Scalable Computing: Practice and Experience

Scientific International Journal for Parallel and Distributed Computing

ISSN: 1895-1767

Universitatea de Vest din Timişoara

Volume 24(4) December 2023

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

Scalable Computing: Practice and Experience

Volume 24, Number 4, December 2023

TABLE OF CONTENTS

Papers in the special issue on Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities :

Papers in the special issue on Cloud Computing for Intelligent Traffic Management and Control:

Papers in the special issue on Sentiment Analysis and Affective computing in Multimedia Data on Social Network:

Papers in the special issue on Scalability and Sustainability in Distributed Sensor Networks :

Research papers:

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 651–661, DOI 10.12694/scpe.v24i4.2067

GENETIC ALGORITHM SERVICE VULNERABILITY MINING TECHNOLOGY OF ANDROID SYSTEM

XIAOYAN GUO *AND YANFENG SUN †

Abstract. The inland ships energy efficiency is significantly influenced by navigational environment, including speed and direction of wind and water depth. In order to solve the problem of low efficiency of conventional fuzzy test mining, a research method of Android system Service Vulnerability mining technology based on Genetic Algorithm (GA) is proposed. An efficient genetic selection operator model based on probability ranking and combination is also presented to improve the sample coverage and fuzzy test efficiency. Through the framework testing on different systems of mobile phones, multiple system service vulnerabilities are excavated. The execution results guide the generation of test cases, which reduces the proportion of invalid parameters in the test process to improve the efficiency of fuzzy testing. It is observe that the fuzzy test based on GA is much better than the conventional fuzzy test method in the vulnerability mining of system services, and has certain effectiveness and superiority. In addition, after using the two-point crossover algorithm to recombine the gene strings of two individuals, the phenotype of the newly generated individual gene string may become meaningless. It is observed that the selection algorithm factor has a very low p-value, while the ANOVA test confirms at least two groups that have statistically-significant difference.

Key words: System service; Vulnerability mining; Binder Genetic algorithm; Android system; Fuzzy test

1. Introduction. With the rapid growth of mobile Internet technology, mobile devices have greatly improved people's life and entertainment. At present, the mainstream operating systems in the market are mainly Android IOS and Windowphoneo. Android is an intelligent operating system released by Google [1]. Android has a large number of applications and developers. Due to the low threshold of Google's application developers, they can easily get the official developer signature of Android applications from Google, and Google has not adopted a strict security review system, resulting in increasingly serious security problems for Android applications. Android system services play an important role in the whole Android system. While Android system services provide functions for mobile phone users, there are also some threats and risks. These security vulnerabilities can cause serious consequences [2]. For example, if an application obtains the SMS system service in the system service, it will get the user's SMS message, and the user's privacy is likely to be exposed. In addition, if special external data is used in the process of using external system services, Android system services may crash, and even serious consequences such as remote code execution and memory damage may occur. Therefore, the security of Android system services needs our attention. These system services are provided by systems or system applications running in the background. These system services encapsulate the basic functions of Android system, and they open the call interface to ordinary applications [3]. These basic functions include Bluetooth, call and so on.

The system service code occupies the main part of the Android framework. In this sense, this is also an important difference between Android system and traditional desktop PC operating system, so the traditional vulnerability mining tools for desktop operating system are not applicable to Android operating system. In the test process, according to the feedback of the results, guide the genetic algorithm to continuously mutate the test parameters, and propose an efficient genetic selection operator model based on probability ranking and combination, so as to improve the sample coverage and fuzzy test efficiency [4]. Through the testing of the framework on different system versions of mobile phones, multiple system service vulnerabilities are excavated.

According to the query of CNVD (China national vulnerability database) and other well-known vulnerability submission platforms, the previously submitted an droid vulnerabilities are concentrated in the application layer, mostly in the types of component exposure, information disclosure, secondary repackaging, privilege promotion

*Yellow River Conservancy Technical Institute, Kaifeng, Henan 475003, China (xiaoyanguo10098@gmail.com).

† China Radio and television Henan Network Co., Ltd. Kaifeng branch, Kaifeng, Henan 475003, China.

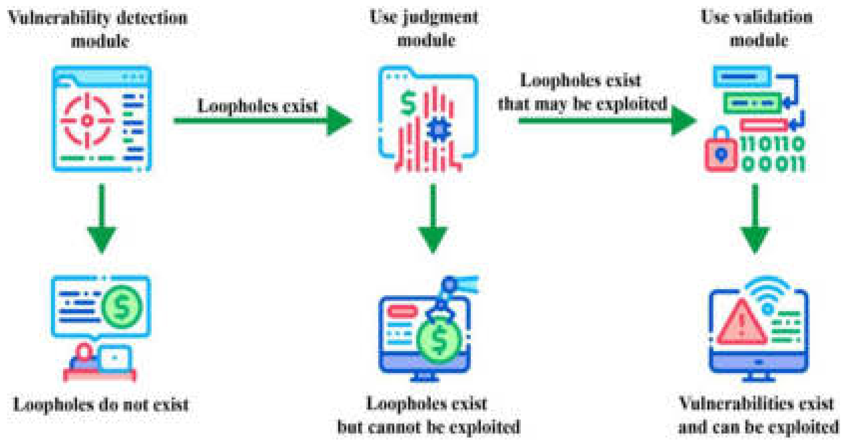


Fig. 1.1: An Android system Service Vulnerability verification method, device and process [7]

and so on, but there is little research and Exploration on the system service level. As the core process running at the bottom of the system, Android system services may lead to privacy disclosure once they are obtained by malicious programs.

In addition, if the service receives and uses the incoming illegal parameters during operation, it may cause unknown results such as system restart, denial of service and even memory damage. At present, the relevant research mainly focuses on the mining of Linux or windows driver vulnerabilities. The driver runs in the system kernel state, and it is difficult to analyze the interaction between device driver and kernel by static analysis; Dynamic analysis needs to run relevant hardware and provide unconventional input, which makes it more difficult to mine driver vulnerabilities. The fuzzy testing technology based on genetic algorithm maps the input data to the gene space through coding, obtains the path conditions by using the white box test method, calculates the fitness of the test cases based on the path coverage, and obtains the test cases that meet the conditions through genetic algorithm [5]. Without the source code, the execution path conditions of the program cannot be known, so this method is inadequate when the source code of Android driver cannot be obtained. How to make the vulnerability mining system generate more effective test cases without source code and dig out more unknown vulnerabilities in less time is the research difficulty of Android driven vulnerability mining technology [6, 7]. Figure 1.1 shows an Android system Service Vulnerability verification method, device and process.

