNet, and U-net outperformed Inception, DenseNet, and Xception. This study also found a significant confidence gap between picture and exam features.

The author [20] presented a novel approach in their study, which aimed to predict outcomes on computed tomography (CT) exams. The methodology employed in this study consisted of implementing a two-stage model that integrated Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The technique that was advised shown a higher level of performance in comparison to both the conventional CNN model and a single stage CNN-LSTM network. This is supported by the Area Under the Curve (AUC) value of 0.95.

The study by [21] presents a mask RCNN model that utilizes a probability-based approach to enhance the local properties of densely populated microscopic particles by initially upsampling the values in the feature map. Additionally, a specific region was selected from the retrieved image based on the probability of pulmonary embolism (PE) occurrence. The extraction of anchors within the candidate zone, as opposed to the entire image, offers time and space efficiency benefits, particularly in light of the expanding feature maps. The author asserts that the quality of the work was enhanced with the elimination of a majority of invalid anchors. The empirical evidence demonstrates that the proposed technique is both effective and efficient.

In [22], the author used an artery-aware 3D fully convolutional network (AANet) to encode arterial data as prior information to recognize arteries and PEs. The author suggests using local and global vascular artery values as soft attentions to distinguish pulmonary embolisms (PEs) from other soft tissues. The CAD-PE dataset was used to test the techniques for the artery.

3. Proposed Model. A collection of Chest CT-Scan images was acquired for our investigation using the Kaggle platform. The collection comprises chest computed tomography scans and associated patient medical data. The dataset has two subfolders. The folders serve as repositories for training and examination materials. The existing data was further adjusted in order to consolidate the final training data by reducing the potential influence of confounding variables and enhancing the distribution of classes. Furthermore, we have been provided with the CSV files containing the train and test datasets. The train folder consists of CT scans from a sample of 100 distinct patients, whereas the test folder exclusively comprises CT scans from a smaller sample of only five patients. In order to optimize the model's ability to create data with a high level of accuracy and efficiency, it is recommended to utilize the train folder and is shown in Fig.3.1.

One of the parameters is frequently given more weight by deep learning algorithms. These two objectives can coexist using Pytorch, a machine learning library that makes it possible to use code as a model, makes debugging easier, and is compatible with a variety of widely used computing libraries while still being effective and GPU-friendly.

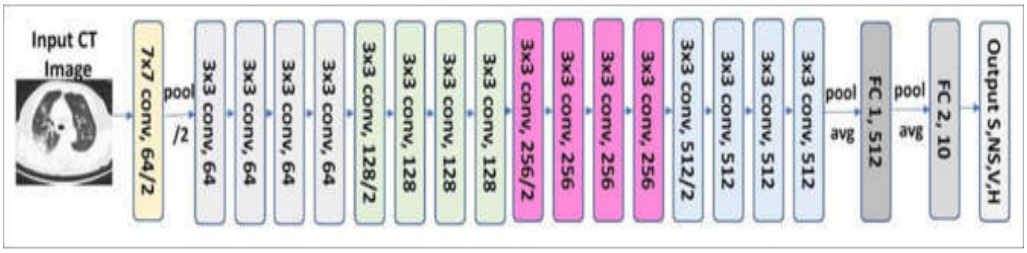


Fig. 3.1: ResNet Architecture

3.1. Transfer Learning Techniques:. A substantial dataset and robust computer resources are necessary to develop a high-performing CNN from the beginning. Here's where TL comes in; it's a super-effective strategy that drastically cuts training time while still needing some information [23, 24]. The goal of this approach is to use crucial information (weights) from CNN models that have already been trained on a big dataset to address problems in other domains. TL improves CNN models' generalization abilities. This has led to many uses for pre-trained CNN models' TL, especially in the medical imaging field, where it can aid in the classification of pictures used in diagnostic procedures. Several widely used CNN models have already been pre-trained. We have utilized ResNet and DenseNet models to differentiate between PE and not PE because of their high performance in picture classification. To begin, we collected 644 CTA photos from both PE and non-PE classrooms to create a well-rounded collection. Then, we've used a method called data augmentation to further enrich the dataset by adding in updated versions of the original input images.Operations like as rotation, resizing, cropping, and zooming have been used. The developed dataset was actually used to feed the already-trained ResNet and DenseNet CNN models.

3.2. RESNET Architecture:. The ResNet, also known as the residual neural network, is a type of architectural design that integrates residual learning into a conventional convolutional neural network. This integration serves as a strategy to address the issues of gradient dispersion and accuracy degradation that are commonly observed in deep networks throughout the training process [25]. The user possesses full control over the accuracy and speed of their development. The methodology for implementing the ResNet model is illustrated in Figure 3. Assume that the variable x represents the input value. If the working weights exhibit negative values, it is advisable to disregard such data. The weights are exclusively accessible to the relu activation function, which lacks the capability to transmit them to other entities.

Dense Net was created because the vanishing gradient has a major impact on the performance of advanced neural networks. In other words, the further apart the input and output levels are, the more likely it is that the information will be lost along the way. The output of the first layer is the input to the second layer whenever the composite function is utilized. Implementation details of the Dense Net model are depicted below. This multi-step method incorporates non-linear activation layers, convolution, pooling, and batch normalization. DenseNet-121 is one of the several Dense Net representations available and is an implementation of the Dense Net Figure format. DenseNet provides three distinct Dense Blocks, with DenseNet-160 and DenseNet-201 being the most popular options. The numbers signify the neural network's layer count. Numbers are $5 + (6 + 12 + 24 + 16) * 2 = 121$. Compact Net Convolution and pooling at Layer 5 Transitional Layers: 3 (6,12,24) Classification is the top layer (6) Block 2 ($Conv1 * 1and3 * 3$). The vanishing gradients issue is resolved by dense networks, allowing for the training of models with fewer parameters. Data moves swiftly as a result of dynamic feature propagation. In the Figure below , there are three dense blocks that make up a deep dense net. Convolution and pooling are used to change feature- map sizes and link two nearby blocks.

Densely Connected Convolutional Network, also referred to as DenseNet, is an alternative nomenclature for this particular network architecture. The concept of having dense blocks in a neural network architecture, where each layer is interconnected with every other layer in a forward feed way, serves several purposes. Firstly, it addresses the issue of vanishing gradient, which is a common challenge in deep learning. Secondly, it enhances the propagation of features throughout the network, allowing for more effective information flow. Lastly, it

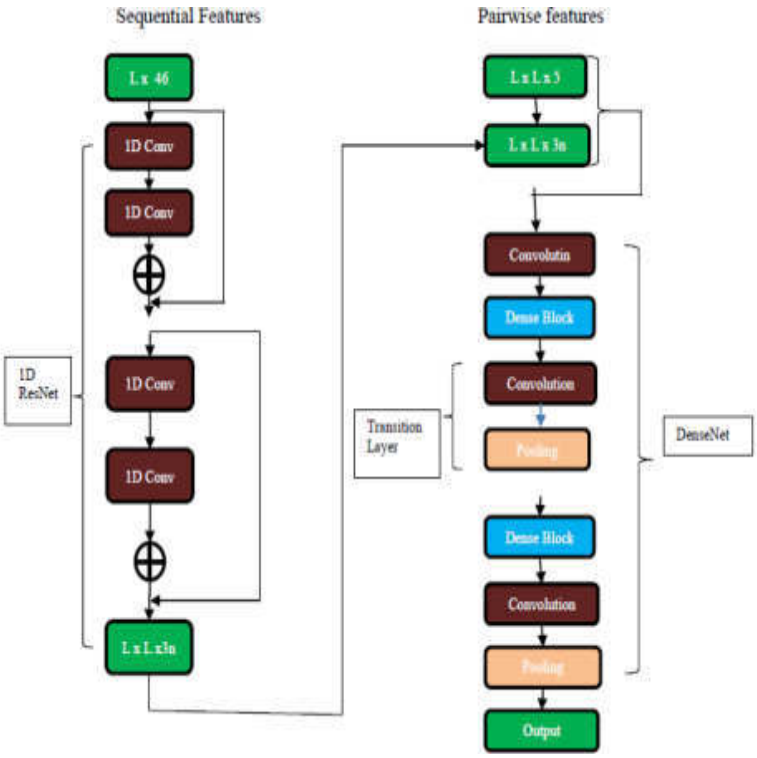


Fig. 3.2: Program flow

promotes the reuse of features, enabling the network to leverage previously learned.

To incorporate residual learning into a DenseNet architecture, original DenseNet design is modified by adding residual connections. Here are the various steps which are involved in the proposed model:

Dense Block. The dense block is the basic building block of DenseNet, consisting of multiple consecutive layers. In each layer, the output feature maps are concatenated with the input feature maps from previous layers.

Transition Layer. After each dense block, a transition layer can be added to downsample the spatial dimensions of the feature maps. This typically involves a convolutional layer and a pooling operation, such as average pooling.

Residual Connection. Skip connections are introduced within the dense block or between the dense blocks. These skip connections can be implemented by adding the input feature maps to the output feature maps of a dense block or a transition layer.

Global Average Pooling and Classifier. At the end of the network, apply global average pooling to aggregate the spatial information into a single feature vector. Connect this feature vector to a fully connected layer or a softmax classifier for the final prediction. By combining the dense connections of DenseNet with residual connections, we can benefit from both the feature reuse and gradient flow properties of DenseNet and the ease of training and optimization provided by residual learning.

This modified architecture, often referred to as DenseNet with residual connections, can improve the model's expressiveness, training efficiency, and overall performance on various tasks.

The framework utilized in this study is founded upon an integrated deep neural network, as depicted in Fig. 3.2. This network is comprised of one-dimensional residual and densely connected convolutional networks. The ResNet architecture possesses the ability to address the issues of gradient vanishing and exploding by virtue of its distinctive identity and residual mapping properties, hence enabling effective training of deep network structures. However, it is important to note that the number of parameters in ResNet is directly related to its

Fig. 3.3: Learning Rate

depth. The DenseNet architecture addresses the issue of gradient vanishing by leveraging its dense connections, which facilitate optimal feature reuse and enhance feature transfer. During the data preparation phase of our system, one-dimensional sequential features are encoded as vectors and subsequently inputted into the ResNet network. Subsequently, the two-dimensional pairwise data are integrated with the one-dimensional characteristics obtained from a one-dimensional residual network, and the combined features are then fed into the DenseNet network.

To incorporate several input features for a residue pair, we concatenate the vectors representing the residues and their corresponding features, namely $=v_1, v_2, ., v_i, ., v_L$. This concatenated vector is then utilized as a single input feature for the residue pair. Next, the pairwise characteristics are integrated with the aforementioned elements to construct the input for the subsequent section of the network. In order to mitigate the issue of overfitting in the network, a dropout method is employed, whereby neurons are randomly deactivated during the training process at a dropout ratio of 80%. An efficient stochastic optimization technique was employed in this study, utilizing the Adam optimization algorithm with a learning rate of 0.01. In our proposed model, the training of model parameters is accomplished through the utilization of the maximum likelihood function. The loss function, which is employed to quantify the discrepancy between predicted and actual values, is formulated as the negative log-likelihood function, specifically referred to as the cross-entropy function. The formula is as follows:

$$E(t, y) = -\sum_i t_i log Y_i \tag{3.1}$$

3.3. Learning Rate:. The learning rate parameter, often ranging from 0 to 1, plays a crucial role in facilitating gradient updating. While its value is often modest, it can reach a maximum of 1 (equivalent to 100%). The attainment of the precise optimum of the loss function can be achieved by modifying the learning rate. Both low and high learning rates lead to inefficiency in terms of time and resource utilization [26,27]. A decrease in the learning rate leads to an extended duration of training, which subsequently results in escalated expenses for cloud GPU usage.A higher rate of occurrence could potentially lead to a model that lacks the ability to reliably forecast outcomes shown in Fig.3.3.

When there is a greater weight disparity between iterations, larger steps are taken [28,29]. The optimal outcome can be quickly attained using this method, but the precise optimal value cannot be. Every iteration is thought of as having smaller steps when the weight differences are smaller. In this method, more epochs are required to achieve the ideal result, but there are very few opportunities to miss the precise optimal value.

4. Results and Discussion. Rapid computation and training are needed for the implementation. Google Tensor Processing Units (TPUs) are employed to speed up training, while Kaggle Cloud TPUs are used to

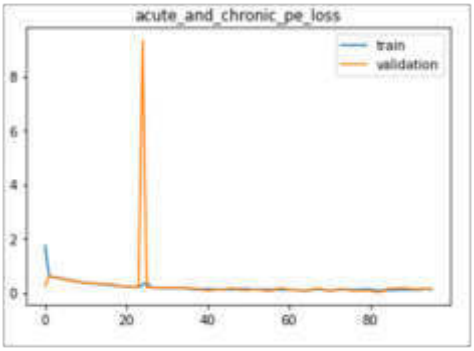


Fig. 4.1: Central PE Loss Of the system

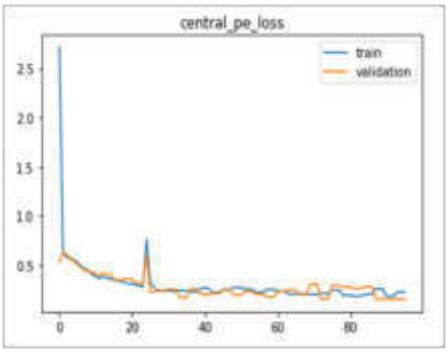


Fig. 4.2: Central PE Loss of the system

complete the work. Using a common training dataset with a batch size of 100 images and an initial learning rate of 0.000010, the PE- DeepNet model was trained across 25 iterations. The optimizer algorithm Adam was utilised, and the loss function was binary cross entropy. The model has a total of 4,399,489 parameters, 4,374,705 of which were trainable, and 24,784 of which were not. Under these circumstances, the model's training accuracy was 94.20 percent and its validation accuracy was 92.30 percent. The following figures show various loses of proposed model

The evaluation or quantification of a model's performance depends in detecting or diagnosing central pulmonary embolism (PE) and Intermediate Pulmonary Embolism. The central PE loss and Intermediate PE loss and acurracy graphs are depicted in Fig.4.1 and Fig.4.2 respectively.

The suggested model's loss and accuracy plots are displayed in the figures above. The validation loss can be lowered to 0.423 by using a higher number of epochs. Area Under Curve of the RoC is 0.567691, Accuracy is 91.2 percent , and Train Loss is 0.513 for the proposed architecture whereas. after 30 epochs, ResNet is 81.54 percent accurate, and the AUC value is found to be 0.853 and training and validation loss are found to be 0.6112 and 0.548 respectively shown in Fig.4.3. Similarly for the same number of epochs for DenseNet we have 83.78 percent accuracy,AUC value is around 0.866 and the training and validation losses are found to be 0.5676 and 0,477 repectively and all the obtained results are depicted in Table. 4.1.

Fig.4.4 is an inferno-themed colormap of CT scans of patients. From the subclass "Perceptually Uniform Sequential," a spectacular visual depiction of the "inferno" is at your disposal shown in Fig.4.5. One of the matplotlib color maps, "The Luminance of Inferno,climbs smoothly and uniformly in brightness over time. The body is black, yet there are attractive purple overtones shown in Fig.4.6.

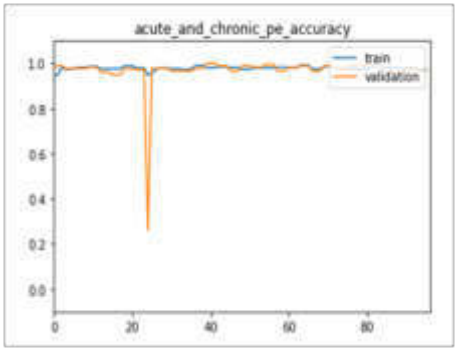


Fig. 4.3: PE Loss and accuracy

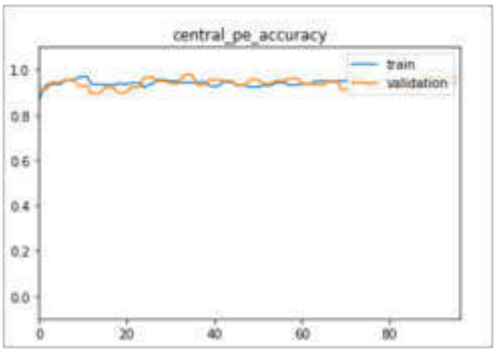


Fig. 4.4: Intermediate PE Loss and accuracy

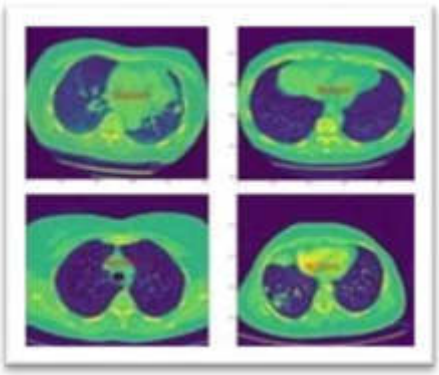


Fig. 4.5: CT scans in Bone Model

5. Conclusion. By combining both ResNet and DenseNet we can leverage the strengths of both the architectures.This hybrid architecture aims to capture the benefits of residual connections in ResNet and dense connections in DenseNet, While we have different ways to combine these techniques ,we have introduced skip connections along with dense connections within the network so as to improve to the gradient flow ,feature reuse and overall performance. The proposed model was applied to the provided dataset to find the presence of Pulmonary Embolism and for calculation of Accuracy, and the associated validation loss. Using these results,

Table 4.1: Comparison Table

	AUC	Accuracy	Training loss	Validation Loss
ResNet	0.853	81.54	0.6112	0.548
DenseNet	0.866	83.78	0.5676	0.477
Proposed system	0.870	91.2	0.513	0.423

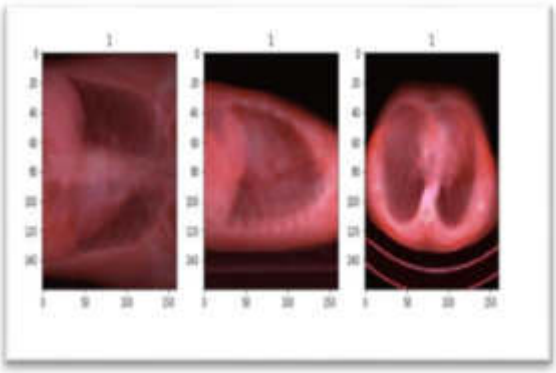


Fig. 4.6: CT scans in Inferno Model

we were able to predict a person's likelihood of having a disease, enabling an early diagnosis. A low validation loss number improves the model's performance. The model network's use of the ReLU activation function allowed it to focus on the most relevant information while ignoring the rest. The code is built in CUDA, a parallel computing environment and language that makes extensive use of the GPU to speed up calculations.

There are numerous restrictions and issues with deep learning technology, despite of its successes in the medical and therapeutic fields. When there is a large amount of annotated data, deep learning typically calls for analysis. This It's difficult to annotate medical records. Images. The specialised knowledge of radiologists is required for the labelling of medical images. Therefore, it takes time to annotate a suitable medical image. Moreover, it takes a lot of time. Medical image annotation is possible, but is currently impossible due to the sheer number of unlabeled medical images. Due to their extensive storage in PACS for a long time, there are a tremendous number of images. A considerable amount of time using Deep learning will be saved as an annotation if it can use unlabeled images and learning techniques.

REFERENCES

[1] Aydin, N., Cihan, Ç., Celik, O., Aslan, A. F., Odabas, A., Alatas, F., and Yıldırım, H. (2022). Segmentation of Acute Pulmonary Embolism in Computed Tomography Pulmonary Angiography Using the Deep Learning Method.

[2] Xu, H., Li, H., Xu, Q., Zhang, Z., Wang, P., Li, D., and Guo, L. (2023). Automatic detection of pulmonary embolism in computed tomography pulmonary angiography using scaled- YOLOv4. Medical Physics.

[3] Fayad, F. H., Zhong, Z., Sollee, J., Jiao, Z., Bai, H., and Atalay, M. (2022). Pulmonary Embolism Mortality Prediction With Deep Learning Based on Computed Tomographic Pulmonary Angiography and Clinical Data. Circulation, 146(Suppl1), A12497-A12497.

[4] Cheikh, A. B., Gorincour, G., Nivet, H., May, J., Seux, M., Calame, P., and Crombé, A. (2022). How artificial intelligence improves radiological interpretation in suspected pulmonary embolism. European Radiology, 32(9), 5831-5842.

[5] Vijayachitra, S., Prabhu, K., Abarana, M., Deepa, A., and Loga Priya, L. (2022). Deep Learning Technique-Based Pulmonary Embolism (PE) Diagnosis. In Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2021 (pp. 695-702). Singapore: Springer Nature Singapore.

[6] Yadlapalli, P., Teja, A. L., Raju, C. M. A., Reddy, K. S. S., Mithra, K., and Dokku, B. (2022, January). Segmentation of Pulmonary Embolism Using Deep Learning. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-5). IEEE.

[7] Ryan, L., Maharjan, J., Mataraso, S., Barnes, G., Hoffman, J., Mao, Q., ... and Das, R. (2022). Predicting pulmonary

embolism among hospitalized patients with machine learning algorithms. Pulmonary circulation, 12(1), e12013.
[8] Karthick Raghunath, K. M., et al. "Utilization of IoT-assisted computational strategies in wireless sensor networks for smart infrastructure management." International Journal of System Assurance Engineering and Management (2022): 1-7..
[9] Kolossváry, M., Raghu, V. K., Nagurney, J. T., Hoffmann, U., and Lu, M. T. (2023). Deep learning analysis of chest radiographs to triage patients with acute chest pain syndrome. Radiology, 221926.
[10] Adleberg, J., Wardeh, A., Doo, F. X., Marinelli, B., Cook, T. S., Mendelson, D. S., and Kagen, A. (2022). Predicting patient demographics from chest radiographs with deep learning. Journal of the American College of Radiology, 19(10), 1151-1161.
[11] Remy-Jardin, M., and Remy, J. (2022). Artificial Intelligence-Based Detection of Pulmonary Vascular Disease. Artificial Intelligence in Cardiothoracic Imaging, 491-500.
[12] Huhtanen, H., Nyman, M., Mohsen, T., Virkki, A., Karlsson, A., and Hirvonen, J. (2022). Automated detection of pulmonary embolism from CT-angiograms using deep learning. BMC Medical Imaging, 22(1), 43.
[13] Varshney, A., Bansal, A., Agarwal, A., Mishra, V. K., and Badal, T. (2022, February). A comparative study of deep learning models for detecting pulmonary embolism. In Advanced Computing: 11th International Conference, IACC 2021, Msida, Malta, December 18–19, 2021, Revised Selected Papers (pp. 82-98). Cham: Springer International Publishing.
[14] Vinothkumar, Veerasamy, et al. "Geraniol modulates cell proliferation, apoptosis, inflammation, and angiogenesis during 7, 12-dimethylbenz [a] anthracene-induced hamster buccal pouch carcinogenesis." Molecular and cellular biochemistry 369 (2012): 17-25.
[15] Long, K., Tang, L., Pu, X., Ren, Y., Zheng, M., Gao, L.,& Deng, F. (2021). Probability-based Mask R-CNN for pulmonary embolism detection. Neurocomputing, 422, 345-353.
[16] Guo, J., Liu, X., Chen, Y., Zhang, S., Tao, G., Yu, H.,& Wang, N. (2022, September). AANet: artery-aware network for pulmonary embolism detection in CTPA images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 473-483). Cham: Springer Nature Switzerland.
[17] Contreras-Luján, E. E., García-Guerrero, E. E., López-Bonilla, O. R., Tlelo-Cuautle, E., López-Mancilla, D., & Inzunza-González, E. (2022). Evaluation of machine learning algorithms for early diagnosis of deep venous thrombosis. Mathematical and Computational Applications, 27(2), 24.
[18] Ruggiero, A., & Screaton, N. J. (2017). Imaging of acute and chronic thromboembolic disease: state of the art. Clinical radiology, 72(5), 375-388.
[19] Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. Metabolism, 69, S36-S40.
[20] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M.& Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.
[21] Becattini, C., Agnelli, G., Germini, F., & Vedovati, M. C. (2014). Computed tomography to assess risk of death in acute pulmonary embolism: a meta-analysis. European Respiratory Journal, 43(6), 1678-1690.
[22] Devarajan, Devikanniga, et al. "Cervical cancer diagnosis using intelligent living behavior of artificial jellyfish optimized with artificial neural network." IEEE Access 10 (2022): 126957-126968..
[23] Avanija, J., et al. "Designing a Fuzzy Q-Learning Power Energy System Using Reinforcement Learning." International Journal of Fuzzy System Applications (IJFSA) 11.3 (2022): 1-12.
[24] Niu, S., Liu, Y., Wang, J., & Song, H. (2020). A decade survey of transfer learning (2010–2020). IEEE Transactions on Artificial Intelligence, 1(2), 151-166.
[25] Thaseen, Aliya, et al. "Breast Cancer Detection Using Deep Learning Model." Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022. Singapore: Springer Nature Singapore, 2023.
[26] Koti, Manjula Sanjay, et al. "Efficient Deep Learning Techniques for Security and Privacy in Industry." Cyber Security and Operations Management for Industry 4.0. CRC Press, 2022. 2-32.
[27] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019). On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265.
[28] Jia, Z., Lin, S., Gao, M., Zaharia, M., & Aiken, A. (2020). Improving the accuracy, scalability, and performance of graph neural networks with roc. Proceedings of Machine Learning and Systems, 2, 187-198.
[29] Ramakrishna, Mahesh Thyluru, et al. "Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data." Entropy 25.2 (2023): 245.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: May 29, 2023
Accepted: Oct 11, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 547–556, DOI 10.12694/scpe.v25i1.2359

MODELING A SMART IOT DEVICE FOR MONITORING INDOOR AND OUTDOOR ATMOSPHERIC POLLUTION

T JEMIMA JEBASEELI*, DEAGEON KIM† AND DONGOUN LEE‡

Abstract. Air pollution is caused by chemical, physical, or biological components alters the fundamental properties of the atmosphere, and contaminates the interior or outdoor settings. With the rapid industrialization activity in recent years, there is an urgent need to monitor air quality. The proposed research provides a mechanism for monitoring air pollution in indoor and outdoor environments. The system consists of an IoT-based Arduino device. The Arduino IDE connects the Temperature and Humidity sensor, Grove light sensor, and air quality sensor to measure the air pollutants such as Carbon dioxide CO_2, Nitrogen oxide (NOx), and Particulate Matter PM2.5. The sensors work efficiently and provide qualitative findings from the environment when they are exposed to CO_2, gasoline, solvents, thinner, formaldehyde, and other harmful chemicals. The Wi-Fi module of the Blues Wireless Notehub is used for secure data routing to the IoT cloud. The Air Quality Index (AQI) measures provide information on whether there is something unsafe in indoor and outdoor environments. For collecting and analyzing the device data, the Notecard is intended to be used with a cloud storage service. Also, the indoor fire detector identifies the incident of fire intimates the users through the alarm, and measures the indoor pollutant at that time. The proposed smart IoT product model would be an excellent device for air quality monitoring because of its long-term consistency and low electricity consumption.

Key words: air pollution; air quality; Arduino; sensors; cloud; humidity; smart IoT; pollutants; temperature.

1. Introduction. The air is polluted due to the development in technology and rapid increase in industry. The government initiated different plans to control this pollution by introducing a sustainable green environment. The air pollution in indoor and external surface areas causes around five million people to die. This affects the GDP of the country [1]. The main source of air pollution is harmful gases and dust produced in the industry, CO_2, Nox, PM 2.5, etc [2]. The ocean has been primarily responsible for more than 90% of global warming over the past fifty years. According to the new research, warming in the upper water is responsible for approximately 63% of all warming. Since pressure increases when heated, rising ocean water raises the global sea level [3]. The annual mean global temperature is predicted to be between 1.1°C and 1.7°C higher than before the levels between 2022 and 2026. The likelihood of global surface temperature reaching 1.5°Cover the pre-industrial level for the year, 2022-2026 is nearly equal to 48%. The five-year mean has a slight 10% probability of reaching this limit. The probability of at least one year topping the hottest year on record, 2016, is 93% in 2022 - 2026. There is a 93% chance that the five-year mean for 2022-2026 will be higher than the five-year mean for 2017-2021[4]. The global temperature of land and surface on January 2022 was 0.89°c ie) 1.60°F. Many accidents happen nowadays due to the occurrence of fire in the kitchen. The number of air pollutants increased due to fire. This fire is caused by indoor and outdoor incidents [5]. When the fire is noticed, then manually it is controlled. The linkage to the smartphone and computer devices is used by the systems to operate the network equipment and enhance identification. So that it could successfully perform pre-disaster identification and warning, it could assess exhaustive fire smoke, humidity, light, as well as other numerous related characteristics. The indoor and outdoor pollutants are considered for analyzing the air quality to model the final device. Flame sensors are vital tools for spotting fires and are employed in diverse application fields, including automotive, firefighter robotic arms, parking safety gear, factories, and commercial and residential cooking spaces [6].

The following are the proposed study's major contributions:

*Associate Professor, Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India (jemima_jeba@karunya.edu),

†Professor, Department of Architectural Engineering, Dongseo University, Republic of Korea. (Corresponding Author: deageonkk@gmail.com, gun43@hanmail.net)

‡Professor, Department of Architectural Engineering, Dongseo University, Republic of Korea (Corresponding Author: Idu21@dongseo.ac.kr)

1. Using smart IoT devices for precise monitoring of indoor air quality.
2. Using cloud computing for real-time indoor and outdoor air quality analysis.
3. Developed a mobile application for the proposed IoT system with any time, anywhere capabilities.
4. The device has been tested for data reliability, and the platform has been implemented in a building to demonstrate its viability.

2. Literature Review. Air pollution is a big problem, particularly in modern cities owing to increased mobility and urban population. It is notable and has an impact on people's health. As a result, there is a need to track and control pollution, as well as to establish a non-toxic atmosphere in which humans, wildlife, trees and plants may live a sustainable life. Thus government and the NGOs made huge efforts [7]. Particulate matter (PM) less than PM 2.5 is harmful to the human body. It penetrates deep into the heart, lungs, and respiratory system [8-9]. AQI is the measure to indicate the number of pollutants in the air. Nitrogen oxide is one of the pollutants that damage the respiratory system. According to the latest findings, the nitrogen oxide in the air is the reason for COVID-19 deaths [10].

Because students and teachers expel carbon dioxide (CO2) when they breathe, measuring the level of this gas is a reliable method of assessing whether the air in the classroom is clean. High CO2 levels increase the likelihood of inhaling air that has previously been exhaled by someone else. As a result, CO2 was regarded as a critical indicator for ventilating educational facilities over the COVID-19 pandemic [11]. Buildings consume energy and emit carbon dioxide all through every stage of their life cycle, such as construction, use, deconstruction, and decommissioning. The most energy is consumed during the operational and building phases, in that order. The energy needed for preserving a comfortable environment is included in the use or functioning phase, with the most pertinent being energy for running Heating Ventilation and Air Conditioning (HVAC) systems, domestic hot water, illumination, and home appliances.The difference in primary energy usage among these two phases was extensively investigated, with the utilization of phases consuming more [12]. To develop accurate forecasting techniques and a predictive AQI model for four different gases—carbon dioxide, sulphur dioxide, nitrogen dioxide, and atmospheric particulate matter—three artificial intelligence algorithms such as Linear Regression, Support Vector Regression, and the Gradient Boosted Decision Tree model and analyze air qualities using various sensors [13].

Many recent studies on air pollution have mostly focused on predicting air quality concentrations [14-15] and measuring air quality [16-19]. Forecasting air quality has been popular, and various approaches have been created for this purpose, including the statistical HMM (Hidden Markov model) [11], grey model first order one variable [12], SVM (Support Vector Machine) [13], and fuzzy time series forecasting model [14-15]. Burkart et al [20] examined the impacts of air pollution, secondhand smoking, and indoor environmental contamination. Meta-Regression-Bayesian, Regularized, the Trimmed (MR-BRT) method is used to calculate the risk of occurrence of diabetes due to this air pollutant. The monitoring system was developed using IoT technology by M.V.C.Caya et al. [21]. IoT systems contain disparate devices from many vendors, which might pose compatibility concerns. Hojaiji et al. [22] developed single sensor nodes to track air pollutants. These portable devices can detect one or two pollutants, but their accuracy is relatively low. Zho Z et al. [23] described that the low-cost sensors provide adequate accuracy of data and a respectable sensing range, but they cannot compete with the accuracy provided by traditional monitoring systems. Akashdeep Sharma et al. [20] proposed a multi-model IoT and deep learning-based identification, distribution, and continuous tracking systems for intense fire areas. Aditya et al. [24] developed a rule-based method to measure carbon monoxide for identifying air pollution. There is a need for a device to measure the atmospheric air quality to find the pollution in the air. The proposed system considers the existing system and its limitations as well as its challenges.

2.1. Challenges. Monitoring air pollution in both indoor and outdoor environments presents several challenges.

1. Indoor and outdoor air pollution are caused by a variety of factors, and the pollutant levels vary significantly over time and space. This makes identifying and quantifying all pollutants present and obtaining a precise assessment of air quality is difficult [21].
2. There are no universally recognized norms or recommendations for indoor air quality, and installing and operating outdoor air quality monitoring stations are costly. This reduces the amount of data to be collected to render and compare the data on air quality is challenging [22].

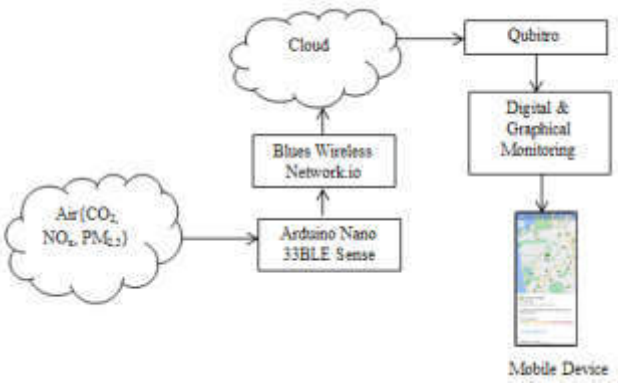


Fig. 3.1: The proposed smart IoT device to monitor indoor and outdoor atmospheric pollution

3. Some kinds of air pollutants, such as particulate matter and gaseous pollutants are difficult to monitor due to their small size [23].

3. Methodology. IoT is an efficient platform for handling scenarios without human intervention. The air is polluted due to indoor and outdoor pollutants. The smart IoT system collects information from the surroundings through various sensors, processes the huge data, and stores it in a cloud for historical review and usage. The proposed smart IoT system uses different sensors to analyze these pollutants and notify the dangerous situation where the pollutants are to be suppressed. There are several methods for tracking environmental data. An SD card in a controller is used for manually recording and transferring data from connected devices. This is to be handled by Wi-Fi or Bluetooth low-energy communication-based controllers. The cloud is a core part of every IoT system. The data collected by individual IoT sensors becomes helpful when combined with data from other relevant sensors. An approach to provide pervasive, effective, on-demand connectivity to the workstations, storage, frameworks, and capabilities of a massive pool of reconfigurable hardware and software is known as cloud computing. It takes little managerial effort or communication with service providers. A cloud system is expected to include features like on-demand consciousness, wide data services, virtualized resources, fast adaptability, and quantifiable services. As shown in Figure 3.1, the smart IoT system connects the APMS (Air Pollution Monitoring System) with Arduino Nano 33 BLE sense header. This smart IoT system can measure the number of pollutants present in the air. The Blue wireless notehub.io connects the devices with the cloud server. The observed data is stored over the cloud and Qubitro does the various statistical analyses of the data. The details of the pollutant are displayed on the user's mobile devices.

3.1. Indoor Pollutants. The indoor monitoring system is capable of identifying significant contaminants. Since these devices may be turned up immediately and the network life span can be prolonged as much as necessary, power is not a major problem for such devices. Carbon dioxide (CO_2) is an inert, tasteless, and colorless gas. Excessive exposures to CO_2 have significant health consequences. When large amounts of CO_2 are utilized, stored, or created, dangerously high concentrations may be the consequence. The harmful quantities of CO_2 are difficult to detect by the human eyes. CO_2 monitors are widely employed to limit the possibility of a CO_2 leak. A CO_2 sensor monitors the interior ambient levels fast and precisely, protecting employees, and facilities from injury or hazard. CO2 exposures have several negative health consequences. Migraine, nausea, agitation, numbness, or pins and needles sensation, chest tightness, restlessness, weariness, a faster heart rate, high blood pressure, unconsciousness, dyspnea, and tremors are some of the symptoms. Table 3.1 shows the standard Co2 measurements and their effects.

3.2. Models for the estimation of pollutants. Inside the building, the occurrence of fire produces hazardous carbon dioxide (Co2). Monitoring air pollution both indoors and outdoors with an Arduino UNO necessitates the use of a variety of sensors and components for measuring a few air quality parameters.

Table 3.1: Atmospheric Co_2 Pollutants level and its effects

Co2 Measurement (ppm)	Effects
400	Average atmospheric concentration
600-800	Indoor air quality acceptable
1,000-1,200	Good atmospheric concentration
5000	8-hour exposure period is the maximum allowed
6000-30,000	Concerning this exposure is just temporary
3-8 %	Increased breathing rate and headache
$\leq$ 10 %	Vomiting, nausea, and unconsciousness
$\geq$ 20%	Deaths due to sudden unconsciousness

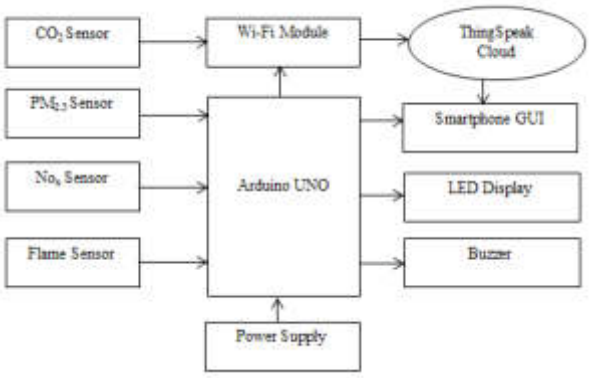


Fig. 3.2: Architectural diagram of the proposed IoT system for indoor air pollutant detection.

The Arduino UNO possesses six analog input pins that can be used to directly connect up to 6 sensors. After connecting the sensors to the Arduino UNO, the program will read the values from the sensors and show them on an LCD panel, send the data to a computer for processing and store it on the cloud, and trigger notifications or control other devices depending on the readings from the sensors.

These pollutants are present in the air for a long time after the incidents happened. Numerous call points and sensors are attached to the automatic fire alarm panels that make up the planned traditional smoke detector. The zonal circuit is used to connect a separate circuit for each level or fire section. There would be many zone bulbs on the fire alarm monitoring panel. The purpose of establishing zones is to provide an overall overview of where a flame has happened. The list of regions on a control panel, and hence the number of circuits wired inside the structure, determine the precision of understanding where a fire originated. The control panel then needs to be connected to at least two sounder circuits, which might include bells, a digital buzzer, and other such audible instruments. As illustrated in Figure 3.2, assemble the components and upload the program to Arduino. Hold a lighted lighter or a matchstick in front of the flame sensor to verify its operation.

The YG1006 NPN Photo Transistor is used in the flame sensor. This has been coated using black epoxy, which makes it capable of detecting infrared radiations, hence it is susceptible to infrared radiation in its wavelength range of 760nm to 1100nm. Using this type of flame sensor, it detects the Infrared Light from a distance of 100cm within its detection angles of 60 degrees.

The flame output frequency is 'high' within typical settings. Whenever the sensor senses fire, the output changes to 'low'. The buzzer is activated when Arduino detects a 'low' signal on its input pin.

3.3. Outdoor Pollutants. Nitrogen oxide (NOx) is a harmful gas molecule. Its chemical compounds of nitrogen and oxygen are an important element of air pollution. NOx pollutant is produced by cars and off-road

Table 3.2: Atmospheric NOx Pollutants concentration and AQI status

Nox Measurement ($\mu g/m^3$)	AQI Category
400+	Severe
281-400	Very Poor
181-280	Poor
81-180	Moderately polluted
41-80	Satisfactory
0-40	Good

Table 3.3: Atmospheric PM2.5 Pollutants concentration and AQI status

PM2.5 Measurement ($\mu g/m^3$)	AQI Category
35	Good
75	Moderate
115	Lightly polluted
150	Medially polluted
250	Heavily polluted
350	Severely polluted

vehicles like construction equipment, ships, and industrial emissions such as power stations, heating systems, cement factories, and windmills. NOx is commonly observed as a brownish smoke. Table 3.2 displays the baseline Nox measures as well as the AQI level.

3.4. PM 2.5. Particulate Matter 2.5 is the name given to tiny particles. They are minute granules or condensation in the atmosphere with a size of two and a half millimeters or less. According to the research, PM 2.5 levels of 12 g/m3 or even below are considered safe, with little to no risk of exposure. If the contribution makes or surpasses 35 g/m3 in 24 hours, the air is considered dangerous and may cause issues for people who already suffer from breathing difficulties. Recent epidemiological studies have linked PM exposure to impaired cognitive density, neurodegeneration, aberrant blood-brain barrier function, brain neuronal activation, and a higher likelihood of Alzheimer's disease, Parkinson's disease, and ischemic cerebral disease. Table 3.3 shows the standard PM 2.5 measurements and its AQI status.

3.5. Models for the estimation of pollutants. The outdoor pollutants are measured through the smart IoT system as shown in Figure 3.3. The device consists of the components of Arduino UNO, IIC/I2C/TWI/SPI serial interface board module for LCD, DHT11temperature and humidity sensor, sharp optical dust sensor (GP2Y1010AU0F), SD Card reader/writer module, RTC clock, and 16X2 LCD. The final display shows the temperature, humidity, time, and dust contents. AQI gives information about the risk level of pollutants in the air from 0 to 500. The proposed system is implemented through a three-step process.

1. It consists of the hardware setup, where the environmental data is collected using a fire sensor, air quality sensor, and light sensors. Arduino UNO connects the entire sensor. It is the controller to process all the data.
2. It includes the Notehub environmental setup. Here, the sensor data will be transferred to the Blues Notehub cloud through the Blues Notecard and Blues Notecarrier.
3. It is the environmental setup for Qubitro. The sensor data stored in the cloud is processed here. Hence, the sensor data is post-processed by Qubitro.

3.6. Implementation details of the proposed indoor and outdoor pollutant detection system.

1. Extract data of all pollutants from the sensors which are connected to the Arduino.
2. Based on its average concentration and the total number of pollutants, assign a risk category to each pollutant. R_1, R_2, ...R_n where 'n' is the number of pollutants.

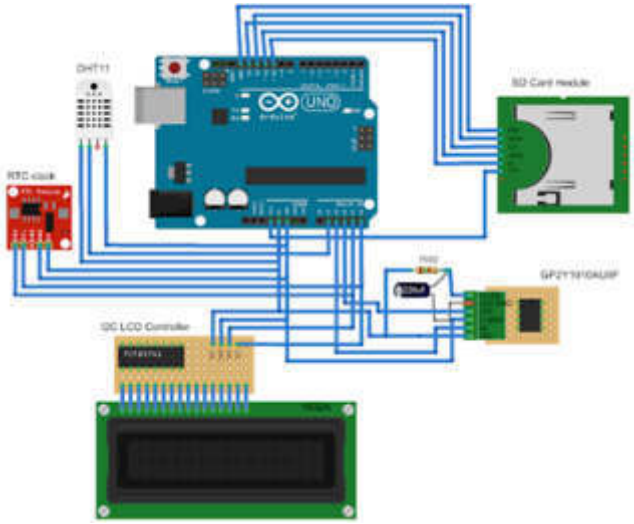


Fig. 3.3: Outdoor air pollutant monitoring system.

Table 3.4: Status of AQI related to health

AQI	Status	Color
0-50	Good	Green
51-100	Moderate	Yellow
101-150	Harmful	Orange
151-200	Unhealthy	Red
201-300	Very unhealthy	Purple
301-500	Hazardous	Maroon

3. Obtain the maximum value after sorting the risk groups in decreasing order as follows:

$$R(max), ..., R(min) \implies R(max)p \tag{3.1}$$

4. Calculate the air quality index is calculated as per the following equation.

$$I_{AQI} = \frac{I_{max} - I_{min}}{A_{upper} - A_{lower}}(AR_{upper} - AR_{lower}) + I_{min} \tag{3.2}$$

where I_{AQI} is the final AQI value of the location, I_{max} is the upper AQI limit, I_{min} is the lower AQI limit, A_{upper} is the upper ambient limit, A_{lower} is the lower ambient limit, AR_{upper} is the real ambient limit.

5. Display the AQI status based on the measured value as shown in Table 3.4.
6. Based on the AQI measures display the scale of pollutants and match them with their corresponding color scale value.

3.7. Pseudocode to measure the air quality. The pseudocode to measure the air quality of indoor and outdoor environments is described as follows.

1. Initialize CO_2 sensor, NOx sensor, and PM2.5 sensor.
2. Initialize variables to store sensor readings (CO_2 reading, NOx reading, PM2.5 reading).
3. Initialize variables for calibration of sensors.
4. LOOP:

5. # Collect sensor data
6. READ CO_2 sensor data
7. READ Nox sensor data
8. READ PM2.5 sensor data
9. # Apply calibration to convert raw sensor values to actual units
10. IF calibration is required for the CO_2 sensor:
11. APPLY CO_2 calibration
12. IF calibration is required for the NOx sensor:
13. APPLY NOx calibration
14. IF calibration is required for the PM2.5 sensor:
15. APPLY PM2.5 calibration
16. # Calculate the Air Quality Index (AQI) for each pollutant
17. CALCULATE CO_2 AQI
18. CALCULATE NOx AQI
19. CALCULATE PM2.5 AQI
20. # Store or transmit the data
21. STORE CO_2 reading, NOx reading, PM2.5 reading, and AQI values
22. locally or transmit to a cloud
23. # Delay for a specified interval (every 'n' minutes)
24. DELAY for a specified time interval
25. END LOOP

The proposed method produced a smart air pollution prototype that purified the air using a flame sensor, PM2.5 sensor, CO_2 sensors, NOx sensor, an Arduino UNO, and multiple filters. The filters are filtering particles ranging in size from about 100 to 0.1 μm. The quantity and concentration of gases, smoke, CO_2, PM2.5, and NOx were measured by the sensors before and following the filtering process. The PM2.5 sensor recognizes Particulate Matter in the 0.2 to 8 micron range. The outputs of flame sensors and PM sensors are converted into particle concentrations by Arduino UNO. The Wi-Fi Module transmits these findings to the ThingSpeak Cloud platform. The sensor outputs are viewed in the ThingSpeak Cloud on any computer or mobile device. The concentrations of various gases and the corresponding particle size concentrations are plotted before and following filtration, and the findings are analyzed. It was discovered that the intended filter successfully filters all particles larger than 0.1 μm. Both the air monitoring system and air purifier systems are tested for reliability and effectiveness.

4. Results and Discussion. The goal of real-time data gathering and analysis in this research is to capture the CO_2, Nox, and PM 2.5 concentrations utilizing sensors. The smart IoT sensor node senses the values from the environmental parameters. The microcontroller board contains a sensor to monitor the sensed values from the sensor nodes. Then the data is transmitted through the Internet and stored in the cloud server. The sensors mounted on the node station send the collection to the online service, and the outcome of the process of determining air quality will be published on the Android devices as real-time information about the pollution. The sample collected input data are shown in Table 4.1.

Figure 4.1 shows the smart web-based interface through which one can know about the environmental air pollutants, their corresponding air quality index, and the status of the particular location on their mobile devices. Based on this information the public and the government may take necessary action to improve the quality of the air. The sensor started observing the pollutants and read the details, deciding to publish the air quality status as shown in Table 4.2.

The proposed system's primary functions include measuring AQI and converting the results into a format that is simple for any user to understand. The proposed technique will assess the aggregate AQI while accounting for the concentrations of all air contaminants. The smart IoT system will then be converted into a color scale to comprehend the findings and gauge the risk of entering a certain region due to air quality.

As shown in Figure 4.2, the proposed system identifies the sensor values accurately while compared with the competitive methods. Also, the AQI value of each location is verified with the manually calculated measurements.

Table 4.1: The sample observed data from the resources (June 2023 to August 2023) on air quality

Id	Date	Day	NOx	CO_2	PM2.5
1	1/7/23 1:00 AM	Saturday	0.02	0.19	7.2
5	2/7/23 1:00 AM	Sunday	0.015	0.16	6.4
241	12/8/23 1:00 PM	Saturday	0.009	0.1	9.2
246	12/8/23 2:00 AM	Saturday	0.015	0.16	6.2
248	12/8/23 3:00 PM	Saturday	0.018	0.16	3.5
...	.....	.....	....	....	....
56950	24/8/2023 3:00 AM	Thursday	0.024	0.08	7.2

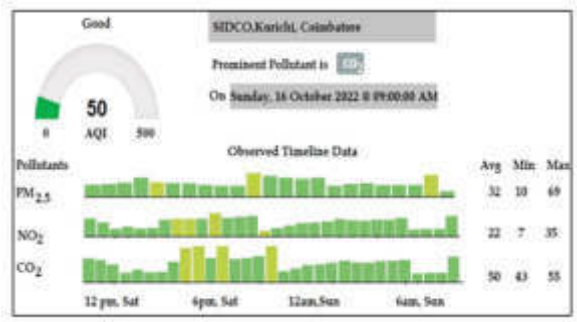


Fig. 4.1: Smart web-based IoT system interfaces to analyze the real-time air pollutants, air quality index, and status.

Table 4.2: Obtained results from the IoT sensors

AQ Sensor Value	16
Pollution Status	Fresh Air
Humidity	73 %
Temperature	28.6°C
Light Sensor	64 %
Dry Level	73 %

5. Conclusions. Air pollution is a huge danger to any country's health, economy, and biodiversity. The proposed smart IoT system discusses the causes and effects of air pollution, as well as a complete examination of air quality monitoring and an effective IoT-based method for air quality index measurement. Despite the existence of several outstanding smart air quality monitoring devices, the research topic remains challenging. This smart IoT-based AMPS model is an attempt to create smart, low-powered, and highly efficient air quality monitoring devices that can monitor continuously and send alerts or notifications about indoor and outdoor air pollution to the relevant people for further processing.

Acknowledgment. This work was supported by Dongseo University, "Dongseo Cluster Project" Research of 2023 (DSU-20230006).

REFERENCES

[1] HaiAnh H, *Does GDP growth necessitate environmental degradation*, 2020.
[2] Air pollution, Available: https://www.who.int/health-topics/air-pollution/#tab=tab_1

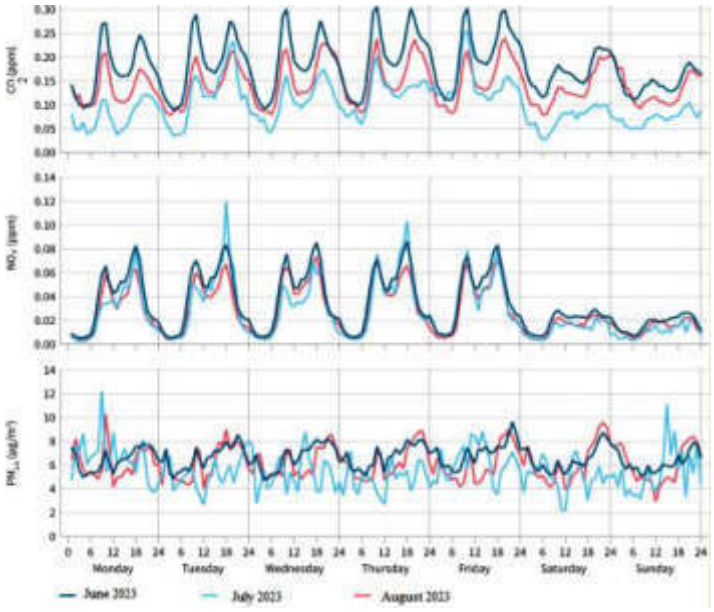


Fig. 4.2: The generated report from July 2023 to August 2023 on air quality based on the proposed IoT system

[3] Climate Change, Ocean Heat Content, Available: https://www.climate.gov/news-features/understanding-climate/climate-change-ocean-heat-content

[4] Global climate report, Available: https://www.ncei.noaa.gov/

[5] Akashdeep Sharma et al, *IoT and deep learning-inspired multi-model framework for monitoring Active Fire Locations in Agricultural Activities*, Computers and Electrical Engineering, vol. 93, 2021.

[6] Idrees Z and Zheng L, *Low cost air pollution monitoring systems: A review of protocols and enabling technologies*, Journal of Industrial Information Integration, vol. 17, 2020.

[7] Wenming Zhao et al, *Design of low-energy buildings in densely populated urban areas based on IoT*, Energy Reports, vol. 8, 2022.

[8] Jhanvi Aroraa, Utkarsh Pandya, Saloni Shaha, and Nishant Doshia, *Survey- Pollution Monitoring using IoT*, Procedia Computer Science, vol. 155, pp. 710-715, 2019.

[9] Air pollution linked with higher Covid-19 death rate: Study, 2020. Available:https://economictimes.indiatimes.com/news/environment/pollution/air-pollution-linked-with-higher-covid-19-death-rate-study/articleshow/75249602.cms

[10] Sun W et al, *Prediction of 24-hour-average PM 2.5 concentrations using a hidden Markov model with different mission distributions in Northern California*, Sci. Total Environ, vol. 443, 2013.

[11] Tabuenca et al, *Affordances and core functions of smart learning environments: A systematic literature review*, IEEE Transactions on Learning Technologies, vol. 14, no. 2, pp.129-145, 2021.

[12] García-Monge M et al, *Is IoT monitoring key to improve building energy efficiency? Case study of a smart campus in Spain*, Energy and Buildings, vol. 285, 2023.

[13] Almalawi A et al, *An IoT based system for magnify air pollution monitoring and prognosis using hybrid artificial intelligence technique*, Environmental Research, vol. 206, 2022.

[14] Dincer N.G and Akkus O, *A new fuzzytime series model based on robust clustering for forecasting of air pollution*, Ecological Informatics, vol. 43, pp. 157–164, 2018.

[15] S. Suryono et al, *A web-based wireless sensor system to measure carbon monoxide concentration*, 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), pp. 1-5,2017.

[16] Nastro V et al, *Passive and active methods for Radon pollution measurements in historical heritage buildings*, Measurement, vol. 114, pp. 526-533, 2018.

[17] Sammarcoa M et al, *Using geosocial search for urban airpollution monitoring*, Pervasive and Mobile Computing, vol. 35, pp. 15-31, 2017.

[18] Noblesa C.J et al, *Ambient air pollution and semen quality*, Environmental Research, vol. 163, pp. 228–236, 2018.

[19] Burkart K et al, *Estimates, trends, and drivers of the global burden of type 2 diabetes attributable to PM 2.5 air pollution, 1990–2019: an analysis of data from the Global Burden of Disease Study 2019*, The Lancet Planetary Health, vol. 6, no. 7, 2022.

[20] M. V. C. Caya et al, *Air pollution and particulate matter detector using raspberry Pi with IoT based notification*, 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, pp. 1-4, 2017.

[21] Hojaiji et al, *Design and calibration of a wearable and wireless research grade air quality monitoring system for real-time data collection*, GHTC 2017, IEEE Global Humanitarian Technology Conference, pp.1 - 10, 2017.
[22] Zho Z et al, *Edge computing based IoT architecture for low cost air pollution monitoring systems: a comprehensive system analysis, design considerations & development*, Sensors, vol. 3021, pp.1 – 23, 2018.
[23] Aditya Akbar Riadi and Muhammad Imam Ghozali, *Web Based Information System of Carbon Monoxide Pollution Wibowo Harry Sugiharto*, E3S Web of conferences, vol. 73, no.1, 2018, 10.1051/ e3sconf/20187305026.
[24] R. Senthilkumar et al *Intelligent based novel embedded system based IoT enabled air pollution monitoring system*, Microprocessors and Microsystems, vol.77, 2020,

Edited by: Mahesh T R
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Jun 22, 2023
Accepted: Oct 5, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 557–563, DOI 10.12694/scpe.v25i1.2397

AN IMPROVED COVERAGE HOLE FINDING SYSTEM FOR CRITICAL APPLICATIONS BASED ON COMPUTATIONAL GEOMETRIC TECHNIQUES

ANITHA CHRISTY ANGELIN* AND SALAJA SILAS†

Abstract. Wireless Sensor Networks (WSNs) contain coverage holes caused by both random node sensor deployment and malfunctioning nodes. Because fixing the battery is challenging, collaborative discovery and assessment of coverage shortfalls, as well as getting rid of these holes, has been recognized as critical in WSNs. While placing nodes for sensors in a large-scale WSN is challenging. This research provides a cost-effective coverage hole detection approach based on collaborative distributed point placement. Create a polygon first by employing an angle estimate approach and neighbor data. Following that, a based on points hole identification technique is used to assess if a coverage issue appears in a large-scale WSN's supplied ROI. Furthermore, the region of the coverage hole is estimated using computational geometry-based polygonal triangulation methods. The accuracy of the method is tested here using statistical data. The results show that it outperforms earlier coverage hole-detecting algorithms. In particular, the method improves coverage rate by 75% when compared to conventional methodologies. It also lowers energy usage by 90%, adding to increased network lifetime. The quantitative favourable results demonstrate the effectiveness of the collaborative distributed point placement technique in detecting and successfully resolving coverage gaps in WSNs. In regards to coverage rate, energy consumption, and network longevity, the system being proposed beats previous coverage hole-detecting techniques.

Key words: Coverage hole; Computational Geometry; Visibility approximation; Triangulation; Energy efficiency.

1. Introduction. Coverage hole identification is required for essential WSN-based applications because of failures of nodes and haphazard implementation [1]. The operation of WSN is impacted by energy scarcity in terms of transmission range, regular or haphazard placement of sensors, localization, planning, coverage, as well as additional issues. Intruder identification in warfare, disaster assistance, medical, and workplace monitoring all employ sensor data [2-5]. Data will be lost, or propagation will be delayed if a node breaks while transmitting. Thus, enhancing coverage requires recognizing coverage holes [6-7]. Probabilistic and computational geometry-based algorithms were used for sensor network coverage hole detection depending on system boundaries and data limitations. By detecting holes in coverage hole detection using node location data. From various studies, It is evident that computational geometry-based algorithms function better than empirical ones [8-9]. This collaborative approach uses computational geometry to classify coverage holes in wireless sensor networks. This method is based on distributed energy-efficient point location-based coverage hole identification. A visibility approximation technique is used to establish where a coverage hole stops, and adjacent nodes begin. Considerations include crossover, sensor categories, and node locations [10]. The coverage hole is identified using a polygon triangulation technique and the one-hop neighbors of the coverage hole nodes. This approach is more energy-efficient.

2. Literature Review. Coverage hole detection strategies based on probabilistic and computational geometry methodologies are surveyed in the literature. Node density calculations using probabilistic approaches for great coverage do not sufficiently support hole identification techniques [11-12]. Each sensor node makes dispersed decisions. The probabilistic technique requires node density but less location-based data [13]. Coverage issues are now undetectable. Computational geometry finds coverage holes [14-15]. The tree-based coverage hole detection method for detecting, shaping, and sizing coverage holes [16]. Coverage holes, on the other hand, are recognized by the relative location of surrounding nodes. The implementation of a sensor network border

*Research Scholar, Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India. (anithachristyangelin1p@gmail.com)
†Professor, Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India. (blessingsalaja@gmail.com)

node structure technique, although higher-density nodes are needed to guess coverage holes [17]. The Distributed Boundary Detection based on the Connected Independent Set (BDCIS) approach precisely recognizes borders and holes. This strategy proposes that nodes aggregate their one-hop neighbor's connection details and create distinct data categories [18-19].

Energy use is substantial while precision is low when compared to other approaches. The distributed virtual force-based hole identification and healing proposed by Zhou et al. [14] reduces unnecessary nodes and holes [20]. The WSNs coverage hole is discovered using a computational geometric analytical approach. The algorithm's network lifetime is shorter than that of others. Se Hang et al. [30] proposed FD-CT (Force-Directed Contour Tracing), a method that uses force-directed algorithms and contour tracing techniques to identify holes in wireless sensor networks without the use of location data or anchor points. The method, however, has problems with false hole detection. Tapas and colleagues [31] proposed using convex hull methods to calculate the circumference of both stationary and portable wireless sensor networks (WSNs). When a border node within a stationary WSN needs to be replaced, this method chooses a new neighbouring node to maximize total coverage along the network's boundary.

Khedr et al. [22] propose classification-distributed Distributed Coverage Hole Detection (DCHD). Unbounded coverage gaps were contained using the WSN algorithm. The approach includes node placement, sensing coverage overlap, and non-overlapping range. Its downside is the high computational overhead. Z. Kang et al. developed a distributed technique that is connectivity-based and coordinate-free [23]. BCPs are used in the technique, which can be done on a single node through neighbor verification. This algorithm is highly difficult and computationally complex. Computational Geometric techniques let neighbors understand their respective positions need localization, the number of nodes, the total number of holes [24-25], hole size, the number of network messages exchanged, the median node degrees, and other factors that influence boundaries identification time, consumption of energy, the precision of detection, and communication costs [26-27]. For such massive networks, distributed protocols assure reliability. Existing methods have several drawbacks, the most notable of which are the high number of messages exchanged the length of time it takes to make a selection and the fact that numerous internal nodes are incorrectly identified as boundary nodes [28-29]. The distributed method is connectivity-based and coordinate-free. BCPs are used in the technique, which can be done on a single node through neighbor verification. The algorithm is sophisticated. In geometric approaches, neighboring nodes can share information to understand the relative coordinates of neighboring nodes.

3. Methodologies. The proposed framework allows the network's nodes that sense to be placed on a 2-D grid. Here, the main nodes inside the goal region are consistently positioned, while border nodes are evenly scattered beyond the intended area's exterior boundaries. Every node lacks precise location data, and the node could be classified as a node within the system or doesn't rely on a starting context. Considering the following circumstances: R_s is the sensing range's radius; R_c does a sensor node's communication range have a radius that allows for $R_c = 2R_s$.

3.1. System model. Every node in the network's hierarchy has a binary sensing design that uses a unique ID that is unique to that node. In Figure 3.1, the Point location-oriented coverage hole detection architecture comprises a Geometrical Visualisation Unit (GVU) for visible estimation to recognize the neighbor node and a minimal cost triangulation approach for finding the polygon's border nodes [18]. The Hole Detection Unit (HDU) is made up of an exact position-based hole detection method that detects the position of the failing node that caused the amount of coverage hole. The hole-related data that the hole detection system evaluates is saved in the hole data repository for use later on.

3.2. Visibility approximation algorithm.

Procedure.

Input: Set of points S enclosing the sensor nodes.

Output: Polygon P with boundary points.

1. Initialize an empty set S enclosing to store the sensor nodes defining the boundary.
2. Initialize a variable m to 0.
   Repeat the following steps:
3. Set E[m] to the current position.

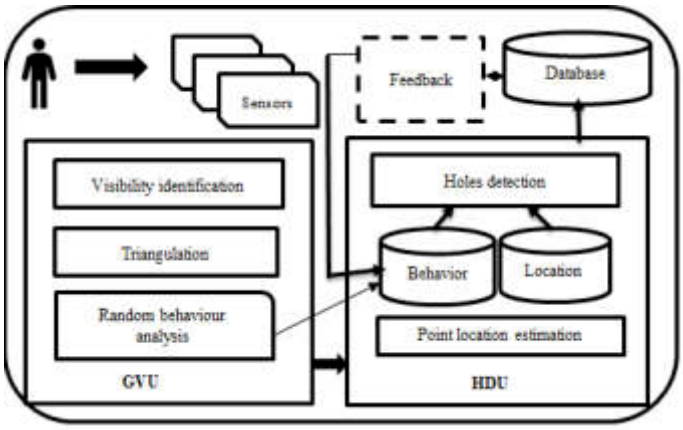


Fig. 3.1: DPHD framework

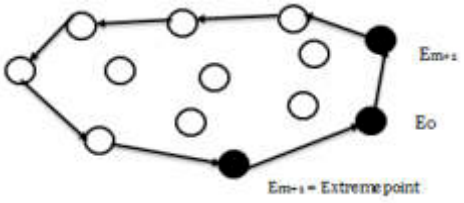


Fig. 3.2: Visibility approximation algorithm

4. Initialize a variable ending position to the position of the first sensor node in set S.
5. For each sensor node S[n] in set S:If ending position is equal to the position of S[n], update ending position to the position of S[n].
6. Increment m by 1.
7. Update the current position to ending position.
8. Repeat the loop until ending position is equal to E[0].

This method processes points in the order they were received. As shown in Figure 3.2, the approximation of visibility begins with m=0 and a polygon point E=0. Take the left-most location and select E_{m+1} so that every point is to the opposite side of the path E_{m+1}. Following that, as illustrated in Figure 5.1, modify the polygon feature set and continue till they hit the left edge. In O(n^1) steps, all polar angles define this point in space as E_m in the polar coordinate centre. The loop within the loop keeps track of every point in the set S, while the outermost loop keeps track of each position in the polygon. As a result, the total run time is O(nh).

3.3. Point location estimation algorithm. Figure 3.2 shows the point location estimation algorithm used to locate the failure node from the monotonous triangle T. Slabs divide monotonous polygons. A slab's S-side is between two successive segments. Find the zone that contains a certain location whenever non-cross segments cross the surface from left to right. Determine whichever vertical block has a sensor node as the horizontal surface splits into vertical slabs to facilitate point localization. Determine the extent of a sensor location when the surface is divided into non-intersecting portions which run from left to right. The method allows for point localization in exponential time. It is simple to implement since each slab may cover a significant portion of its sections. The area required to build the slabs and the region inside the slabs may be as huge as O(n2). The polygon construction process is critical in identifying the coverage hole borders. The time complexity O(n log n) shows that it scales well with the total number of sensor nodes. The memory need for maintaining intermediate information during polygon building is illustrated by the space complexity of O(n).

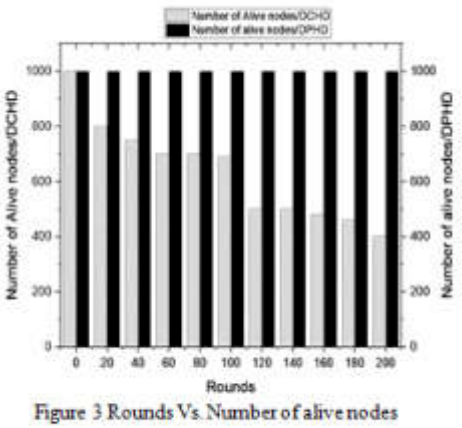


Figure 3 Rounds Vs. Number of alive nodes

Table 1 Number of Alive nodes in DPHD Vs. DCHD

Rounds	Number of Alive nodes /DCHD	Number of alive nodes /DPHD
0	1000	1000
20	800	1000
40	750	1000
60	700	1000
80	700	1000
100	690	1000
120	500	1000
140	500	1000
160	480	1000
180	460	1000
200	400	1000

Fig. 5.1: Modify the polygon feature set

3.4. Coverage hole identification. The hole regions are computed by choosing the nodes in the sensor network that are nearest to the sector dimensions and determining the exact location of the idle sensor node inside the specified segment. The hole size is estimated employing a hole detection technique in which the node power is contrasted to the threshold value obtained through a probabilistic methodology. Nodes having a lower energy consumption below the acceptable value are detected as failing nodes that cause coverage gaps. The hole detection step entails determining whether or not coverage holes exist in the region of interest. The algorithm's time complexity of $O(n^2)$ indicates that the time required for execution rises exponentially with a given number of sensor nodes. The memory overhead during hole detection is shown by the space complexity of O(n).

4. Experimental Setup. As it is expected that the transmitter and receiver distance is twice that provided by the sensors, a 40-meter range is employed in the resultant experimental case. The specs of the model are based on standard IEEE 802.15.4 MAC and PHY criteria. The baseline energy budget for the sensor node is 50 J, and 0.14 J of energy is lost due to the transmission of each control packet. To facilitate the formation of unique clusters of unconnected nodes, a hole is arbitrarily generated between the one-hop and entirely attached node. Every node receives a set of parameters to locate the sensor array of its failing neighbor to obtain information about the other node's choices via its one-hop neighbors.

5. Results and Discussions. Simulation findings reveal that the proposed method for DPHD has a lower node mortality rate than traditional protocols. The proposed method DPHD attempts to automatically alter the status of the network nodes to maintain the entire network as operational as feasible. As a result, the average lifespan of nodes inside the designed algorithm is quite sensitive.

5.1. Number of alive nodes. Figure 5.1 depicts the evaluation of living nodes for multiple rounds of transmission of packets while using the coverage hole identification technique. According to the basic fitting study, the proposed approach fits perfectly when contrasted with the existing distributed coverage hole detection technique. Figure 5.2 depicts a basic fitting evaluation that describes the variation in the aggregate amount of living nodes.

The mean model's quantitative element is R-Squared. The closer R-squared is to one, the more accurately it depicts the range of responses that lie around the median. As a result, higher values for R-squared suggest better predictions. Adjusted R-squared will always be smaller than R-squared, however, it is generally fairly small when assessing inadequate test coefficients amid excessive noise. The R-Squared value for DPHD is one,

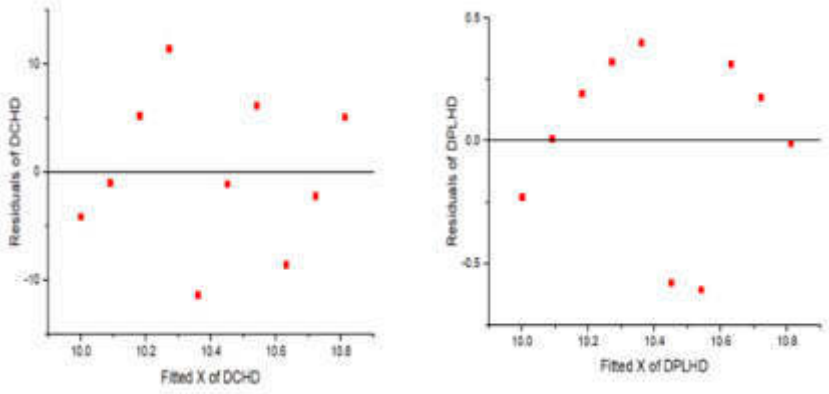


Fig. 5.2: Simple fitting analysis for Rounds Vs Number of alive nodes

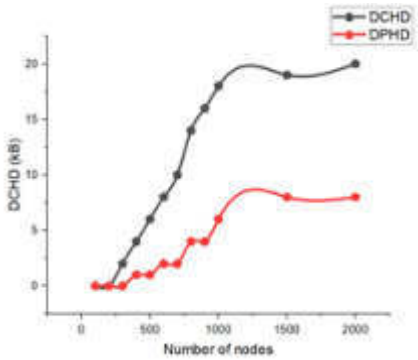


Table 2 Analysis of Control Packet Overhead for DCHD and DPHD

Plot	DCHD	DPHD
Number of Points	12	12
Degrees of Freedom	10	10
Residual Sum of Squares	105.855	8.659
R-Square	0.8282	0.9116
Adj. R-Square	0.81105	0.9028

Fig. 5.3: Control packet overhead VS number of nodes

which is larger than the conventional DCHD technique [25].

5.2. Control packet overhead. Figure 5.3 shows the control portion of the overlay packet, which defines the path between the sender and the recipient as well as the total number of application-specific information bytes transferred. DPHD control packet overhead is compared to DHCD for various node counts [25]. Our protocol outperforms DHCD with fewer availability holes. The simulation reveals no control packet overhead for sensors under 100. When the number of sensors exceeds 200, sensor density increases and control packets can lead to more redundant sensors on the network.

5.3. Average energy consumption. Figure 5.4 depicts a study of the median energy usage for numerous holes with multiple sensors. Let n be the count of neighbors of a sensor Si, and Et and Er denote the amount of energy expended by the sensor when sending and receiving data from its neighbors. Because node Si must broadcast its position data as well as collect location data from its k neighbors, a node's energy usage in the DPHD method may be E = Et + nEr. As seen in Figure 6, the overall energy usage is influenced by the variety of sensors put in the location. The corresponding values are shown in Table 3. This approach employs a huge number of sensor nodes with a variety of holes. Each additional pair of holes increases the amount of energy consumed. This condition is due to the increased number of holes between neighbors that include covering holes.

Descriptive statistics are used to characterize the energy usage data in Figure 5.3. The sample and observations are straightforward. The proposed framework is the foundation for almost every aspect of quantitative analysis as well as basic visual analysis. The corresponding values are shown in Table 2.

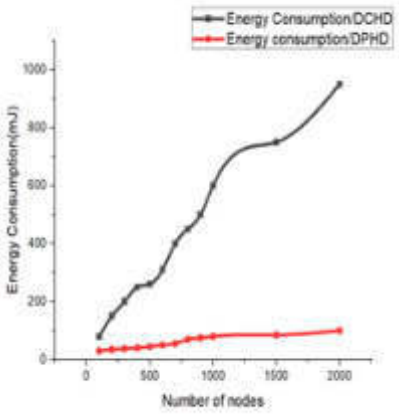


Table 3 Analysis of Energy Consumption for DCHD and DPHD

Equation	y = a + b*x	
Plot	Energy Consumption/DCHD	Energy consumption/DPHD
Weight	No Weighting	
Intercept	60.42289 ± 16.54572	28.99751 ± 3.44326
Slope	0.46388 ± 0.01803	0.03945 ± 0.00375
Residual Sum of Squares	10896.21891	471.89502
R-Square (COD)	0.98511	0.91699
Adj. R-Square	0.98362	0.90869

Fig. 5.4: Average energy usage against node count

5.4. Complexity analysis. The visibility estimation phase simply eliminates the ability to identify nodes and edges that are associated with boundaries. As a result, the maximum computing complexity of this phase is O (nh). To avoid network costs resulting from signal transport among sensor nodes, the phase of hole identification uses just locally stored data to compute the minimal crucial thresholds, hence the complexity of its computation is O(1). Both the triangulation and point position estimate stages continuously decide the result based on the number of nodes, however, a little increases the level of difficulty that can be lowered utilizing later approaches.

6. Conclusions. The detection of coverage holes supports the idea that there is a greater need for mission-critical software to locate coverage holes. The reclamation of holes and the optimization of the recovery following the appearance of holes is another possible area of research that might be conducted in the future. To prevent hole formation, movable sinks or numerous sinks can only make use of a very limited number of ways. If the mobile sink is nearby, nodes will transmit their data to it, preventing an excessive quantity of wasted energy supply during multi-hop delivery. Managing protocols that were developed specifically for these intricate networks is necessary to provide a flexible infrastructure. In comparison to probabilistic methods, geometrical methods require a greater amount of resources while finding holes and extracting data, because they have a greater number of nodes than probabilistic methods. Effective hole healing approaches on the other end of the spectrum, might be used to enhance coverage hole detection concerns.

REFERENCES

[1] M. Verma and S. Sharma, "A Greedy Approach for Coverage Hole Detection and Restoration in Wireless Sensor Networks," Wirel. Pers. Commun., vol. 101, no. 1, 2018.
[2] J. Park, S. Lee, and S. Yoo, "Time slot assignment for converge cast in wireless sensor networks," J. Parallel Distrib. Comput., vol. 83, 2015.
[3] G. Hua, Y. X. Li, and X. M. Yan, "Research on the wireless sensor networks applied in the battlefield situation awareness system," in Communications in Computer and Information Science, 2011, vol. 144 CCIS, no. PART 2.
[4] Di. Pant, S. Verma, and P. Dhuliya, "A study on disaster detection and management using WSN in Himalayan region of Uttarakhand," in Proceedings - 2017 3rd International Conference on Advances in Computing, Communication and Automation (Fall), ICACCA 2017, 2018, vol. 2018-January.
[5] P. Kumar and H. J. Lee, "Security issues in healthcare applications using wireless medical sensor networks: A survey," Sensors, vol. 12, no. 1. 2012.
[6] A. Saipulla, C. Westphal, B. Liu, and J. Wang, "Barrier coverage with line-based deployed mobile sensors," Ad Hoc Networks, vol. 11, no. 4, 2013.
[7] W. Abdellatief, H. Abdelkader, and M. Hadhoud, "An energy-efficient coverage hole detection technique for randomly deployed

wireless sensor networks," in Proceedings of 2016 11th International Conference on Computer Engineering and Systems, ICCES 2016, 2017.
[8] X. Feng, X. Zhang, J. Zhang, and A. A. Muhdhar, "A coverage hole detection and repair algorithm in wireless sensor networks," Cluster Comput., vol. 22, 2019.
[9] B. Khalifa, Z. Al Aghbari, A. M. Khedr, and J. H. Abawajy, "Coverage Hole Repair in WSNs Using Cascaded Neighbor Intervention," IEEE Sens. J., vol. 17, no. 21, 2017.
[10] M. A. Razzaque and S. Dobson, "Energy-efficient sensing in wireless sensor networks using compressed sensing," Sensors (Switzerland), vol. 14, no. 2. 2014.
[11] P. Chanak, I. Banerjee, and H. Rahaman, "Load management scheme for energy holes reduction in wireless sensor networks," Comput. Electr. Eng., vol. 48, 2015.
[12] H. P. Gupta, S. V. Rao, A. K. Yadav, and T. Dutta, "Geographic routing in clustered wireless sensor networks among obstacles," IEEE Sens. J., vol. 15, no. 5, 2015.
[13] P. Si, C. Wu, Y. Zhang, Z. Xia, and L. Cheng, "A hole detection and redeployment strategy in wireless sensor network," J. Inf. Comput. Sci., vol. 8, no. 13, 2011.
[14] S. M. Koriem and M. A. Bayoumi, "Detecting and measuring holes in Wireless Sensor Network," J. King Saud Univ. - Comput. Inf. Sci., 2018.
[15] T. Amgoth and P. K. Jana, "Coverage hole detection and restoration algorithm for wireless sensor networks," Peer-to-Peer Netw. Appl., vol. 10, no. 1, 2017.
[16] W. Li and Y. Wu, "Tree-based coverage hole detection and healing method in wireless sensor networks," Comput. Networks, vol. 103, 2016.
[17] S. P. Fekete, A. Kröller, D. Pfisterer, S. Fischer, and C. Buschmann, "Neighborhood based topology recognition in sensor networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3121, 2004.
[18] R. Beghdad and A. Lamraoui, "Boundary and holes recognition in wireless sensor networks," J. Innov. Digit. Ecosyst., vol. 3, no. 1, 2016.
[19] S. Lee, M. Younis, and M. Lee, "Connectivity restoration in a partitioned wireless sensor network with assured fault tolerance," Ad Hoc Networks, vol. 24, no. PA, 2015.
[20] P. K. Sahoo, M. J. Chiang, and S. L. Wu, "An efficient distributed coverage hole detection protocol for wireless sensor networks," Sensors (Switzerland), vol. 16, no. 3, 2016.
[21] J. Wang, C. Ju, H. jin Kim, R. S. Sherratt, and S. Lee, "A mobile assisted coverage hole patching scheme based on particle swarm optimization for WSNs," Cluster Comput., vol. 22, 2019.
[22] A. M. Khedr, W. Osamy, and A. Salim, "Distributed coverage hole detection and recovery scheme for heterogeneous wireless sensor networks," Comput. Commun., vol. 124, 2018.
[23] Z. Kang, H. Yu, and Q. Xiong, "Detection and recovery of coverage holes in wireless sensor networks," J. Networks, vol. 8, no. 4, 2013.
[24] C. Qiu, H. Shen, and K. Chen, "An Energy-Efficient and Distributed Cooperation Mechanism for k-Coverage Hole Detection and Healing in WSNs," IEEE Trans. Mob. Comput., vol. 17, no. 6, 2018.
[25] J. Zhang, H. Chu, and X. Feng, "Efficient Coverage Hole Detection Algorithm Based on the Simplified Rips Complex in Wireless Sensor Networks," J. Sensors, 2020.

Edited by: Mahesh T R
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Jul 4, 2023
Accepted: Oct 2, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 565–576, DOI 10.12694/scpe.v25i1.2400

SECURE STEGANOGRAPHY MODEL OVER CLOUD ENVIRONMENT USING ADAPTIVE ABC AND OPTIMUM PIXEL ADJUSTMENT ALGORITHM

SE-JUNG LIM* AND AMBIKA UMASHETTY†

Abstract. An effective security model with low computational complexity, minimal quality-compromised, and improved security robustness is essential due to the rapid expansion of multimedia data communication through different cloud services. An image steganography model has been proposed for the security of multimedia data in an unreliable cloud environment. This study's main goal is to provide efficient steganography with barely hidden information. A secure data communication model has been developed over a cloud environment using the Adaptive Artificial Bee Colony Algorithm and the Optimum Pixel Adjustment Algorithm Based Image Steganography Method. An adaptive ABC (artificial bee colony) technique is applied to select the optimal pixel positions in the cover image because the goal of this investigation is to increase PSNR. The Optimal Pixel Adjustment approach is used to minimise embedding errors while maintaining the stego image's appearance identical to the cover image after embedding. The MATLAB platform is used to implement the proposed method. The results show that the proposed AABC-based OPA is more efficient across all measures investigated during the embedding and extraction processes.