In Android system, application is composed of four components, among which activity and service are two important components. Components may be in the same process or in different processes. When activities or services are in different processes, the cross process communication between them is realized through binder. Android's cross process communication mechanism is based on binder mechanism, not any of the mechanisms mentioned above. Binder cross process communication mechanism is not a communication mechanism created by Android system from 0 to 1.

1.1. Contribution.

1. This paper focuses on a Service Vulnerability mining technology of Android system based on genetic algorithm, which combines feedback mechanism and data optimization through genetic algorithm.
2. In order to solve the problem of low efficiency of conventional fuzzy test mining Android system service vulnerabilities, a research method of Android system Service Vulnerability mining technology based on genetic algorithm is proposed.
3. This paper proposes a research on Service Vulnerability mining technology of Android system based on genetic algorithm. Combining genetic algorithm with fuzzy testing technology, the variation of

parameters is guided by fitness function to ensure the diversity of parameters, and the corresponding combined variation operation is carried out for different parameters according to the variation priority table of data type.

4. This method reduces the influence of combination explosion on parameter genetic variation to a certain extent, and improves the coverage of test cases.

The rest of the paper is organized as follows. The related work is reviewed and discussed in Section 2 followed by the section 3 which explains the research methodology utilized in this work. Section 4 gives the result analysis and Section 5 concludes the paper.

2. Literature Review. At present, many scholars have proposed some new vulnerability mining methods: vulnerability mining methods based on symbolic execution. These methods have the problem of low degree of automation [8].Authors used symbolic execution method to study driver vulnerability mining, and used hardware virtualization technology to solve the problem of requiring specific equipment to test driver vulnerabilities. However, there is a path explosion problem in symbol execution, and the mining type is single. What's more, there is no mature Android full symbolic system, and this technology can not be effectively transplanted to Android driven vulnerability mining. In terms of vulnerability mining, fuzzy testing is one of the fast and simple mining tools, but the traditional fuzzy testing has the shortcomings of lack of understanding of the target program, random and blind testing. How to overcome these shortcomings has always been the focus of fuzzy testing research [9].

Palazzolo, N. and others used the fuzzy test method in the process of Android system Service Vulnerability mining, but the test data type was single and did not construct the complex parameter types in the communication process, which led to the incomplete coverage of parameter use cases to some extent [10]. Zhang, S. Q. and others used fuzzy testing to mine vulnerabilities in system services, but did not reasonably control the variation of multidimensional parameters in system services, which may lead to problems such as combination explosion. How to solve the problems of low use case coverage and multi-dimensional parameter variation in the process of fuzzy testing has also been a loophole [11].

Peng, D. and others proposed a method to test Android system based on binder mechanism adopted by Android system. This method uses a third-party application to destroy the Android system kernel memory by passing abnormal numbers into the system service, so as to obtain the permission to manipulate the kernel space. Then import the shared library into this space, bypass a series of Android system security mechanisms such as SELinux, and achieve the purpose of improving application permissions. This method takes advantage of the negligence of Android system service in parameter checking, tests on Android system, successfully obtains the permission of Android system server through media player system service, and rebounds shell successfully [12].

Martowibowo, S. Y., and others proposed another test method. They analyze the special input verification of Android system services at the system framework layer to find input verification vulnerabilities. They first analyzed the exceptions thrown by the service interface when the Android system service performs input verification on different inputs, looked for the exploitable interface, analyzed the interface, and finally designed the application to scan the similar interfaces that may exist in all Android devices. For the Android system, the vulnerability detection method was used to detect the vulnerability of [90] of the 13 service parameters of the Android system, and finally sent to the Android system. Zhang, B. and others believe that the smooth operation of Android terminal is inseparable from the support of system services. For example, SMS manager is required for receiving and sending SMS, and win dowmanager is required for opening and closing windows. Various service managers provide access interfaces to the bottom layer, which facilitates the call of upper applications [14].

Meng, Y. and others first carried out systematic research on the security of Android customized content and invented the attached system. By dynamically analyzing the relevant files of sensitive operation of the device and comparing them with the relevant files of the native system, we can see the difference between its security protection and the native system. After several platform tests, it is found that the customized systems of different manufacturers have different degrees of security problems. Although some customization related driver vulnerabilities have been tested, there is no specific research on Android driver vulnerability mining [15].

Author evaluates the K-Nearest Neighbor (K-NN) supervised algorithm performance in determining students' learning styles [16]. Edeh Author presented the entrepreneurship education across the globe. Author examines cyber-security awareness among undergraduate students from crime on the cyber-space [17].

Author in this paper discuss the problem of firmware vulnerability mining and the traditional method of vulnerability mining research based on fuzzing test which is not efficient. A noval mining vulnerabilities method in industrial firmware is proposed. This method constructs test cases for the variables for triggering the vulnerabilities [18]. The presented method can reduce about 23% of test cases and can effectively improve test efficiency.

Author in this paper presents a novel solution for detecting rare malware programs and provides the scarcity of datasets for modeling these malware [19]. Author's analysis system includes an internet simulator and a human emulator to successfully execute them and prevent system halting. An objective function is used to optimize the vital indicators and tracking rate with a linear time complexity. Real-world malware samples were used for the performance evaluation and comprehensive scenarios were involved to evaluate the proposed strategy performance. The results demonstrate the improvement in detection accuracy and the results also demonstrated an enhancement in true positive rate for the presented deep-learning algorithm. Author in this work presented an Android malware detection framework GA-Stacking which employs stacking to compose five different base classifiers [20]. The GA is applied to optimize the hyper parameters of the framework and experiments show that stacking could improve malware detection accuracy as compared with single classifier. The presented technique achieves accuracies of 98.43% and 98.66% on CIC-And Mal and CIC MalDroid datasets, which shows the efficiency and feasibility of the presented method. Author in this paper presented a machine learning-based detection approach by utilizing hybrid analysis-based particle swarm optimization (PSO) and an adaptive genetic algorithm (AGA) [21]. The feature selection is performed by applying PSO in the dataset. Further, the XGBoost and random forest (RF) machine learning classifiers performance is optimized utilizing the AGA. With the random forest classifier, an accuracy of 98.72% and F-score were achieved. Our results present that the PSO application and an AGA greatly increases the classification performance of the information obtained from the hybrid analysis. Author in this paper details GA-based feature selection which helps Android malware detection [22]. The machine learning algorithms with GA-based feature selection for 1104 static features included in the Andro-AutoPsy dataset is used. The comparative analysis is done by the author and showed that the GA performed better than the information gain-based method, which is generally used as a feature selection method. Moreover, machine learning using the presented GA-based feature selection has an absolute advantage in terms of time as compared to ML without feature selection. Further, it is useful to apply GA-based feature selection to improve malware detection performance.

2.1. Research Gaps. The inland ships energy efficiency is significantly influenced by navigational environment, including speed and direction of wind and water depth. The inland navigational environment complexity makes it difficult for determining the optimal speeds under different environmental conditions for the best energy effectiveness. In order to solve the problem of low efficiency of conventional fuzzy test mining, an efficient research method of is needed.

3. Research Methodology.

3.1. Research on Android vulnerability mining. Security vulnerabilities refer to some problems and defects in some security schemes, which are embodied in the detailed implementation of hardware, software and protocols. The damage to the system can be completed by the attacker without authorization. These security vulnerabilities are usually some vulnerable people left by carelessness. These entries may exist in computer hardware and computer components, applications or some online resources. Binder is an inter process communication mechanism in Android system. In Android system, generally speaking, different applications run in different processes. For the same application, different system components can also run in different processes [23]. When a process wants to provide services for other processes, it needs to provide services through inter process communication. In Android system, application is composed of four components, among which activity and service are two important components. Components may be in the same process or in different processes. When activities or services are in different processes, the cross process communication between them is realized through binder. Android's cross process communication mechanism is based on binder mechanism. Binder cross process communication mechanism is not a communication mechanism created by Android system from 0 to 1. It is developed on the basis of open binder project. As an IPC mechanism, binder's architecture is a distributed

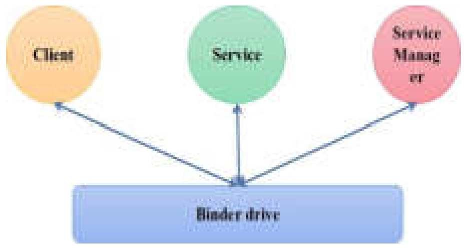


Fig. 3.1: Binder architecture

component architecture, which can provide remote calling functions. Binder mechanism is composed of four parts, including user space and kernel space. The four components are client, server, binder driver and service manager. The client, server and service manager run in user space and the driver runs in kernel space, as shown in Figure 3.1. Binder mechanism can effectively combine the above four parts. Binder driver component is the core member, client server interaction in user space is completed through driver, and service manager is responsible for auxiliary management of system services. The interaction between client and server is completed through the underlying driver. The Android system has helped us realize the driver and service manager, and the rest of the client and server need to be implemented by Android developers themselves [24].

3.2. Genetic Algorithm.

3.2.1. Basic flow of genetic algorithm. The search of genetic algorithm begins with a potential solution set of the problem, which is also the parameter space of the actual problem. For different problems, there are many methods to generate the parameter space of practical problems. In fuzzy testing, random algorithm is usually used to generate the parameter space of practical problems. Gene code each input parameter in the parameter space according to the predetermined coding rules to generate the initial population, and then perform the following iterative process until the predetermined iterative threshold is reached or the required optimal solution has been found: The fitness value of each individual is calculated through the fitness function, and then according to the fitness value of the individual, the better individual is selected according to a certain selection algorithm to be inherited to the next generation [25]. The selected individuals are paired in pairs, and the paired two individuals are cross operated according to a certain cross probability to exchange some genes. Then, an individual is randomly selected from these individuals to randomly change the value of one or some loci in the individual gene coding string with a certain mutation probability. Usually, the value of the locus is replaced by alleles. In the iterative process, individuals with poor fitness can be appropriately eliminated according to specific conditions. The basic flow of genetic algorithm is shown in Figure 3.2. After the iteration of genetic algorithm, the individual encoded by gene needs to be decoded to obtain the optimized solution of the actual problem.

3.2.2. Gene Coding. When using genetic algorithm to solve specific problems, we first need to solve the problem of parameter coding and decoding in the actual problem solution space. The parameters in the solution space of practical problems can only be processed by genetic algorithm after they are transformed into individuals (also known as chromosomes) in the genetic space represented by gene strings using certain coding rules [26]. This process is called the coding process from individual phenotype to gene in genetic algorithm. On the contrary, the conversion process from individual genotype to phenotype is called decoding process. The conversion process is shown in Figure 3.3.

The common gene coding methods of genetic algorithm include binary coding, gray coding and floating point coding. Binary coding is the most commonly used coding method. The gene encoded by it is a binary string composed of 0 or 1. Using binary coding, encoding and decoding operations and genetic operations such as mutation and crossover are easy to implement, and also comply with the coding principle of minimum character set. However, due to the randomness of binary coding, its local search ability is poor. After mutation, the individual's phenotype changes greatly and is easy to be far away from the optimal solution. Gray code

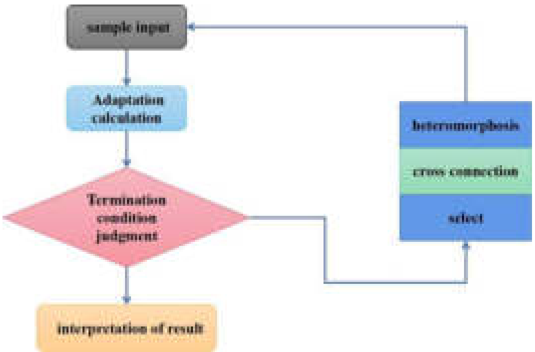


Fig. 3.2: Basic flow chart of genetic algorithm

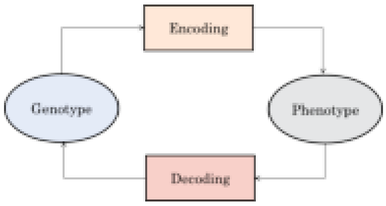


Fig. 3.3: Schematic diagram of genetic algorithm encoding and decoding

is an improvement of binary code. Only one code point is different between the codes corresponding to two adjacent integers encoded by gray code, and the other code points are exactly the same. Assuming that there is binary code , and its corresponding gray code is , the binary code can be converted into the corresponding gray code using formulas (1) and (2), and the gray code can be converted into the corresponding binary code using formulas (3) and (4), where i=1,2......n-1. For the individual encoded by gray code, the phenotype before mutation and the phenotype after mutation are continuous, so it has better local search ability. At the same time, because gray code is still a binary code in essence, gray code still has the advantages of easy implementation of genetic operations such as crossover and mutation, and conforms to the coding principle of minimum character set. Floating point number coding method is to represent each gene of chromosome with a real number. The coding length of individual depends on the number of decision variables. Floating point coding is often used to solve the continuous function optimization problem with multi-dimensional and high precision requirements. When dealing with individuals with such problems, binary coding will produce gene coding individuals with large length, which will lead to a sharp increase in search space. Floating point coding method has the advantages of improving the accuracy of genetic algorithm, improving the computational complexity of genetic algorithm, and facilitating genetic search in a large space.

$$g_n = b_n \tag{3.1}$$

$$g_i = b_i * \oplus * b_{i+1} \tag{3.2}$$

$$b_n = g_n \tag{3.3}$$

$$b_i = b_{i+1} * \oplus * b_i \tag{3.4}$$

3.3. Genetic manipulation. In order to make the individuals in the population approach the optimal solution in the process of generation by generation evolution, in the process of genetic iteration, it is necessary to perform certain genetic operations on the individuals according to the fitness value. There are three common genetic operations: selection, crossover and variation. These three genetic operations are also called genetic operators in genetic algorithms.