Key words: Optimal Pixel Adjustment, Integer Wavelet Transform, Adaptive Artificial Bee Colony, Steganography, Uncertain conditions, Cloud environment.

1. Introduction. As the internet evolves, so does the prevalence of online crime. Today, information security is a serious concern, and steganography offers a number of solutions. Steganography is the practise of hiding information under a cover image [1-2]. To prevent the revealing of the secret image, the originality of the cover image must be retained even after the information has been inserted. Four categories of steganography are capable of sustaining the preceding standards, including a) Image steganography, in which the secret message is kept hidden within the image and no change in image quality is noticeable, b) Audio steganography, in which the secret message is kept hidden within the audio and no change in audio quality is perceptible. With c) Text steganography, the secret message is inserted without changing the text's meaning, and with d) Video steganography, the hidden message is inserted without affecting the video's quality [3-4].

The secret information or message hidden under a cover image should be recovered without damaging the integrity of the original image in audio, text, video, and other formats [5-6]. Steganography differs from cryptography in that it hides the message's actual presence while hiding its contents [7-8]. Steganography has been chosen over cryptography because it prevents a third party from reading the message if they get their hands on it. Steganographic messages, as opposed to cryptographic messages, doesn't attract the attention of a third party. In this aspect, steganography has an advantage over cryptography because it provides both security and encryption. Steganography, which hides data, and cryptography, which safeguards data, are contrary to one another.

In steganography, it might be difficult to recover data without a known technique because of invisibility or hidden components [9]. The image distortion seen in the stego image is the result of inserting hidden messages in to cover image [10-11]. There are two drawbacks in this; a) Because the size of the cover image is fixed, adding additional messages may cause image distortion; therefore, a compromise must be made between the adding capacity offered in any given cover image. b) The second drawback is that the stego image has some very slight distortions that interfere with the feature of the cover image.

Steganography can be divided into the following categories: (a) Spatial domain, which primarily consists of Least Significant Bit (LSB) Steganography and the Bit Plane Complexity Slicing (BPS) method. It is frequently

*AI Liberal Arts Studies, Honam University, 120, Honamdae-gil, Gwangsan-gu, Gwangju-si, 62399, South Korea (Sejunglim321@gmail.com).
†Sharnbasva University Kalaburagi, Karnataka, India (ambika.umashetty@gmail.com)

used because of its great capacity for hidden data and ease of deployment. (ii) Transform domain: The secret information is enclosed with the transform coefficient of the cover image. As a result, the makeover elements of the cover image contain the secret information. DCT, DFT, and DWT are three examples of wide area Steganography [12].

The spatial domain is challenging in this case because modifications to the image content may result in visually or statistically identifiable features. By using statistical analysis to determine the embedding depth, one may improve the safety and volume by hiding an adequate amount of bits in dissimilar pixels; this technique is known as "Group of Bits Substitution" (GBS). There are two systems; the 1-bit GBS strategy hides one bit per pixel, and the 2-bit GBS technique hides two bits per pixel. Normally, one pixel corresponds to one byte of the image. So for now, embedding is done by substituting out a group of bits in a pixel with a different set of bits that are positioned identically [13].

2. Literature review. The readapting stage of EA Image Stegnography on the basis of LSB. The histogram of the absolute difference of the adjacent pixels when similar is revisited reveals a pulse distortion to the lengthy exponential tail [14]. Employing this observation, a complex steganalytic method based on B-Spline fitting was applied. Additionally, it could accurately determine the threshold value needed to incorporate secret data as well as isolate the block size and stego image from those with block sizes greater than one.

According to [15], the LSB-based technique was not a widely used steganography algorithm in the spatial domain. Instead, most methods focus on pseudorandom number generators to determine the hiding positions within the cover image, without taking into account the relationship between the size of the secret message and the image's actual content. The secret message size and variance between two neighbouring pixels in the cover image would be taken into consideration while choosing the embedding regions for edge adaptive and LSB matching revisiting image steganography. Low security and low embedding rate are the main factors that have a significant impact on the system. Therefore, in order to increase the embedding rate, they have chosen sharp edge regions.

A paper [16] used an iterative strategy to enhance the steganography system by enhancing the image quality. The hidden message is embedded while the image quality is optimised using evolutionary algorithm. MSE, HVS deviation, and other parameters were considered for evaluation.

Quantum steganography has been defined by [17] and might even address several problems with conventional steganography that make it ineffective at hiding data. Anonymity, quantum image digital blocking, and a few other categories can be found in quantum image data hiding. Because many image data hiding algorithms were built on the LSB data hiding model, it plays a significant role. They experimented with adding clustering method to increase embedding rate without losing the secret information or image quality after using LSB to hide information in the cover image.

A HVS has been defined by [18] that is unsafe for steganography analyzers. A binary image steganography method was proposed with the aim of reducing texture embedding distortion. They have therefore assessed the distortion of the visual quality by validating the binary image and the generated image.

In order to minimize distortion while still retrieving the secret message or data, [19] proposed a steganography system. In order to reduce the risk of recognition via steganalysis, a method that might determine the interactions between embedding variations was used. To increase security, CMD, or clustering modification direction, is used [20, 21].

3. Proposed AABC-OPA based Image Steganography Method. As security demands increase, encryption alone is no longer sufficient; consequently, steganography is an advancement to encryption. It contains no further encryption. Steganography and encryption, on the other hand, improve information security. For selecting the best results, optimisation algorithms are utilised. With the rise of internet communication, data must be protected even when transported from sender to receiver via an insecure route. In order to secure sensitive data via another media, the steganography technology plays an important role in the field of information hiding. Due to its higher level of accuracy, image is regarded as a significant important medium among various cover media. Cover images, which can be coloured or grayscale, are used to hide hidden information in image steganography.

The main objective of this work is to manage the optimal pixel values, where secret information is embedded. The colour image and the secret information that must be hidden are read in the first step of the proposed

approach, which employs the Adaptive ABC algorithm. The blue components of the image are subjected to an integer wavelet transformation, and the AABC algorithm—also known as the Adaptive Artificial Bee Colony Algorithm—is applied to those modified coefficients to obtain the best value for hiding data. OPA is also used to increase image quality.

3.1. Color Plane Separation and Integer Wavelet Transform. The RGB cover/original image is divided into R, G, and B colour components in the proposed steganography technique. As the Human Visual System (HVS) has the least impact on the blue component, only the blue component of these is separated. The blue components used to be transformed by the IWT.

Commonly, wavelet domain permits us to hide data in regions that the human visual system (HVS) is less delicate to, for instance, the high resolution detail bands (HL, LH and HH), Hiding information in these arenas enable us to increase the robustness although keeping up good visual quality. Integer wavelet convert maps an unabridged number informational index into another complete number informational index. IWT stands forward than DWT as with DWT transformation, there is fortuitous of losing of data throughout reconstruction.

The Integer Wavelet transform usages Haar Wavelet decomposition filter that can be written as equations 3.1 to 3.2:

$$l_{1,d} = \left\lfloor \frac{l_{0,2d} + l_{0,2d+1}}{2} \right\rfloor \tag{3.1}$$

$$h_{1,d} = l_{0,2d+1} - l_{0,2d} \tag{3.2}$$

where $l_{1,d}$ and $h_{1,d}$ are the low and high frequency outputs at time.

Also, the Inverse of the Haar Wavelet decomposition filter can be given as in equations 3.3 to 3.4:

$$l_{0,2d} = l_{1,d} - \left\lfloor \frac{h_{1,d}}{2} \right\rfloor \tag{3.3}$$

$$l_{0,2d+1} = l_{1,d} + \left\lfloor \frac{h_{1,d} + 1}{2} \right\rfloor \tag{3.4}$$

The 2-Dimensional Integer Wavelet transform implemented on image, results in four frequency constituents given as in equation 3.5 to 3.8:

$$A_{p,q} = \left\lfloor \frac{(M_{2p,2q} + M_{2p+1,2q})}{2} \right\rfloor \tag{3.5}$$

$$H_{p,q} = M_{2p,2q+1} - M_{2p1,2q} \tag{3.6}$$

$$V_{p,q} = M_{2p,2q+1} - M_{2p1,2q} \tag{3.7}$$

$$D_{p,q} = M_{2p,2q+1} - M_{2p1,2q} \tag{3.8}$$

where $A_{p,q}, H_{p,q}, V_{p,q} and D_{p,q}$ are the approximation, horizontal, vertical and diagonal coefficients.

Supplementary, the inverse of 2-Dimensional Integer Wavelet transform can be attained as in equations 3.9 to 3.12:

$$M_{2p,2q} = A_{p,q} - \left\lfloor \frac{H_{p,q}}{2} \right\rfloor \tag{3.9}$$

$$M_{2p,2q+1} = A_{p,q} + \left\lfloor \frac{H_{p,q} + 1}{2} \right\rfloor \tag{3.10}$$

$$M_{2p+1,2q} = A_{2p,2q+1} + V_{p,q} - H_{p,q} \tag{3.11}$$

$$M_{2p+1,2q+1} = A_{2p+1,2q} + D_{p,q} - V_{p,q} \tag{3.12}$$

After getting the frequency coefficients, it is essential to detect the optimal pixel points (i.e. mapping points) in which the embedding is to be done. The optimal pixel points are attained from the AABC process from the arbitrarily produced mapping points.

3.2. Adaptive Artificial Bee Colony Algorithm for Optimal Pixel points. In the projected Adaptive ABC, the scout bee phase is gifted using the position updation of particle via PSO algorithm. Moreover, the Flowchart of AABC algorithm is given as in the Fig 3.1.

The steps involved in the AABC algorithm is specified as below:

Step 1: Population Initialization

The algorithm is recognized by subjectively generating optimal pixel location that communicates to the result in the search space. Let the arbitrarily generated initial pixel location is provided by, $P_x(x = 1, 2, \cdots n)$ where n designates the number of pixel points.

Step 2: Fitness evaluation

With the help of fitness function, the fitness value of the solution is intended to get the best pixel point. It's exposed in below equation 3.13:

$$\text{fitness}\,(P_x) = \text{Max}(P\,\text{NR}) \tag{3.13}$$

Now, the objective function is selected as the maximum of PSNR (Peak to Signal Noise Ratio). The chief idea behind this is to detect the optimal pixel points with that the image quality must not be devastated.

Step 3: Employed bee phase

In the employed bee's stage, every engaged bee determines a novel pixel points P_{xy}^{E} in the locality of its available food source P_{xy} by using equation 3.14:

$$P_{xy}^{E} = P_{xy} + \text{rand}\,(P_{xy} - P_{zy}) \qquad \text{where } z = (1, 2, \cdots n); z \neq x \tag{3.14}$$

where P_{xy} is the y^{th} pixel location of the x^{th} employed bee; P_{xy}^{E} is a novel solution for P_{xy} in the dimension; P_{xy} is the neighbor bee of P_{xy} in engaged bee population; is a number arbitrarily selected in the range of [-1,1].

Step 4: Fitness evaluation for new food source

Fitness values are recognized for each new pixel point and select the best pixel point.

Step 5: Probability based selection of Employed bee food source (Onlooker bee stage:)

After determining the optimal pixel points, the technique consequently uses probability based selection of pixel locations found from employed bee phase. Utilizing the equation 3.15 determine the probability of the designated pixel point is intended:

$$\text{Probability}_{\,x} = \frac{\text{fitness}_{\,x}}{\sum_{x=1}^{n} \text{fitness}_{\,x}} \tag{3.15}$$

where $fitness_x$ is the fitness value of x^{th} employed bee food source. For the result designated on the basis of probability, newer solution is engendered from its neighborhood on the basis equation 3.15. Again, the fitness is assessed for the solution generated from onlooker bee phase and the neighborhood onlooker food sources. On the basis of the fitness function, the outstanding pixel point is designated.

Step 6: Scout bee stage

In a cycle, after all engaged bees and onlooker bees complete their searches, the algorithm forms to detect new solution from the unrestricted food sources. In case, if the same food source comes recurrently (i.e. more than three times), the scout bee phase is originated. In the projected AABC algorithm, the scout bee phase solution updation is done by means of the particle's position updation technique of PSO. The new solution is produced via the particle's position updation process for subsequent iteration can be performed using subsequent representation shown in equations 3.16 and 3.17:

$$S_a^{t+1} = S_a^t + \mu_1 \left(p^t - m_a^t\right) + \mu_2 \left(g_a^t - m_a^t\right) \tag{3.16}$$

$$m_a^{t+1} = m_a^t + S_a^{t+1} \tag{3.17}$$

where μ_1 and μ_2 are unsystematic variables disseminated erratically in $[0, \omega_1]$ and $[0, \omega_2]$.Furthermore, p^t and g_a^t are the individual best and universal finest component at t^{th} iteration.

Step 8: Stop Criteria

Repeat step 2, up until a better fitness or maximum number of iterations is met and the solution that is holding the best fitness value is designated and it is quantified as optimal pixel point for embedding.

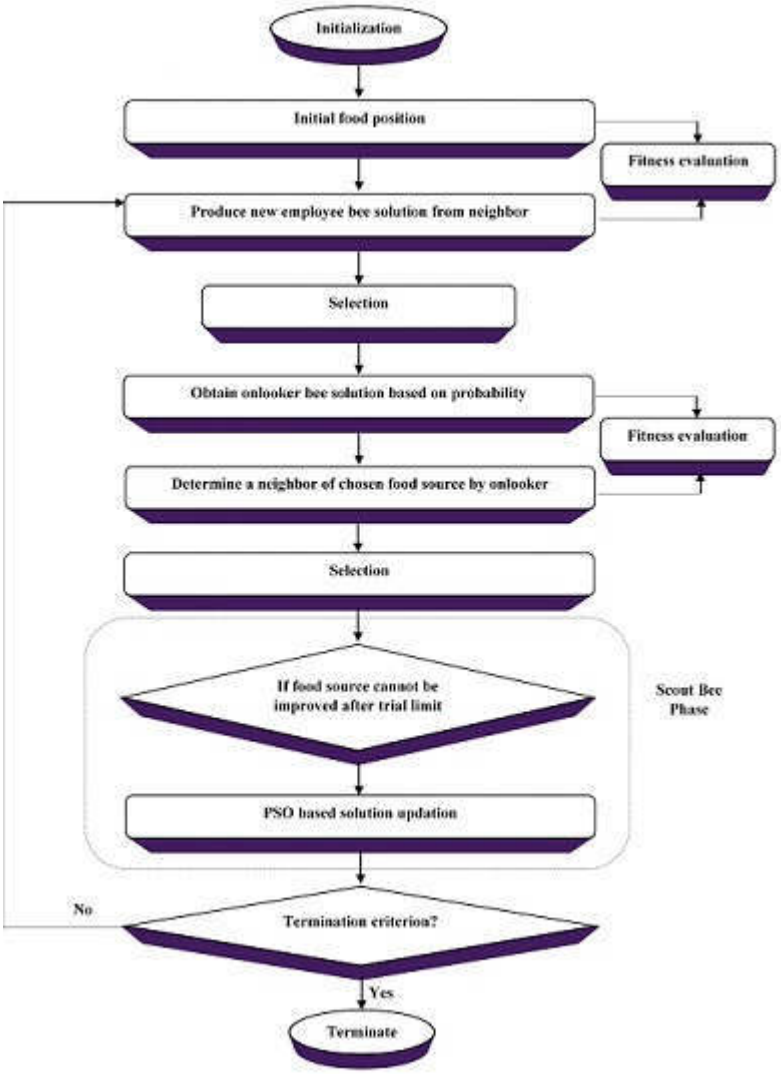


Fig. 3.1: Flowchart of AABC algorithm

3.3. Embedding Phase. It is possible to insert hidden bits in addition to the ideal pixel positions. The secret image is now converted to a binary image and forwarded to the embedding service. After embedding, on the basis of the embedding error, the OPA (Optimal Pixel Adjustment) procedure is completed to progress the image quality of the stego image. The proposed embedding process is clearly presented in the Fig 3.3.

3.4. Image Quality improvement by Optimal Pixel Adjustment Algorithm (OPA). The Optimal Pixel Adjustment method is now applied to the Stego image in order to improve image quality. Additionally, the OPA reduces the differences between the stego image and the cover/original image. As a result, the OPA produces higher hiding capacity, higher PSNR, and lower distortions. Furthermore, the stego image's invisible property remains preserved.

Presumptuous the host image be 'C' and the stego image be 'S'. And the pixel values, and being the pixel values of the host image and also the stego image. Now, the embedding error can be designed as, . On the basis of the embedding error, pixel adjustment technique is performed.

Input: Random pixel points
Output: Optimal pixel points

Start
Initialize the population of random pixel points,
Assess the fitness using equation (5),
Repeat
{// Produce Employed bee to select the new pixel points
{ **For** x=1,2….n
Do

$$P_{xy}^{E} = P_{xy} + rand\left[-1,1\right](P_{xy} - P_{zy})$$

$$\forall z \in (1,2,\cdots n);\ z \neq x$$

Calculate $fitness\left(P_{xy}^{E}\right)$

If ($fitness\left(P_{xy}^{E}\right) \leq fitness\left(P_{xy}\right)$) then

$$P_{x} = P_{xy}^{E}$$

End if
End for}
// Produce Onlooker bee to select the new optimal pixel points
{**For** x=1,2….n
do
select solution based on probability value,

$$Probability_{x} = \frac{fitness_{x}}{\sum_{x=1}^{n} fitness_{x}}$$

Produce solution using,

$$P_{xy}^{O} = P_{xy}^{E} + rand\left[-1,1\right](P_{xy}^{E} - P_{zy}^{E})$$

Calculate $fitness\left(P_{xy}^{O}\right)$

If ($fitness\left(P_{xy}^{E}\right) \leq fitness\left(P_{xy}^{O}\right)$) then

$$P_{x} = P_{xy}^{O}$$

End if
End for }// Produce Scout bee to select the new optimal pixel points
If food source is not improved further
{Check trial limit;
If (trial limit >3)
Update solution using,

$$S_{a}^{t+1} = S_{a}^{t} + \mu_{1}(p^{t} - m_{a}^{t}) + \mu_{2}(g_{a}^{t} - m_{a}^{t})$$ // velocity updation

$$m_{a}^{t+1} = m_{a}^{t} + S_{a}^{t+1}$$ // position updation
End if}
Iteration=Iteration+1;
}**End**

Fig. 3.2: Pseudo code for projected artificial bee colony algorithm

3.5. Extraction Phase. The IWT coefficients for the Stego image are initially extracted during the reconstruction of secret information. The mapping function obtained through the AABC process is now retrieved. In addition to this, the number of LSBs substituted during the OPA method is also inspected. The hidden bits can currently be recovered one by one from the designated mapping locations. The embedded secret data is provided by the aggregate of retrieved secret bits.

The proposed extraction technique is given in the Fig 3.5.

4. Results and Discussion. This section includes the outcome and discussion of the proposed reversible steganography technology using the Adaptive Artificial Bee Colony algorithm and the Optimal Pixel Adjustment technique. The proposed approach is implemented by MATLAB software, and the experiment is carried out

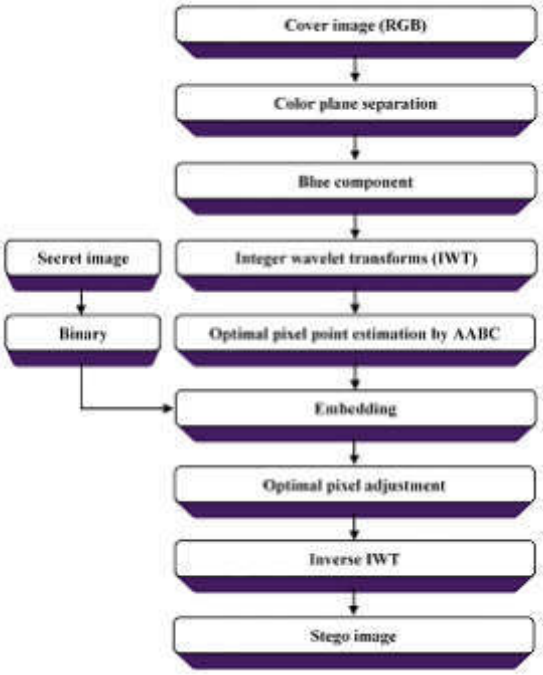


Fig. 3.3: Flowchart of Proposed Embedding procedure

Input: Cover Image:'C' of $(U \times V)$ pixel values
Embedded Stego Image: 'S'
Binary Secret image:'B'
Output: Improved Stego Image: '$\hat{S}$'

Read cover image, 'C' of $(U \times V)$ pixels
Read the Binary Secret image, 'B'

$$B = \{b_m \mid 0 \le m \le v, b_m \in [0,1]\}$$

Rearrange the Binary Secret image pixels to form a $T - bit$ virtual image

$$\hat{B} = \{\hat{b}_m \mid 0 \le \hat{m} < \hat{v}, \hat{b}_m \in [0,1,\cdots 2^t - 1]\}$$

Where, $\hat{v}$ is of $(U \times V)$

For embedding, three cases are followed for replacing the LSB's of stego image

Case (i): $2^{t-1} < \nabla < 2^t$

$$\hat{P}_S = \begin{cases} P_C - 2^t; P_C \ge 2^t \\ P_C; otherwise \end{cases}$$

Case (ii): $-2^{t-1} < \nabla < 2^{t-1}$

$$\hat{P}_S = P_C;$$

Case (iii): $-2^t < \nabla < -2^{t-1}$

$$\hat{P}_S = \begin{cases} P_C + 2^t; P_C < 256 - 2^t \\ P_C; otherwise \end{cases}$$

Fig. 3.4: Optimal Pixel Adjustment Algorithme

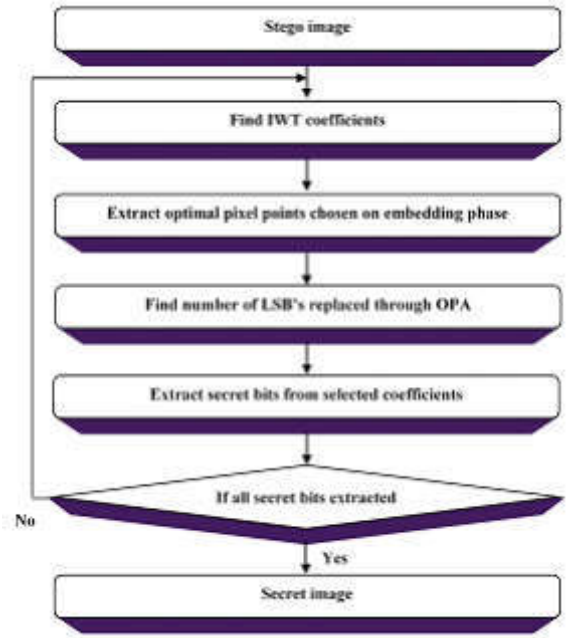


Fig. 3.5: Flowchart of Proposed Extraction procedure

on a system with 4 GB of RAM and a 2.10 GHz Intel i-3 processor.

For investigation, six standard test images, (a) Lena (b) Lake (c) Barbara (d) Gold hill (e) Tiffany (f) Peppers were occupied as cover/original image. This developed image data was distorted to frequency domain and optimal pixel points were extracted from AABC technique. Furthermore, to progress the image quality of stego image OPA method is presented. The experimental results for the projected AABC and existing ABC were contrasted with various image quality parameters and investigated in this section.

4.1. Quality Analysis Parameters. Measures such as BER, MSE and PSNR are utilized to evaluate the quality of embedded image with the original cover image. While NC and NAE are used to evaluate the quality of extracted and original secret image.

Mean Square Error (MSE): Mean Square Error is well-defined as the squared difference between the cover image and the stego-image. Using below equation, MSE can be designed as in Equ 4.1,

$$MSE = \frac{1}{PQ} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} [M(p,q) - \hat{M}(p,q)]^2 \tag{4.1}$$

where $M(p,q)$ and $\hat{M}(p,q)$ are the original and reconstructed cover images.

Peak Signal to Noise Ratio (PSNR): Peak Signal to Noise Ratio is well-defined as the peak error within cover image and stego-image. The Peak Signal to Noise Ratio (PSNR) is a measure utilized to measure the quality of the watermarked image shown in Equ 4.2. The higher the PSNR value, quality image will be better. While, the lower value of PSNR designates the poor quality image.

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right) \tag{4.2}$$

Bit Error Rate (BER): BER calculates the actual number of bit positions that are altered in the stego-image associated with cover image. **Normalized Absolute Error (NAE):** The NAE measures the distortion

Table 4.1: Performance Analysis using Existing ABC based OPA

	ABC based OPA				
	PSNR	MSE	BER	NC	NAE
Barbara	29.5129	0.887031	0.989689	0.305295	1.304358
Boat	29.87128	0.823445	0.934756	0.231959	1.231959
Foreman	28.53606	1.109306	0.975289	0.28538	1.28538
House	30.20251	0.760269	0.931733	0.227976	1.227976
Lena	32.07021	0.511243	0.989156	0.303655	1.303655
Peppers	29.29086	0.94384	0.992178	0.308575	1.307638

Table 4.2: Performance Analysis using Adaptive ABC based OPA

	Adaptive ABC based OPA				
	PSNR	MSE	BER	NC	NAE
Barbara	33.28491	0.865703	0.968925	0.328107	1.028107
Boat	32.70261	0.823341	0.901472	0.254883	1.022484
Foreman	35.01686	1.09051	0.95679	0.302999	1.026903
House	33.4493	0.751351	0.912071	0.234489	1.020134
Lena	35.16847	0.48623	0.96888	0.331865	1.028032
Peppers	34.53792	0.920343	0.968791	0.324836	1.029201

within the secret image and the extracted secret image. Low value of NAE designates the lower distortion. It is calculated as follow in Equ 4.3

$$NAE(p,q) = \frac{\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1}|B(p,q) - B^*(p,q)|}{\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1}|B(p,q)|} \tag{4.3}$$

where $B(p,q)$ and $B^*(p,q)$ are the original and extracted secret data.

Normalized Correlation (NC): The difference between the original and the extracted secret image is restrained by Normalized Correlation (NC), connecting to the numerical investigation of efficiency performance. Normalized Correlation (NC) is well-defined as in Equ 4.4,

$$NC(p,q) = \frac{\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1}B(p,q)B^*(p,q)}{\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1}[B(p,q)^2]} \tag{4.4}$$

4.2. Performance Analysis. In this segment, the performance valuation of the proposed Adaptive ABC based OPA steganography technique is associated with the available ABC based OPA steganography method. Table 4.1 provides the performance outcome of available ABC based OPA steganography method. Table 4.2 provides the performance outcome of proposed Adaptive ABC based OPA steganography technique.

From the Tables 4.1 and 4.2,

- PSNR is 35.16847 for 'Lena' image that is the highest value attained for the proposed technique while the PSNR of the same image for existing technique is 32.07021.

- Similarly, it is well known that the values obtained for MSE, BER, and NAE parameters for the proposed technique are lower when compared to the existing methodologies. This demonstrates the proposed technique's minimum error occurrence during both embedding and extraction phases.

- Highest NC is 0.331865 for the proposed technique and the Highest NC for existing technique is 0.308575 that is less than the proposed method's result.

- It is clear that the values obtained for the proposed AABC-based OPA are better throughout all measures investigated during the embedding and extraction processes.

Additionally, each metric's values are presented independently in the following Figures 4.1 to 4.5.

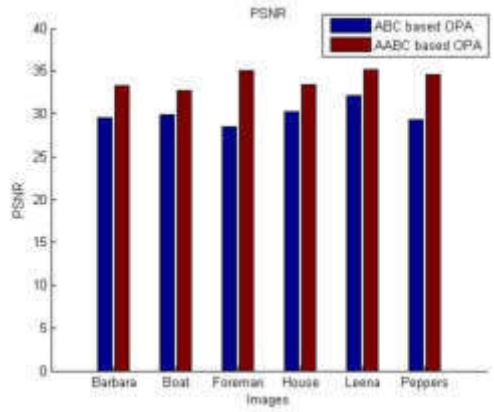


Fig. 4.1: PSNR comparison plot for proposed AABC and ABC

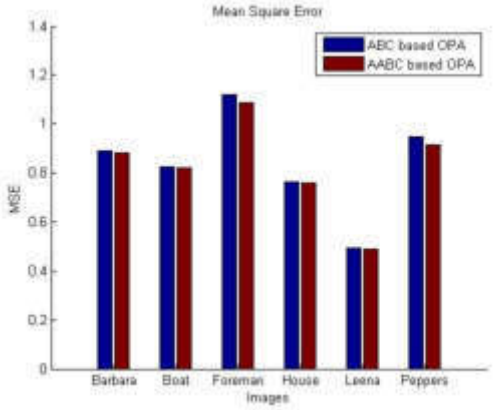


Fig. 4.2: MSE comparison plot for proposed AABC and ABC

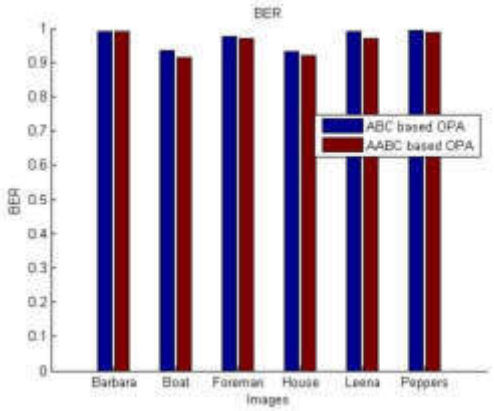


Fig. 4.3: BER comparison plot for proposed AABC and ABC

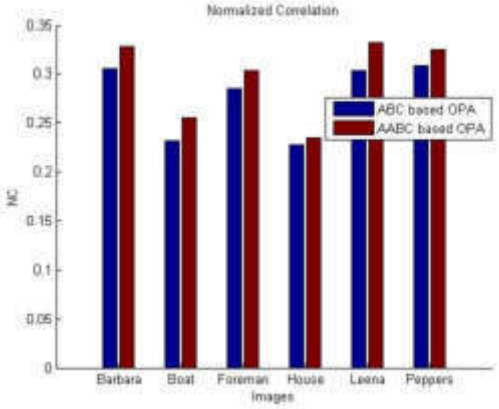


Fig. 4.4: NC comparison plot for proposed AABC and ABC

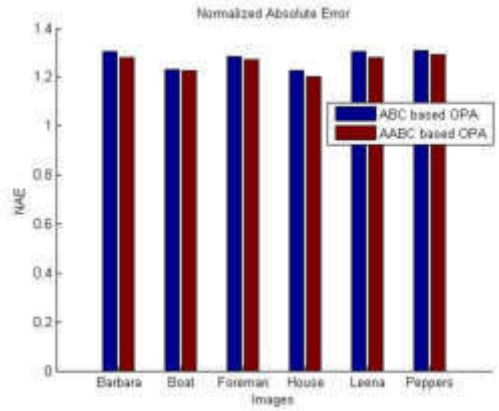


Fig. 4.5: NAE comparison plot for proposed AABC and ABC

5. Conclusion. This study employs an efficient data embedding method based on an Adaptive Artificial Bee Colony based Optimal Pixel Adjustment algorithm. For experimentation, the proposed Adaptive ABC based OPA method is compared with the ABC based OPA method. Additionally, a number of metrics such as BER, MSE, and PSNR are used to assess the quality of the embedded image in comparison to the original cover image; NC and NAE are used to assess the quality of the extracted and original secret image. According to the results, the proposed Adaptive ABC-based OPA algorithm gives a good results when compared with the existing ABC-based OPA technique. Using the proposed method increases the size of the cover image, which might grab the attention of attackers; therefore, future work should focus on minimising the size of the cover image while retaining the hidden information.

REFERENCES

[1] Junhui He,Weiqiang and Shaohua, "A secure image sharing scheme with high quality stego-images based on steganography", International Journal of Multimedia Tools and Applications, 2016.

[2] H. Arif and H. Hajjdiab, "A comparision between steganography software tools," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, pp. 423-428, 2017. doi: 10.1109/ICIS.2017.7960030.

[3] Xiangyang, Fenlin, Shiguo, Chunfang, and Stefanos, "On the Typical Statistic Features for Image Blind Steganalysis", IEEE Journal on Selected Areas in Communications, Vol. 29, No. 7, Pp. 1404-1422, 2011.

[4] Konstantinos Karampidis, Ergina Kavallieratou, Giorgos Papadourakis,"A review of image steganalysis techniques for digital forensics",Journal of Information Security and Applications,Vol 40,Pp 217-235, 2018. https://doi.org/10.1016/j.jisa.2018.04.005.
[5] Swain, "Digital Image Steganography Using Variable Length Group of Bits Substitution", Procedia Computer Science, vol. 85, pp. 31-38, 2016.
[6] Veerashetty, Sachinkumar. "Secure communication over wireless sensor network using image steganography with generative adversarial networks." Measurement: Sensors 24: 100452, 2022.
[7] Giriprakash, "Image Steganography by Variable Embedding and Multiple Edge Detection using Canny Operator", International Journal of Computer Applications, Vol. 48, No. 16, Pp. 15-19, 2012.
[8] Muthukumaran, V., Vinoth Kumar, V., Joseph, R. B., Munirathnam, M., Beschi, I. S., Niveditha, V. R. (2022, November). Efficient Authenticated Key Agreement Protocol for Cloud-Based Internet of Things. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3 (pp. 365-373). Singapore: Springer Nature Singapore.
[9] Jindal and Partap Sin, "Image Steganography with Multilayer Security Using Moderate Bit Substitution", International Journal of Applied Sciences, Vol. 14, No. 8, Pp. 738-747, 2014.
[10] Zhang and Dan, "Detection of LSB Matching Steganography in Decompressed Images", IEEE Signal Processing Letters, Vol. 17, No. 2, Pp. 141-144, 2010.
[11] Bin, Shunquan, Wang, and Huang, "Investigation on Cost Assignment in Spatial Image Steganography", IEEE Transactions on Information Forensics and Security, Vol. 9, No. 8, Pp. 1264-1277, 2014.
[12] Bin, Xingming, Lingyun, Haijun and Yang, "Detection of LSB Matching Steganography using Neighborhood Node Degree Characteristics", International Journal of Information Technology, Vol. 10, No. 8, Pp. 1601-1607, 2011.
[13] Santos and Jorge, "Artificial Neural Networks Applied to Image Steganography", IEEE Latin America Transactions, Vol. 14, No. 3, Pp. 1361-1366, 2016.
[14] J. Anitha, Sirmathi and Meenakshi, "Steganography Based Secure Data Storage and Intrusion Detection for Cloud Computing Using Signcryption and Artificial Neural Network", International Journal of Applied Sciences, Engineering and Technology, Vol. 13, No. 5, Pp. 354-364, 2016.
[15] Luo, and Huang, "Edge Adaptive Image Steganography Based on LSB Matching Revisited", IEEE Transactions on Information Forensics and Security, Vol. 5, No. 2, Pp. 201-214, 2010.
[16] Maithili, K., Vinothkumar, V., Latha, P. (2018). Analyzing the security mechanisms to prevent unauthorized access in cloud and network security. Journal of Computational and Theoretical Nanoscience, 15(6-7), 2059-2063.
[17] Yahya and feng Lu, "Quantum Image Steganography and Steganalysis Based On LSQu-Blocks Image Information Concealing Algorithm", International Journal of Theoretical Physics, Vol. 55, No. 8, Pp. 3722-3736, 2016.
[18] Bingwen and Wei Sun, "Secure Binary Image Steganography Based on Minimizing the Distortion on the Texture", IEEE Transactions on Information Forensics and Security, Vol. 10, No. 2, Pp. 243-255, 2015.
[19] Bin, Wang, Xiaolong, Tan and Huang, "A Strategy of Clustering Modification Directions in Spatial Image Steganography", IEEE Transactions on Information Forensics and Security, Vol. 10, No. 9, Pp. 1905-1917, 2015.
[20] Velliangiri, S., Karthikeyan, P., and Vinoth Kumar, V. Detection of distributed denial of service attack in cloud computing using the optimizationbased deep networks. Journal of Experimental and Theoretical Artificial Intelligence, 33(3), 405-424, 2021.
[21] Uplaonkar, Deepak S., and Nagabhushan Patil. "Ultrasound liver tumor segmentation using adaptively regularized kernel-based fuzzy C means with enhanced level set algorithm." International Journal of Intelligent Computing and Cybernetics 15, no. 3, 438-453, 2021.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Jul 5, 2023
Accepted: Nov 7, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 577–591, DOI 10.12694/scpe.v25i1.2405

VULNERABILITY DETECTION IN CYBER-PHYSICAL SYSTEM USING MACHINE LEARNING

BHARATHI V*AND C. N. S. VINOTH KUMAR†

Abstract. The cyber-physical system is a specific type of IoT communication environment that deals with communication through innovative healthcare (medical) devices. The traditional medical system has been partially replaced by this application, improving healthcare through efficiency, accessibility, and personalization. The intelligent healthcare industry utilizes wireless medical sensors to gather patient health information and send it to a distant server for diagnosis or treatment. The healthcare industry must increase electronic device accuracy, reliability, and productivity. Artificial intelligence (AI) has been applied in various industries, but cybersecurity for cyber-physical systems (CPS) is still a recent topic. This work presents a method for intelligent threat recognition based on machine learning (ML) that enables run-time risk assessment for better situational awareness in CPS security monitoring. Several machine learning techniques, including Nave Bayes (65.4%), Support Vector Machine (64.1%), Decision Tree (89.6%), Random Forest (92.5%), and Ensemble crossover (EC) XG boost classifier (99.64), were used to classify the malicious activities on real-world testbeds. The outcomes demonstrate that the Ensemble crossover XG boost enabled the best classification accuracy. When used in industrial reference applications, the model creates a safe environment where the patient is only made aware of risks when categorization optimism exceeds a specific limit, minimizing security managers' pressure and efficiently assisting their choices.

Key words: Cyber-Physical Systems, Trustworthy Artificial Intelligence, Cybersecurity, Healthcare, Machine Learning, Critical Infrastructures.

1. Introduction. The healthcare landscape has changed due to the assumption of developing automation like the Internet of Things (IoT), intelligent bio-medical sensors (BMSs), and the cloud that have increased life expectancy rates. As a result, it raised people's living standards. The fifth industrial revolution (industry 5.0) is based on a cyber-physical system that connects digital diagnostic products, like computers and the Internet, to physical processes[33]. Healthcare professionals can use H-CPS to process the sensed data and make wise decisions. Medical practitioners must adhere to H-CPS-based procedures to provide better treatment for less money. IoMT smart devices can gather, evaluate, and broadcast various data in a healthcare setting that uses H-CPS.

Additionally, these wearable sensors continuously monitor the patient's health characteristics, such as blood pressure, temperature, and pulse rate, and communicate the information to nearby access systems for computation and feature selection. Using Artificial Intelligence (AI)-enabled technology, the pre-processed data is sent to remote computing equipment for disease detection or prognosis [18]. The medical sector has recently been exposed to more complex and extensive cyber risks, drawing attention to the lack of cybersecurity skills. For instance, the healthcare supply chain has just been exposed to a new cyber threat, becoming more widespread and stable each year. This threat revealed the industry's overall inadequate cybersecurity architecture. Cyber risk is related explicitly to two concurrent advancements: First, the increasingly pervasive incorporation of technologies, modernization, and novel healthcare systems [34], including automated treatment pathways, electronic health records, individualized therapies, and widely scattered IoMT (Internet of Medical Things) equipment.

On the contrary side, cybersecurity practice upgrading and invention procedures find it challenging to keep up with the rate of advancements in technology. Because of the intersection of these two tendencies, the healthcare industry is highly vulnerable to cyber threat, which has increased in both severity and frequency in

*Department of Networking and Communications, College of Engineering and Technology (CET), SRM Institute of Science and Technology, Kattankulathur Chennai, India (bv3994@srmist.edu.in)

†Corresponding Author, Department of Networking and Communications, College of Engineering and Technology (CET), SRM Institute of Science and Technology, Kattankulathur Chennai, India. (vinothks1@srmist.edu.in).

recent years. The availability of the data, therefore making it impossible for legitimate owners to access the data to make it susceptible to exploitation, as well as the integrity, correctness, and alteration of the accuracy, are three potential targets for cyberattacks in the health sector. Knowing an individual's or a portion of the population's health figure could have financial implications. The requirement to strike a balance between the necessity for security and information privacy and the accessibility of information to maintain the essential benefit of the person's health adds another layer of complexity to many crucial professions, such as healthcare. The above explains why it can be challenging to put stringent cybersecurity controls in place that hamper healthcare, especially in times of need and urgency[11].