3.3.1. Selection operator. Selection, also known as replication, is the first step of genetic operation. It eliminates the fittest according to the fitness value of each individual calculated by the fitness function: the higher the probability that the individual with higher fitness will be inherited to the next generation, and the lower the probability that the individual with lower fitness will be inherited to the next generation. In this way, the individuals in the population can continuously approach the optimal solution. The main function of selection operation is to avoid losing useful genetic information in genetic iteration and improve global convergence and computational efficiency. Therefore, the choice of algorithm design will affect the final result of the algorithm. Common selection operators include roulette selection, random competition selection, best reservation selection and random selection without playback.

3.3.2. Crossover operator. The crossover operation of genetic algorithm simulates the process of two homologous chromosomes forming new chromosomes through mating and recombination. Crossover is also called recombination. Its basic operation is to select two parent individuals according to a certain probability and form two new offspring individuals by exchanging some gene strings of the two parent individuals. The two offspring individuals inherit some genes of the parent individuals. Through the exchange of gene strings, new chromosomes are generated, which improves the diversity of the population. The key to the design of crossover operator of genetic algorithm mainly lies in: 1. The determination of intersection position; 2. How to carry out partial gene exchange. Crossover operator is the main method of generating new individuals in genetic algorithm, which plays an important role in the correct implementation of genetic algorithm. Before the crossover operation, the individuals in the population need to be paired in pairs. The more common pairing strategy is random pairing, that is, randomly assign n individuals in the population to [n / 2] pairing groups. The crossover operation is completed on two individuals in each paired group.

3.3.3. Mutation operator. In genetic algorithm, the mutation operation is completed by replacing some gene values in the individual gene coding string with their alleles. For example, in binary coding, change ”0” to ”1” or ”1” to ”0”. Using mutation operation in genetic algorithm can avoid the loss of some information caused by selection and crossover operation. Crossover operation is the main method of genetic algorithm to generate new individuals, which determines the global search ability of genetic algorithm. Although mutation operation is only an auxiliary method to generate new individuals in genetic algorithm, it can avoid premature phenomenon and improve the local search ability of genetic algorithm, which determines the local search ability of genetic algorithm. Cross operation and genetic operation cooperate with each other, so that genetic algorithm can obtain better search performance when solving optimization problems. There are two main purposes of using mutation operator in genetic algorithm: 1. Improve the local search ability of genetic algorithm; 2. Maintain the diversity of the population and prevent the occurrence of precocity.

4. Results and Discussion. For the initial test data set, the traditional simple genetic algorithm and the optimized genetic algorithm are used for optimization. The primary optimization result of a population is shown in Figure 4.1. In Figure 4.1, the horizontal axis represents the population evolution algebra, and the vertical axis represents the average fitness value of individuals in each generation of population.

As can be seen from Figure 4.1, although the traditional simple genetic algorithm converges rapidly and the population has achieved high fitness value in the early stage of evolution, the average fitness value after population convergence is lower than that obtained by the genetic algorithm optimized in this paper. In order to verify the performance of ASFuzzer in actual vulnerability mining, this paper uses ASFuzzer to mine vulnerabilities in practical applications in the Internet, and compares the mining results with those of WFBGA and spike. This experiment digs the official website of a university and the official website of an organ. The comparison of vulnerability mining results of a university is shown in Figure 4.2 .

From the analysis of experimental results, the number of vulnerabilities mined by ASFuzzer framework is more than the number of system service vulnerabilities mined by conventional fuzzy test under the same system version, and some results have been achieved in the test on the customized system of third-party manufacturers. Genetic algorithm mutation generates highly diversified test cases, which greatly improves the possibility of triggering vulnerabilities. By analyzing the interface and its parameters that generate exceptions, 15 of the 20 vulnerabilities are caused by the variation of multi parameter combination, that is, exceptions will be triggered only when specific data is filled.The experimental results show that the fuzzy testing based

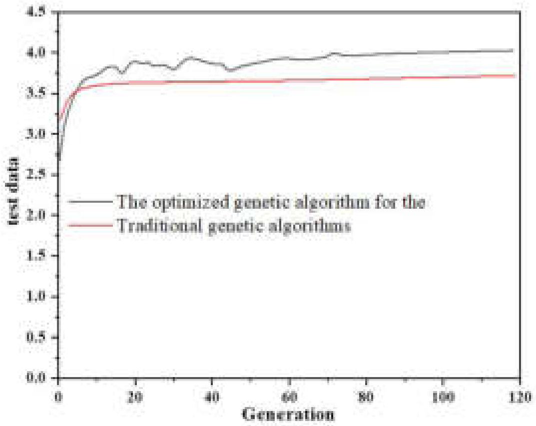


Fig. 4.1: Genetic optimization results

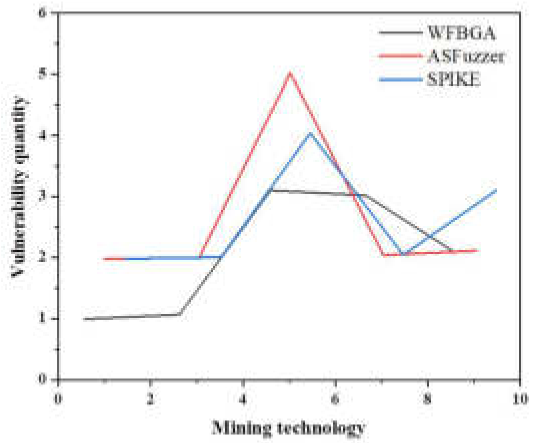


Fig. 4.2: Comparison of vulnerability mining results

on genetic algorithm is much better than the conventional fuzzy testing methods in the vulnerability mining of system services, and has certain efficiency and superiority. In the mining strategy of genetic algorithm, the individual in the current population is selected randomly in the selection operator stage. As individuals about to enter the next round, random selection may miss excellent individuals, mislead the direction of testing and affect the mining efficiency. This paper makes full use of the variability range of a single parameter and the quantitative relationship of parameters, and combines the arrangement and combination method to guide the selection operation in the process of genetic algorithm, so as to ensure the timely input of excellent individuals, so as to guide the development of mining testing in the direction of high efficiency.