The use of AI as a decision support tool while leaving ultimate decision-making in the hands of people was agreed upon by machine learning (ML), artificial intelligence, and cybersecurity. The multiple contributions to this topic proposed some ML 'anomaly-based' approaches. Unfortunately, there are fewer comparative studies to determine the best effective learning method for improved detection capacity and fewer false alarms than for ML approaches. The possibilities of boosting ML approaches' dependability and the extent to which they apply to real-world situations are also major open questions. Even though many security strategies for Critical Infrastructure (CI) have been presented, there are still many obstacles to overcome to autonomously spot risks in the face of complexity, uncertainty, and change, especially when considering phenomena like sensors are compromised [30]. Additionally, a novel perspective on CIs has emerged recently, which they have seen as complicated Medical Cyber-Physical Systems (HCPS).

Modern CPS comprises real and intangible elements, including databases and software algorithms for data elaboration and electro-mechanical devices, sensors, and actuators. Threats can be both tangible and intangible because of the nature of CPS. Cyberthreats, for example, could directly affect the physical components' integrity and indirectly affect the environment's and the related parties' overall health. These threats may also have a chain of associated repercussions. This paper's essential contribution is as follows: We offer a vulnerability assessment technique for CPSs that considers cyber-physical and physical-cyber interdependencies to derive goal-oriented attack routes. The suggested procedure:

- Artificial intelligence is a paradigm for coordinating the efforts of many machine learning algorithms to detect and prevent harmful or malicious occurrences.
- Exposes sophisticated cyber-physical assaults by using vulnerability analysis techniques to deduce the motivations of adversaries.
- Attack route analysis is made more efficient by switching from a blind analysis to an algorithmic analysis with clear end goals.
- It is a practical approach to computing risk and evaluating Likelihood and Impact based on security-relevant criteria.

The remainder of the paper is organized as follows: This document is organized for the remaining portions: The relevant work's outline is presented 2, while the material and methods are presented in Section 3. In Section 4, experimental analysis is presented. Section 5 serves as the paper's conclusion.

2. Related Work. The CPS explosive expansion, security, and privacy are necessary for reliable communication in innovative healthcare [37]. Sun et al. [24] addressed the privacy and security concerns with IoT in the healthcare sector and remarked on the potential routes for further study. Hu et al. [16] used attribute-based encryption to address the issue of safe communication between a BAN and its data consumer (end-user). According to Chandrasekaran et al. [7], the technique [16]is ineffective for repeated data transfer, and they provide a novel system for safely transmitting data in WBAN. Blockchain technology was used by Egala et al. [10] to create a safe and decentralized platform for sharing health data records without jeopardizing system privacy. Kumar and Chand [22] presented a blockchain-based privacy-privacy data-sharing system for the healthcare sector, where an Identity-based broadcast group encryption technique protects each transaction. As a result, interest in managing complex cybersecurity systems increased, and AI techniques were incorporated to assist with automation [4][13]. AI is revolutionizing cybersecurity due to extensive analysis of data, faster reaction times, and effective customization of threat detection for limited records. Further, Artificial Intelligence has already-existing and synergistic applications for pattern recognition and computer vision to identify physical threats [12]. The authors of [15] have created a hybrid IoT generator, a framework for estimating cellular network performance. This platform was combined with big data and Machine Type Communications traffic

models.[6][1] provides information on the various ML-based strategies. The authors systematically explain how machine learning approaches operate and offer their assessments. An overview of ML algorithms in IoT of healthcare data is provided in [5]. This study uses supervised learning, semi-supervised learning, and unsupervised learning ML model types to classify data from the healthcare industry and show the work on the data. The threat modeling tools STRIDE [1], Factor Analysis of Information Risk (FAIR)[5], and OCTAVE [9] have all been utilized in the process of assessing the level of risk present in CPSs across a variety of application areas. Another prevalent strategy[35] combines two or more methodologies: STRIDE and CVSS. It is possible that the "traditional" impact criteria of confidentiality, integrity, and availability will not be sufficient for CPSs; consequently, the methods used to assess the cyber risk posed by these systems must typically be industry-specific. This is why research on the safety and security of CPSs is carried out simultaneously. In [20], we comprehensively analyze different approaches to co-engineering of safety and security. In [23], we summarize risk assessment techniques applicable to the smart grid scenario. Kandasamy et al. [19] presented a general overview of the methods for assessing the Internet of Things risks. A rundown of a few methods for determining how vulnerable SCADA systems are to attack is provided for us in reference [8][26] delves into the various approaches that can be taken to perform risk assessments in the automotive sector. Recent research, such as that presented in [25], examines various risk assessment strategies for CPS from the perspectives of safety, security, and the integration of all three and proposes specific categorization criteria. Current approaches to risk assessment for CPSs, which primarily focus on either one or the other of these two types of interdependencies, ignore, for the most part, cyber-physical and physical-cyber interdependencies. When researching the system, the authors of[29] focused on its physical components, whereas Homer et al. [14] investigated only the system's cyber components. It has been demonstrated by Krotofil et al. [27] that this is not the case, despite the fact that attackers may use the physics of the mechanism that is behind a CPS. When it comes to developing security policies, these same authors argued that the physical process layer should be considered. According to[28], research into cyber-physical systems needs to adopt a more comprehensive methodology because of the complex intertwining of computer networks and physical processes.

Regrettably, to the best of our knowledge from Table 3.1, no risk assessment method that satisfies this criterion has ever been made public. This study covers a knowledge gap with its recommended practices. The recommended strategy facilitates research of the entire cyber-physical system for each undesired event, in contrast to present methodologies. As a result, the review shows some of the disadvantages over existing methods; hence, in this work, three possibilities were evaluated to find the optimum strategy that will build up the existing research gaps.

Possibility 1: This work uses binary classification to alert the customer whenever an abnormality has been found by identifying its kind or context. Knowing the nature of a threat is necessary to take adequate preventative measures, even though its identification is crucial.

Possibility 2: Given the four components that make up the system, this instance seeks to tell the operator about the one that the anomaly has affected. As a result, it is an inter-categorization. The classes investigated are 5: pulse rate, temperature, SpO2, blood pressure, and the scenario in which an anomaly impacts no sensor.

Possibility 3: The most recent experiment aimed to categorize the incidents into the following categories: failure, damage, accident/damage, cyberattack, failure/damage, and, ultimately, the lack of abnormalities. The response time could be significantly shortened by resorting to relevant risk management by providing the user only with the known malicious scenario.

3. Materials and Methods. This paper examines five machine-learning techniques to discover trends in PCA information. They categorize strange events using the selected models, including hardware (sensor issues), cyberattacks, and sabotage. Naive Bayes, SVM, DT, RT, and Ensemble crossover XG boost classifiers are the models that have been chosen depending on the latest research and taking into account the FPR rates attained by prior research. The overall system architecture was shown in Figure 3.1.

Data are periodically gathered from various patients. These data comprise multiple situations structured in CSV files across a range of time—the length of the file changes according to the circumstance and malfunctioning element. Typical operational occasions and strange events include physical malfunctions and cybersecurity issues. For example, a decision-maker may need to understand these scenarios or become aware of them when malicious activity occurs. Attack analysis are crucial because poorly handled circumstances may result in highly

Table 2.1: Comparative analysis of existing methodology

Tag	Year	Protocol	Attack	Entrypoint	Control	Evaluation metric	Type
1	2022 Khadr et al. (2022)[21]	ZigBee	Jamming	S-S	Parallel-Channel Security-aware Medium Access Control (PCS-MAC) algorithm	Throughput	Physical
2	2022 Yu and Park (2022)[39]	IEEE 802.15.6	Eavesdropping, brute force, service disruption, masquerading	E-S, S-C	Authentication protocol based on blockchain technology and PUFs	Computation time, communication overhead	Simulated
3	2022 Pu et al. (2022)[32]	IEEE 802.15.6	Eavesdropping, data manipulation, replay, service disruption, masquerading	E-S, S-C	Lightweight, anonymous authentication and key agreement protocol	Communication overhead, computation time, energy consumption, CPU time, CPU cycles	Simulated
4	2021 Alzahrani et al. (2021)[3]	IEEE 802.15.6	Eavesdropping, brute force, replay, masquerading	E-S, S-C	Authenticated key agreement based on Burrows-Abadi-Needham (BAN) Burrows et al. (1990) logic	Computation time, communication overhead, energy consumption	Simulated
5	2021 Wang et al. (2021)[38]	IEEE 802.15.6	Eavesdropping, data manipulation, replay, service disruption, masquerading	E-S, S-C	Authentication protocol based on blockchain technology and PUFs	Computation time, communication overhead	Simulated
6	2021 Hussain et al. (2021)[17]	IEEE 802.11	Eavesdropping, data manipulation	S-C, W-C	Physical layer scheme (Gray code)	N/A	Physical
7	2021 Surminski et al. (2021)[36]	IEEE 802.11	Eavesdropping, buffer overflow	C-A	Remote attestation	Runtime, energy consumption, communication overhead, race conditions	Hybrid
8	2020 Alladi et al. (2020)[2]	IEEE 802.15.6	Eavesdropping, data manipulation, masquerading, ARP spoofing, replay	S-C, W-C	Two-way, two-stage authentication protocol using PUFs	Computation time	Simulated

negative operating costs.

The following stages can be used to breakdown the suggested method:

3.1. Data Collection. The dataset was obtained from iot-healthcare-security-dataset. The provided dataset includes regular and malicious traffic data for IoT healthcare use cases. A use case was developed for an Internet of Things (IoT)-based Intensive Care Unit (ICU) consisting of two beds. Each bed has nine patient monitoring devices (See Figure 3.2), sensors, and one control unit known as the Bedx-Control-Unit. All of these devices were developed with the IoT-Flock tool.

The proposed ICU system is based on the Internet of Things (IoT) technology and has a capacity of two beds. Each bed is equipped with nine patient monitoring devices, often called sensors, along with one control unit. The term used to refer to this entity is the Bedx-Control-Unit, where 'x' is the number assigned to each bed, ranging from Bed1 to Bed2. The responsibility of the Bedx-Control-Unit encompasses several tasks,

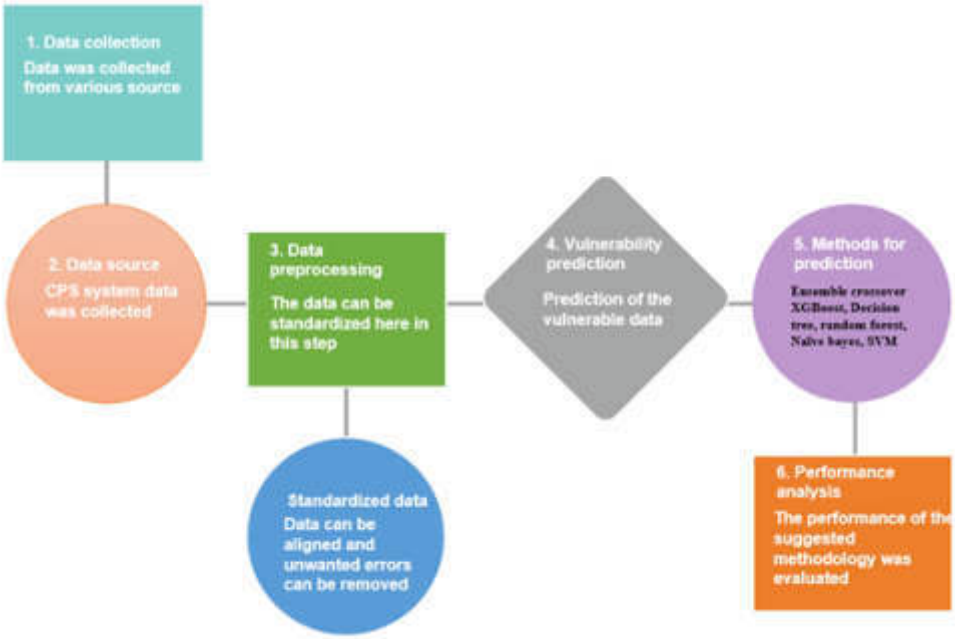


Fig. 3.1: Suggested Architecture of Cyber-Physical System

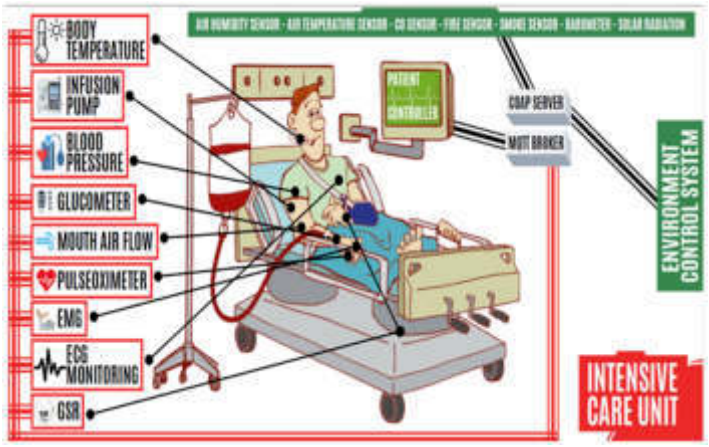


Fig. 3.2: Dataset description

including but not limited to configuring the time profile, determining the dosage administered by an infusion pump, and activating emergency alarms. These activities are contingent upon the patient's physical status, as monitored by the patient monitoring devices.

In a similar vein, an additional control unit was included to facilitate the monitoring of environmental equipment, which was that named the Environment-Control-Unit. The Environment-Control-Unit is tasked with regulating the environmental conditions inside the Intensive Care Unit (ICU), including maintaining specific temperature and humidity levels, detecting the presence of smoke, and activating an emergency alert in

Table 3.1: Patient monitoring sensors

Device Name	**Description**	**Data Profile**	**Time Profile**
Remote Electrocardiogram (ECG) monitoring	Test the electrical and muscular functions of the heart	Pulse Rate (0-200 bpm)	1.0 s
Infusion Pump	A generic device is used to deliver the nutrients and drugs to the patients at a controlled amount	Dose (10-100 mL)	10.0 min
Pulsoximeter (SPO2)	A device that tells the oxygen saturation (i.e.,amount of oxygen dissolved) in blood	Oxygen in blood (35-100%)	1.0 s
Nasal/Mouth AirFlow Sensor	Provides the (breathing) respiratory rate of a patient	Device Respiratory rate (0-60ppm peaks/min)	1.0 s
Blood Monitor Sensor	Measure the pressure of the blood in the arteries when the heart beats	Systolic & diastolic pressure (0-300 mmHg)	2.0 s
Glucometer	A device used to determine the amount of glucose in the blood.	Glucose in Blood (10-150 mg/dL)	10.0 min
Body Temperature	Sensor Measures the temperature of the body	Temperature (0-120 F)	10.0min
Electro-myography (EMG) Sensor	Measures the electric potential produced by the body's muscles	Muscle rate(contractions/min)(0-60cpm)	5.0min
Galvanic skin response (GSR) Sensor	Measures the electrical conductance of skin	Conductance(0-20uS)(micro Siemens)	5.0 min

the event of critical situations in order to uphold the necessary ICU environment. In our specific scenario, both the devices used for patient monitoring operate on the MQTT protocol. The MQTT protocol is characterized by its connection-oriented nature and ability to guarantee the appropriate transmission of packets. Table 3.1 presents an overview of the use case for IoT-based intensive care units.

The attacks identified using this dataset include an MQTT distributed denial-of-service, MQTT publish flood, brute force, and SlowITE attack. The following sections describe the types of attacks that IoT-Flock supports.

- MQTT Publish Flood— A Distributed Denial of Service (DDoS) attack has the potential to deplete the available network bandwidth and exhaust the resources of the targeted victim system. Due to the implementation of more effective mitigation strategies at the network and transport layers, DDoS attackers have shifted their focus towards attacking the application layer. Internet of Things (IoT) devices adhere to either the periodic or event-driven paradigm when transmitting data via application layer protocols. The systematic model device transmits data at regular intervals, such as the temperature sensor sending temperature data to the server every five seconds. In the context of event-driven models, devices transmit data only in response to specific events. For instance, inside an intensive care unit, a motion sensor will only transmit data to the server upon detecting activity in the designated area. According to recent literature, it has been observed that the act of publishing messages at a rapid rate using the MQTT protocol might potentially lead to a denial of service attack. These assaults can potentially impede data transmission significantly and pose substantial risks, particularly in critical sectors such as industrial operations, smart hospitals, and smart transport systems. The potential consequences of data transmission delays may result in the destruction of assets and pose significant risks to human life.
- MQTT Authentication Bypass Attack —To establish a connection with the MQTT broker, which necessitates authentication, MQTT clients transmit MQTT connect requests that include fields for username and password. The discovery was made that the authentication mechanism of MQTT may be circumvented by omitting the password field from the MQTT packet and just supplying a valid username. Despite the mitigation measures used in recent versions of MQTT brokers, the processing

of erroneous packets by an MQTT broker may still result in operational delays, particularly when such packets are sent in substantial quantities. Hence, using IPS to block such an unauthorized packet can mitigate the latency problem associated with the MQTT broker.

- MQTT Packet Crafting Attack —The present assault involves deliberately manipulating MQTT packets to cause a targeted application to malfunction or cease functioning entirely. The assailant initiated a connection with the MQTT broker at the Transport layer and started publication without first issuing a connection request to the MQTT broker.
- COAP Replay Attack— During this attack, an unauthorized individual does an initial network scan to get the addresses of COAP clients and servers and payload information. Subsequently, the unauthorized individual modifies the payload by substituting it with inaccurate data and transmits it to the COAP server, using a deceptive technique that mimics the Sensors 2021, 21, 3025 12 of 19 COAP client IP. The magnitude of this assault becomes apparent when examining instances in which environmental sensors use COAP protocols to relay ambient data to the COAP server. This may be illustrated when the temperature sensor transmits fluctuations in the intensive care unit's temperature. Subsequently, the condition is established by using the data mentioned above. In the event that an assailant employs IP spoofing techniques, they may transmit a manipulated ICU temperature reading, including anomalous values, hence instigating severe and detrimental consequences inside the ICU environment.

3.2. Data pre-processing. This step aims to clarify the information more understandable for the user. The first three steps in pre-processing are: a) Data arrangement: The information must be shown in a logical manner. b) Data scrubbing: Any corrupted or missing data must be removed, replaced, or added to the data. c) Sampling Data: Data must have been sampled regularly before being transferred via communication channels to eliminate redundancy without compromising information.

Transform the data following the algorithm and your understanding of the issue. Feature scaling, deconstruction, or aggregation are all examples of transformation. Features can be aggregated to merge numerous instances into a single element or decomposed to retrieve the valuable components embedded in the data. Three distinct yet interconnected steps can be used to describe this process:

3.3. Vulnerability prediction. For the prediction of the CPS vulnerability in which the visualized data can be split up into train and test data, are separately given as input for the Nave Bayes, Support Vector Machine, Decision Tree, Random Forest, and Ensemble crossover XG boost classifier listed below,

a.Decision Tree (DT) Classifier. A DT is a simple classifier that may be used to put data into groups. In DT, the data is continuously segmented according to a predetermined criterion. Well-known in the field of supervised classification are the DTs. They are effective at categorization tasks, have straightforward decision-making processes, and can be created (trained) quickly and easily thanks to an efficient algorithm. Since it was one of the first elite regression analysis techniques taught to those studying predictive modeling, it has become one of the most well-known approaches in the field.

$$\mathrm{X} = [D_x, D_y] \tag{3.1}$$

where D_x and D_y are the factors that go into the equation,

$$D_x = \frac{1}{3} \frac{\sum_{i=0}^{n-1} (X_i + X_{i+1}) (Y_i X_{i+1} - Y_{i+1} X_i)}{\sum_{i=0}^{n-1} (Y_i X_{i+1} - X_{i+1} X_i)} \tag{3.2}$$

and

$$C_y = \frac{1}{3} \frac{\sum_{i=0}^{n-1} (X_i + X_{i+1}) (\mathrm{X} Y_{i+1} - X_{i+1} Y_i)}{\sum_{i=0}^{n-1} (X_i x_{i+1} - X_{i+1} Y_i)} \tag{3.3}$$

b.Random Forest (RT) Classifier. The supervised ML model includes the ML approach known as Random Forest. There are several different types of DTs that make up the RF classifier. The predicted accuracy is increased by averaging the subsets of all trees. Instead of relying on only one set of decision trees,

RF takes the average of all the votes to determine the outcome. Each branch of the decision tree responds to a question regarding the current state of affairs.

Possible values for the X_i property of a nominal (divided) data set are $L_i, \ldots, L_j$. To get the Gini Index for this characteristic, use the following Equation (3.4) formula.

$$G\left(X_i\right) = \sum_{j=1}^{j} \Pr\left(Y_i = L_j\right)\left(1 - \Pr\left(Y_i = L_j\right)\right) = 1 - \sum_{j=1}^{j} \Pr\left(Y_i = L_j\right)^2 \tag{3.4}$$

c.Naïve Bayes (NB) Classifier. It is easy to estimate conditional probabilities using the Bayes' theorem. The Equation (3.5) looks like this:

$$P(A \mid R) = \frac{P(R \mid A) * P(A)}{P(R)} \tag{3.5}$$

where R and A are random variables, $P(A \mid R)$ is the probability that Y if X is true,$P(R \mid A)$ is the probability that X if R is true, P(R) is the probability of X, and P(A) is the probability of Y if A is true.

d.Support Vector Machine.Data vulnerability Classification and Estimation Using a Support Vector Machine Model. In this investigation, we focus on the classification of signal quality, which is often a two-classification issue. In several cases involving categorization into two groups, the SVM-based model performed well. For a given training set $\{x_i, y_i\}, i = 1 \ldots, K$, where x_i is a feature vector of length$x_i \in R^d$, and y_i is the label, it is possible to train a classifier. Therefore, the SVM-based model may be used for both estimating and classifying signal quality. The quality estimate label is $y_i \in \{1, 0\}$, where excellent and terrible represent extremes. $y_i \in \{1, -1\}$ is the categorization label for abnormal and normal cases. The goal of a support vector machine (SVM) classifier is not only to differentiate between the classes, but also to create a hyperplane between them. It is also possible to build the ideal hyperplane by solving Equation (3.6),the following optimisation issue.

$$\min \phi(\mathbf{V}) = \frac{1}{2}\left(\mathbf{V}^T\mathbf{V}\right) + C\sum_{i=1}^{K} \xi_i \tag{3.6}$$

subject to

$$y_i\left(\left(\mathbf{V}^T\varphi\left(\mathbf{x}_i\right)\right) + b\right) \geq 1, i = 1, \cdots, K.$$

Here ξ_i is a error relaxation variable and $\xi_i \geq 0$, C is a factor of penalty, and w is the coefficient vector. $\varphi(x_i)$ is presented in order to construct a nonlinear SVM. Converting the optimisation issue into Equation (3.7).

$$\max L(\boldsymbol{\alpha}) = \sum_{i=1}^{K} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{K} \alpha_i\alpha_j y_i y_j \kappa\left(\mathbf{Y}_i, \mathbf{Y}_j\right) \tag{3.7}$$

subject to

$$\sum_{i=1}^{K} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \cdots, K$$

where $k\left(x_i, y_i\right)$ is a kernel function. In this work, the RBF kernel function is used. Furthermore, sigma has been determined experimentally to be 14.

e. Ensemble crossover XGBoost. This DT ensemble uses gradient boosting, which allows it to scale very well. Similar to gradient boosting, XGBoost maximises an objective function by minimising a loss function. Due to XGBoost's exclusive reliance on DTs as base classifiers, a modified loss function is used to regulate the tree complexity, as shown in Equations (3.8) and (3.9).

$$\begin{aligned} L_{\text{xgb}} &= \sum_{i=1}^{N} L\left(y_i, F\left(Y_i\right)\right) + \sum_{m=1}^{M} \Omega\left(h_m\right), \\ \Omega(h) &= \gamma T + \frac{1}{-\lambda}\|\omega\|^2. \end{aligned} \tag{3.8}$$

where the leaf output scores are indicated by the symbol ω and T is the number of leaves on the tree. A prepruning method can be created by incorporating this loss function into the split criterion for decision trees. Higher values result in simpler trees. The amount of loss reduction gain needed to split an internal node is determined by γ. In XGBoost, shrinkage is a further regularisation parameter that lowers the additive expansion step size. Lastly, other strategies, like tree depth, can be employed to keep the complexity of the trees to a minimum. Reduced tree complexity has the added benefit of accelerating model training and requiring less storage space.

The number of records classified as normal, uncertain, or abnormal in each of the reference categories is used to calculate the overall score.

These numbers are denoted by $Nn_k, Nq_k, Na_k, An_k, Aq_k, andAa_k$. The various categories of risk level are represented as follows (based on the distribution of the complete test set):

$$\begin{aligned} Va_1 &= \frac{\text{Particular attack abnormal records}}{\text{total abnormal records}}, \\ Va_2 &= \frac{\text{Particular attack abnormal records}}{\text{total abnormal records}}. \\ Vn_1 &= \frac{\text{Dataset cluster normal records}}{\text{total normal records}}, \\ Vn_2 &= \frac{\text{Dataset cluster normal records}}{\text{total normal records}}. \end{aligned} \tag{3.9}$$

The sensitivity and specificity ratio are defined as (based on a subset of the test set) Equations (3.10) and (3.11).

$$\text{SeVa}_1 \frac{Aa_1}{Aa_1 + Aq_1 + An_1} + Va_2 \frac{Aa_2 + Aq_2}{Aa_2 + Aq_2 + An_2} \tag{3.10}$$

$$\begin{aligned} Sp = & Vn_1 \frac{Nn_1}{Na_1 + Nq_1 + Nn_1} \\ & + Vn_2 \frac{Nn_2 + Nq_2}{Na_2 + Nq_2 + Nn_2}. \end{aligned} \tag{3.11}$$

The overall risk score is then the average of these two values:

$$\text{Overall vulnerable score } = \frac{(Se + Sp)}{2}$$

4. Experimental Analysis. The software Matlab has been used to implement the algorithms. The outcomes for each scenario are displayed below.

The simulated output is illustrated in Figure 4.1. The vulnerability was classified as normal and abnormal depending on the obtained risk score. Then, the attack type was identified as MQTT publish flood.

When applied to all of the information that makes up an entire epoch, the loss function yields a numeric estimate of the loss during that time. While developing an iterative curve, some data will inevitably be lost. The resultant curve shows that training and testing the classifier took much less time and effort when compared to previous approaches. Our model is underfitted if there is a substantial gap between the training and validation losses. The training loss may be reduced if more data were included in the sample. (either the overall number of layers or the number of neurons in each layer). Figure 4.2 displays the data we used to calculate the validation loss. However, when evaluating a model's performance on the validation set, the validation loss statistic is the statistic of choice. The validation set is a subset of the data used to evaluate the performance of the model. The sum of all false positives in both the validation set and the training set is the testing loss. The proposed EC-XG boost strategy results in much lower amounts of level loss than the currently available mechanisms.

The simulated output of the vulnerable values in the dataset by the suggested algorithm was demonstrated using a sample illustrated in Figure 4.3.

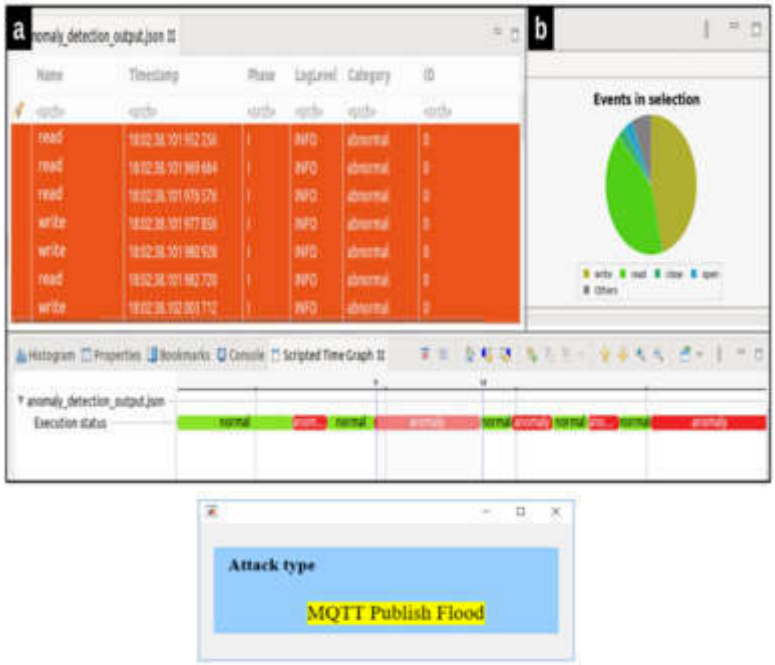


Fig. 4.1: Simulated output

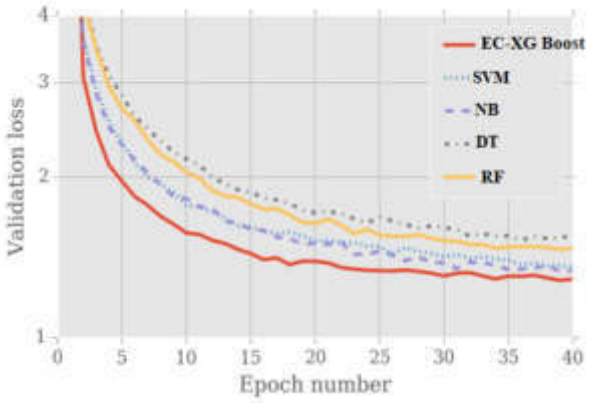


Fig. 4.2: Epoch Vs. Loss

As of from Figure 4.4, training and validation accuracy was calculated. Here, a high level of training and testing accuracy was obtained, showing the mechanism's efficiency.

Some performance measures are shown below that may be used to verify the effectiveness of the proposed technique. The following metrics have been calculated using the equations (4.1) (4.2) (4.3) to evaluate the trained models:

Accuracy: It counts how many potential exploits were accurately identified. How well the findings mirror the

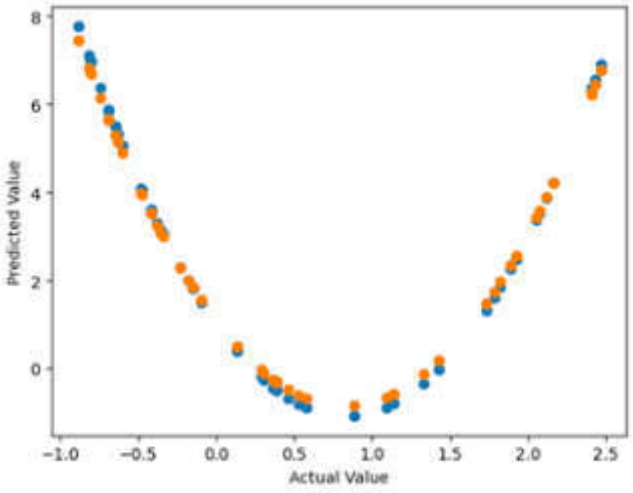


Fig. 4.3: Simulated vulnerability data prediction output

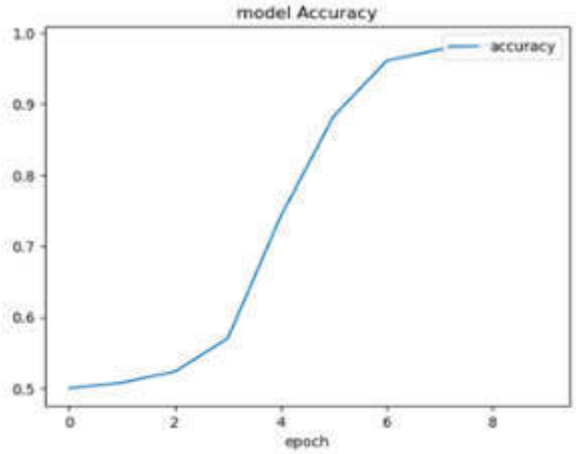


Fig. 4.4: Epoch Vs. accuracy

actual outcomes is determined by this factor.

$$\text{Accuracy} = \frac{(TP + TN)}{(FN + FP + TN + TP)} \tag{4.1}$$

Precision: It determines how accurate the suggested technique behavior is by separating required vulnerable code from the dataset

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{4.2}$$

Recall: The ratio of correctly predicted instances and all instances

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{4.3}$$

Table 4.1: Binary classification comparision

Methodology	**Accuracy**(%)
SVM	64.1 %
NB	65.4 %
DT	89.6%
RF	92.5 %
EC-XG boost	99.64 %

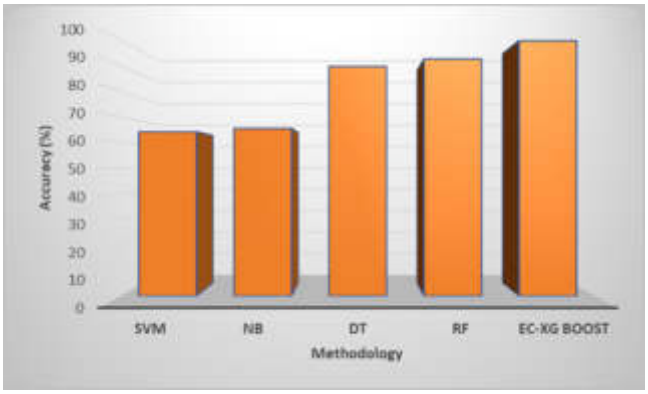


Fig. 4.5: Classification outcome of an algorithm.

Table 4.2: Comparative analysis of the different classifiers

Classes	Decision Tree		SVM		RF		NB		Ensemble XG Boost	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Class 1	87.1	86.2	6.2	68.5	91.7	92.4	14.1	50.2	98.5	99.1
Class 2	84.3	86	9.7	16.5	90.5	91.2	73	10.3	97.2	98.3
Class 3	86.2	87	32.1	10.4	91.3	92.7	25.5	10.1	99.4	98.9
Class 4	87.4	87.3	23.5	11.5	90.4	92.5	80.3	10.5	99.5	99.3

where
True Positive (TP) : actual = 1, predicted = 1
True Negative (TN) : actual = 0, predicted = 0
False Positive (FP) : actual = 1, predicted = 0
False Negative (FN) : actual = 0, predicted = 1

The classification results of the five algorithms applied to the dataset are shown in Figure 3.2. For Ensemble crossover XG boost, RF, and DT, the maximum accuracy was attained at 99.64%, 92.5%, and 89.6%, respectively.

As of from the table 4.1 and figure 4.5 Ensemble crossover XG boost, RF performs the best, with 99.64% and 92.5% accuracy. The importance of highlighting that both techniques exhibit high Recall and Precision per class, as well as increased sensitivity and a low number of false positives, cannot be overstated and is highlighted in Figure 4.6. 99.64% and 92.5% accuracy are obtained using XG boost and RF, respectively. Due to a correlation between the various cases, which prevents a clear differentiation, the remaining algorithms under examination perform worse than those previously analyzed.

The comparative analysis of the different classifiers over precision and recall was done in Figure 4.6 and Table 4.2. From the analysis, the crossover XG boost classifier overcomes the other methodology by obtaining

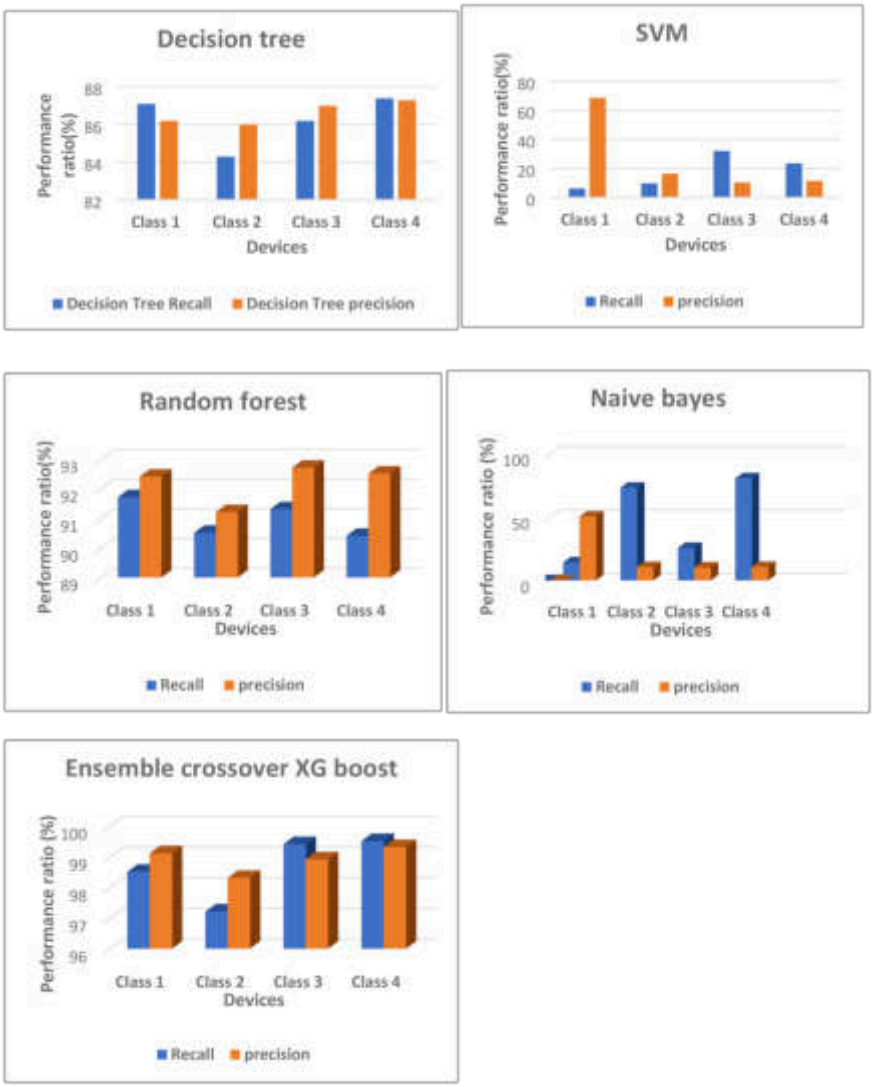


Fig. 4.6: Comparison of Recall and Precision.

a high range of precision and recall.

To prove the efficiency of the suggested methodology it can be compared with the existing methodologies [31].

From the result obtained from the above analysis (See Table 4.3), it was revealed that the suggested EC-XGboost methodology expresses more satisfied results than other existing mechanisms by getting a higher range of performance ratio over CPS vulnerability prediction than the other mechanism in use.

Table 4.3: Comparative performance analysis

S.No	Methods	AUC	TN	FP	FN	TP	Accuracy(%)	F1	Time (s)
1	ResNet	0.8490	82.270	11.320	3.490	2.920	85.190	28.290	6330
2	Inception	0.9610	93.500	0.200	4.100	2.310	95.710	51.760	9760
3	FCN	0.9550	88.160	5.430	3.920	2.490	90.650	34.760	10160
4	MLP	0.7580	72.220	21.370	4.860	1.550	73.770	10.550	1130
5	GC-LSTM + Resnet	0.9740	93.290	0.310	3.270	3.140	96.420	63.770	10560
6	GC-LSTM + Inception	0.9760	92.100	1.490	3.350	3.060	95.160	55.870	14090
7	GC-LSTM + FCN	0.9720	92.280	1.30	3.680	2.730	95.010	52.260	1342. 0
8	GC-LSTM + MLP	0.9370	93.400	0.190	6.130	0.280	93.680	8.140	765. 0
9	CyResGrid	0.9840	93.470	0.130	3.420	2.990	96.450	65.030	714. 0
10	Ensemble crossover XG Boost	0.990	94.0	0.0010	2.000	3.200	99.60	98.90	50

5. Conclusion. To enhance the maintenance of the integrity of CI based on CPS, this work aimed to construct computational mathematics on a pertinent case analysis with appropriate information. The overall evaluation revealed that Ensemble crossover XG boost, RF, and DT outperformed SVM and Naive Bayes in performance. Ensemble crossover XG boost demonstrated the best performance across all algorithms, with 99.64% accuracy in scenario classification. The dataset's instance count should be increased to improve Ensemble crossover XG boost accuracy. Cyber-physical security significantly impacts society, business, and the economy and is crucial to safeguarding vital infrastructure. Protection against cyber threats can be considerably enhanced by awareness of the most recent technology and dangers. To improve scenario identification, rescue operations, and strategic planning, it will be essential for CPS security in the future to automate threat detection and the activation of suitable remedies using Security Orchestration, Automation, and Response (SOAR) systems. Our proposed approach will eventually be used for virtual and distributed Linux deployments. We also aim to use a caching method and batch processing to boost our solution's speed. Each microservice in our architecture will use a "PROSPECT" secure data container, allowing for granular role-based and attribute-based access control to be applied to the settings stored within. A relational database management system would be combined with this.

REFERENCES

[1] A. Aldahiri, B. Alrashed, and W. Hussain, *Trends in using iot with machine learning in health prediction system*, Forecasting, 3 (2021), pp. 181–206.

[2] T. Alladi, V. Chamola, et al., *Harci: A two-way authentication protocol for three entity healthcare iot networks*, IEEE Journal on Selected Areas in Communications, 39 (2020), pp. 361–369.

[3] B. A. Alzahrani, A. Irshad, A. Albeshri, and K. Alsubhi, *A provably secure and lightweight patient-healthcare authentication protocol in wireless body area networks*, Wireless Personal Communications, 117 (2021), pp. 47–69.

[4] N. B. Amor, S. Benferhat, and Z. Elouedi, *Naive bayes vs decision trees in intrusion detection systems*, in Proceedings of the 2004 ACM symposium on Applied computing, 2004, pp. 420–424.

[5] ———, *Naive bayes vs decision trees in intrusion detection systems*, in Proceedings of the 2004 ACM symposium on Applied computing, 2004, pp. 420–424.

[6] M. Binkhonain and L. Zhao, *A review of machine learning algorithms for identification and classification of non-functional requirements*, Expert Systems with Applications: X, 1 (2019), p. 100001.

[7] B. Chandrasekaran, R. Balakrishnan, and Y. Nogami, *Secure data communication using file hierarchy attribute based encryption in wireless body area networks*, Journal of Communications Software and Systems, 14 (2018), pp. 75–81.

[8] Y. Cherdantseva, P. Burnap, A. Blyth, P. Eden, K. Jones, H. Soulsby, and K. Stoddart, *A review of cyber security risk assessment methods for scada systems*, Computers & security, 56 (2016), pp. 1–27.

[9] C. A. A. Dorofee, *Managing information security risks: the octave (sm) approach*, 2002.

[10] B. S. Egala, A. K. Pradhan, V. Badarla, and S. P. Mohanty, *Fortified-chain: a blockchain-based framework for security and privacy-assured internet of medical things with effective access control*, IEEE Internet of Things Journal, 8 (2021), pp. 11717–11731.

[11] A. Fiaschetti, F. Lavorato, V. Suraci, A. Palo, A. Taglialatela, A. Morgagni, R. Baldelli, and F. Flammini, *On the use of semantic technologies to model and control security, privacy and dependability in complex systems*, in Computer Safety, Reliability, and Security: 30th International Conference, SAFECOMP 2011, Naples, Italy, September

19-22, 2011. Proceedings 30, Springer, 2011, pp. 467–479.
[12] G. Garibotto, P. Murrieri, A. Capra, S. De Muro, U. Petillo, F. Flammini, M. Esposito, C. Pragliola, G. Di Leo, R. Lengu, et al., *White paper on industrial applications of computer vision and pattern recognition*, in Image Analysis and Processing–ICIAP 2013: 17th International Conference, Naples, Italy, September 9-13, 2013, Proceedings, Part II 17, Springer, 2013, pp. 721–730.
[13] M. A. M. Hasan, M. Nasser, B. Pal, and S. Ahmad, *Support vector machine and random forest modeling for intrusion detection system (ids)*, Journal of Intelligent Learning Systems and Applications, 6 (2014), pp. 45–52.
[14] J. Homer, A. Varikuti, X. Ou, and M. A. McQueen, *Improving attack graph visualization through data reduction and attack grouping*, in Visualization for Computer Security: 5th International Workshop, VizSec 2008, Cambridge, MA, USA, September 15, 2008. Proceedings, Springer, 2008, pp. 68–79.
[15] W.-H. Hsu, Q. Li, X.-H. Han, and C.-W. Huang, *A hybrid iot traffic generator for mobile network performance assessment*, in 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), IEEE, 2017, pp. 441–445.
[16] C. Hu, H. Li, Y. Huo, T. Xiang, and X. Liao, *Secure and efficient data communication protocol for wireless body area networks*, IEEE Transactions on Multi-Scale Computing Systems, 2 (2016), pp. 94–107.
[17] A. M. Hussain, K. Abualsaud, E. Yaacoub, T. Khattab, A. Gehani, and M. Guizani, *A testbed for implementing lightweight physical layer security in an iot-based health monitoring system*, in 2021 International Wireless Communications and Mobile Computing (IWCMC), IEEE, 2021, pp. 486–491.
[18] M. A. Jan, M. Usman, X. He, and A. U. Rehman, *Sams: A seamless and authorized multimedia streaming framework for wmsn-based iomt*, IEEE Internet of Things Journal, 6 (2018), pp. 1576–1583.
[19] K. Kandasamy, S. Srinivas, K. Achuthan, and V. P. Rangan, *Iot cyber risk: A holistic analysis of cyber risk assessment frameworks, risk vectors, and risk ranking process*, EURASIP Journal on Information Security, 2020 (2020), pp. 1–18.
[20] G. Kavallieratos, S. Katsikas, and V. Gkioulos, *Cybersecurity and safety co-engineering of cyberphysical systems—a comprehensive survey*, Future Internet, 12 (2020), p. 65.
[21] M. H. Khadr, H. B. Salameh, M. Ayyash, H. Elgala, and S. Almajali, *Jamming resilient multi-channel transmission for cognitive radio iot-based medical networks*, Journal of Communications and Networks, 24 (2022), pp. 666–678.
[22] M. Kumar and S. Chand, *Medhypchain: A patient-centered interoperability hyperledger-based medical healthcare system: Regulation in covid-19 pandemic*, Journal of Network and Computer Applications, 179 (2021), p. 102975.
[23] V. Lamba, N. Šimková, and B. Rossi, *Recommendations for smart grid security risk management*, Cyber-Physical Systems, 5 (2019), pp. 92–118.
[24] D. Liu, J. Shen, Y. Chen, C. Wang, T. Zhou, and A. Wang, *Privacy-preserving data outsourcing with integrity auditing for lightweight devices in cloud computing*, in Information Security and Cryptology: 14th International Conference, Inscrypt 2018, Fuzhou, China, December 14-17, 2018, Revised Selected Papers 14, Springer, 2019, pp. 223–239.
[25] X. Lyu, Y. Ding, and S.-H. Yang, *Safety and security risk assessment in cyber-physical systems*, IET Cyber-Physical Systems: Theory & Applications, 4 (2019), pp. 221–232.
[26] G. Macher, E. Armengaud, E. Brenner, and C. Kreiner, *Threat and risk assessment methodologies in the automotive domain*, Procedia computer science, 83 (2016), pp. 1288–1294.
[27] M. Malatji, *Industrial control systems cybersecurity: Back to basic cyber hygiene practices*, in 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), IEEE, 2022, pp. 1–7.
[28] J. Mendel et al., *Smart grid cyber security challenges: Overview and classification*, e-mentor, 68 (2017), pp. 55–66.
[29] S. Pan, T. Morris, and U. Adhikari, *Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data*, IEEE Transactions on Industrial Informatics, 11 (2015), pp. 650–662.
[30] P. Perrone, F. Flammini, and R. Setola, *Machine learning for threat recognition in critical cyber-physical systems*, in 2021 IEEE International Conference on Cyber Security and Resilience (CSR), IEEE, 2021, pp. 298–303.
[31] A. Presekal, A. Ştefanov, V. S. Rajkumar, and P. Palensky, *Attack graph model for cyber-physical power systems using hybrid deep learning*, IEEE Transactions on Smart Grid, (2023).
[32] C. Pu, H. Zerkle, A. Wall, S. Lim, K.-K. R. Choo, and I. Ahmed, *A lightweight and anonymous authentication and key agreement protocol for wireless body area networks*, IEEE Internet of Things Journal, 9 (2022), pp. 21136–21146.
[33] H. Qiu, M. Qiu, M. Liu, and G. Memmi, *Secure health data sharing for medical cyber-physical systems for the healthcare 4.0*, IEEE journal of biomedical and health informatics, 24 (2020), pp. 2499–2505.
[34] I. C. Reinhardt, J. C. Oliveira, and D. T. Ring, *Industry 4.0 and the future of the pharmaceutical industry*, Pharm Eng, (2021), pp. 1–11.
[35] N. Shevchenko, B. R. Frye, and C. Woody, *Threat modeling for cyber-physical system-of-systems: Methods evaluation*, Software Engineering Institute: Pittsburgh, PA, USA, (2018).
[36] S. Surminski, C. Niesler, F. Brasser, L. Davi, and A.-R. Sadeghi, *Realswatt: remote software-based attestation for embedded devices under realtime constraints*, in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 2890–2905.
[37] S. Verma, S. Kaur, M. A. Khan, and P. S. Sehdev, *Toward green communication in 6g-enabled massive internet of things*, IEEE Internet of Things Journal, 8 (2020), pp. 5408–5415.
[38] W. Wang, Q. Chen, Z. Yin, G. Srivastava, T. R. Gadekallu, F. Alsolami, and C. Su, *Blockchain and puf-based lightweight authentication protocol for wireless medical sensor networks*, IEEE Internet of Things Journal, 9 (2021), pp. 8883–8891.
[39] S. Yu and Y. Park, *A robust authentication protocol for wireless medical sensor networks using blockchain and physically unclonable functions*, IEEE Internet of Things Journal, 9 (2022), pp. 20214–20228.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Jul 7, 2023
Accepted: Oct 17, 2023


Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 593–605, DOI 10.12694/scpe.v25i1.2406

NOVEL AUTHENTICATED STRATEGY FOR SECURITY ENHANCEMENT IN VANET SYSTEM USING BLOCK CHAIN ASSISTED NOVEL ROUTING PROTOCOL

ANAND N PATIL*AND SUJATA V MALLAPUR†

Abstract. In recent days, the fast growth of Vehicular Ad-Hoc Network has made smart driving practicable. The VANET performs communication with other vehicles with the intention of improving traffic flow and eliminating accidents and road hazards. VANET is nonetheless susceptible to security attacks by malicious users because of the open wireless nature of the communication channels. As a result, in this proposed system, the novel authenticated strategy for security enhancement in VANET system using block chain assisted novel routing protocol is implemented to prevent from vulnerable attacks. When vehicle enters a new roadside zone, it must continue the reauthentication procedure using the current RSU, which lowers the VANET system's overall effectiveness. Consequently, those mentioned problems are solved via blockchain technology. Reauthentication is effectively accomplished through safe authentication code transmission between the succeeding RSUs thanks to the decentralized nature of blockchain technology. The proposed system's reliability robustness against numerous harmful threats assures that it provides superior security. Furthermore, the Horse Optimization Algorithm based routing protocol assisted with blockchain technology used in this approach significantly reduce the computational and communication cost when compared to traditional approaches. To evaluate the effectiveness of blockchain assisted routing protocol, performance parameters such as PDR, routing overhead, throughput, communication and computational cost with varying scalabilities are measured. According to the findings, the proposed system works better than other existing approaches. In addition to this, our proposed framework substantially guarantees a secure and trustworthy vehicular environment with user privacy preserved.

Key words: Vehicular Ad-hoc Network (VANET), Blockchain, Roadside units (RSU), Horse Optimization Algorithm (HOA)

1. Introduction. VANET is gaining great attention in the exchange of information on smart cars and smart public transit systems. The advent of Vehicular Ad-hoc Networks (VANETs) offers consumers a more practical as well as comfortable driving experience [1]. However, with VANETs, authentication and user privacy continue to be important challenges. Internal vehicle broadcasting of bogus messages must be stopped in order to safeguard vehicles' confidentiality from sneaky attacks. As a result of the interacting channels between wireless nodes being public and unprotected by any form of security measures, VANET is subject to a variety of safety and transparency problems. Furthermore, the traditional manner of transactional data storage lacks distributed and decentralised security, making it possible for a third party to start being dishonest. In order to prevent assaults like physical and eavesdropping attacks, an individual's confidentiality and delicate details, such as actual identity-related data, must be securely protected [2-6]. One of the most harmful attacks is Data Poisoning attacks, which act to reduce the trustworthiness of data interchange [7]. Hence to enhance the trustworthiness, secure and efficient message authentication protocol is used [8]. It tries to achieve mutual authentication across vehicles and roadside units (RSUs). But they fail to meet the need of authenticating hundreds of messages each second in VANETs [9]. In general, the architecture of VANET consist of two layers. The bottom layer involves onboard units and roadside units. The onboard units in the top layer affixed to the vehicles enables wireless communication and the roadside units act as an intermediary-nodes put on roadsides, which act as a link from the top layer and the OBUs. Each vehicle utilizes the OBU to periodically transmit messages about the flow of traffic, including its speed, position, and current road conditions. The RSU gets traffic-related informations from vehicles within its communication range and transmits valid messages to the application server for further analysis.

In an effort to meet the VANETs security standards, several security solutions or systems have been offered in numerous studies, particularly in the previous ten years. Even while such initiatives can meet many

*Assistant Professor, Dept. of CSE, Sharnbasva University, Kalaburagi, Karnataka 585102, India (corresponding author: patilanand1990@gmail.com)
†Professor, Dept.of AI& ML, Sharnbasva University, Kalaburagi, 585102, India (sujatamallapur@gmail.com)

of the security and effectiveness requirements for VANETs, they are prone to a variety of serious reliability, safety, and confidentiality issues. Additionally, most of the strategies proposed are based on the assumption that perpetual delicate data should be stored in a perfect tamperproof device (TPD), that shouldn't ever be breached, physically cloneable, or vulnerable to attacks via side channels by an adversary party [10]. But in reality, this assumption's viability could not be feasible. The confidentiality and safety aspects of the information distribution process in VANETs are significantly impacted by the authentication system. With its plentiful storage and processing resources, certain common technologies, such as cloud computing [11, 12] is utilized which enhances the efficiency. The OBUs that store sensitive data typically validate 1000 to 5000 messages per second for 100 to 500 automobiles are within their contact radius. In this situation, cloud-assisted VANET has substantially helped OBUs to handle the significant processing task load while enhancing traffic efficiency and road safety. However, relational databases, which are typically used by cloud providers to store metadata, are open to privacy violations from the perspective of users' data [13]. In order to solve this, encryption is utilized to safeguard the secrecy of the user, but the inquiry may reveal the user's information and location [14]. Hence, the soft computing approach such as machine learning and deep learning approaches were used. Despite the fact that these two approaches function well in intrusion detection systems [15], the machine learning strategy fails to identify attacks when there are a large number of vehicle nodes in the VANET system [16], and the most intrusion detection system employing deep learning has a longer detection time [17]. To address this aforementioned problem, Block chain is used in this system. The blockchain network's main advantage is that, due to its extremely cheap computational costs, it can successfully address problems with unidentified authentication [18].

To perform optimal route selection to transfer the packet from source to destination node. Routing protocol is used in the VANET system. Due to the resilient nature of vehicle mobility, traditional optimisation algorithms are locked with limited optimal paths [19]. To manage this IoT-based route-discovery process, many optimisation techniques such as the genetic algorithm (GA), particle swarm optimisation (PSO), and cuckoo search optimisation (CSO) [20-23] have been converted from continuous to discrete space. Algorithms' poor convergence has been produced by parameter tweaking and the conversion from continuous to discrete space, which are out of balance with intensification and diversity. Furthermore, its search procedure fails to have exclusivity and selectivity. So, an effective optimization-based routing technique is used in this approach to sort out this problem. Several privacy preserving authentication schemes [24-26] are used to improve security in the VANET system. However, it provides better security, it possess high computational and communication cost.

To sort out all these problems, in this proposed system, a novel authenticated strategy for security enhancement in VANET system using blockchain assisted novel routing protocol is employed. The contribution of our work are as follows:

1. An authentication scheme followed by vehicle initialization and registration is established for the security in VANET system
2. Blockchain-enabled exchange protocol that makes it possible to transfer possession of a vehicle in a distributed, safe manner
3. Detection of malicious node in VANET is developed with enhanced security
4. Lightweight pseudonym management scheme for VANET legitimacy via block chain is developed

The paper is constructed accordingly: An authentication scheme followed by vehicle initialization and registration is done. After authentication, an algorithm is developed to detect the malicious node with enhanced security and finally the section 2 is finished with the pseudonym management scheme The findings and analysis are finished in part 3, and the section's conclusion is made in section 4.

2. Proposed system. The motivation of the proposed system is to facilitate secure and reliable mechanisms for data forwarding. In the proposed system, initially drivers of the vehicles must enter their confidential details about the vehicle directly to the closest Trusted Authority (TA). Only the TA maintains the data with the highest level of security in its database. The TA will track the genuine identity of the vehicle from the fictious identities using this sensitive information in the event disputes. The RSUs and motor units effectively finish their first-time verification with the TA to get an authentication key and the pseudonym identification. After the first step of authentication in the TA is complete according to the authentication code, the RSU

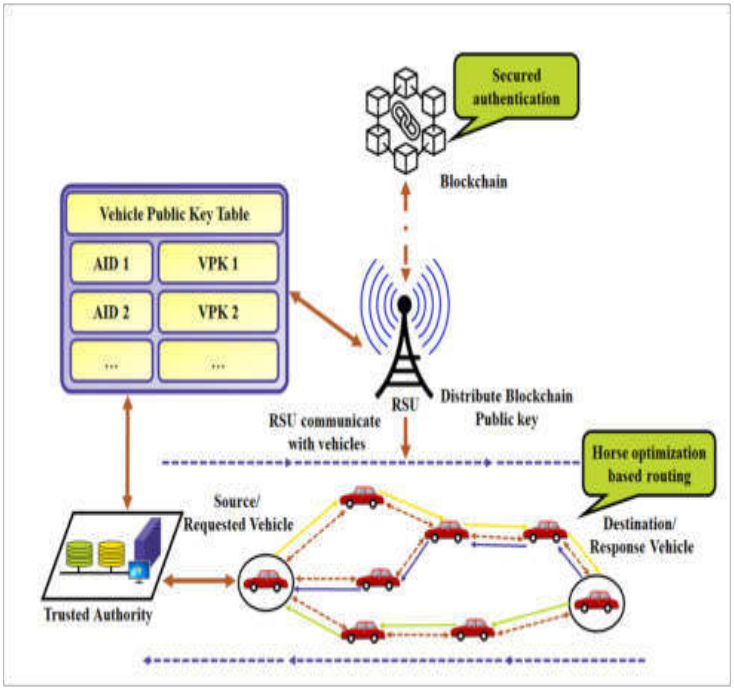


Fig. 2.1: Block diagram of proposed system

executes authentication process in the vehicle using the blockchain network as users enter the service zone of RSU. Then, when the vehicles enter into the communication range of current RSU, the current RSU performs authentication depending upon the handover certificate provided by the former RSU. The vehicle will be given the authentication token after the procedure is accomplished.

After the completion of authentication procedure, routing is performed to select the optimal path when transmission takes place from one user to another. To perform transmission in a secured way, a pseudonym management scheme is performed after the malicious node detection. Hence, in this proposed system, the high level security is achieved with better efficiency.

2.1. Authentication . This section describes the three steps of the envisioned authentication strategy, initialization, registration, and authentication.

2.1.1. Initialization Phase. The system seeks to generate private keys for the trusted authority during its initial setup. The TA results in two looping groupings which is represented as G_1 and G_2. It is generated with order p which fulfills the bilinear map relation $e : G_1 * G_2 \rightarrow G_T$. g_1 and g_2 are considered as the producers of the cyclic groups G_1, G_2 respectively. Furthermore, the trusted authority selects three cryptographics such as H_1, H_2 and H_3, where H_1, H_2 and H_3 is represented by:

$$H_1 : G_1 \rightarrow \{0,1\}^* \tag{2.1}$$

$$H_2\{0,1\}^* \rightarrow Z_q^* \tag{2.2}$$

$$H_3 : \{0,1\}.^* \rightarrow Z_q^* \tag{2.3}$$

The TA generates its public key $(Q_T A)$ by randomly selecting the private key β, where $Q_T A$ is computed as

$$Q_T A = g_1^{\beta} \tag{2.4}$$

Also, these TA generates system parameters such as $p, g_1, g_2, G_1, G_2, G_T, e, H_1, H_2, H_3, Q_T A$.

2.2. Registration phase. For vehicle and RSUs, in-person registration is completed in the TA as detailed below. To perform registration, users of the vehicle have to provide all required personal details, comprising of their contact details, email and postal address. Following the successful submission and validation of the private information by the TA, which selects a random integer $u_i \epsilon Z_q^*$ for every vehicle owner and estimates $PK_v = g_1^(u_j)$ to be the user's public key. Following the automobile's successful registration in the TA, The vehicle user is then covertly given both the private and public key while offline. Additionally, the TA creates a code of authentication for every single user according to the time as given below:

$$AC(t) = \frac{\sum_{i=1}^{n} v_i a_i}{nn} \tag{2.5}$$

Here $v_i \epsilon Z_q^*$ indicates the value of identity selected by the trusted authority for each vehicle user, $a_1 \epsilon Z_q^*$ denotes the identity value which is provided by automobile owner to the TA at the period of registration, and n represents the total amount of vehicles. Once the Authentication code is produced, the TA encrypts the vehicle transaction information with session keys that have been mutually agreed upon by the TAs before broadcasting it to all the trusted authorities. After receiving the transmitted message from the neighbourhood TA, every TA is committed to resolving the riddle. A confirmed authentication code along with the pseudonym is then added as a new block to the end of the block chain when the minor has successfully solved the riddle. In order to protect each vehicle's genuine identity from other network entities, the TA also generates a pseudonym identity for it. The TA aids in performing updation and storing the every vehicle's owner authorization code in the blockchain. Finally, each TA and RSU maintains a duplicate of the most current blockchain in its database as soon as the verified AC has been connected to the block chain in order to examine the new RSUs in the VANET. These values for each vehicle are updated in the block chain.

2.2.1. Authentication phase. This stage of the authentication procedure is done in the vehicle by the RSU and this process is done by using the authentication code which TA generates while the registration procedure is underway. In the authentication phase, As the automobile approaches the RSUs network coverage region, an RSU is needed to verify a vehicle user's validity in an anonymous manner. The vehicle user will provide the pseudonym ID and node number to the RSU once he has arrived at the first RSU region. RSU validates the pseudonym's authentication code and identity. This phase creates a session key between the RSU and the vehicle, as seen below. The car proprietor calculates session key as given in equation

$$SK_v = PK_r^{u_i} \tag{2.6}$$

where $PK_r = g_2^k i$ and k_i represents the public and private key of the RSU. In the same way, the RSU computes session key as given in equation:

$$SK_{r,1} = PK_v^{K_i} \tag{2.7}$$

where $PK_v = g_1^{u_i}$ indicates the vehicle user's public key. After this process, the RSU chooses a master key $r_i \epsilon Z_q^*$ and computes:

$$K_r = g_1^{r_i} \tag{2.8}$$

The value of SK_r is maintained secretly by the RSU. Moreover, the RSU computes:

$$PK_{r,1} = g_2^{r_i} \tag{2.9}$$

The calculated value is delivered to the owner along with a timestamp value T_1.

The driver of the vehicle then confirms the youthfulness of the time-stamp T_1 thereby calculating $SK_{v,1}$ accordingly:

$$SK_{v,1} = PK_{r,1}.SK_V^{(H_2(ID|(|AC|)|T_1||H_1(SK_v)))} \tag{2.10}$$

where, ID identifies the user of the vehicle using a pseudonym. Afterwards, the session key created using the vehicle is computed as follows:

$$SK = e(SK_{v,1}, g_1) \tag{2.11}$$

Following that, the RSU determines the session key accordingly:

$$SK_{r,2} = SK_r.SK_{r,1}^{H_2(ID|(|AC|)|T_1||H_1(SK_r,1))} \tag{2.12}$$

The vehicle is deemed to be outlawed and the RSU fails to provide services if the vehicle's Authentication Code is classified as zero. The blockchain's tamper-resistance capability prevents revoked vehicle consumers from passing for regular vehicles and forbids them exchanging information involving the RSU. The RSU generates the session key as shown below:

$$SK = e(g_2, SK_{r,2}) \tag{2.13}$$

Verification

$$SK = e(g_2, SK_{r,2}) = e(SK_{v,1}, g_1) \tag{2.14}$$

After performing authentication, routing is performed to select the most efficient way to transfer the data between sources and destinations. In this system, HOA based routing technique is used with blockchain technology to perform efficient routing approach that occurs during a safe and evenly dispersed handover of control from one automobile user to another.

2.3. Blockchain assisted HOA based routing protocol. To carry out the routing process, HOA is implemented. The hierarchical structure of horse herds serves as the primary source of inspiration for HOA. Horses favour living in herds. Since many animals coexist in big groups, it is crucial to establish a stable hierarchy in order to promote social cohesion and reduce aggressiveness. The conduct of the horses adopts a hierarchical structure after they form a herd; the animals in the herd with the highest status tend to drink and eat first. Low-status herd members eat later, and occasionally some may not receive enough food. High-ranking horses may have prohibited lower-ranking horses from eating at all in the event that there was little available feed. Horse herds contain a dominant mare and the hierarchy of the horses within a herd determines which horses have priority access to resources. The primary stage of the strategy determines a horse's position in a herd by taking into account its fitness value for that particular herd. Assume there are k horses in the herd, and P stands for a function.

$$Herd = H_1, , H_k \tag{2.15}$$

$$P = Herd \rightarrow 1, , K \tag{2.16}$$

If fitness$(H_x) \leq$ fitness (H_y) where $x \neq$y and $x, y\epsilon 1, ..., k$ then $P(H_x) > P(H_y)$ If fitness H_x=fitness(H_y) where $x \neq y$ and $x, y\epsilon 1, ...k$ then $P(H_x) - P(H_y)](x - y) > 0$ For each horse, the rank is determined as follows:

$$H_x - Rankofeachhorse = (P(H_x))/K \tag{2.17}$$

Every herd has a centre, which is the weighted average of the horse's placement within the herd. The status of the horse is thus represented by the weight. The following formula is used to determine the herds' centre:

$$Herd_{center} = \frac{\sum_{x=1}^{k} z_x H_x.rank}{\sum_{x=1}^{k} H_x.rank} \tag{2.18}$$

Euclidean distance is computed to determine the location between the stallion and the horse herd's center:

$$Dim(stallion, herd) = \sqrt{\sum_{y=1}^{D} im(stallion_y - Herd_{center})^2} \tag{2.19}$$

where Dim is the totality of search space dimensions. If a horse is a member of a herd, its velocity is modified as follows:

$$Vel_{x,y}^{T+1} = Vel_{x,y}^{T} + H_{x.rank} * (Herd_{center,y}^{T} - Z_{x,y}^{T}) \tag{2.20}$$

$$Vel_{x,y}^{T+1} = Vel_{x,y}^{T} + H_{Rand} * (Herd_{center,y}^{T} - Z_{x,y}^{T}) \tag{2.21}$$

Rand is a random value ranging from 0 to 1. The current iteration is denoted by T, while the next iteration is denoted by $T+1$. The horse memory (Mem) is a matrix with the same number of rows as the HMP values in the D column and the horse.

$$Mem_{x}^{T+1} = \begin{bmatrix} Mem_{1,x,1}^{T+1} & .... & Mem_{1,x,D}^{T+1} \\ .... & .... & .... \\ Mem_{HMP,x,1}^{T+1} & .... & Mem_{Hmp,x,1}^{T+1} \end{bmatrix}$$

The memory matrix's cells are updated using this equation.

$$Mem_{k,x,y}^{T+1} = Z_{x,y}^{T+1} * N(0, SD) \tag{2.22}$$

where N stands for a normal distribution (SD) with a mean of zero. Every possible solution or path is assigned a fitness value by HOA. The HOA-based routing technique's objective is to select the route that requires the least amount of energy and travel time. The fitness function, which is illustrated as follows, can be thought of as a minimization function.

$$F - t = minRE_i * DIST_i \tag{2.23}$$

where F_i represents the fitness function of population i, RE_i is the energy required by the i^{th} population. $Dist_i$ represents the overall distance of i^{th} route. The pathways that already exist are initialized as principal populations in HOA. Following are the potential routes.

$$Sol = P_i, i = 1, 2, , N \tag{2.24}$$

The initial set of population is represented by Sol and N indicates the route count. The path's distance and overall energy expenditure are provided by,

$$P = RE, DIST \tag{2.25}$$

where DIST stands for the overall distance and RE stands for the node's remaining energy in the path.

$$RE = f_1 = \sigma_R E = \sqrt{(1/N) \sum_{i=1}^{n} \mu_{RE} - e(node_j)^2} \tag{2.26}$$

$$\mu_{RE} = 1/n \sum_{i=1}^{n} E(node_i) \tag{2.27}$$

when calculating the value of uniform load dispersion across sensors, the standard deviation for RE (σ_{RE}) is used. A routing technique determined by the Horse Optimisation Algorithm (HOA) is used for identifying the best routes depending on fitness function. Additionally, to enable shared memory across network points, the HOA-based routing approach uses blockchain technology.

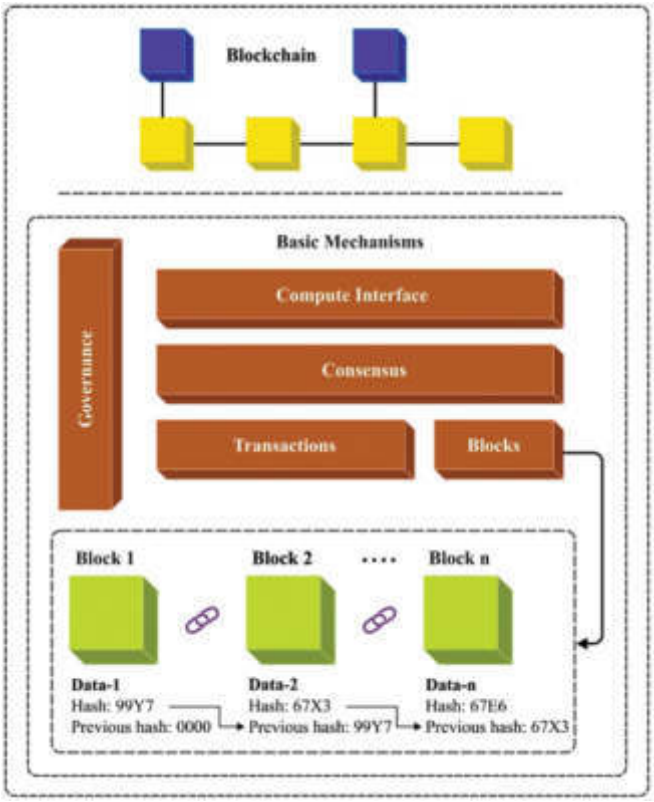


Fig. 2.2: Framework of Blockchain

2.3.1. Blockchain as a shared memory. The blockchain system relies on a ledger that records each transaction moving through a network. As a result, a specific method is needed to identify the nodes that are being moved and the path that they take. On this method, it is possible to save the routes of live, ongoing blockchain transactions. In order to carry out this, the network nodes are treated as coins. To be more precise, a message is transferred from a source to Base Station (BS) by particular nodes, and the originating node influences its proprietorship. Each node has a base Station at the beginning, and all the base station are considered to be inactive. In other words, none of the nodes have an active base station. Then it requests ownership's route node from the BS. The nodes are transferred via the chosen path once the transactions have been registered on the blockchain. The transferring nodes return possession of the route node to the BS, including themselves, once the data has been successfully transported to the BS. This is simply done to let the peer's network know that the communication was successful. Following this, these nodes are released. An origin node, in their opinion, has the capacity to be the owner of u nodes where $u \leq n$. Figure 2.2 illustrates the blockchain's structural components.

2.4. Detection of malicious node in VANET with enhanced security. A node serves as a source which is known to be information generator in VANET communication. Other intermediary nodes between the source and destination operate as relay nodes, while one more node serves as the message's destination. There are third parties in VANETs known as Trusted Authorities (TAs) who offer security. The administration of the network's vehicle identities and the verification of misbehaviour reports received by the verifier nodes fall within the purview of the TA. If the reports are determined to be true, the TAs will adjust the nodes' mistrust values appropriately. Every vehicle has a black list given by the TA that contains an inventory of dangerous nodes, as well as a white list provided by the cluster head for that particular vehicle. To detect the malicious node with enhanced security, following is a description of the Algorithm 1.

Algorithm 1 Algorithm for detection of malicious nodes

1: Initially, vehicle V_N enters the vehicular network.
2: Obtain the cluster keys
3: The parameters such as load, distrust value and distance of nodes in area V_n is to be computed for selecting the verifier.
4: For selecting the verifier, Compute the decision parameter D_p. $D_p = W_1 * L_D + W_2 * D_v + W_3 * D_s$
5: Where $W_1 + W_2 + W_3 = 1$
6: W_1, W_2, W_3 represents weight factors of parameters L_D, D_v, D_s which are Load, Distrust value and Distance accordingly.
7: The nodes are detected with the decision parameter value which is smaller than selection threshold i.e., $(D_p < T_{vs})$.
8: Distribute the nodes which is obtained as verifiers from step 5 to the newly entered vehicle V_N.
9: The vehicle's behaviour is monitored by verifiers
10: If (verifier notices vehicle V_N acting strangely)
11: Notify to the cluster head (CH)
12: jump to step 9;
13: else
14: jump to step 7;
15: Compute cluster head of a new distrust value of vehicle V_N.
16: If the distrust value falls below or equals to detection threshold i.e.,
17: If $(D_v \leq T_M D)$ then
18: the whitelist gets updated and jump to 7
19: else
20: jump to 11
21: alert message is transmitted to all other nodes
22: Perform updation in black list about the entry of vehicle V_N.
23: Remove the discovered malicious vehicle from the network

By this process, if any vehicle found to be acting strangely inside the network, a warning message is sent to all the nodes and update it in the black list about the vehicle's entry. So that the malicious vehicle gets detected and the other vehicle inside the network zone gets alerted.

2.5. Lightweight pseudonym management scheme. The network's vehicle V attaches its public key and uses pseudo id to denote the message's sender, digital signature of the message which was created using its private key and a local table is maintained along with it. This table comprises of valid (PID_v, PK_v)pairs and the termination time for each entry. We point out that performance is impacted by how long the data in a local table is valid for. If the period of time is very brief, the frequency at which the vehicles must check the RSU is increased, increasing the delay. if the time period is lengthy, the local table might have outdated or incorrect information. So, to get rid of this problem, at least once every five minutes, vehicles are required to change their pseudonyms. In order to ensure the freshness of data, a validity period in the LT is to be reasonable. We utilised a validity period of 5 s in our simulations. Every time a message is received, the authentication method is conducted, and the result is a binary value that indicates whether the message is legitimate or not. A message with a valid digital signature sent by an authorized sender is deemed "VALID". The message can be used. If not, the message is deemed false and is immediately deleted (Algorithm 2).

When the RSU receives a request for a (PID_v, PK_v) pair, it searches for the PIDv in the blockchain to confirm the validity of the current public key. The keys and associated pseudo-IDs for each vehicle may vary, and in order to maintain secrecy, these pseudonyms must be constantly altered. Because each item in the LTs has a limited validity period, it is important to make sure that the expired IDs are not being utilised. Additional details about a vehicle, such as misbehaviour complaints and reputation ratings, may also be seen on the blockchain. The proposed solution performs vehicle authentication as an immediate search in the blockchain for its PID. In order to reduce the computational effort significantly, we design an easy-to-use pseudonym

Algorithm 2 Algorithm for detection of malicious nodes

Input: Message retrieved from vehicle v, local table (LT) at receiving vehicle
Output: verification status of received message from v.
Verify the status of PID_v and PK_v in local table
if (PID_v, PK_v) ϵ LT and not terminated then
Sender v is authenticated
else
Fix random wait time t_w and wait
if status message is not retrieved for (PID_v, PK_v) during the time period t_w then
send request message to perform validation and wait for response.
end if
process RSU response
if response reveals 'the message is authentic' then
a. Add (PID_v, PK_v) to LT
b. displays the message 'sender v is authentic'
else
sender v is not authentic
displays received message is not VALID
end if
end if
if sender v is authenticated then
a. Validate digital signature using PK_v
if signature is valid then
received message is VALID
else
received message is not VALID
end if
end if

authentication method.

3. Results and Discussion. The effectiveness of the proposed approach is examined in this section under various circumstances. In the following section, it provides the detailed comparison of proposed system with existing approaches considering packet delivery ratio, throughput, routing overhead. In addition to this, computational cost, communication and storage cost analogization is made in the graph to determine the enhancement in the proposed system.

3.1. Packet delivery ratio (PDR). Table 3.1 and graph representation in figure 3.1 depicts the results attained by proposed routing technique assisted with blockchain on the basis of PDR with various node counts. The novel routing assisted with blockchain shows maximum PDR value whereas, the techniques such as PSO, GA and GWO shows minimum PDR values. As an illustration, during 500 nodes, a maximum PDR of 0.971 is achieved with assist of proposed protocol and other routing protocols shows a minimum PDR value compared to proposed approach.

3.2. Throughput. The suggested approach's throughput analysis under different node counts is shown in Figure 3.2. The findings reveal that the suggested paradigm produces successful results with high throughput.

3.3. Routing overhead. Figure 3.3 compares the routing overhead of different approaches such PSO, GA, GWO and HOA. The overhead incurred in the transmission and reception of packets for the packet transfer rates inside the network is considered in this approach. GWO shows high level of overhead compared to other techniques whereas, the routing protocol used in this proposed approach incur lesser routing overhead. As this approach achieves low overhead, it is efficient in attaining maximum effective throughput.

Table 3.1: Ratio of Packet delivery

Number of nodes	PSO	GA	GWO	HOA
100	0.949	0.961	0.989	0.995
200	0.938	0.956	0.980	0.989
300	0.929	0.949	0.974	0.982
400	0.921	0.938	0.963	0.976
500	0.916	0.925	0.956	0.971

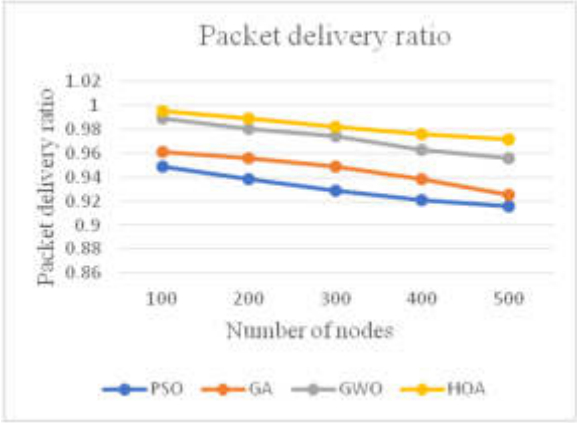


Fig. 3.1: Packet delivery ratio

Table 3.2: Throughput analysis

Number of nodes	PSO	GA	GWO	HOA
100	83.12	87.63	96.54	99.32
200	76.15	81.24	95.46	98.65
300	67.95	75.45	91.00	97.64
400	61.35	68.96	88.96	94.54
500	55.09	60.75	86.75	89.96

3.4. Computation cost. Figure 3.4 makes it clear that the proposed system requires less computational time for authentication than related schemes, as it only needs around 400 ms to authenticate 80 users compared to more than 450 ms for the other existing approaches to verify 80 vehicle users. Additionally, the computation time of the recommended method increases linearly as the number of vehicles increases.

3.5. Communication cost. Figure 3.5 represents the communication cost of proposed system. Comparison is made against existing approaches with varying number of users. The proposed system shows the minimum cost of communication compared to other existing approaches

The proposed system performs better than the conventional approaches with regard to PDR, throughput, routing overhead, communication and computational cost. In addition to this, it provides high level of security with efficient authentication scheme and routing protocol assisted with blockchain. In this study, the blockchain is used to store the nodes' credentials in order to protect network privacy and assure tamper resistance but the limitation of using blockchain is that it uses a lot of resources because there are more transactions happening simultaneously. With the assistance of blockchain based routing protocol, the optimal path is selected which enhances the efficiency of routing as well as the communication and computation cost gets reduced.