4.1. Comparison among evolutionary configurations. To determine the performance of the configuration in the evolutionary, every combination of the 3 fitness functions are considered and applied these combinations on each of the system services, for evolutionary fuzzing campaigns. Number of covered basic

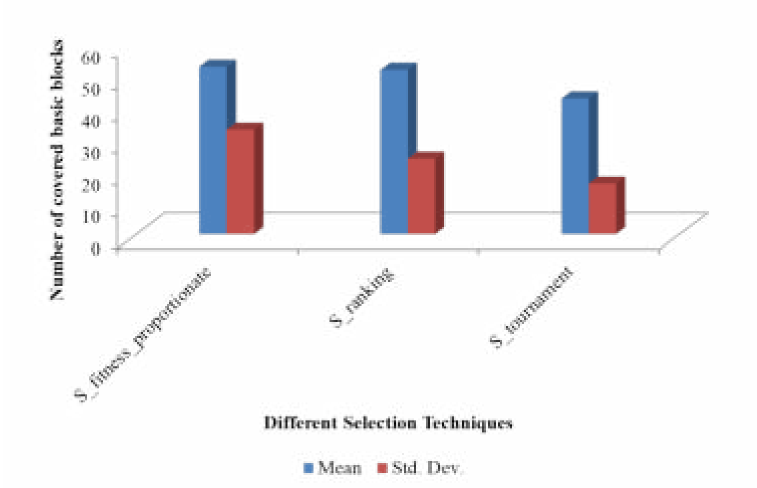


Fig. 4.3: Number of covered basic blocks, with respect to different selection algorithms

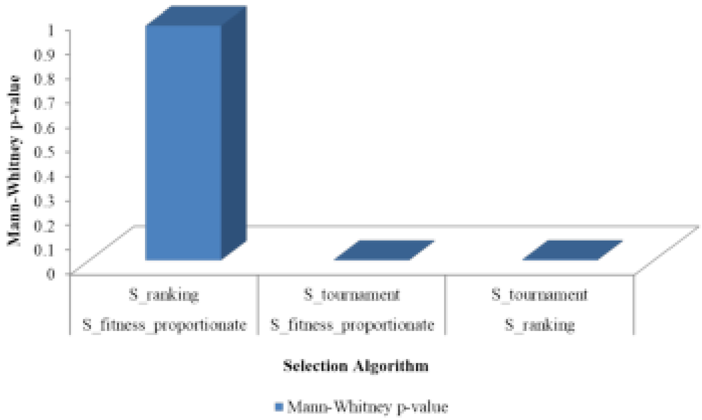


Fig. 4.4: Pairwise analysis of selection algorithms

blocks w.r.t different selection algorithms are shown in Figure 4.3.

Figure 4.3 show the number of basic blocks grouped by the fitness functions and the selection algorithms. The selection algorithm factor has a very low p-value: the selection algorithm has a significant effect, however, while the ANOVA test confirms at least two groups that have statistically-significant difference, it does not point out the specific group to detect the best selection algorithm for fuzzing. The pairwise tests and measure effect size are performed in each case of the results are in Figure 4.4.

The null hypotheses are considered that fitness function choices and the selection algorithm have no effect on the testing, and the non-parametric Kruskal-Wallis tests are performed since data are not distributed normally. The selection algorithm factor has the choice of a selection algorithm having significant effect. The best selection algorithm detection is performed for evolutionary fuzzing. The pairwise tests are performed and the effect size is measured.

5. Conclusion. This paper introduces a research on Service Vulnerability mining technology of Android system based on genetic algorithm, which combines feedback mechanism and data optimization through genetic algorithm. The execution results are used to guide the generation of test cases, which reduces the proportion of invalid parameters in the test process and improves the efficiency of fuzzy testing. After being applied to the Android driver vulnerability mining practice, several unpublished driver vulnerabilities are found, such as binder, camera and other exploitable denial of service attacks. Through the continuous optimization of use cases by genetic algorithm, it can generate more comprehensive and diversified use cases to test the objective function, and then trigger the vulnerability. By analyzing the interface and its parameters that generate exceptions, 15 of the 20 vulnerabilities are caused by the variation of multi parameter combination, that is, exceptions will be triggered only when specific data is filled. The test results on different systems show that ASFuzzer test framework can effectively mine the vulnerabilities of system services, and then find some potential security problems; At the same time, it also shows the efficiency superiority of this scheme compared with the conventional fuzzy test mining method. The current research work still has some shortcomings as follows: in the multiple execution process of genetic optimization algorithm, it still converges to the local optimal solution. This may be due to the insufficient continuity of gray code encoding for string data. In addition, after using the two-point crossover algorithm to recombine the gene strings of two individuals, the phenotype of the newly generated individual gene string may become meaningless. These may lead to poor optimization results of genetic algorithm. In future, this work can be continuing on a project to determine the mobile technology influence on the menace of cybercrimes in Nigeria.The selection algorithm factor has the choice of a selection algorithm having significant effect. The best selection algorithm detection is performed for evolutionary fuzzing. The pairwise tests are performed and the effect size is measured. The future research will focus on the present energy efficiency optimization to provide the foundation for energy efficiency.

REFERENCES

[1] Guo-Hong, S., Application development research based on android platform, In 2014 7th International Conference on Intelligent Computation Technology and Automation, IEEE, 579-582, 2014.

[2] Arabo, A., & Pranggono, B., Mobile malware and smart device security: Trends, challenges and solutions, IAIC Transactions on Sustainable Digital Innovation, 1(2), 526-531, 2013.

[3] Alam, T, Cloud Computing and its role in the Information Technology, IAIC Transactions on Sustainable Digital Innovation, 1(2), 108-115, 2020.

[4] Atabati, A., Alizadeh, M., Schuh, H., & Tsai, L. C., Ionospheric scintillation prediction on s4 and roti parameters using artificial neural network and genetic algorithm, Remote Sensing, 13(11), 2092-2096, 2021.

[5] Ponticelli, G. S., Lambiase, F., Leone, C., & Genna, S., Combined fuzzy and genetic algorithm for the optimisation of hybrid composite-polymer joints obtained by two-step laser joining process, Materials, 13(2), 283-288, 2020.

[6] [4] Duan, Li, Ruizheng, Shi, Ni, & Yao, et al., Real-time patient-specific ecg arrhythmia detection by quantum genetic algorithm of least squares twin svm, Journal of Beijing Institute of Technology, v.29(01), 32-40, 2020.

[7] Hraiech, S. E., Chebbi, A. H., Affi, Z., & Romdhane, L., Genetic algorithm coupled with the krawczyk method for multi-objective design parameters optimization of the 3-upu manipulator, Robotica, 38(6), 1138-1154, 2020.

[8] Zameni, M., Rezaei, A., & Farzinvash, L., Two-phase node deployment for target coverage in rechargeable wsns using genetic algorithm and integer linear programming, The Journal of Supercomputing, 77(12), 1-29, 2021.