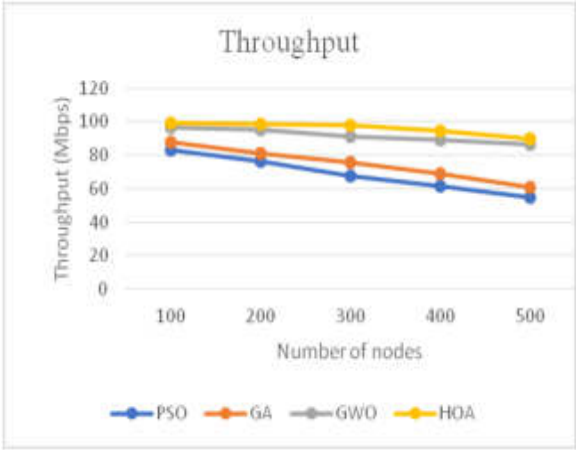


Fig. 3.2: Throughput

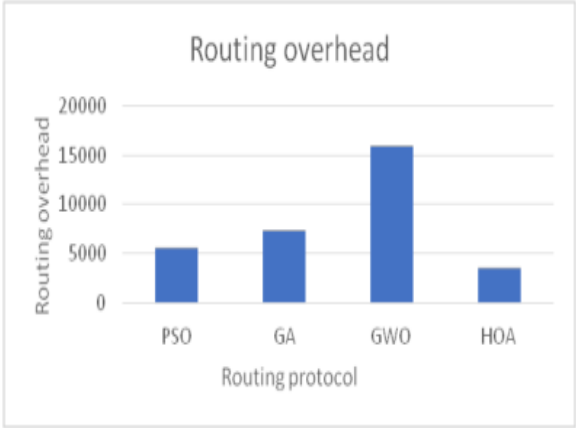


Fig. 3.3: Routing overhead

4. Conclusion. To achieve the best route selection and security in the vehicular ad-hoc network, a unique authenticated technique using block chain assisted routing protocol was developed in the current study. The proposed method involves two main phases, including the authentication and HOA-based routing processes. In order to share network information in real-time situations, the transactions are also kept in the blockchain. A thorough experimental investigation was performed, and the findings were evaluated using different metrics, in order to confirm the presented approach's superiority. The outcomes of the trial indicate that the proposed method outperformed other existing strategies significantly and offers high level security at a more affordable rate for computing and communication. As a part of future scope, we will simulate our proposed framework mechanism on real-time traffic data of vehicle information sharing scenarios.

REFERENCES

[1] Othman S. Al-Heety, Zahriladha Zakaria, Mahamod Ismail, Mohammed Mudhafar Shakir, Sameer Alani, Hussein Alsariera 2020, "A Comprehensive Survey: Benefits, Services, Recent Works, Challenges, Security, and Use Cases for SDN-VANET", IEEE Access, vol. 8, no. 5, pp. 91028-91047.

[2] Sagheer Ahmed Jan, Noor Ul Amin, Mohamed Othman, Mazhar Ali, Arif Iqbal Umar, Abdul Basir 2021, "A Survey on Privacy-Preserving Authentication Schemes in VANETs: Attacks, Challenges and Open Issues", IEEE Access, vol. 9, no.

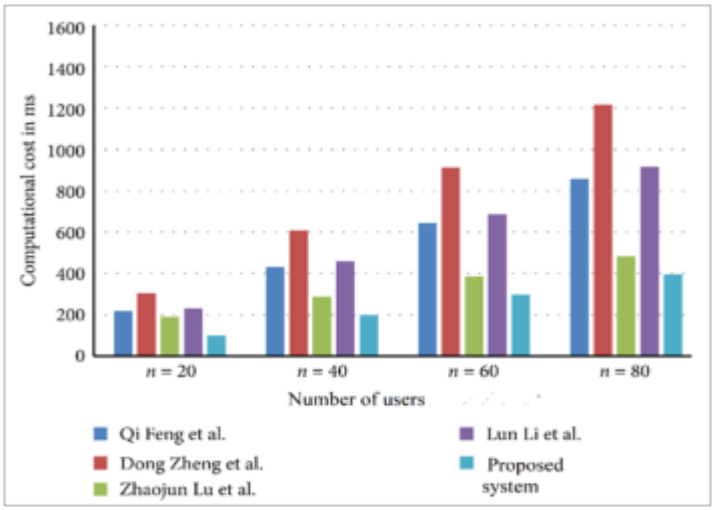


Fig. 3.4: Computational cost

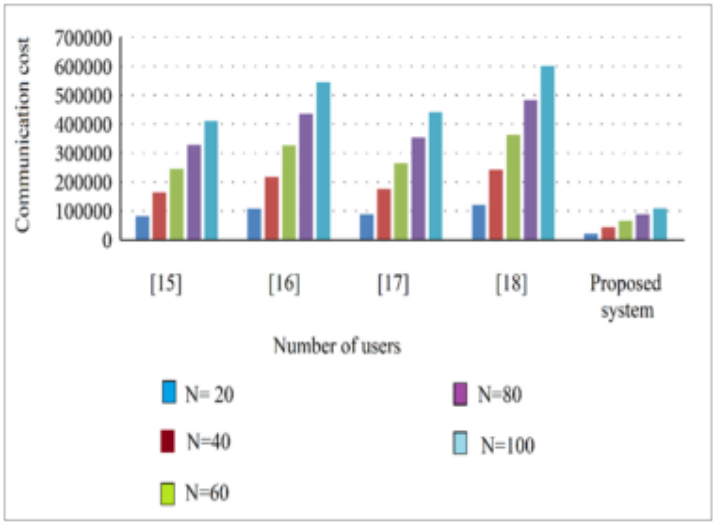


Fig. 3.5: Communication cost

11, pp. 153701-153726.
[3] Waqar Khalid, Naveed Ahmed, Suleman Khan, Zahid Ullah, Yasir Javed 2023, "Simulative Survey of Flooding Attacks in Intermittently Connected Vehicular Delay Tolerant Networks", IEEE Access, vol. 11, no. 7, pp. 75628-75656.
[4] Abdullahi Chowdhury, Gour Karmakar, Joarder Kamruzzaman, Alireza Jolfaei, Rajkumar Das 2020, "Attacks on Self-Driving Cars and Their Countermeasures: A Survey", IEEE Access, vol. 8, no. 11, pp. 207308-207342.
[5] Zhaojun Lu, Gang Qu, Zhenglin Liu 2019, "A Survey on Recent Advances in Vehicular Network Security, Trust, and Privacy", IEEE Transactions on Intelligent Transportation Systems, vol. 20, no.2, pp. 760-776.
[6] Fengzhong Qu, Zhihui Wu, Fei-Yue Wang, Woong Cho 2015, "A Security and Privacy Review of VANETs", IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 12, pp. 2985-2996.
[7] Carlos Pedroso, Thiago S. Gomides, Daniel L. Guidoni, Michele Nogueira, Aldri L. Santos 2022, "A Robust Traffic Information Management System Against Data Poisoning in Vehicular Networks", IEEE Latin America Transactions, vol. 10, no. 12, pp. 2421-2428.
[8] Peng Wang, Yining Liu 2021, "SEMA: Secure and Efficient Message Authentication Protocol for VANETs", IEEE Systems Journal, vol. 15, no. 1, pp. 846-855.

[9] Kyung-Ah Shim 2023, "Security Analysis of Conditional Privacy-Preserving Authentication Schemes for VANETs", IEEE Access, vol. 11, no. 4, pp. 33956-33963.
[10] Tao Zhang, Quanyan Zhu 2018, "Distributed Privacy-Preserving Collaborative Intrusion Detection Systems for VANETs", IEEE Transactions on Signal and Information Processing over Networks, vol. 4, no. 3, pp. 148-161.
[11] Yujue Wang, Yong Ding, Qianhong Wu, Yongzhuang Wei, Bo Qin, Huiyong Wang 2019, "Privacy-Preserving Cloud-Based Road Condition Monitoring With Source Authentication in VANETs", IEEE Transactions on Information Forensics and Security, vol. 14, no. 7, pp. 1779-1790.
[12] Muthukumaran, V., Kumar, V. V., Joseph, R. B., Munirathanam, M., & Jeyakumar, B. (2021). Improving network security based on trust-aware routing protocols using long short-term memory-queuing segment-routing algorithms. International Journal of Information Technology Project Management (IJITPM), 12(4), 47-60.
[13] Ruba Awadallah, Azman Samsudin 2021, "Using Blockchain in Cloud Computing to Enhance Relational Database Security", IEEE Access, vol.9, no. 10, pp. 137353-137366.
[14] Sultan Almakdi, Brajendra Panda, Mohammed S. Alshehri, Abdulwahab Alazeb 2021, "An Efficient Secure System for Fetching Data From the Outsourced Encrypted Databases", IEEE Access, vol. 9, no. 5, pp. 137353-137366.
[15] Edivaldo Pastori Valentini, Geraldo Pereira Rocha Filho, Robson Eduardo De Grande, Caetano Mazzoni Ranieri, Lourenço Alves Pereira Júnior, Rodolfo Ipolito Meneguette 2023, "A Novel Mechanism for Misbehavior Detection in Vehicular Networks", IEEE Access, vol.11, no. 7, pp. 68113-68126.
[16] Aekta Sharma, Arunita Jaekel 2022, "Machine Learning Based Misbehaviour Detection in VANET Using Consecutive BSM Approach", IEEE Open Journal of Vehicular Technology, vol. 3, no.12, pp. 1-4.
[17] Natarajan, Rajesh, Gururaj Harinahallo Lokesh, Francesco Flammini, Anitha Premkumar, Vinoth Kumar Venkatesan, and Shashi Kant Gupta. "A Novel Framework on Security and Energy Enhancement Based on Internet of Medical Things for Healthcare 5.0." Infrastructures 8, no. 2 (2023): 22..
[18] Zhaojun Lu, Qian Wang, Gang Qu, Haichun Zhang, Zhenglin Liu 2019, "A Blockchain-Based Privacy-Preserving Authentication Scheme for VANETs", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 12, pp. 2792-2801.
[19] Maithili, K., Vinothkumar, V., & Latha, P. (2018). Analyzing the security mechanisms to prevent unauthorized access in cloud and network security. Journal of Computational and Theoretical Nanoscience, 15(6-7), 2059-2063.
[20] Meie Shen, Zhi-Hui Zhan, Wei-Neng Chen, Yue-Jiao Gong, Jun Zhang, Yun Li 2014, "Bi-Velocity Discrete Particle Swarm Optimization and Its Application to Multicast Routing Problem in Communication Networks", IEEE Transactions on Industrial Electronics, vol. 61, no. 12, pp. 7141-7151.
[21] Shahid Abbas, Nadeem Javaid, Ahmad Almogren, Sardar Muhammad Gulfam, Abrar Ahmed, Ayman Radwan 2021, "Securing Genetic Algorithm Enabled SDN Routing for Blockchain Based Internet of Things", IEEE Access, vol. 9, pp. 139739-139754.
[22] Muthukumaran, V., Vinoth Kumar, V., Joseph, R. B., Munirathnam, M., Beschi, I. S., & Niveditha, V. R. (2022, November). Efficient Authenticated Key Agreement Protocol for Cloud-Based Internet of Things. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3 (pp. 365-373). Singapore: Springer Nature Singapore..
[23] Q. Feng, D. He, S. Zeadally, and K. Liang 2020, "BPAS: blockchain-assisted privacy-preserving authentication system for vehicular ad hoc networks," IEEE Transactions on Industrial Informatics, vol. 16, no. 6, pp. 4146–4155.
[24] D. Zheng, C. Jing, R. Guo, S. Gao, and L. Wang 2019, "A traceable blockchain-based access authentication system with privacy preservation in VANETs," IEEE Access, vol. 7, pp. 117716–117726.
[25] L. Li, J. Liu, L. Cheng et al 2018, "a privacy-preserving blockchain-based incentive announcement network for communications of smart vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 7, pp. 2204–2220.
[26] Z. Lu, Q. Wang, G. Qu, H. Zhang, and Z. Liu, "A blockchain-based privacy-preserving authentication scheme for VANETs," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 12, pp. 2792–2801, 2019.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Jul 7, 2023
Accepted: Oct 9, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 607–613, DOI 10.12694/scpe.v25i1.2970

CONFIGURATION OF CONTAINER DEPLOYMENTS ON THE COMPUTE CONTINUUM USING ALIEN4CLOUD

ADRIAN SPĂTARU* AND JULEN APERRIBAY†

Abstract. This paper presents the drawbacks and benefits of using Alien4Cloud as a platform for deploying container-based applications on the Compute Continuum. To achieve this, a plugin has been developed to deploy container-based applications in multiple Kubernetes clusters and to configure the containers based on their dependencies. More specifically, our implementation has been validated using a system of two Kubernetes clusters (one in Romania, and one in Spain) and a machine-learning application that has been successfully deployed using this system.

Key words: Compute Continuum, Service Orchestration

AMS subject classifications. 68M14, 68M15

1. Introduction. Alien4Cloud [9] is a mature platform built for the Development and Operations (DevOps) paradigm. It uses TOSCA (Topology Specification for Cloud Applications), an OASIS specification for modelling a complete application stack and automating its deployment and management in Cloud environments [6]. Using TOSCA, the administrator uploads component specifications and uses these definitions to create an application topology.

Alien4Cloud (A4C) offers an easy-to-use interface to Drag&Drop TOSCA components and link requirements of one with the capabilities of another. Nevertheless, this is not the main feature of the platform. Some components can register as services in the A4C Service Registry, which allows other applications to connect to these registered services. An example is explained in detail in Section3.2. Moreover, the platform's functionality can be extended through plugins. An Orchestrator plugin is responsible for deploying and monitoring an application topology, and is able to configure the parameters of the components before they start.

We use an extension of TOSCA, namely *TUFA* [13] and implement an Orchestrator plugin to manage Kubernetes deployments for the Container definitions in the TUFA specification. We validate our plugin using an industrial application that detects anomalies in the manufacturing process.

The key contributions of this paper are:

- the implementation of the Orchestrator plugin that transforms TUFA specifications in Kubernetes descriptions and interacts with the Kubernetes Orchestrator to deploy and monitor applications.
- the validation of the developed plugin using two Kubernetes clusters and a machine learning application
- an experience-driven assessment of the benefits and drawbacks of using a platform like Alien4Cloud for the management of deployments in the Compute Continuum setting

The paper is structured as follows. Section 2 presents the work behind the popularity of the TOSCA specification and the alternatives to Alien4Cloud. Section 3 describes the system architecture and details the use case scenario. Section 4 presents the benefits and drawbacks of Alien4Cloud as a platform for developing future-proof Orchestration plugins. Finally, Section 5 draws the conclusion.

2. Related Work. Numerous specifications in specific domains address the modelling of Cloud infrastructure, services, relationships, and policies for monitoring and scaling. A comprehensive review examining Cloud Modelling Languages (CML), as documented by Bergmayr et al. [4], highlighted TOSCA as the predominant

*Department of Computer Science, West University of Timişoara, Romania (Corespondig author: adrian.spataru@e-uvt.ro).

†IDEKO, Spain (japerribay@ideko.es)

one. Although certain CMLs [2, 8] predate the TOSCA standard, and some emerged concurrently [1, 3, 5, 12], no new modelling languages appeared since 2014. Furthermore, extensions of existing CMLs aim to integrate with the TOSCA specification, resulting in a convergence of cloud modelling semantics.

With the rise of Fog and Edge computing, the TOSCA model has attracted various proposals for extensions, focusing on diverse models for these resources. One approach involves using these resources as a backend for serverless computing [14]. The serverless paradigm allows users to define the topology in terms of application fragments, which are distributed across the resource fabric using load-balancing techniques.

SODALITE [7, 11] employs a static service optimization process based on prior application runs using different configurations.

In the realm of deploying TOSCA-based applications, several alternatives stand out as viable options to Alien4Cloud. Cloudify[1] is an open-source orchestration platform compliant with TOSCA standards. It offers capabilities for modeling, deployment, and management across hybrid cloud environments. Heat[2], a component-based orchestration tool for OpenStack environments, utilizes TOSCA templates to describe the necessary infrastructure and services for applications.

The previously mentioned alternatives are mature developments that provide extreme granularity with respect to the configuration of Virtual Machines and infrastructure in general. Our approach focuses on the DevOps perspective which shifts fast to the development of microservices which run in Containers. The most popular container orchestration engines Kubernetes, which is a resource generally managed outside of the DevOps pipeline, by a system administrator.

An alternative solution is provided by the developers of Alien4Cloud, more specifically the Kubernetes plugin[3]. One drawback that we identified with this plugin is the fine granularity for modelling applications. The application developer is required to exceed his expertise and configure Kubernetes entities, which may not be possible for all application developers. Our approach simplifies this knowledge gap and aids the application developer in defining the relations between software components, not Kubernetes infrastructure.

The TUFA specification [13] has definitions for microservices packaged as containers, and this paper presents their integration with the Kubernetes orchestration engine and the use of A4C services for the automatic configuration of containers based on their dependencies.

3. System Architecture. Various components play essential roles in managing the Application life cycle, as depicted in Figure 3.1. The Alien4Cloud platform serves as the runtime environment, supplemented with TUFA definitions and the Orchestrator plugin. The **User Interface** facilitates service definition management and application deployments. Additionally, the **Components Repository** serves as the storage hub for *Component Definitions*, encompassing requirements and dependencies. The *Application Developer* uses this component repository to store such definitions, which subsequently aids the *Application Operator* in creating a new Application Topology from scratch or starting from one created by the developer in the **Topology Repository**.

The **Orchestrator plugin**[4] represents our main contribution and is developed using the Alien4Cloud plugin development guidelines[5]. The plugin allows for the creation of multiple Orchestrators, each connected to a corresponding Kubernetes cluster. The *fabric8*[6] java library is used for communicating with the Kubernetes clusters via HTTP REST. The **Service Registry** allows the administrator to register external services that provide different functionalities like storage services or message queuing services. The administrator authorises each Orchestrator to use a given service. If the service is internal to the Kubernetes Cluster where it runs, then the administrator will only allow the corresponding Orchestrator to use it. It makes little sense to allow other Orchestrators to use this because the deployed components cannot access the service. By default, no Orchestrator is authorised to use any of the services registered with the platform.

[1] https://docs.cloudify.co
[2] https://docs.openstack.org/heat
[3] https://github.com/alien4cloud/alien4cloud-kubernetes-plugin
[4] https://github.com/adispataru/tufa-a4c-orchestrator
[5] https://alien4cloud.github.io/#/developer_guide/plugin.html
[6] https://github.com/fabric8io/kubernetes-client

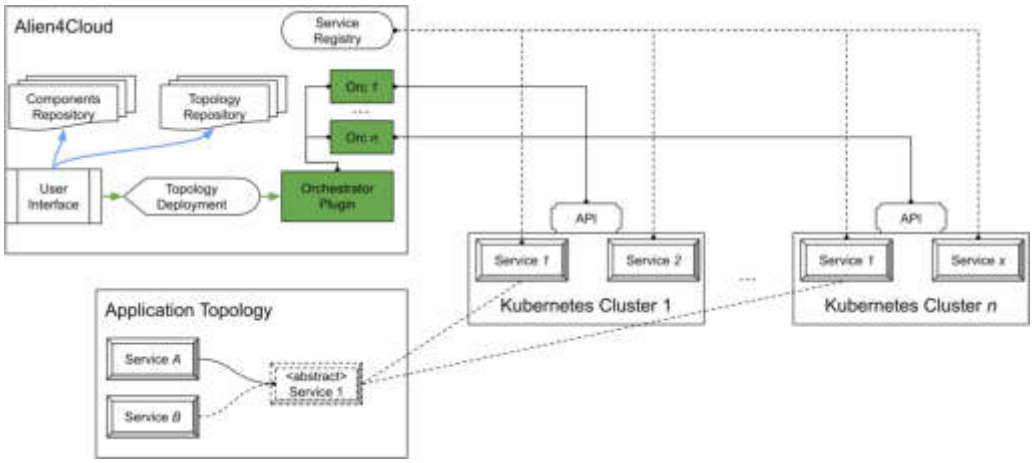


Fig. 3.1: System architecture

Fig. 3.2: Application definition using TUFA definitions in Alien4Cloud Topology Editor

3.1. Anomaly detection in manufacturing use-case. To better illustrate the advantages of the system, we present an example application. The application monitors data coming from a ball-screw component of a manufacturing machine and detects anomalies based on the position and acceleration of the ball-screw.

Figure 3.2 presents the application modelled using TUFA specification and viewed in the Alien4Cloud Topology Editor screen.

The application consists of 5 components:

- Data Manager component – responsible for locally persisting the data sent from the manufacturing machine, and uploading it to Secure Storage using the S3 interface. Additionally, this component monitors the accuracy of predicted anomalies. When the accuracy falls below a given threshold, the Data Manager uploads the local data to Secure Storage and contacts the Training component to use it and train a new model. This component uses a persistent volume to persist data locally and use it in case of recovery (e.g., container restart).
- Training component – responsible for training the machine-learning models and storing them using the S3 interface.
- Inference component – responsible for applying the trained machine learning model on new data.
- Data Broker – this component is an MQTT [10] service which directs messages. The industrial machine connects to this component and send data related to the position and acceleration of the ball-screw. The messages are redirected to the Data Manager component that schedules inference operations in

batch.

- Secure Storage – this component exposes an S3-compatible interface that allows users to store data. This component stores data used for training machine-learning models, the machine learning models used for inference, and the predicted anomalies.

All of the first three components (Data Manager, Training and Inference) depend on the other two for transferring messages and sending data. More concretely, the first three need to know the endpoint and credentials for routing messages and upload data. This configuration is achieved by binding two pieces of information: the *Config Map* data of a component and the inputs declared in the TUFA definition. More specifically, for the Data Manager component, the Config Map data consists of two YAML documents, one for configuring the application parameters and one for configuring the application logging mechanism. The training and inference component use similar configuration. An excerpt from the configuration file related to application parameters is presented in Listing 1.

Listing 1: Example definition of fields in the Config Map of the Data Manager component

```
mqtt:
  host: INPUT_DATABROKER_IP
  port: INPUT_DATABROKER_PORT
  keepalive: 60
  credentials:
    username: INPUT_DATABROKER_USER
    password: INPUT_DATABROKER_PASSWORD
storage:
  type: s3
  bucket_name: ideko-uc3-anomaly-detection
  credentials:
    gateway_url: INPUT_GATEWAY_URL
    skyflok_token: INPUT_GATEWAY_PASSWORD
```

The developer of the component will need to declare the values that require replacement (e.g. *INPUT_DATABROKER_IP*) as inputs of the *create* operation of the TUFA definition. An example is presented in Listing 2 for the Training component, but the Data Manager and the Inference components have the same inputs. The three only differ in the implementation chosen for creating the container, more specifically, the docker image used for deployment.

Listing 2: Example definition of inputs for the fields present in the Config Map of the Training component

```
interfaces:
  Standard:
    create:
      inputs:
        INPUT_DATABROKER_IP: { get_property: [REQ_TARGET, mqtt, ip_address]}
        INPUT_DATABROKER_PORT: { get_property: [REQ_TARGET, mqtt, port]}
        INPUT_DATABROKER_USER: { get_property: [REQ_TARGET, mqtt, user]}
        INPUT_DATABROKER_PASSWORD: { get_property: [REQ_TARGET, mqtt, password]}
        INPUT_GATEWAY_URL: { get_property: [REQ_TARGET, storage, url]}
        INPUT_GATEWAY_PASSWORD: { get_property: [REQ_TARGET, storage, password]}
      implementation:
        file: serrano/acceleration-classifier-training:0.1
        repository: serrano
        type: tufa.art.Deployment.Image.Container
```

For the *INPUT_DATABROKER_IP* parameter, the *get_property* function receives 3 arguments:

1. *REQ_TARGET* specifies that the property is to be examined in the target which satisfies a requirement of the current component
2. *mqtt* specifies the name of requirement
3. *ip_address* specifies the name of the property found in the requirement target.

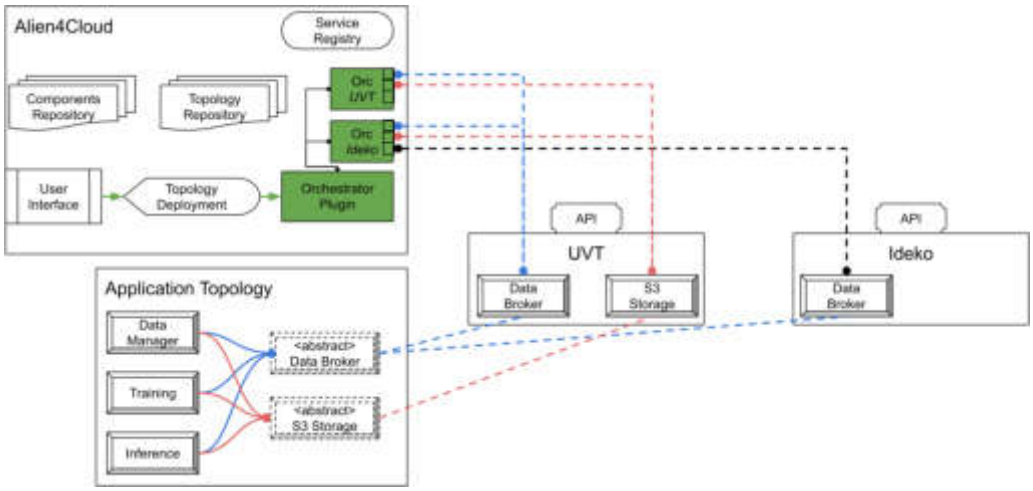


Fig. 3.3: System configuration for the validation of the plugin

The target that satisfies the *mqtt* requirement is the Data Broker component. Therefore, the IP address of the Data Broker will replace the *INPUT_DATABROKER_IP* text in the Config Map of the corresponding service. To achieve this during run-time, the dependency (i.e. Data Broker) is deployed before the dependent component (e.g. Training component) such that the IP of the dependency is known and can be injected in the configuration of the dependent component. However, the Data Broker and Secure Storage services presented in Figure 3.2 are abstract, and the Alien4Cloud platform will find a match for them using the Service Registry.

It is essential to observe the green letter "a" that appears in the boxes corresponding to the Data Broker and Secure Storage services from Figure 3.2. The green letter "a" means that the two services are abstract: they don't need to be deployed together with the other three components. Rather, they should be *matched* with Services that have been registered with Alien4Cloud.

3.2. System under test. For a more comprehensive understanding of the mechanism that matches components, we present the system under test. We have installed Alien4Cloud and have uploaded the Orchestrator plugin and the TUFA specifications for the use case.

Figure 3.3 offers a concrete view on the system. Two Kubernetes clusters are registered the with the Orchestrator. One Kubernetes cluster is installed at a university, Universitatea de Vest din Timisoara (UVT), in the west of Romania, consisting of three workers using HPE ProLiant DL385 Gen10 hardware. The second cluster is installed at a company, IDEKO, in the north of Spain, and consists of 3 Edge devices developed by IDEKO and installed on milling machines. The specific performance of the clusters is not the scope of this paper since performance of the application is not under scrutiny. Nevertheless, it's worth mentioning that an IDEKO worker has 4 cores and 8 GB of RAM, a pretty powerful edge device. The UVT worker has four times the number of cores and eight times the size of RAM, framing this as a Fog infrastructure provider.

There are two Orchestrator instances, one for UVT and one for IDEKO. In the UVT cluster, we have the two services that our use-case application depends on: the Data Broker and the Storage Service. The services have public access based on authentication and thus both Orchestrators have access to the two services.

The Data Broker instance deployed on Ideko premises also uses an authentication mechanism but does not have a public IP, thus only containers deployed on the Ideko Kubernetes cluster can reach it. In this case, only the Orchestrator instance corresponding to the Ideko cluster is authorised by the Alien4Cloud administrator to use the Ideko Data Broker.

3.3. Application Deployment. After using the Topology Editor as presented in Figure 3.2, the Alien4-Cloud interface leads the user through several decisions before deployment, out of which two are of interest:

1. Locations – The user needs to select the location for the deployment of the application (i.e., UVT or

Table 3.1: Configurations and functional tests

Test	Explanation
Container Image	Container image can be downloaded from a private repository. If the credentials for the repository do not exist in the Kubernetes cluster, they are created.
Persistent Volume Claim creation	Persistent Volume Claims (PVC) can be created. In the case of UVT, persistent volume claims are managed dynamically by a storage class. In the case of Ideko, local Persistent Volumes are created by our plugin apriori to satisfy the PVC.
Container binding to PVC	If a component requests a volume attachment, then the PVC is created and bound to the container at start-up
Config Map configuration	The Config Map has been successfully updated to replace the placeholder information with actual IP and credentials for the services matched in Alien4Cloud.
Matching services connection	The deployed components can access and interact with the services selected during the matching process.
Cleanup	All deployments, Config Maps and ephemeral volume claims are deleted from the Kubernetes cluster.

Ideko)

2. Matching – Abstract components are matched against the Service Registry; the user needs to select which service to use for each corresponding abstract component

Depending on the selected location, Alien4Cloud filters out matching nodes. In our experiment, if the user chooses the UVT location, the only options are to use both the Data Broker and the Storage services deployed at UVT. If the user selects the Ideko location, then both the UVT and Ideko instances of the Data Broker are available for the user to choose from. The user will select the Ideko instance to speed up the communication between these components. However, for the Storage service no instance is available on the Ideko premises. Thus, the components will use the instance deployed at UVT (since it is publicly available).

Finally, the application has been successfully deployed using all possible configurations and all functional tests presented in Table 3.1 have been passed.

4. Discussion. During the development of the plugin, we have encountered many circumstances, and we reckon the reader is interested in the details. This section presents the drawbacks and benefits of using Alien4Cloud to deploy applications on the Compute Continuum.

Benefits:

- Fit for purpose – in the context of the current investigation, Alien4Cloud solved an important problem in our use-case, most importantly the matching of external services
- Easy development – The plugin development process is relatively easy for an user which has experience with TOSCA and Java syntax. Documentation is accessible and well written.
- Built-in authorisation – The platform provides a complex mechanism for authorisation based on users and groups, handling access to all resources: applications, services, locations.

Drawbacks:

- No official maintenance – during the development of our plugin, the developers of Alien4Cloud announced the final software release with no future maintenance.
- Hard maintenance – the platform has been originally developed using Java 1.8, and several soft breaking changes have been introduced since the introduction of modules in Java 1.9. The breaking changes have not been solved until the final release, and the most recent Java version that can be used to compile a plugin for the platform is version 1.15.
- Design limitations – Some limitations are inherited from the original purpose of deploying applications in Virtual Machines. This did not impose any limits for deploying Containers, but future standards can not be accommodated since maintenance reached its end.

Overall, the platform offers the intended functionality with few efforts.We successfully extended its functionality through TOSCA based definitions and our Orchestrator plugin. The extended functionality allowed

us to achieve our goal of deploying containers in two locations on the Compute Continuum: Edge and Fog.

However, the end of official maintenance for the Alien4Cloud project imposes hard effort to invest in migrating the codebase to a more recent version of Java.

5. Conclusion. This paper investigated the suitability of the Alien4Cloud platform in managing the configuration of container deployments on the Compute Continuum. In oder to achieve our goal, the TUFA extension to the TOSCA specification has been used to define container-based applications, and an Orchestrator plugin has been developed to interact with the Kubernetes API in order to deploy an application topology.

Our experimental setup involved two Kubernetes clusters, one with higher performance and internet bandwidth, considered a Fog node, and one with lower performance, consisting of three Edge devices. A machine learning application for anomaly detection in the manufacturing industry has been successfully been deployed to both clusters. Moreover, the components of an application can access services deployed across clusters with the help of the Orchestrator plugin developed.

If an official maintainer is found for the Alien4Cloud platform, it can be an important piece in the puzzle of deploying applications on the Compute Continuum.

Acknowledgement. This work was partially supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-P4-ID-PCE-2020-0407, within PNCDI III. The work has been partially supported by a grant of European Union's Horizon 2020 Research and Innovation programme under grant agreement No 101017168, acronym SERRANO.

REFERENCES

[1] V. Andrikopoulos, A. Reuter, S. Gómez Sáez, and F. Leymann, *A gentl approach for cloud application topologies*, in European Conference on Service-Oriented and Cloud Computing, Springer, 2014, pp. 148–159.

[2] D. Ardagna, E. Di Nitto, P. Mohagheghi, S. Mosser, C. Ballagny, F. D'Andria, G. Casale, P. Matthews, C.-S. Nechifor, D. Petcu, et al., *Modaclouds: A model-driven approach for the design and execution of applications on multiple clouds*, in 2012 4th International Workshop on Modeling in Software Engineering (MISE), IEEE, 2012, pp. 50–56.

[3] T. Benson, A. Akella, A. Shaikh, and S. Sahu, *Cloudnaas: a cloud networking platform for enterprise applications*, in Proceedings of the 2nd ACM Symposium on Cloud Computing, 2011, pp. 1–13.

[4] A. Bergmayr, U. Breitenbücher, N. Ferry, A. Rossini, A. Solberg, M. Wimmer, G. Kappel, and F. Leymann, *A systematic review of cloud modeling languages*, ACM Computing Surveys (CSUR), 51 (2018), pp. 1–38.

[5] A. Bergmayr, J. Troya Castilla, P. Neubauer, M. Wimmer, and G. Kappel, *Uml-based cloud application modeling with libraries, profiles, and templates*, in CloudMDE 2014: 2nd International Workshop on Model-Driven Engineering on and for the Cloud co-located with the 17th International Conference on Model Driven Engineering Languages and Systems (MoDELS 2014)(2014), p 56-65, CEUR-WS, 2014.

[6] T. Binz, U. Breitenbücher, O. Kopp, and F. Leymann, *Tosca: portable automated deployment and management of cloud applications*, in Advanced Web Services, Springer, 2014, pp. 527–549.

[7] E. Di Nitto, J. Gorroñogoitia, I. Kumara, G. Meditskos, D. Radolović, K. Sivalingam, and R. S. González, *An approach to support automated deployment of applications on heterogeneous cloud-hpc infrastructures*, in 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), IEEE, 2020, pp. 133–140.

[8] X. Etchevers, T. Coupaye, F. Boyer, and N. De Palma, *Self-configuration of distributed applications in the cloud*, in 2011 IEEE 4th International Conference on Cloud Computing, IEEE, 2011, pp. 668–675.

[9] FastConnect, *Application lifecycle enablement for cloud*. online, 2016. Accessed July 25, 2017.

[10] U. Hunkeler, H. L. Truong, and A. Stanford-Clark, *Mqtt-s—a publish/subscribe protocol for wireless sensor networks*, in 2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08), IEEE, 2008, pp. 791–798.

[11] I. Kumara, G. Quattrocchi, D. Tamburri, and W.-J. V. D. Heuvel, *Quality assurance of heterogeneous applications: The sodalite approach*, in European Conference on Service-Oriented and Cloud Computing, Springer, 2020, pp. 173–178.

[12] G. C. Silva, L. M. Rose, and R. Calinescu, *Cloud dsl: A language for supporting cloud portability by describing cloud entities.*, in CloudMDE@ MoDELS, 2014, pp. 36–45.

[13] A. Spătaru, G. Iuhasz, and S. Panica, *Tufa: A tosca extension for the specification of accelerator-aware applications in the cloud continuum*, in 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, 2022, pp. 1178–1183.

[14] A. Tsagkaropoulos, Y. Verginadis, M. Compastié, D. Apostolou, and G. Mentzas, *Extending tosca for edge and fog deployment support*, Electronics, 10 (2021).

Edited by: Dana Petcu
Research papers
Received: Dec 3, 2023
Accepted: Dec 19, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 615–636, DOI 10.12694/scpe.v25i1.2097

CONVOLUTION NEURAL NETWORKS FOR DISEASE PREDICTION: APPLICATIONS AND CHALLENGES

SNOWBER MUSHTAQ AND OMKAR SINGH*

Abstract. More people are using Deep Learning techniques in the healthcare field as a result of the quick development in domains like Computer Vision, Graphics Processing Technology, and the accessibility of medical imaging datasets. Convolutional Neural Networks (CNNs), in particular, have quickly emerged as the preferred technique for processing clinical data. CNN-based designs have been embraced by the diagnostic imaging group to assist physicians with disease identification. Since AlexNet's enormous success in 2012, CNNs have indeed been employed more and more in the analysis of medical images to boost the effectiveness of physicians. This article summarises various CNN architectures for predicting medical diseases and their challenges. We examine the utilization of Deep Learning for the prediction of various diseases, including Brain diseases, Diabetic Retinopathy, and Lung cancer. This research also provides a survey of datasets available for analysis.

Keywords: Convolutional Neural Networks (CNNs), AlexNet, ResNet, Brain diseases, Diabetic Retinopathy, Lung disease, Alzheimer Disease

1. Introduction. Human lives are impacted by health complications. When a patient is receiving medical care, healthcare professionals gather clinical evidence about that individual and use information about the general community to decide how to treat that individual. Therefore, data is key to solving health problems, and better information is essential for enhancing clinical outcomes. Medical imaging is a crucial part of modern medicine. Because it allows for detailed exploration inside the human body in a non-invasive fashion. Deep Learning has reported promising results in medical image analysis. The major reason behind this is the advent of deep Convolutional Neural Networks (CNNs).

Huge phenotyping from observational data [78], autism subtyping [21] by clustering comorbidity, lymph node metastases from breast pathology [34], and the diagnosis of Diabetic Retinopathy [35] are just a few instances of the work being carried out in Deep Learning for healthcare. Deep Learning problems well adapted for healthcare [26], the requirement for visibility [115], also utilizing big data for targeted therapy [8] have been the focus of previous studies of deep learning in the medical field, that have focused entirely on biological applications [7]. In this paper, we review various CNN architectures and their application in disease prediction. Figure 1.1 explains the structure of the study.

The key contributions of this study are as follows:

- The study provides a thorough description of the different CNN architectures. Moreover, their complexity and challenges are also presented.
- The study reviews the literature pertaining to Diabetes diagnosis using CNN. Moreover, the literature on diagnosis of Diabetic Retinopathy using CNN is also reviewed.
- The study reviews the literature pertaining to diagnosis of brain diseases like Alzheimer's disease and Parkinson's disease using CNN.
- The study also presents a review on the diagnosis of lung cancer using CNN.

Image Analysis and Artificial Intelligence. Artificial Intelligence (AI) is not a novel idea. Renowned intellectuals like Leonardo Da Vinci [124] attempted to build automata that mimicked human actions. These days, it appears that this is already the case. Though there are many self-adjusting intelligent systems already, AI has grown exponentially, particularly in the field of health informatics [84]. AI in healthcare is indeed a rapidly expanding discipline that inspires enthusiasm and raises baffling concerns. AI is the capability of a machine to simulate biological mental capabilities. The term "AI" refers to a wide variety of technologies. One of the

* Department of Electronics and Communication, National Institute of Technology Srinagar (J&K), 190006 India (Snowbernitsri@gmail.com, omkar.parihar@gmail.com)

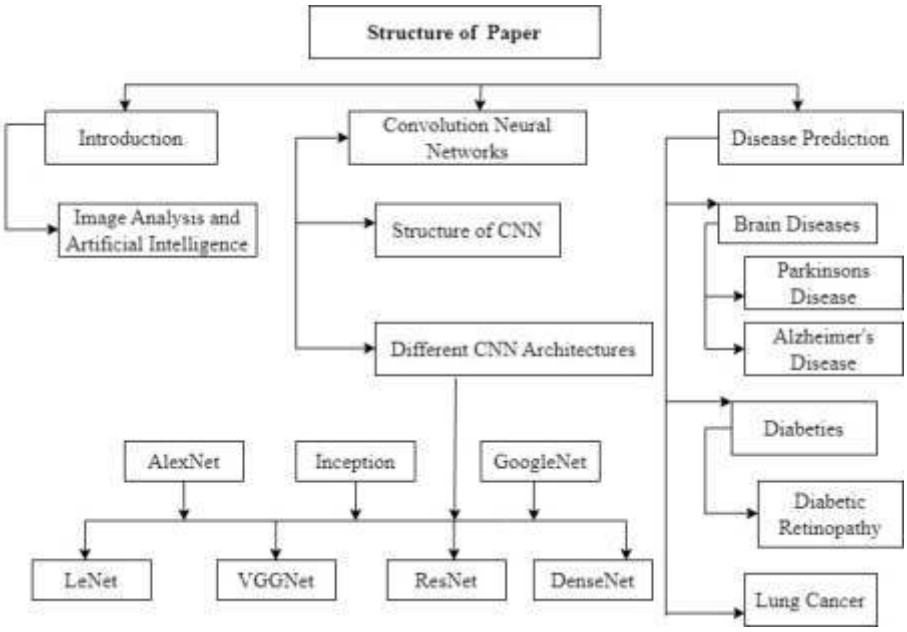


Fig. 1.1: Organizing and Visualizing Map of the Survey

most widely applicable methods in healthcare professionals among them is Machine Learning. Its techniques in medicine have been made possible by three overlapping technological advancements:
a) The emergence of "big data" worldwide and analysis of exceedingly massive databases.
b) The extraordinary rise in CPU computational capabilities.
c) The invention of novel Deep Learning methods.

One of the most well-established sub-fields in Computer Engineering, Deep Learning, has improved performance in many areas, particularly in the analysis and categorization of pathological images. The revolutionary change towards Deep Learning systems, which most researchers find appealing due to the effectiveness and "clarity" of the present models, is predominantly to blame for the expanding innovations. In practice, it is sufficient to think of Deep Learning systems as a black box toward which we supply data for input and output as a baseline for the intended training in most implementations (supervised learning) [95]. CNN, one of the Deep Learning techniques has latterly been proven to be an assuring approach in biomedical image analysis.

2. Convolutional Neural Network. The sub-field of the Machine Learning described as Deep Learning is focused on Artificial Neural Networks, a group of methods that are based on the composition and function of the brain. It is typically a neural network with three or more layers and belongs to the Machine Learning category. These Artificial Neural Networks attempt to replicate how the human brain functions but fall far short, allowing it to understand using enormous amounts of data. Various metrics could provide a response to the question, why Deep Learning? These are:

Commitment to Global Learning: Deep Learning is widely termed ubiquitous learning since it can function in nearly all application fields.

Robustness: In general, Deep Learning approaches do not need carefully planned features. However, the optimum attributes are automatically learned in connection with the task under consideration. However, robustness to the source data's typical variations is gained.

Generalization: Different applications or types of data can employ the same Deep Learning method, known as transfer learning. Additionally, it is a beneficial method for issues in which the data is insufficient.

Scalability: Deep Learning is very extensible. ResNet [39], created by Microsoft, has 1202 levels and therefore is extensively used in high-performance computing environments. Deep Learning consists of a number

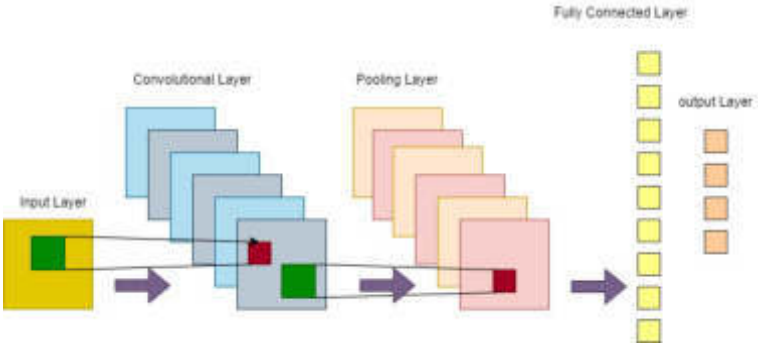


Fig. 2.1: Components of a CNN [69]

of architectures. These include; CNNs, Recurrent Neural Networks, Long-Short Term Memory (LSTM), Auto-Encoders (AEs), and Deep Belief Networks [88]. CNNs are one of the most widely used Deep Learning architectures [89]. Medical Images are one of the areas of Image Processing, where CNN, a subclass of Artificial Neural Networks [123] has gained leadership.

To handle data with a grid pattern, such as images, CNN is a Deep Learning model that is based on the structure of the animal visual cortex [47]. CNN is intended to be dynamic and is able to acquire spatial hierarchies of characteristics, from low-level to high-level structures. CNN for Deep Learning is well-liked for three key reasons:

a) CNNs do not require feature extraction manually because they understand the characteristics independently.
b) Outcomes from CNNs for identification are extremely precise.
c) It can be expanded on pre-existing networks by using CNNs that can be retrained for new recognition tasks.

CNNs are meant to automatically analyze and are able to adapt and learn spatial feature hierarchy by training algorithms that employ a wide range of construction blocks, such as pooling, convolution, and fully connected layers. The extraordinary outcomes have been disclosed in the object recognition contest considered as the ImageNet's Large Scale Visual Recognition Competition (ILSVRC) in 2012 [87], which is the most founded methodology among diverse Deep Learning Models. In several subjects, including medical technology, CNN has performed at an extremely high level. Deep Learning the prospect for diagnosing lymph node metastases, straining for diabetic retinopathy, and categorizing skin lesions, was founded by Gulshan et al. [35], Ehteshami Bejnordi et al. [9], and Esteva et al. [25] correspondingly. Knowing such cutting-edge approaches will benefit clinical radiologists as well as academics, which use CNN for their jobs in radiology and medical imaging as Deep Learning could soon impact clinical practice.

2.1. Structure of CNN. Comparable to a standard Neural Net, there are three layers in the CNN: input, hidden, and output. The distinction is that the image intake for CNN is the pixel matrix, and the image feature attained by the convolution estimation is the output [95]. The convolution kernel, from which the phrase "Convolution Neural Network" emanates, is the most crucial segment of CNN. Each pixel in the two-dimensional matrix n x n of the Convolution Kernel has a proportional weight. A CNN is a specific type of Artificial Neural Network with very few associations between the layers that strive to keep spatial relationships within the data. Every layer function in a CNN, on a small area of the preceding layer, with the input organized in a grid structure and handed through layers that preserve those relationships. CNNs are competent in creating a highly effective model for input data, making them intent for jobs involving images. A CNN is trained via backpropagation and gradient descent, just like classic Artificial Neural Networks. Figure 2.1 explains the layers of a CNN. These are:

a) Convolutional layers. The activations from the preceding layer are connected in the convolutional layers with several small parameterized filters, normally of size 3x3, and then kept in a tensor called W(J, I), where the filter number is represented by J and the layer number is represented by I. One drastically reduces the number of weights that need to be understood, i.e. translational equivariance at each layer.

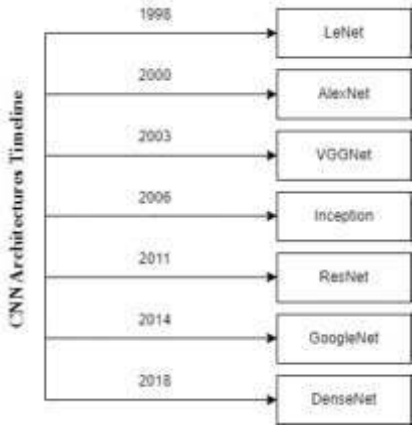


Fig. 3.1: CNN Timeline

This weight-sharing is necessary because characteristics that occur in one portion of the image will probably also occur in other areas. If a filter is competent in glimpsing horizontal lines, then it can detect them wherever they occur. A tensor feature map is generated after filters are applied at every input point in a convolution layer.

b) Activation layer. Nonlinear activation functions generate the feature maps from a convolutional layer. It enables nearly every nonlinear function to be approximately replicated by the neural network as a whole [62]. The extremely basic sigmoid, tanh rectified linear units or ReLUs, and its variations such as leaky ReLUs or parameterized ReLUs, are often the activation functions [38]. When the feature maps are fed through an activation function, new tensors—often also referred to as feature maps—are generated.

c) Pooling layer. CNNs use the pooling technique to generalize the features that the convolution filters have extracted, allowing the network to identify features regardless of where they are in the image. Small grid areas are indeed the input for pooling operations, which further yield single integers for every area. The max-pooling or average-pooling are commonly used to calculate the number. Utilizing convolutions having longer strides is another method for obtaining the pooling's downsampling impact. The network architecture can be clarified by eliminating the pooling layer without compromising production [101].

3. Different CNN Architectures. Different CNN architectures have already been presented over the past ten years [98]. A crucial component in improving the efficiency of many applications is model design. Since 1989 to the present, CNN's architecture has gone through a number of changes. These changes comprise regularisation, optimization techniques, and structural restructuring. On the other hand, it needs to be emphasized that major improvement in CNN effectiveness, is primarily the result of the rearrangement of the processing elements and the introduction of new blocks. The use of network depth was among the most innovative breakthroughs in CNN. Figure 3.1 gives the timeline of various CNN architectures.

3.1. LeNet. One of the inaugural CNNs, LeNet-5, contributed to the evolution of Deep Learning. In the year 1998 paper, "Gradient-Based Learning applied to Document Recognition," [61] introduced LeNet. For the image classification process from the MNIST dataset, they used LeNet-5 CNN. They were the first to use the backpropagation technique in real-world settings and thought that introducing limitations from the task's domain would significantly improve the capacity to learn complexity. The LeNet-5 CNN model has seven layers. This model's uncomplicated structure was the primary factor in its success.

Architecture of LeNet. The network is referred to as Lenet-5 because it comprises 5 layers with learnable parameters. It has 3 pairs of Convolutional Layers with an average pooling mixture. So have two fully connected layers succeeding the convolution and average pooling layers. Finally, a Softmax predictor arranges the images in the appropriate class. Figure 3.2 explains the architecture of Lenet.

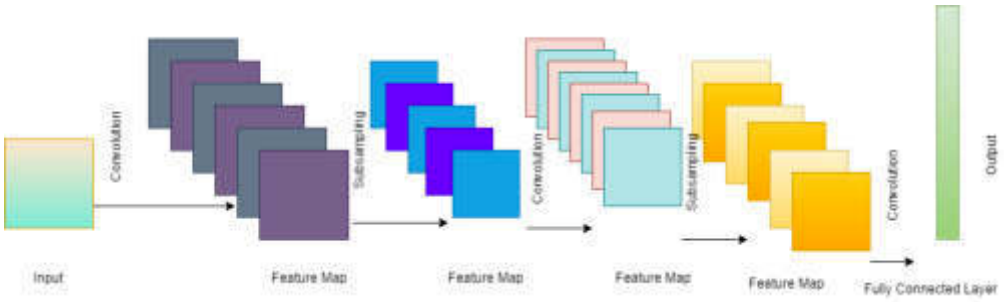


Fig. 3.2: Architecture of LeNet [61]

Complexity and challenges. This network is quite easy to grasp and served as a fantastic foundation for the field of neural networks. With character recognition images, it performs effectively. The system suffers to scan for all properties because it isn't very deep, resulting in simulations that perform poorly. It would be challenging for the neural network model to adapt and generate a precise model if it wasn't provided with just enough characteristics from the training images.

3.2. AlexNet. A model as sophisticated as AlexNet is able to produce high precision on very difficult datasets. But taking out any of the convolutional layers will severely damage AlexNet's effectiveness. It is a well-known architecture for almost any object-detection task, and it may have numerous applications in the computer vision field of artificial intelligence problems. LeNet's [5] debut signalled the beginning of deep CNNs. Those CNNs could only be used for recognition of handwritten digits tasks, which are not easily scalable to all image classes. AlexNet is well-regarded in deep Network architectures [56] because it produced ground-breaking achievements in the areas of image recognition and classification. AlexNet was first introduced by Krizhevesky et al.[56], who then increased CNN's learning capacity by deepening it and adding a number of variable optimization algorithms. The technique developed by Krizhevesky et al. periodically runs across a number of structural units during the development phase to ensure that the features the algorithm learned are extra resilient. ReLU [120] could also be used as a non-saturating activation function to speed up converge [43] by lessening the gradient vanishing problem.

Architecture of AlexNet. The very first CNN to employ a GPU to optimize effectiveness was AlexNet. Five convolutional layers, three max-pooling layers, two normalization layers, two fully connected layers, and one softmax layer make up its architecture. Convolutional filters and then a nonlinear activation function called ReLU make up every convolutional layer. The pooling layers are used to carry execute Max Pooling. Due to the presence of fully connected layers, the intake size is set. The intake dimension is typically stated as 224x224x3, however, because of padding, it actually comes out to be 227x227x3. There are 60 million elements in AlexNet in total. Figure 3.3 explains the architecture of AlexNet.

Complexity And Challenges. A system as sophisticated as AlexNet is able to achieve highly accurate on really difficult datasets. But taking out any of the convolutional layers will negatively affect AlexNet's efficiency. For any object-detection operation, AlexNet is a prominent design, and it has numerous uses inside the field of computer vision of machine intelligence challenges. AlexNet may also be credited with introducing Deep Learning to related domains like Language processing and analysis of medical images as just a significant step toward rendering it more broadly usable.

3.3. VGGNet. Visual Geometry Group (VGG) is a complex CNN architecture that is typical and contains numerous layers. In 2014, scientists from the University of Oxford, Karen Simonyan and Andrew Zisserman, presented the VGGNet framework for CNNs [100]. The term "deep" refers to the number of layers, with VGG-16 or VGG-19 having 16 and 19 neural network layers, respectively. VGG architecture operates as the footing for innovative visual recognition techniques. The VGGNet, designed as a deep neural network, outperforms benchmarks on many tasks and databases outside ImageNet. It also remains among the most often used computer vision architectures today.

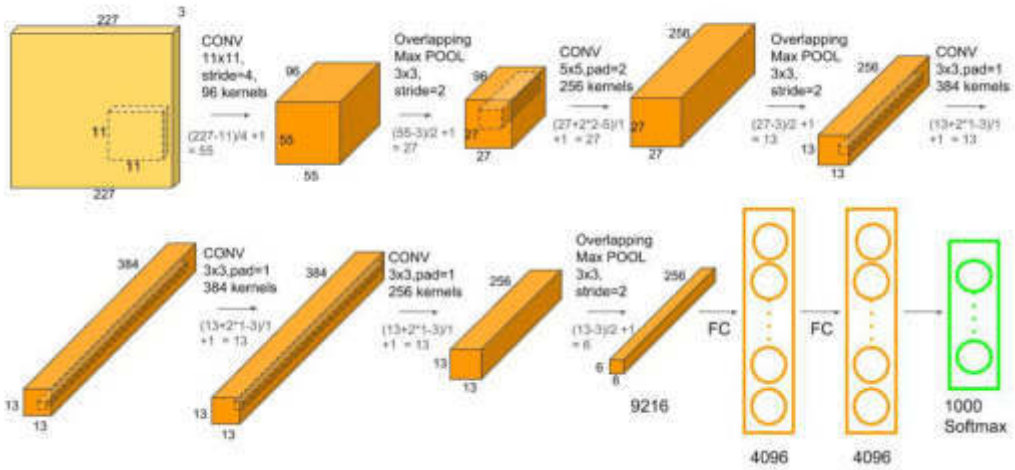


Fig. 3.3: Architecture of AlexNet [56]

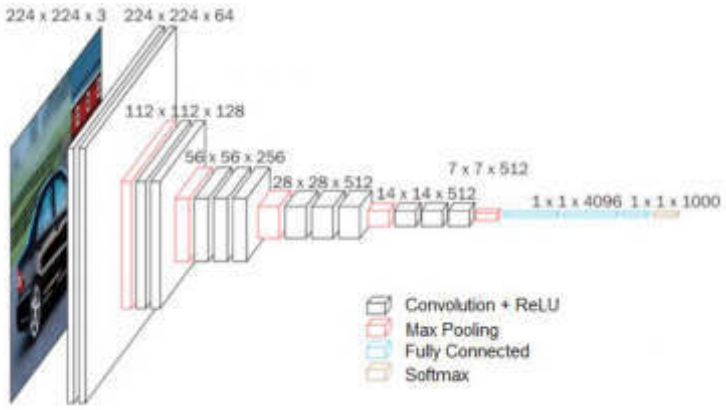


Fig. 3.4: Architecture of VGGNet [100]

Architecture of VGGNet. VGG's input is configured to an RGB image with a 224x244 resolution. The training set image's mean RGB values are determined, and the image is then used as an input to the VGGNet. The convolution phase is fixed, as well as 3x3 or 1x1 filters are employed. There are 3 completely connected layers and the number of convolutional layers plus fully connected layers determines their value, which ranges from VGG11 to VGG19. The minimum standard VGG11 consists of 3 fully connected layers and 8 convolutional layers. There are 16 convolutional layers in the maximal VGG19 plus three fully connected layers. The VGGNet also does not have a pooling layer following every convolution layer, a number of 5 pooling layers, spread behind convolutional layers. Figure 3.4 depicts the architecture of VGGNet.

Complexity and challenges. With each level of the convolution layer, the quantity of filters doubles. This fundamental idea underlies the architecture of VGG16. The VGG16 model is greater than 533MB due to its depth and quantity of completely connected layers. Because of this, building a VGGNet is a gradual task. Many Deep Learning image classification issues use the VGG16 model, however, simpler network topologies like GoogleNet and SqueezeNet are frequently chosen. In either case, the VGGNet is a wonderful basic foundation

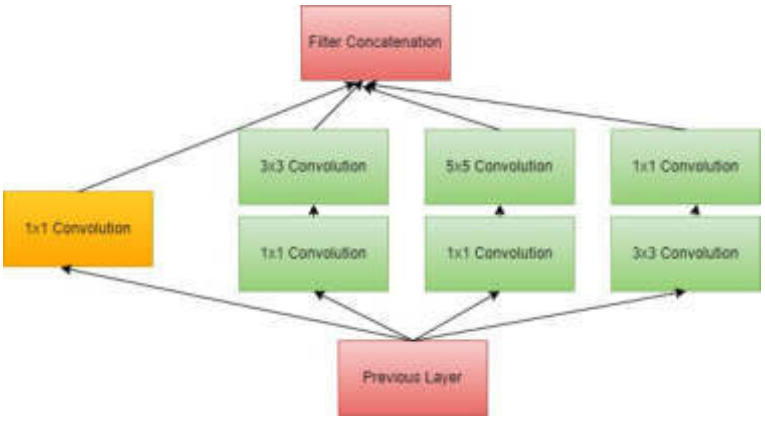


Fig. 3.5: Architecture of Inception Module [108]

for educational reasons since it is simple to set up.

3.4. Inception Network. A CNN with an organizational layout made up of repetitive elements known as Inception modules [108] is considered an inception network. Convolutional layers are enclosed within modules or blocks that are stacked as opposed to stacking convolutional layers themselves.

Architecture of Inception Network. Figure 3.5 depicts the architecture of the Inception module. It uses parallel processing and extracts the features concurrently. This is the prime characteristic of the Inception network that differs it from other CNN architectures. Figure 3.5 depicts that the Inception module simultaneously performs convolution operations of different sizes and then concatenates the outputs from all the operations and creates the next feature.

Complexity and Challenges. It is having the capacity to use different convolution filter sizes and extract features from input data at different scales. In order to improve the network's overall ability to extract features, 1x1 conv filters learn cross-channel patterns. And has effective utilization of computational resources with little rise in workload for an Inception network's outstanding performance output. Once the Inception section is split into its constituent parts, it is simple to break down and comprehend. The issue of overfitting, which happens when the quantity of input features is high throughout training, will be increasingly prevalent as our model grows in size (more layers). The total amount of layers will expand along with the number of variables, thus must also prepare to beef up our processing resources before we can execute the computation on these parameters. Therefore employing an Inception network will reduce computing costs while simultaneously expanding the width and depth of the system, instead of expanding the computing resource.

3.5. ResNet. ResNet (Residual Network), the ILSVRC 2015 winner, was created by He et al. [39]. In contrast to earlier systems, the goal was to create an ultra-deep network immune to the vanishing gradient problem.

Architecture of ResNet. The network employs a VGG19-inspired 34-layer plain network topology, to which the bypass link is introduced. The structure is subsequently changed into a residual network by these short-cut links. Figure 3.6 explains the architecture of ResNet. ResNet-34 was the initial ResNet architecture, and it included inserting shortcut interconnections to transform a simple net into an equivalent residual network [39]. In this instance, the CNN included 33 filters, whilst the simple network was influenced by VGGNets (VGG16, VGG19). ResNets, though, are simpler and require fewer filters than VGGNets.

Complexity and Challenges. ResNet is a significant advancement that altered the process of learning deep CNNs for tasks involving computer vision. Whereas the initial ResNet had 34 layers and 2-layer restriction blocks, more sophisticated models, such the Resnet50, used 3-layer restriction blocks to assure high efficiency and shorter training durations.

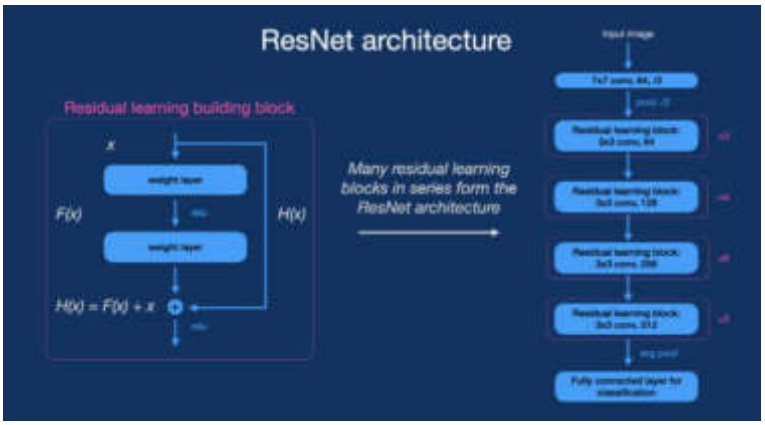


Fig. 3.6: Architecture of ResNet [11]

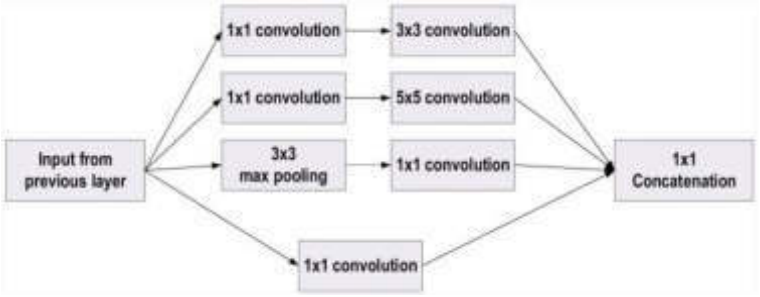


Fig. 3.7: Architecture of Googlenet Module [108]

3.6. GoogleNet. The winning entry in the 2014-ILSVRC competition was GoogleNet, also known as Inception-V1 [108]. The primary goal of the GoogleNet design is to achieve top-level precision with reduced processing expense. Since it combines multiple-scale convolutional transformations by using merging, modification, and splitter algorithms for extracting features, it suggested a new inception block (module) idea in the framework of CNN. In comparison to previous winners AlexNet (Winner of ILSVRC 2012) and ZF-Net (Winner of ILSVRC 2013), with a significantly lower error rate over VGGNet, it really has delivered a marked decline in the failure rate (2014 runner-up).

Architecture of GoogleNet. Figure 3.7 shows how the inception component architecture is organized. This design requires filters of various sizes including 5×5, 3×3, and 1×1 to record channel information as well as spatial information at various spatial levels of resolution. Small modules that implement the very same idea of Network-in-Network (NIN) architectures [65], that substituted every level with a micro-neural network, are used to substitute the common convolutional layer of GoogleNet.The GoogLeNet merge, transform, and split principles have been used, backed by focusing on a problem associated with various types of learning of variants present inside a class of multiple images that are comparable to each other. Figure 9 depicts the architecture of GoogleNet.

Complexity and Challenges. The goals of Google Learning Network were to increase learning ability and improve the efficiency of CNN characteristics. Additionally, it controls the processing by adding a blockage layer of a 1×1 convolutional filter before employing large-size kernels. Sparse connections were used by GoogleNet to solve the duplicate content issue. By skipping those useless channels, it reduces expenses. The number of interconnections was reduced by using a GAP layer as the end layer instead of an FC layer. The utilization of regularisation and RMSProp as an optimizer were 2 extra consistency considerations [17]. The primary

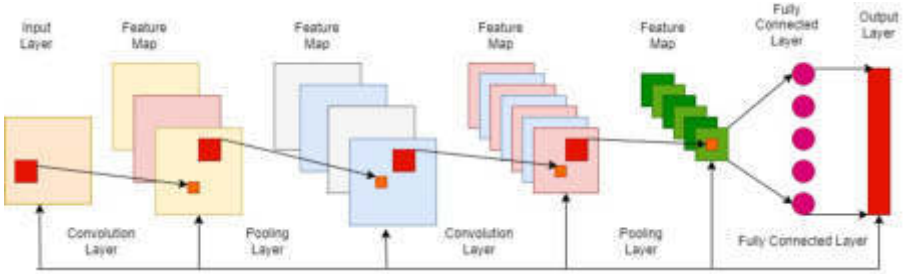


Fig. 3.8: Architecture of DenseNet [129]

drawback of GoogleNet, on the other hand, was its diverse architecture, which necessitates adaption from one component to the other. The representational jam, which significantly reduced the feature space in the layer below and sometimes led to the loss of critical info, is one of GoogleNet's major flaws.

3.7. DenseNet. DenseNet is among the most current revelations in neural networks for visual object detection. ResNet and DenseNet are roughly comparable, however, there are a few key distinctions. DenseNet, which Huang et al. citehuang2017densely designed to assure the greatest information flow between layers in the networks, earned the best paper prize at CVPR2017. Every level in DenseNet receives extra inputs from all layers that came before it and transmits its own extracted features to all layers that came after it. Huang et al. [45] presented CondenseNet as a solution to the issue of DenseNet's high memory utilization. The network structures are often gradually hierarchical. The input of the i^{th} layer in such a network structure is comprised of the feature maps from the $(i-1)^{th}$ layer. Every layer in the system is tied directly to the front layers, which is the fundamental concept behind DenseNet. Figure 3.8 depicts the architecture of DenseNet.

Architecture of DenseNet. The first convolutional layer, which receives the input, is the only one in a traditional feed-forward CNN that acquires the output of the convolutional layer before it. This convolutional layer then produces an output of extracted features, which is then passed on to the subsequent convolutional layer. As a result, there are L direct connections for each layer, one from one to the next. Figure 10 describes the architecture of DenseNet. Figure 3.8 explains the architecture of DenseNet.

Complexity and Challenges. By altering the typical CNN architecture and streamlining the connection among layers, DenseNet addresses the challenge of Vanishing Gradient. Each layer in a DenseNet architecture is connected to each other layer directly, giving rise to the densely connected CNN. There are $L(L+1)/2$ close links among both L layers as shown in Figure 10. DenseNets, as opposed to its regular CNN or ResNet equivalents, has acquired state-of-the-art capabilities and improved results among comparable datasets because they require fewer parameters and permit feature reuse, resulting in more compact models.

4. Disease prediction and analysis. Modern people suffer from a range of illnesses as a result of both their environment and lifestyle habits. Therefore, predicting disease sooner has become a crucial challenge. The most difficult challenge is to predict disease accurately. Deep Learning is crucial in predicting the disease in order to solve this issue.

4.1. Diagnosis of Diabetes. Diabetes is a metabolic condition that affects a lot of individuals all over the world. Each year, its incidence rates are frighteningly rising. Diabetes-related problems in several of the body's major organs could be lethal if not treated [107]. Early diabetic diagnosis is crucial for prompt treatment that can prevent the condition from escalating to severe problems. Deep Learning techniques have shown promising results in diagnosing diabetes at its onset.

Diabetes occurs in Cardiac Autonomic Neuropathy (CAN), a total neural system disturbance that reduces heart rate variation. Consequently, Heart Rate Variability (HRV) is a sign to detect the presence of diabetic neuropathy[77]. To achieve a more objective evaluation and diabetes diagnosis employing iris images, a combined Deep Learning and image processing technique has been proposed by Onal et al.[71]. The proposed methodology initially recognized the iris boundary in the iridology chart, after which it automatically identified

Table 4.1: Summary of research works corresponding to CNN for Diabetes diagnosis

Work	**Method**	**Accuracy**
[107]	LSTM, CNN and its combinations	95.7%
[71]	Hybrid method + VGG16	80%
[106]	CNN + CNN-LSTM + heart rate signals	90.9% using CNN-LSTM, 93.6% using 5 fold cross-validation, 95.1% using CNN-LSTM
[50]	SVM + CNN-LSTM + IF-CNN	96.26%
[68]	CNN-Bi-LSTM	98%
[33]	CNN	97.3%
[111]	AlexNet+VGG-16+SqueezeNet	AlexNet 93.46%, VGG-16 91.82% and SqueezeNet 94.49%

the pancreatic region. CNNs were then used to diagnose diabetes on images, and the outcomes were contrasted with other CNN models. It was determined that an efficiency of 80% was achieved using the suggested strategy in conjunction with the VGG16 architecture and automatic pancreatic area partitioning. Wang et al.[117] proposed a model for forecasting improvements in diabetic symptoms using an enhanced CNN technique. The model can aid doctors to forecast the probability of recurrence in patients after discharge and use case records of inpatient diagnosis and treatment to rate the patient's effectiveness of the treatment. Pal et al. [75] give an overview of the current Deep Learning methods that are used to forecast diabetes in its beginning stages. It can help researchers in this field by giving them knowledge of the most advanced techniques for earlier diabetes detection. Fufurin et al.[27] proposed a technique for detecting type 1 diabetes using infrared imaging spectrometry of exhaled human breath. The strategy can be employed in everyday clinical practice, but the results need to be confirmed on a bigger database and in subsequent biomedical studies. Swapna et al. [106]employed Deep Learning networks of CNN-LSTM and CNN combination to automatically identify the irregularity. Approaches to Deep Learning do not need feature extraction, in contrast to the standard analytical techniques that have been used up to this point. Kamalraj et al.[50] proposed the Pet Dog-Smell Sensing (PD-SS) method and Interpretable Filter-based CNN (IF-CNN) prediction model, that could effectively diagnose diabetes using PIMA Indian diabetes databases. This could improve the general approach to disease forecasting in the patient database, perhaps handling difficulties with older Deep Neural Network-based algorithms. Leveraging the publicly accessible PIMA Indian diabetes database, Madan et al.[68] developed a continuous monitoring hybrid Deep Learning-based model to detect and diagnose Type 2 diabetes mellitus. The research provided four contributions. Initially, they conduct an evaluation of various Deep Learning algorithms. In order to identify (and diagnose) Type 2 diabetes, they subsequently proposed integrating two models, CNN-Bi-LSTM. The proposed approach proved better than previous approaches. Goel et al. [33] provided a comparison of the CNN model's effectiveness in predicting sugar levels by employing the four non-linear activation functions sigmoid, tanh, ReLU, and ELU. According to the research observations, CNN offers a maximum accuracy of 97.3% when used in conjunction with the ELU activation function. Table 4.1 gives a summary of research works on Diabetes diagnosis using CNN variants.

Diagnosis of Diabetic Retinopathy. Diabetic Retinopathy (DR) is a common complication of diabetes mellitus that harms vision. It can lead to blindness, if not detected early. It is not a reversible process, and treatment only sustains vision. Early detection and treatment can remarkably diminish the risk of vision failure. The hand-operated diagnosis process of retinal images by doctors is time, energy, cost-consuming, and prone to misdiagnosis, unlike automated diagnosis using AI. In recent years, DR classification and detection have made extensive use of Deep Learning. Even with the integration of numerous diverse sources, it can effectively acquire the properties of the provided data [13]. There have been numerous Deep Learning-based approaches, including AEs, CNNs, Restricted Boltzmann Machines, and Sparse coding that have been used for the diagnosis of DR [37]. Unlike Machine Learning approaches, the effectiveness of these approaches improves as the amount of training data rises [20] because the variety of discovered attributes expands.

Gayathri et al. [29] demonstrate a new CNN model to automatically extract from retinal fundus images for enhanced classification results. In the proposed approach, different machine learning classifiers are fed the CNN output characteristics as input. The evaluation findings demonstrate that the J48 classifier and

the recommended feature extraction approach surpass all other learners. Kwasigroch et al.[58] suggest a Deep Learning strategy to simplify the detection of DR. The most widely used class of Deep Learning algorithms, deep CNNs, succeeded at image recognition and analysis. Qomariah et al. [81] presented a support vector machine-based Deep Learning algorithm for feature extraction and categorization. As input features for classification utilizing the support vector machine, they utilize the high-level attributes of the final fully-connected layer depending on transfer learning via CNN. By employing this technique, it was observed that the classification process using CNN with fine-tuning required less computation time. Gayathri et al. [30] offer a technique for computerized DR grading in which characteristics from fundus images may well be retrieved and classified according to seriousness by employing Deep Learning and Machine Learning technologies. The identification of global and local characteristics from visuals is accomplished using a Multipath-CNN (M-CNN).To investigate fundus images and automatically differentiate among controls (i.e., no DR), moderate DR (i.e., a combination of mild and moderate Non-Proliferative DR (NPDR)), and severe DR (i.e., a group of severe NPDR, and Proliferative DR), a deep CNN of 18 convolutional layers as well as 3 fully connected layers is proposed by Shaban et al.[96]. The suggested method dramatically improves the availability of retinal care by eliminating the requirement for a retina expert and precisely diagnosing and evaluating diabetic retinopathy. Chen et al. [12] findings demonstrate that deep CNN-based algorithms are successful in facilitating autonomous DR detection by identifying patients' retinal images. To support their CNN learning, similar methods generally rely on an extremely large dataset made up of retinal images with predetermined categorization labels. Comparing the proposed approach to contemporary representative integrated CNN learning models, the classification accuracy can be increased by 3%. Hemanth et al. [40] suggest a different, hybrid method of using retinal fundus images for the diagnosis of DR. The hybrid approach, particularly, is built on combining both image processing and Deep Learning for better outcomes. Zeng et al.[128] By categorizing color retinal fundus images into two grades, a computer-aided diagnosis methodology development of deep learning algorithms is suggested by Zeng et al.[128] to accurately diagnose the referable DR. This study uses a transfer learning technique to create a distinctive CNN model with such a Siamese-like structure. The proposed method achieves an Area Under the receiver-operating characteristic curve (AUC) of 0.949 using a training dataset of only 28104 images as well as a test set of only 3510 images. Gangwar et al. [28] use pre-trained Inception-ResNet-v2 with transfer learning, and construct a customized set of CNN layers on top of Inception-ResNet-v2 to create the hybrid version. The model outperformed other results that have been reported. Automated identification of the DR stage is presented by Qureshi et al. [83] using a novel multi-layer framework of Active Deep Learning (ADL). The CNN model was used to develop the ADL system to dynamically feature extracted as contrasted to manually created attributes. CNNs are recommended by Wu et al. [119] as an automated clinical tool for identifying five stages of DR seriousness categories as a hierarchically Coarse-to-Fine network (CF-DRNet). The CF-DRNet greatly improves the categorization effectiveness of five-class DR grading while adhering to the hierarchical character of DR marking. Liu et al. [66] offer a novel approach driven by ensemble learning, the WP-CNN, which incorporates several weighted pathways into CNNs. Backpropagation is used in WP-CNN to optimize various path weight coefficients, and the return features are averaged enabling quick convergence. Pao et al. [76] proposed that the green element of a retina image was utilized to compute the entropy image. They trained this network on the publicly accessible Kaggle dataset that used a high-end graphics processing unit (GPU), and showed outstanding results, especially for a high-level classification problem. Figure 4.1 gives us the sample of a severe non-proliferative DR (NPDR) fundus image that shows the severity and likelihood, of the presence of microaneurysm, hemorrhage, and exudate. Figure 4.2 outlines the general measures taken by a CNN model to categorize fundus images into 5 severity categories. Table 4.2 presents the summary of research works corresponding to DR diagnosis using CNN. Table 4.3 presents the datasets available for DR diagnosis.

4.2. Brain Diseases. The operations center of our body is the brain. The brain is impacted by a wide range of conditions and abnormalities. The buildup of aberrant proteins in our brain is a prevalent trigger of neurodegenerative illnesses. They comprise, among others, ALS (Amyotrophic Lateral Sclerosis), Parkinson's disease, Alzheimer's disease, and others.

4.2.1. Parkinson's Disease. A neurological condition that affects voluntary muscle movement is Parkinson's Disease (PD). Identifying PD and its root causes is essential for developing its treatment and prevention plan. Traditional PD diagnostic techniques suffer from subjectivity as they rely on the evaluation of movements

Fig. 4.1: Sample of a severe NPDR fundus image that shows the severity and likelihood, the presence of microaneurysm, hemorrhage, and exudate [109]

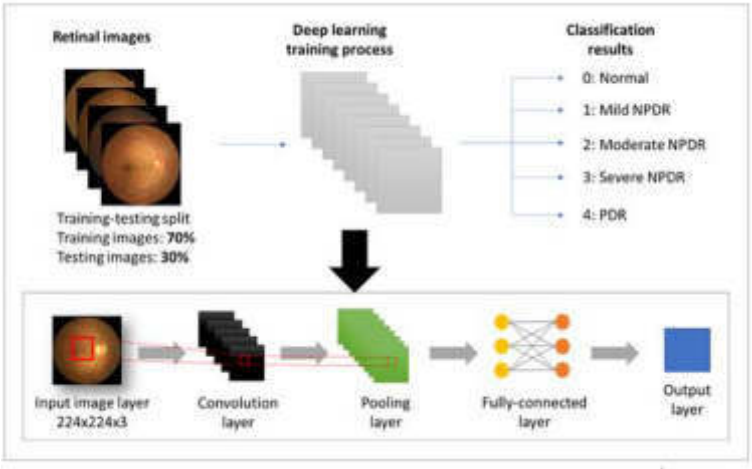


Fig. 4.2: Deep Learning process for classifying images to 5 severity levels [109]

that are sometimes subtle to human eyes and therefore make the correct classification difficult. This makes early diagnosis of PD challenging. To combat this challenge, Deep Learning has been identified as the potential solution. Several researches have reportedly been conducted to analyze the potential of Deep Learning in PD diagnosis. Taleb et al. [110] investigated how various Deep Learning architectures, such as the CNN and the CNN-BLSTM, can be utilized to diagnose PD via time series analysis. Hire et al.[42] presented a group of CNNs for the detection of PD using speech recordings from 50 patients suffering from the condition and 50 healthy people from the PC-GITA database. Kurmi et at.[57] proposed a collection of Deep Learning to diagnose PD utilizing DaTscan images. To begin with, they categorized PD by applying four DL models: VGG16, ResNet50, Inception-V3, and Xception.To improve the classification model's overall effectiveness, they used a fuzzy fusion logic-based ensemble technique in the subsequent steps. Yousif et al.[125] proposed a global standard for the diagnosis of PD utilizing voice signals and/or handwritten drawings. To diagnose PD using handwriting images,

Table 4.2: Summary of research works corresponding to DR diagnosis using CNN

Work	Dataset	Method	Accuracy
[58]	88000 retina images(own dataset)	CNN	82%
[81]	Messidor	SVM+CNN	95.83%
[30]	IDRiD, Kaggle, and MESSIDOR	M-CNN	99.62%
[40]	MESSIDOR	image processing + deep learning	94%
[28]	APTOS+Messidor-1	transfer learning+Inception-ResNet-v2	72.33% for Messidor-1 and 82.18% for APTOS
[83]	Own dataset(54,000 images)	ADL-CNN	98%
[66]	-	WP-CNN	94.23%
[80]	Kaggle (80,000)	CNN	75%
[121]	Kaggle (80,000)	CNN	94.5%
[23]	Kaggle (35000)	CNN-ResNet34	85%
[116]	Kaggle (35,126)	CNN (AlexNet, VggNet, GoogleNet and ResNet)	95.68%
[1]	DiaretDB0 (130), DiaretDB1 (89), and DrimDB (125)	CNN	99.17 (DiaretDB0), 98.53 (DiaretDB1), 99.18 (DrimDB)
[52]	MESSIDOR (1200)	CNN (AlexNet, VggNet16, custom CNN)	98.15%
[130]	Own dataset	CNN (ResNet50, InceptionV3, InceptionResNetV2, Xception and DenseNets)	96.5%
[113]	HRF (45) and DRIVE (40)	CNN	93.94%

8 pre-trained CNNs using transfer learning were optimized by Aquila Optimizer. Features from the MDVR-KCL dataset are extracted numerically for the speech signals using 16 feature extraction methods and fed to four different machine learning models tuned by the Grid Search algorithm, as well as pictorially utilizing five different methods and fed to eight pre-trained CNN frameworks. Vyas et al.[114] presented two cutting-edge methods that make use of Deep Learning approaches. CNNs in 2D and 3D that were learned on axial-plane MRI data are employed. Alissa et al.[4] proposed a technique focusing on using drawing tasks to identify patient movement abnormalities. Additionally, their research examines the superiority of the spiral pentagon over the wire cube as a categorization tool. Zhao et al. [131] proposed greedy methodology, integrates the concepts from different regions into a sophisticated one. Every region was trained and tested for this prototype. To categorize the presented participants into PD and healthy utilizing neuroimaging (T1 weighted MRI scans and SPECT) and biologic (CSF) parameters as the database, two frameworks—feature-level and modal-level—are proposed by Ahuja et al. [74]. All of these parameters are combined in the feature-level framework to produce a heterogeneity database that is later provided to two Deep Learning models to diagnose PD. A summary of such research used for PD diagnosis using CNN is presented in Table 4.4.