[9] Atabati, A., Alizadeh, M., Schuh, H., & Tsai, L. C., Ionospheric scintillation prediction on s4 and roti parameters using artificial neural network and genetic algorithm, Remote Sensing, 13(11), 2092-2096, 2021.

[10] Palazzolo, N., Peres, D. J., Bordoni, M., Meisina, C., & Cancelliere, A., Improving spatial landslide prediction with 3d slope stability analysis and genetic algorithm optimization: application to the oltrepò pavese,Water, 13(6), 801-806, 2021.

[11] Zhang, S. Q., & Zhang, Y., Harmonic detection method based on permutation entropy and variational modal decomposition optimized by genetic algorithm, Review of Scientific Instruments, 92(2), 025118-025122, 2021.

[12] Peng, D., Tan, G., Fang, K., Chen, L., & Zhang, Y., Multiobjective optimization of an off-road vehicle suspension parameter through a genetic algorithm based on the particle swarm optimization, Mathematical Problems in Engineering, 2021(9), 1-14, 2021.

[13] Martowibowo, S. Y., & Damanik, B. K., Optimization of material removal rate and surface roughness of aisi 316l under dry turning process using genetic algorithm, Manufacturing Technology, 21(3), 373-380, 2021.

[14] Zhang, B., Zhou, X., Liu, Y., Yang, B., & Zhang, Y., ACombining application of wavelet analysis and genetic algorithm in wind tunnel simulation of unidirectional natural wind field near a sand ground surface, Review of Scientific Instruments, 92(1), 015123-015126, 2021.

[15] Meng, Y., Liang, Y., Zhao, Q., & J Qin., Research on torsional property of body-in-white based on square box model and multiobjective genetic algorithm, Mathematical Problems in Engineering, 2021(41), 1-13, 2021.

[16] Onyema, E.M; Elhaj, M.A.E; Bashir, S.G; Abdullahi, I; Hauwa, A.A; Hayatu, A.S., Evaluation of the Performance of K-Nearest Neighbor Algorithm in Determining Student Learning Styles, Int. J. of Innovative Sci., Eng. & Techn., 7 (1), 91-102, (2020).

[17] Onyema, E.M.., Edeh, C. D.., Gregory, U.S..,. Edmond, V.U., Charles, A.C. and Richard-Nnabu, N.E. Cybersecurity Awareness Among Undergraduate Students in Enugu Nigeria, International Journal of Information Security, Privacy and Digital Forensic (Nigeria Computer Society), 5 (1), 34 -42, (2021).
[18] Li, Y., Liu, X., Tian, H., & Luo, C. (2018, November). Research of Industrial Control System Device Firmware Vulnerability Mining Technology Based on Taint Analysis. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) (pp. 607-610). IEEE.
[19] Javaheri, D., Lalbakhsh, P., & Hosseinzadeh, M., A novel method for detecting future generations of targeted and metamorphic malware based on genetic algorithm IEEE Access, 9, 69951-69970, 2021
[20] Xie, N., Qin, Z., & Di, X., GA-StackingMD: Android Malware Detection Method Based on Genetic Algorithm Optimized Stacking Applied Sciences, 13(4), 2629, 2023.
[21] Hammood, L., Doğru, İ. A., & Kiliç, K., Machine Learning-Based Adaptive Genetic Algorithm for Android Malware Detection Applied Sciences, 13(9), 5403, 2023.
[22] Lee, J., Jang, H., Ha, S., & Yoon, Y., Android malware detection using machine learning with feature selection based on the genetic algorithm Mathematics, 9(21), 2813, 2023.
[23] Senanayake, J., Kalutarage, H., Al-Kadri, M. O., Petrovski, A., & Piras, L., Android source code vulnerability detection: a systematic literature review ACM Computing Surveys, 55(9), 1-37, 2023.
[24] Li, G., Source Code Vulnerability Mining Method Based on Graph Neural Network International Journal of Frontiers in Engineering Technology, 4(4).
[25] Sarker, I. H., Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems SN Computer Science, 3(2), 158, 2022.
[26] Zhang, H., Song, R., Yang, J., Dan, W. U., & Wang, Y., Connection damage detection of double beam system under moving load with genetic algorithm, Mechanika, 27(1), 80-87, 2021.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Nov 13, 2022
Accepted: Sep 9, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 663–672, DOI 10.12694/scpe.v24i4.2069

SECURE STORAGE OF COMPUTER NETWORK DATA BASED ON CLOUD COMPUTING

HONGWEI JIANG*

Abstract. Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources. User's data is stored in large database. The stored data can be accessed and modified by the clients over the Internet. The data is monitored by the Third Party Auditor (TPA) on behalf of the client. Therefore, integrity is lacked by the data stored on the servers. The data integrity is ensured by the cloud services that provide trust to the privacy of users. Aiming at the urgent problem of network data security in cloud computing operations, this paper proposed a security situation assessment system to grasp the security situation in real time. A set of cloud computing network data storage security is proposed. Through this model, the extraction of network data storage security situation elements, the design of network data storage security situation assessment scheme and the calculation method of network data storage security situation value are completed. The experimental results show the error of the predicted value obtained by the network data storage security situation assessment system. The effectiveness of the system model and the superiority of the improved algorithm is verified by the experimental results. This paper uses the cloud model to predict the cloud computing network security situation. On the other hand, the security situation value obtained by the situation assessment process can be used directly without training the original situation value. Performance improvement by the proposed technique over existing technique is seen and it is observe that the proposed technique is 23% and 34% better than the existing techniques.

Key words: Cloud computing; Network security; Situational assessment; Situational prediction; Secure storage; Third Party Auditor; Computing resources

1. Introduction.

The cloud storage system's development and its application in complex environment are increasing rapidly, so, the security of the data has been more and more consideration. With the birth of computers and the rapid development of the Internet, the current research focus is to combine computers that run alone to deal with problems, improve processing power and processing efficiency, and achieve effects similar to "supercomputers". Cloud computing is one of the most popular research directions in the current field. As a new type of network architecture and network computing model, all sectors of society have paid special attention. The improvement of cloud computing service technology to realize multi-tenant technology ensures that they can have some customized functions. The Internet is the carrier for cloud computing to realize resource sharing. However, the Internet also has heterogeneity and openness. Therefore, cloud computing operations may be attacked by the network all the time, such as malicious tampering of user information, interception or deletion of user data, etc. It can be seen that in the cloud computing system, safe and effective network protection and monitoring are very necessary for the entire security system. The research on network security management based on cloud computing is shown in Figure 1.1.