4.2.2. Diagnosis of Alzheimer's Disease. The seventh largest leading cause of death in the world is cognition, including Alzheimer's Disease (AD) [118]. The most widespread form of dementia, accounting for 60% to 80% of cases, is AD. A condition known as dementia is characterized by a decline in mental capacity that goes beyond what may be anticipated with the aging process. It impairs consciousness and damages memory, reasoning, orientation, understanding, computation, learning ability, communication, and the capacity to distinguish. Synapse weakness, synaptic loss, and neurodegeneration are all brought on by alterations in Amyloid Precursor Protein (APP) breakage and synthesis of the APP component beta-amyloid (A), as well as hyperphosphorylated protein aggregation. Key elements of the disease include metabolic, vascular, and inflammatory alterations as well as associated conditions. A healthy brain and an AD-affected brain are contrasted in Figure 4.3.

Table 4.3: DR Datasets

Dataset	No. of Images	Resolution	Comments
Kaggle	88,702 high-resolution images	433 × 289 pixels to 5184 × 3456 pixels	Many of the images on Kaggle are of inadequate grade and have erroneous labels [63, 59, 82, 121, 22, 116]
DIARETDB1	89 publicly available retina fundus images	1500 × 1152 pixels	It has 5 normal images and 84 DR images with annotations from four medical professionals [51, 82, 72]
E-ophtha	463 images	-	The E-ophtha EX and E-ophtha MA are included in this publicly accessible dataset [18, 15]
DRIVE	40 images acquired at 45-degree	565 × 584 pixels	It includes images of a max normal retina images, and there are only seven mild DR images [103]
DDR	13,673 fundus images acquired at a 45-degree	-	757 images of DR lesions [63]
Messidor	1200 fundus color images	-	Images are acquired at a 45-degree FOV [19]
Messidor-2	1748 images	-	Images are acquired at a 45-degree FOV [19]
CHASE DB1	28 images	1280 × 960 pixels	Images acquired at a 30-degree FOV [73]
STARE	20 images	700 × 605 pixels	The freely accessible dataset is used to segment blood vessels [44]
Indian Diabetic Retinopathy Image Dataset (IDRiD)	516 fundus images	-	Contains images of normal retinal structures and diabetic retinopathy lesions [79]
ROC	100 publicly available retina images	768 × 576 to 1389 × 1383 pixels	There are just training reality on the ground [16]
DR2	435 publicly available retina images	857 × 569 pixels	98 images are classified as references [53]

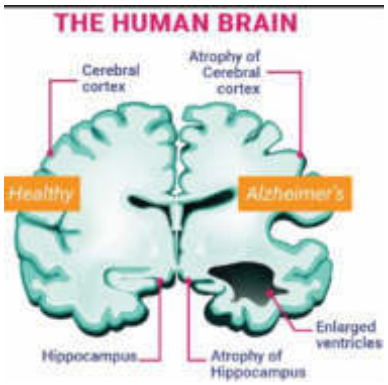


Fig. 4.3: Difference between healthy brain vs severe AD [92]

AD has no cure, however, early diagnosis of AD is crucial to inhibit its progression. The advanced neuroimaging strategies combined with Deep Learning have the potential to diagnose AD in the early stages. Non-invasive neuroimaging techniques are available to understand the pharmacology, function, or structure of the brains [41]. The two categories of imaging technology are typically structural imaging and functional imaging. The anatomy of the brain, including its neurons, synapses, glial cells, etc., can be learned through effective segmentation [41]. The following are the neuroimaging methods more frequently in use for AD:

Magnetic Resonance Image (MRI): The medical imaging technique described as magnetic resonance imag-

Table 4.4: Summary of research works corresponding to PD diagnosis using CNN

Work	Dataset	Method	Accuracy
[110]	HandPDMultiMC	CNN-BLSTM	97.62%
[42]	PC-GITA, a publicly available database	Ensemble of CNNs	99%
[57]	Parkinson's Progression Markers Initiative (PPMI)	VGG16, ResNet50, Inception-V3, and Xception	98.45%
[125]	NewHandPD, MDVR-KCL	VGG19, KNN, SVM	99.75% using the VGG19, 99.94% using the KNN, 100% using the combined the mel-specgram graphical features + VGG19
[114]	318 MRI scans	2D-CNN and a 3D-CNN	3D-CNN 88.9%, 2D-CNN 72.22%
[4]	drawing task	CNN	93.5%
[131]	Three retrospective investigations included 305 Parkinson's patients (aged 59.9–9.7 years) and 227 healthy control individuals (aged 61.0–7.4 years)	CNN	94.1±3.2%
[74]	SPECT	CNN	93.33%
[55]	NTUA	CNN-RNN	98%
[24]	PPMI	3D-CNN	100%
[97]	NIMHANS	CNN	80%
[54]	NTUA	CNN-RNN	98%

ing (MRI), which creates exact images of human tissues and organs, uses a magnetic field and radio waves generated by the computer. Ogawa et al.[70] discovered that operational knowledge about the brain can be obtained via MRI in 1990.

Positron Emission Tomography (PET): FDG-PET is extensively and frequently used in the examination of persons with possible neurodegenerative illnesses, notably AD, to confirm the diagnostic accuracy [90]. It stands for impaired neural function or synaptic degeneration. It was once believed that lower FDG-PET values were really a symptom of neuronal hypo-metabolism caused forward by neuro-degeneration. Instead of reflecting neurons' glucose absorption, it has been discovered to correlate with astrocytes. However, there is proof that anomalies in blood-brain barrier (BBB) transport could be detected by PET by decreased FDG brain absorption.

According to [36], the hippocampal, cortex, and ventricle are the three key brain regions associated with AD. They used layered AEs using a patch-based and ROI-based approach and utilized CNN to diagnose AD. In order to evaluate the proposed CNN model, various image morphological operations and datasets were used. The outcomes of such studies point to the significance of early AD diagnosis utilizing image processing and Deep Learning methods. Suk and Shen [105] suggest a hybrid model for identifying AD built on CNN and Sparse Regression Networks. Multiple Sparse Regression Networks were employed by the model to produce different targeted images. Next, CNN merged these target-level descriptions best determined the output label. The 16-layered VGGNet was altered by Billones et al. [10] for dividing the patients into three categories according to structural MRI: AD, MCI, and HC scanning. Testing carried out for the study showed that the researchers were successfully carrying out categories accurately. According to the writers, this was accomplished. without doing MR image segmentation. Sarraf and Tofighi used LeNet architecture [94] to distinguish AD patients from healthy ones using functional MRI. The findings showed that because of CNN has vast potential in relation to a shift-invariant and scale-invariant feature. Imaging in medicine [67] surpassed a patch-based and voxel-based DBM hybrid [104]. Sarraf and Tofighi [93] used LeNet and GoogleNet architectures in another experiment for diagnosing AD based on both architectural and operational MR images. These suggested and placed-into-use pipelines show a substantial increase in categorization performance over other investigations.

Following presents the datasets available for AD diagnosis:

OASIS: Information from the data sets from neuroimaging called the Open Access Series of Imaging Studies (OASIS) is accessible to the general public for analysis and research. The current MRI data set is made up of a longitudinal collection of 150 participants, ranging in age from 60 to 96, all of whom were recorded employing the same machine and the same procedures. A total of 373 imaging examinations were performed on each patient over the course of at least 2 sessions separated by at least one year [33].

DARS: The Virginia Department for Aging and Rehabilitative Services (DARS) has been compiling information on people with AD as well as other forms of dementia and their carers since 2012 in collaboration with employees from other State Health and Human Resources (HHR) departments [34].

ADNI: The ADNI collection can be used to detect AD, which is typically seen in senior citizens [49], which contains details of MRI scans for 843 subjects with scanner intensity fields ranging from 1.5 T to 3 T. It has been noted that people with mild cognitive impairment(MCI) tend to have reduced intellectual capabilities, particularly reasoning, and loss of memory

IBSR: Brain image features extraction methods are tested and developed using the IBSR dataset [48]. In addition to the MRI data, the dataset additionally includes expert segmentation findings that were carefully supervised. The ground Truth is made up of 20 actual T1-Weighted (T1-W) MRI scans with an expert segmented image that was carefully steered.

MICCAI: The MICCAI-2012 dataset [60] was received via Neuromorphometrics, Inc., Scotts Valley, California, USA, it comprises 35 T1-w MRI volumes and manual segmentation of 134 features. It is mostly employed to segment tissues, tumors, and formations. In 2012, this dataset began with 80 authentic and artificial examples. The quantity of training and testing data has grown over time. Subcortical structure segmentation is done using the MICCAI 2012 task in multi-atlas labeling.

MCSA: MCSA [85] is a population-based controlled trial with the goal of determining the prevalence of MCI along with its causes and risk factors, also include dementia. On October 1, 2004, Olmsted Counties, Minnesota's inhabitants aged 70 to 89 were tallied using the Rochester Epidemiology Survey. The original study participants were randomly selected from among the eligible subjects.

4.3. Diagnosis of Lung Cancer. Among the most prevalent malignancies, Lung Cancer accounts for approximately 3 million cases, more than 1 million fatalities, and 12$ billion in annual spending on health care in the United States [14]. Being one of the fatal tumors, just 17 percent of those diagnosed with lung cancer in the United States remain five years following detection, and life expectancies are worse in emerging economies. Anyone can develop pulmonary cancer, particularly people who smoke or breathe in hazardous, toxic components [46]. Smoking can cause lung cells to mutate. Lung-threatening development [31] was the second-leading death cause in 2015, and it is now the fifth scenario in 2017, as determined by the World Health Organization (WHO) evaluation.

In comparison to the trained radiologists, the latest cyber-physical technologies and computer-aided detection with Deep Learning has shown promising results in lung cancer diagnosis [127, 126, 32]. Many different types of Deep Learning [102] architectures have been studied to understand better how to diagnose lung illness. In [102], a 3D multipath VGG-like system with two setups is suggested.

The two groups are kind knobs and dangerous knobs, and lung knobs and non-knobs, respectively. So various architectural designs are suggested and evaluated in various studies. CNN [100, 112] and its derivatives were primarily covered. CNNs can be used to analyze both 3D and 3D data referred to as 3D CNN/ConvNet and C3D/3D ConvNet, respectively [86]. In blockchain material using extended CNN[102], the lung knobs could be categorized and arranged using lung CT imaging, but their hazard level may also be determined. Although ECNN has time complexity and accuracy limitations, its development is much less sensitive than that of earlier approaches. In [91], CNN was used for lung cancer diagnosis. In [3], 3D-CNNs were used for classifying the CT scans as true (lung cancer) or false (no lung cancer), a result of 86.6 % accuracy was produced on the test set. Lin et al. [64] suggested using a 2D-CNN using the Taguchi optimization procedure to detect lung cancer using CT scans automatically. To increase the accuracy rate of lung cancer using the Taguchi technique, the appropriate parameters for the 2D-CNN architecture were found through the selection of 36 experiments and 8 control factors of varied levels. Sibille et al. [99] proposed that it is possible to

Table 4.5: Summary of research works corresponding to Lung Cancer diagnosis using CNN

Work	**Method**	**Accuracy**
[102]	Blockchain + extended CNN	96.88%
[91]	CNN	Recognize and detect the lung cancer
[3]	3D-CNN	86.6%
[64]	2D-CNN with Taguchi parameter optimization	6.86% and 5.29% more accurate than the original 2D CNN on the two datasets
[99]	Deep-CNN	96.4%
[6]	3D-Deep Learning	94.4%
[122]	Deep Learning on CT images	95% Confidence Interval (CI)
[2]	AlexNet	93.548%

achieve high diagnostic productivity when both CT and PET images are being used, to automate anatomic identification and categorization of fluorine 18-fluorodeoxyglucose PET accumulation pattern in foci suggestive and non-suspicious for cancer in lung tumor patients and lymphoma. Table 4.5 Summarizes research work on Lung Cancer using CNNs.

5. Discussion. Although there are numerous encouraging findings from earlier investigations, there remain a number of challenges to be addressed before Deep Learning can be implemented in diagnostic imaging. Initially, the degree of learning dataset's quality and quantity, as well as its propensity for overfitting and bias, must be taken into account. A Deep Learning generalization should be given, considering the variations in illness occurrence, diagnostic techniques, and medical centers around the world. Therefore, developing assessment methods to measure each technique's effectiveness is necessary. Additionally, as the effect would be strongly influenced by the information quality, there may be legal and ethical concerns around the use of clinical images acquired for marketing. Furthermore, it is crucial to consider the Deep Learning's black-box character. While the Deep Learning-based approach produces outstanding results, it is often complicated or even unattainable to articulate the reasoning behind the judgment. Finally, if we deploye a Deep Learning system in a particular clinical practice process without the instruction of a doctor, legal liability concerns would arise. The inherent constraints of Deep Learning, implementation logistics, and evaluation of acceptance hurdles as well as required socio-cultural or route adjustments are major obstacles to the application of AI systems into healthcare. The following are the main challenges of Deep Learning in medical imaging:

- Low accuracy is sometimes the result of inadequate data. Deep Learning models need ample amount of data to attain high accuracy. In medical imaging, there is dearth of labeled data and this poses a challenge.
- Disease-specific information about rare disorders is limited. This limits the use of Deep Learning in the diagnosis of such dis-orders.
- Small modifications to the input samples can easily fool Deep Neural Network, leading to misinterpretation.
- Heterogeneity of data is another challenge. Nowadays, one barrier to widespread Deep Learning usage in medical imaging is the variety of data. Thousands of handwritten documents scanned combined with broken, redundant, and incomplete information can produce insufficient conclusions and impair decision-making.
- The lack of skilled data scientists and modelers is yet another obstacle to utilizing Deep Learning in medical imaging.

6. Conclusion. Medical imaging is possibly the most appropriate and attractive topic for AI technologies in the biomedicine and healthcare systems sectors. We provided a thorough overview of Deep Learning architectures and reviewed the applications of CNN in medical imaging. It was observed that Deep Learning techniques based on CNNs are becoming more widely accepted in all areas of early disease detection including DR, lung cancer, AD, etc. Data augmentation and transfer learning are examples of methods used to solve problems with Deep Learning methods caused by inadequate data and labels. Improved Deep Learning architectures and

much more computation power are making it possible to function better on massive datasets. This achievement could eventually lead to enhanced computer-assisted detection and treatment systems. Given the recent achievements, Deep Learning approaches would significantly advance clinical disease analysis. However, there are a number of challenges that are yet to be resolved in order to utilize the full potential of Deep Learning techniques in medical diagnosis.

REFERENCES

[1] K. Adem, *Exudate detection for diabetic retinopathy with circular hough transformation and convolutional neural networks*, Expert Systems with Applications, 114 (2018), pp. 289–295.

[2] H. F. Al-Yasriy, M. S. AL-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, *Diagnosis of lung cancer based on ct scans using cnn*, in IOP Conference Series: Materials Science and Engineering, vol. 928, IOP Publishing, 2020, p. 022035.

[3] W. Alakwaa, M. Nassef, and A. Badr, *Lung cancer detection and classification with 3d convolutional neural network (3d-cnn)*, International Journal of Advanced Computer Science and Applications, 8 (2017).

[4] M. Alissa, M. A. Lones, J. Cosgrove, J. E. Alty, S. Jamieson, S. L. Smith, and M. Vallejo, *Parkinson's disease diagnosis using convolutional neural networks and figure-copying tasks*, Neural Computing and Applications, 34 (2022), pp. 1433–1453.

[5] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, Y. Duan, and S. R. Oleiwi, *Towards a better understanding of transfer learning for medical imaging: a case study*, Applied Sciences, 10 (2020), p. 4523.

[6] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al., *End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography*, Nature medicine, 25 (2019), pp. 954–961.

[7] P. Baldi, *Deep learning in biomedical data science*, Annual review of biomedical data science, 1 (2018), pp. 181–205.

[8] J. S. Beckmann and D. Lew, *Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities*, Genome medicine, 8 (2016), pp. 1–11.

[9] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al., *Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer*, Jama, 318 (2017), pp. 2199–2210.

[10] C. D. Billones, O. J. L. D. Demetria, D. E. D. Hostallero, and P. C. Naval, *Demnet: a convolutional neural network for the detection of alzheimer's disease and mild cognitive impairment*, in 2016 IEEE region 10 conference (TENCON), IEEE, 2016, pp. 3724–3727.

[11] J. Boschman, *Available online.* `https://medium.com/one-minute-machine-learning/deep-residual-learning-for-image-recognition-2015-one-minute-summary-aa94949b8fcf`, 2021. [Online; accessed on 15 may 2021].

[12] W. Chen, B. Yang, J. Li, and J. Wang, *An approach to detecting diabetic retinopathy based on integrated shallow convolutional neural networks*, IEEE Access, 8 (2020), pp. 178552–178562.

[13] X.-W. Chen and X. Lin, *Big data deep learning: challenges and perspectives*, IEEE access, 2 (2014), pp. 514–525.

[14] W.-J. Choi and T.-S. Choi, *Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach*, Entropy, 15 (2013), pp. 507–523.

[15] P. Chudzik, S. Majumdar, F. Calivá, B. Al-Diri, and A. Hunter, *Microaneurysm detection using fully convolutional neural networks*, Computer methods and programs in biomedicine, 158 (2018), pp. 185–192.

[16] R. dataset, *Available online.* `http://roc.healthcare.uiowa.edu`. [Online;Google Scholar].

[17] Y. Dauphin, H. De Vries, and Y. Bengio, *Equilibrated adaptive learning rates for non-convex optimization*, Advances in neural information processing systems, 28 (2015).

[18] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, et al., *Teleophta: Machine learning and image processing methods for teleophthalmology*, Irbm, 34 (2013), pp. 196–203.

[19] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al., *Feedback on a publicly distributed image database: the messidor database*, Image Analysis & Stereology, 33 (2014), pp. 231–234.

[20] L. Deng, D. Yu, et al., *Deep learning: methods and applications*, Foundations and trends® in signal processing, 7 (2014), pp. 197–387.

[21] F. Doshi-Velez, Y. Ge, and I. Kohane, *Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis*, Pediatrics, 133 (2014), pp. e54–e63.

[22] S. Dutta, B. Manideep, S. M. Basha, R. D. Caytiles, and N. Iyengar, *Classification of diabetic retinopathy images by using deep learning models*, International Journal of Grid and Distributed Computing, 11 (2018), pp. 89–106.

[23] M. T. Esfahani, M. Ghaderi, and R. Kafiyeh, *Classification of diabetic and normal fundus images using new deep learning method*, Leonardo Electron. J. Pract. Technol, 17 (2018), pp. 233–248.

[24] S. Esmaeilzadeh, Y. Yang, and E. Adeli, *End-to-end parkinson disease diagnosis using brain mr-images by 3d-cnn*, arXiv preprint arXiv:1806.05233, (2018).

[25] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, nature, 542 (2017), pp. 115–118.

[26] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun,

and J. Dean, *A guide to deep learning in healthcare*, Nature medicine, 25 (2019), pp. 24–29.
[27] I. Fufurin, P. Berezhanskiy, I. Golyak, D. Anfimov, E. Kareva, A. Scherbakova, P. Demkin, O. Nebritova, and A. Morozov, *Deep learning for type 1 diabetes mellitus diagnosis using infrared quantum cascade laser spectroscopy*, Materials, 15 (2022), p. 2984.
[28] A. K. Gangwar and V. Ravi, *Diabetic retinopathy detection using transfer learning and deep learning*, in Evolution in Computational Intelligence, Springer, 2021, pp. 679–689.
[29] S. Gayathri, V. P. Gopi, and P. Palanisamy, *A lightweight cnn for diabetic retinopathy classification from fundus images*, Biomedical Signal Processing and Control, 62 (2020), p. 102115.
[30] S. Gayathri, V. P. Gopi, and P. Palanisamy, *Diabetic retinopathy classification based on multipath cnn and machine learning classifiers*, Physical and engineering sciences in medicine, 44 (2021), pp. 639–653.
[31] H. Gen and R. Controllers, *Hewlett-packard enterprise development lp*, 2015.
[32] J. George, S. Skaria, V. Varun, et al., *Using yolo based deep learning network for real time detection and localization of lung nodules from low dose ct scans*, in Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, SPIE, 2018, pp. 347–355.
[33] S. Goel, S. Sharma, and R. Tripathi, *Predicting diabetes using cnn for various activation functions: A comparative study*, in 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), IEEE, 2021, pp. 665–669.
[34] J. A. Golden, *Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen*, Jama, 318 (2017), pp. 2184–2186.
[35] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs*, Jama, 316 (2016), pp. 2402–2410.
[36] K. Gunawardena, R. Rajapakse, and N. Kodikara, *Applying convolutional neural networks for pre-detection of alzheimer's disease from structural mri data*, in 2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP), IEEE, 2017, pp. 1–7.
[37] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, *Deep learning for visual understanding: A review*, Neurocomputing, 187 (2016), pp. 27–48.
[38] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.
[39] ———, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[40] D. J. Hemanth, O. Deperlioglu, and U. Kose, *An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network*, Neural Computing and Applications, 32 (2020), pp. 707–721.
[41] N. L. Hill and J. Mogle, *Alzheimer's disease risk factors as mediators of subjective memory impairment and objective memory decline: protocol for a construct-level replication analysis*, BMC geriatrics, 18 (2018), pp. 1–8.
[42] M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, and D. K. Kumar, *Convolutional neural network ensemble for parkinson's disease detection from voice recordings*, Computers in biology and medicine, 141 (2022), p. 105021.
[43] S. Hochreiter, *The vanishing gradient problem during learning recurrent neural nets and problem solutions*, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6 (1998), pp. 107–116.
[44] A. Hoover, V. Kouznetsova, and M. Goldbaum, *Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response*, IEEE Transactions on Medical imaging, 19 (2000), pp. 203–210.
[45] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, *Condensenet: An efficient densenet using learned group convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2752–2761.
[46] X. Huang, J. Shan, and V. Vaidya, *Lung nodule detection in ct using 3d convolutional neural networks*, in 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 379–383.
[47] D. H. Hubel and T. N. Wiesel, *Receptive fields and functional architecture of monkey striate cortex*, The Journal of physiology, 195 (1968), pp. 215–243.
[48] IBSRDataset, *Available online*. `https://www.nitrc.org/projects/ibsr/`, 2020. [Online; accessed on 4 June 2020].
[49] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, et al., *The alzheimer's disease neuroimaging initiative (adni): Mri methods*, Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 27 (2008), pp. 685–691.
[50] R. Kamalraj, S. Neelakandan, M. R. Kumar, V. C. S. Rao, R. Anand, and H. Singh, *Interpretable filter based convolutional neural network (if-cnn) for glucose prediction and classification using pd-ss algorithm*, Measurement, 183 (2021), p. 109804.
[51] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, *The diaretdb1 diabetic retinopathy database and evaluation protocol.*, in BMVC, vol. 1, Citeseer, 2007, p. 10.
[52] S. H. Khan, Z. Abbas, S. D. Rizvi, et al., *Classification of diabetic retinopathy images based on customised cnn architecture*, in 2019 Amity International conference on artificial intelligence (AICAI), IEEE, 2019, pp. 244–248.
[53] S. M. Khan, X. Liu, S. Nath, E. Korot, L. Faes, S. K. Wagner, P. A. Keane, N. J. Sebire, M. J. Burton, and A. K. Denniston, *A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability*, The Lancet Digital Health, 3 (2021), pp. e51–e66.
[54] I. Kollia, A.-G. Stafylopatis, and S. Kollias, *Predicting parkinson's disease using latent information extracted from deep neural networks*, in 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[55] D. Kollias, A. Tagaris, A. Stafylopatis, S. Kollias, and G. Tagaris, *Deep neural architectures for prediction in healthcare*, Complex & Intelligent Systems, 4 (2018), pp. 119–131.
[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 25 (2012).
[57] A. Kurmi, S. Biswas, S. Sen, A. Sinitca, D. Kaplun, and R. Sarkar, *An ensemble of cnn models for parkinson's disease detection using datscan images*, Diagnostics, 12 (2022), p. 1173.
[58] A. Kwasigroch, B. Jarzembinski, and M. Grochowski, *Deep cnn based decision support system for detection and assessing the stage of diabetic retinopathy*, in 2018 International Interdisciplinary PhD Workshop (IIPhDW), IEEE, 2018, pp. 111–116.
[59] C. Lam, D. Yi, M. Guo, and T. Lindsey, *Automated detection of diabetic retinopathy using deep learning*, AMIA summits on translational science proceedings, 2018 (2018), p. 147.
[60] B. A. Landman and S. Warfield, *Miccai 2012: grand challenge and workshop on multi-atlas labeling*, in Proc. international conference on medical image computing and computer assisted intervention, MICCAI, vol. 2012, 2012.
[61] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.
[62] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural networks, 6 (1993), pp. 861–867.
[63] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, *Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening*, Information Sciences, 501 (2019), pp. 511–522.
[64] C.-J. Lin, S.-Y. Jeng, and M.-K. Chen, *Using 2d cnn with taguchi parametric optimization for lung cancer recognition from ct images*, Applied Sciences, 10 (2020), p. 2591.
[65] M. Lin, Q. Chen, and S. Yan, *Network in network*, arXiv preprint arXiv:1312.4400, (2013).
[66] Y.-P. Liu, Z. Li, C. Xu, J. Li, and R. Liang, *Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network*, Artificial intelligence in medicine, 99 (2019), p. 101694.
[67] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, *Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images*, Scientific reports, 8 (2018), pp. 1–13.
[68] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh, A. Gehlot, M. Rashid, S. S. Alshamrani, and A. S. AlGhamdi, *An optimization-based diabetes prediction model using cnn and bi-directional lstm in real-time environment*, Applied Sciences, 12 (2022), p. 3989.
[69] S. M. Murray, *An exploratory analysis of multi-class uncertainty approximation in bayesian convolutional neural networks*, master's thesis, The University of Bergen, 2018.
[70] S. Ogawa and T. Lee, *Nayak as, glynn p. oxygensensitive contrast in magnetic resonance image of rodent brain at high magnetic fields*, Magn Reson Med, 14 (1990), pp. 68–78.
[71] M. N. Önal, G. E. Güraksin, and R. Duman, *Convolutional neural network-based diabetes diagnostic system via iridology technique*, Multimedia Tools and Applications, (2022), pp. 1–22.
[72] J. I. Orlando, E. Prokofyeva, M. Del Fresno, and M. B. Blaschko, *An ensemble deep learning based approach for red lesion detection in fundus images*, Computer methods and programs in biomedicine, 153 (2018), pp. 115–127.
[73] C. G. Owen, A. R. Rudnicka, R. Mullen, S. A. Barman, D. Monekosso, P. H. Whincup, J. Ng, and C. Paterson, *Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program*, Investigative ophthalmology & visual science, 50 (2009), pp. 2004–2010.
[74] G. Pahuja and B. Prasad, *Deep learning architectures for parkinson's disease detection by using multi-modal features*, Computers in Biology and Medicine, (2022), p. 105610.
[75] S. Pal, N. Mishra, M. Bhushan, P. S. Kholiya, M. Rana, and A. Negi, *Deep learning techniques for prediction and diagnosis of diabetes mellitus*, in 2022 International Mobile and Embedded Technology Conference (MECON), IEEE, 2022, pp. 588–593.
[76] S.-I. Pao, H.-Z. Lin, K.-H. Chien, M.-C. Tai, J.-T. Chen, and G.-M. Lin, *Detection of diabetic retinopathy using bichannel convolutional neural network*, Journal of Ophthalmology, 2020 (2020).
[77] M. A. Pfeifer, D. Cook, J. Brodsky, D. Tice, A. Reenan, S. Swedine, J. B. Halter, and D. Porte Jr, *Quantitative evaluation of cardiac parasympathetic activity in normal and diabetic man*, Diabetes, 31 (1982), pp. 339–345.
[78] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, *Learning probabilistic phenotypes from heterogeneous ehr data*, Journal of biomedical informatics, 58 (2015), pp. 156–165.
[79] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, *Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research*, Data, 3 (2018), p. 25.
[80] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, *Convolutional neural networks for diabetic retinopathy*, Procedia computer science, 90 (2016), pp. 200–205.
[81] D. U. N. Qomariah, H. Tjandrasa, and C. Fatichah, *Classification of diabetic retinopathy and normal retinal images using cnn and svm*, in 2019 12th International Conference on Information & Communication Technology and System (ICTS), IEEE, 2019, pp. 152–157.
[82] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, *Deep image mining for diabetic retinopathy screening*, Medical image analysis, 39 (2017), pp. 178–193.
[83] I. Qureshi, J. Ma, and Q. Abbas, *Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning*, Multimedia Tools and Applications, 80 (2021), pp. 11691–11721.
[84] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, *Deep learning for health informatics*, IEEE journal of biomedical and health informatics, 21 (2016), pp. 4–21.
[85] R. O. Roberts, Y. E. Geda, D. S. Knopman, R. H. Cha, V. S. Pankratz, B. F. Boeve, R. J. Ivnik, E. G. Tangalos,

R. C. Petersen, and W. A. Rocca, *The mayo clinic study of aging: design and sampling, participation, baseline measures and sample characteristics*, Neuroepidemiology, 30 (2008), pp. 58–69.
[86] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, *Erfnet: Efficient residual factorized convnet for real-time semantic segmentation*, IEEE Transactions on Intelligent Transportation Systems, 19 (2017), pp. 263–272.
[87] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., *Imagenet large scale visual recognition challenge*, International journal of computer vision, 115 (2015), pp. 211–252.
[88] T. J. Saleem and M. A. Chishti, *Deep learning for internet of things data analytics*, Procedia computer science, 163 (2019), pp. 381–390.
[89] T. J. Saleem and M. A. Chishti, *Deep learning for the internet of things: Potential benefits and use-cases*, Digital Communications and Networks, 7 (2021), pp. 526–542.
[90] T. J. Saleem, S. R. Zahra, F. Wu, A. Alwakeel, M. Alwakeel, F. Jeribi, and M. Hijji, *Deep learning-based diagnosis of alzheimer's disease*, Journal of Personalized Medicine, 12 (2022), p. 815.
[91] B. K. Samhitha, S. C. Mana, J. Jose, R. Vignesh, and D. Deepa, *Prediction of lung cancer using convolutional neural network (cnn)*, International Journal, 9 (2020).
[92] C. Saraiva, C. Praça, R. Ferreira, T. Santos, L. Ferreira, and L. Bernardino, *Nanoparticle-mediated brain drug delivery: overcoming blood–brain barrier to treat neurodegenerative diseases*, Journal of controlled release, 235 (2016), pp. 34–47.
[93] S. Sarraf, D. D. DeSouza, J. Anderson, G. Tofighi, et al., *Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri*, BioRxiv, (2017), p. 070441.
[94] S. Sarraf and G. Tofighi, *Classification of alzheimer's disease structural mri data by deep learning convolutional neural networks*, arXiv preprint arXiv:1607.06583, (2016).
[95] J. Schmidhuber, *Deep learning in neural networks: An overview*, Neural networks, 61 (2015), pp. 85–117.
[96] M. Shaban, Z. Ogur, A. Mahmoud, A. Switala, A. Shalaby, H. Abu Khalifeh, M. Ghazal, L. Fraiwan, G. Giridharan, H. Sandhu, et al., *A convolutional neural network for the screening and staging of diabetic retinopathy*, Plos one, 15 (2020), p. e0233514.
[97] S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P. K. Pal, and M. Ingalhalikar, *Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri*, NeuroImage: Clinical, 22 (2019), p. 101748.
[98] A. Shrestha and A. Mahmood, *Review of deep learning algorithms and architectures. ieee access, 7, 53040-53065*, 2019.
[99] L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, and M. Schäfers, *18f-fdg pet/ct uptake classification in lymphoma and lung cancer by using deep convolutional neural networks*, Radiology, 294 (2020), pp. 445–452.
[100] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, (2014).
[101] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, *Striving for simplicity: The all convolutional net*, arXiv preprint arXiv:1412.6806, (2014).
[102] A. Srinivasulu, K. Ramanjaneyulu, R. Neelaveni, S. R. Karanam, S. Majji, M. Jothilingam, and T. R. Patnala, *Advanced lung cancer prediction based on blockchain material using extended cnn*, Applied Nanoscience, (2021), pp. 1–13.
[103] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, *Ridge-based vessel segmentation in color images of the retina*, IEEE transactions on medical imaging, 23 (2004), pp. 501–509.
[104] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, et al., *Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis*, NeuroImage, 101 (2014), pp. 569–582.
[105] H.-I. Suk and D. Shen, *Deep ensemble sparse regression network for alzheimer's disease diagnosis*, in International Workshop on Machine Learning in Medical Imaging, Springer, 2016, pp. 113–121.
[106] G. Swapna, S. Kp, and R. Vinayakumar, *Automated detection of diabetes using cnn and cnn-lstm network and heart rate signals*, Procedia computer science, 132 (2018), pp. 1253–1262.
[107] G. Swapna, R. Vinayakumar, and K. Soman, *Diabetes detection using deep learning algorithms*, ICT express, 4 (2018), pp. 243–246.
[108] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[109] N. M. A. Tajudin, K. Kipli, M. H. Mahmood, L. T. Lim, D. A. Awang Mat, R. Sapawi, S. K. Sahari, K. Lias, S. K. Jali, and M. E. Hoque, *Deep learning in the grading of diabetic retinopathy: A review*, IET Computer Vision, (2022).
[110] C. Taleb, L. Likforman-Sulem, C. Mokbel, and M. Khachab, *Detection of parkinson's disease from handwriting using deep learning: A comparative study*, Evolutionary Intelligence, (2020), pp. 1–12.
[111] M. ur Rehman, S. H. Khan, Z. Abbas, and S. Danish Rizvi, *Classification of diabetic retinopathy images based on customised cnn architecture*, in 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 244–248.
[112] R. Van Kranenburg, *The Internet of Things: A critique of ambient technology and the all-seeing network of RFID*, Institute of Network Cultures, 2008.
[113] S. K. Vengalil, N. Sinha, S. S. Kruthiventi, and R. V. Babu, *Customizing cnns for blood vessel segmentation from fundus images*, in 2016 International Conference on Signal Processing and Communications (SPCOM), IEEE, 2016, pp. 1–4.
[114] T. Vyas, R. Yadav, C. Solanki, R. Darji, S. Desai, and S. Tanwar, *Deep learning-based scheme to diagnose parkinson's disease*, Expert Systems, 39 (2022), p. e12739.
[115] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, *Deep learning in biomedicine*, Nature biotechnology, 36 (2018), pp. 829–838.

[116] S. Wan, Y. Liang, and Y. Zhang, *Deep convolutional neural networks for diabetic retinopathy detection by image classification*, Computers & Electrical Engineering, 72 (2018), pp. 274–282.
[117] R. Wang, P. Li, and Z. Yang, *Analysis and recognition of clinical features of diabetes based on convolutional neural network*, Computational and Mathematical Methods in Medicine, 2022 (2022).
[118] WHO, *The top 10 causes of death*. `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`, 2020.
[119] Z. Wu, G. Shi, Y. Chen, F. Shi, X. Chen, G. Coatrieux, J. Yang, L. Luo, and S. Li, *Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network*, Artificial Intelligence in Medicine, 108 (2020), p. 101936.
[120] B. Xu, N. Wang, T. Chen, and M. Li, *Empirical evaluation of rectified activations in convolutional network*, arXiv preprint arXiv:1505.00853, (2015).
[121] K. Xu, D. Feng, and H. Mi, *Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image*, Molecules, 22 (2017), p. 2054.
[122] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, and H. J. Aerts, *Deep learning predicts lung cancer treatment response from serial medical imaginglongitudinal deep learning to track treatment response*, Clinical Cancer Research, 25 (2019), pp. 3266–3275.
[123] R. Yamashita, M. Nishio, R. Do, and K. Togashi, *Convolutional neural networks: An overview and application in radiology. insights into imaging*, (2018).
[124] D. R. Yates, C. Vaessen, and M. Roupret, *From leonardo to da vinci: the history of robot-assisted surgery in urology*, BJU international, 108 (2011), pp. 1708–1713.
[125] N. R. Yousif, H. M. Balaha, A. Y. Haikal, and E. M. El-Gendy, *A generic optimization and learning framework for parkinson disease via speech and handwritten records*, Journal of Ambient Intelligence and Humanized Computing, (2022), pp. 1–21.
[126] X. Yue, H. Cai, H. Yan, C. Zou, and K. Zhou, *Cloud-assisted industrial cyber-physical systems: An insight*, Microprocessors and Microsystems, 39 (2015), pp. 1262–1270.
[127] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks*, in European conference on computer vision, Springer, 2014, pp. 818–833.
[128] X. Zeng, H. Chen, Y. Luo, and W. Ye, *Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network*, IEEE Access, 7 (2019), pp. 30744–30753.
[129] J. Zhang, C. Lu, X. Li, H.-J. Kim, and J. Wang, *A full convolutional network based on densenet for remote sensing scene classification*, Mathematical Biosciences and Engineering, 16 (2019), pp. 3345–3367.
[130] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, *Automated identification and grading system of diabetic retinopathy using deep neural networks*, Knowledge-Based Systems, 175 (2019), pp. 12–25.
[131] H. Zhao, C.-C. Tsai, M. Zhou, Y. Liu, Y.-L. Chen, F. Huang, Y.-C. Lin, and J.-J. Wang, *Deep learning based diagnosis of parkinson's disease using diffusion magnetic resonance imaging*, Brain Imaging and Behavior, (2022), pp. 1–12.

Edited by: Katarzyna Wasielewska-Michniewska
Review papers
Received: Jan 21, 2023
Accepted: Dec 10, 2023

AIMS AND SCOPE

The area of scalable computing has matured and reached a point where new issues and trends require a professional forum. SCPE will provide this avenue by publishing original refereed papers that address the present as well as the future of parallel and distributed computing. The journal will focus on algorithm development, implementation and execution on real-world parallel architectures, and application of parallel and distributed computing to the solution of real-life problems. Of particular interest are:

Expressiveness:

- high level languages,
- object oriented techniques,
- compiler technology for parallel computing,
- implementation techniques and their efficiency.

System engineering:

- programming environments,
- debugging tools,
- software libraries.

Performance:

- performance measurement: metrics, evaluation, visualization,
- performance improvement: resource allocation and scheduling, I/O, network throughput.

Applications:

- database,
- control systems,
- embedded systems,
- fault tolerance,
- industrial and business,
- real-time,
- scientific computing,
- visualization.

Future:

- limitations of current approaches,
- engineering trends and their consequences,
- novel parallel architectures.

Taking into account the extremely rapid pace of changes in the field SCPE is committed to fast turnaround of papers and a short publication time of accepted papers.

INSTRUCTIONS FOR CONTRIBUTORS

Proposals of Special Issues should be submitted to the editor-in-chief.

The language of the journal is English. SCPE publishes three categories of papers: overview papers, research papers and short communications. Electronic submissions are preferred. Overview papers and short communications should be submitted to the editor-in-chief. Research papers should be submitted to the editor whose research interests match the subject of the paper most closely. The list of editors' research interests can be found at the journal WWW site (`http://www.scpe.org`). Each paper appropriate to the journal will be refereed by a minimum of two referees.

There is no a priori limit on the length of overview papers. Research papers should be limited to approximately 20 pages, while short communications should not exceed 5 pages. A 50–100 word abstract should be included.

Upon acceptance the authors will be asked to transfer copyright of the article to the publisher. The authors will be required to prepare the text in LaTeX 2$_\varepsilon$ using the journal document class file (based on the SIAM's `siamltex.clo` document class, available at the journal WWW site). Figures must be prepared in encapsulated PostScript and appropriately incorporated into the text. The bibliography should be formatted using the SIAM convention. Detailed instructions for the Authors are available on the SCPE WWW site at `http://www.scpe.org`.

Contributions are accepted for review on the understanding that the same work has not been published and that it is not being considered for publication elsewhere. Technical reports can be submitted. Substantially revised versions of papers published in not easily accessible conference proceedings can also be submitted. The editor-in-chief should be notified at the time of submission and the author is responsible for obtaining the necessary copyright releases for all copyrighted material.