Another important function of the cloud computing platform is data storage. Because the cloud computing platform stores a large amount of data of each node, and the services of the cloud computing platform are open and extensive, the cloud computing platform is subject to considerable attacks and harms. It is required to carry out the necessary security protection while the data is stored. To sum up, for the cloud computing platform, we need to formulate corresponding security protection assessment strategies, monitor the security status of the entire cloud computing system in real time, and provide detailed security protection logs. The entire platform is analyzed as a whole, resulting in a final analysis report. In this way, the combination of intelligent analysis platform and manual analysis can timely and accurately detect network attacks and make relevant security policy adjustments.

The great convenience is offered by moving huge data into cloud since it reduces hardware management

*Jilin Sport University, Jilin, Changchun, 130022, China (hongweijiang4321@yahoo.com).

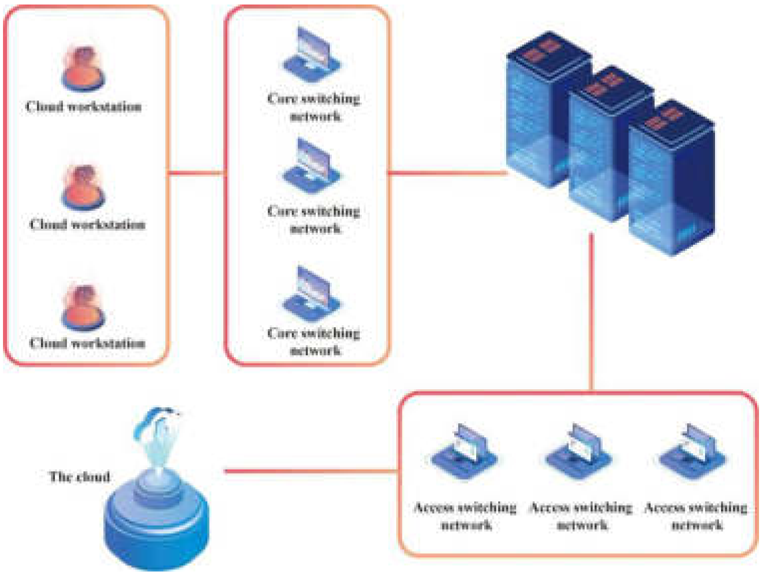


Fig. 1.1: Research on network security and management application based on cloud computing

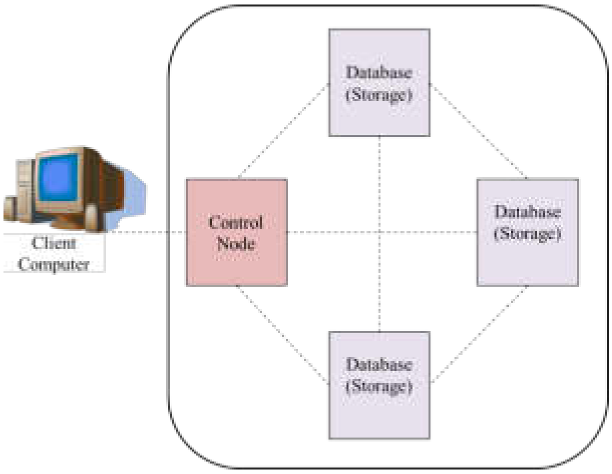


Fig. 1.2: Architecture of cloud storage system

and data maintenance burden. Many types of services such as data as a service and infrastructure are provided to users. Scalable, secure, for clients at low cost is offered by it. The CSP could discard the data which has not been accessed so to overcome the security threats, CSP require a mechanism for the users' data integrity insurance. The third party auditing is emerged to resolve the problem. A cloud storage system architecture includes a control server and storage servers, as shown in figure 1.2.

1.1. Contribution. 1. A research idea of adopting security situation assessment to grasp the security situation of the overall system in real time is presented.

2. The great convenience to users is offered by moving huge data into cloud since it reduces hardware management and data maintenance burden at local machines. Many types of services such as data as a service and infrastructure are provided to users. The third party auditing is emerged to resolve the problem.

Rest of the paper is organized as follows. The related work is reviewed and discussed in Section II followed by the section III which explains the research methodology utilized in this work. Section IV gives the result analysis and Section V concludes the paper.

2. Literature Review. With the rapid development of cloud computing technology, there are several problems that need to be solved urgently, such as: cloud security problems are becoming more and more prominent, showing a diversified and complex trend, what security technology is used to protect the tenant's environment, etc. [1]. The research results of the Cloud Security Alliance listed in Literature [2] show that Internet-based cloud computing servers have been attacked by a large number of hackers. attack. Literature [3] pointed out that the concept of cloud computing originated from European and American countries. As the birthplace of this technology, they have mastered many leading theories and technologies, and as a result, many very famous companies in the field of cloud computing have emerged, such as Yahoo, Yahoo! Technology companies such as Microsoft and Google. In the literature [4], many operators around the world have also launched cloud computing-based services, such as BT in the UK and Verizon in the United States.

These countries have started to conduct research on network security in cloud computing very early, and the current technical level is in the leading position in the world. Taking the United States as an example, it already has a complete security infrastructure, security assessment criteria, certified encryption standards and related regulations, and has formed a sound information security industry chain [5]. Literature [6] pointed out that "Cloud Security Alliance" is a non-profit, non-profit organization, its task is to solve the security problems existing in the cloud computing network, propose solutions, and improve the security index of cloud computing. After the Cloud Security Alliance was established in the United States, cloud computing providers such as Microsoft and Google have joined the alliance. The alliance currently consists of 34 members. Reference [7] pointed out that the research in the field of cloud computing security in the United States has always been at the forefront of the world. American researchers have proposed a variety of cloud security frameworks, and many methods have been applied to actual cloud computing, such as: user data Encryption techniques, secure network connections, secure computers, etc. The giants of cloud computing service providers in the world have already formed an independent system in cloud security, such as increasing investment in cloud computing, attaching importance to research on cloud computing security, continuously strengthening their cloud computing platforms, and launching new cloud computing platforms. security mechanism, etc. [8].

Cloud computing technology started late in China and has not yet been applied on a large scale. Document [9] records the details of a research report called "Cloud Computing in China", which reflects the development of cloud computing in China well, and is a good example for the further development and innovation of cloud computing in China. Reference is provided. Literature [10] shows that, in order to promote the development of cloud computing, the government has established cloud computing R and D centers in Guangzhou, Shanghai, Beijing and other places, and actively conducts research on cloud computing-based applications and cloud computing security. In addition, many Internet companies have also joined the research team of cloud computing [11]. According to the analysis, the current research on the security situation assessment strategy of the cloud computing platform is still lacking and formulate the network security situation assessment of cloud computing platform [12].

The author in this paper detailed the cloud computing based computing model to support the shared pool of computing resources access. Due to the data outsourcing, integrity and data security, it becomes challenging [13]. Author in this paper discussed the process of fighting against network security and the traditional defense has been difficult. To meet the computer network security needs under cloud computing and for the cloud computing high-quality system can be optimized gradually. In this paper, the cloud computing is utilized for security storage design and to ensure the reliability and data upload storage security [14]. Program utilizes the boot password for the existing data encryption security in the management; system design by correcting Tornado data redundancy code. A secure cloud storage prototype system is also implemented by the author based on Cassandra. It is observe that the system can provide the ability of data loss recovery and effectively resist the fault [15]. The author in this article discussed the accessibility of the resources obtainable from the cloud whenever users want, therefore, users purchase the IT service that they do not have maintain things. The data storage model is computing which considers as a web-based generation which utilizes remote servers. Author gives new designing for the information security storage construction where information encrypted and divided into many blocks and distributed between services suppliers instead of relying on one supplier for information storage [16].

Author in this paper discussed a new method of cloud computing that has brought great convenience to network life, but with some security risks. The concept and characteristics of cloud computing are described analyzes the significance for the computer network security in the cloud computing environment and analyzes the security vulnerabilities according to the cloud computing characteristics [17]. Author in this paper discussed a new method of cloud computing that has brought great convenience to network life, but with some security risks. Author in this paper presented a neuro-fuzzy approach for the user behaviour classification and prediction. The analysis is complicated by each user's feedback and the various rules have been implemented for addressing the company's policy to determine the precise behaviour of a user [18]. A Gaussian Radial Basis Function Neural Network (GRBF-NN) is trained for prediction on the basis of set generated by a Fuzzy Rule Based System (FRBS). The scheme is found to be promising in prediction accuracy. Author presented a resource-based task algorithm which is implemented and analyzed to understand the heterogeneous multi-cloud network performance [19]. Author in this paper, a heterogeneous integrated network resource management algorithm is presented. The algorithm adopts the information security transmission technology advantage to collect resources in heterogeneous integrated network and establishes a resource management algorithm model on the basis of information security transmission [20]. The resource management algorithm effectiveness is determined which reduces resource management errors and improve security performance in the resource management process. The main data encryption technology and intelligent collection process of the Internet of Things (IoT) is also discussed by the author. Author in this paper presented a new Chinese Remainder Theorem (CRT)-based data storage mechanism for the user data storage [21]. The CRT-based secured storage scheme adopts encryption schemes which use formulas for performing the encryption and also introduced a new formula for data decryption. In addition, a new formula is introduced for accessing the encrypted cloud data from the cloud database. The security models have been evaluated by analyzing the results and finally, it is proven that the proposed data security model is better comparatively. The author aims to develop a basic cloud-based design for ICC laboratories improvement. The proposed design is built via using "Software as a Service" model [22]. Cloud computing is designed by utilizing the Private Cloud Computing. The presented design provided flexibility to ICC and allows computer network capabilities improvement and managing the resources easily.

2.1. Problem Statement. User's data is archieved in large data centers. The stored data can be accessed and modified by the clients. The data is monitored by the Third Party Auditor on behalf of the client. Therefore, integrity is lacked by the data stored on the servers. The data integrity is ensured by the cloud services that provide trust to the privacy of users.

3. Proposed Method. The role of network security posture assessment in this paper is to inform the system of what dangers may occur, so as to realize the safe storage of network data. The specific process is to preprocess the collected original safety data information, extract the characteristic information of system safety events, and obtain an estimated probability value by using certain mathematical models and calculation methods to determine whether certain safety events occur. As shown in Figure 3.1, this model mainly has three layers from high to low: situational prediction, situational assessment and situational awareness.

The first layer is the situational awareness layer, which is the basis of the entire situational assessment model. At present, there are very mature technical means, which can obtain enough data through the situational awareness layer. By processing the collected data, all information about the current network operating status can be obtained. In order to complete the assessment of the security situation, the situation information is usually transformed into a form that is easier for people to understand, such as XML [23]. The second layer is the situation assessment layer, which is the core of the entire situation assessment model. Security identification is performed on the data obtained by the upper layer, the correlation between security events is mined, the security situation value is calculated, and a security situation curve graph is generated to reflect the security situation of the entire system. The third layer is the situation prediction layer, which judges and predicts the future security situation according to the past and present network security situation, and makes early response strategies and processing method. A significant feature of cloud computing is big data. A large amount of network data is a must for security situation assessment. Moreover, the redundancy between data and false information make the calculation method of situation assessment very complicated. Security situation assessment is a comprehensive research topic, which includes data processing methods, network modeling requirements, and so

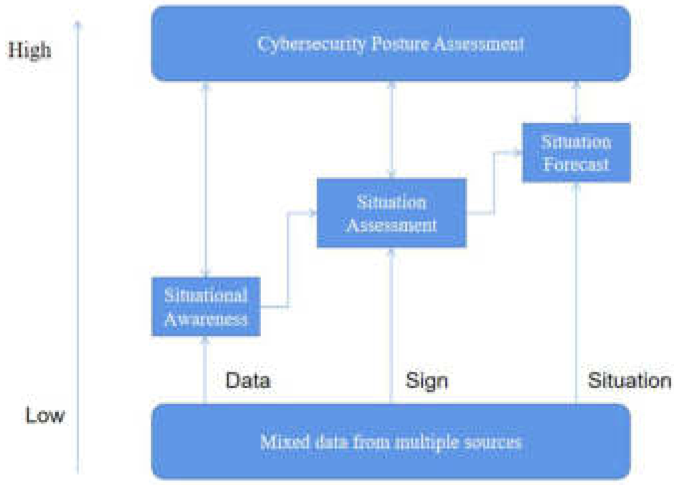


Fig. 3.1: Basic Model of Network Security Situation Assessment

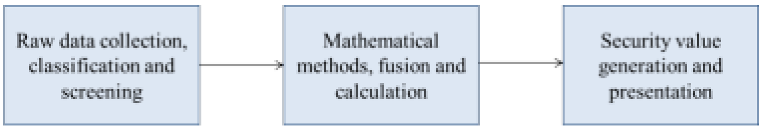


Fig. 3.2: The generation process of network security situation value

on. In the current technology and theory, the two major technologies of data mining and data fusion are the main methods for processing large amounts of data.

3.1. Calculation of security situation value. The size of the network security situation value can clearly characterize the operation of the network. The larger the situation value, the more unstable and dangerous the network operation is. After a series of mathematical calculations, after preprocessing the collected network data, the data is converted into one or several groups of data to obtain the network security situation value. The size of the network security situation value will change with the different network operating conditions. For example, the network has been attacked and suffered different types of attacks. By observing the changes in data, network security managers can judge the security situation of the network, and then judge whether the network is threatened. Figure 3.2 shows the process of generating a network security situation value.

This algorithm not only reduces the complexity of the original mapping algorithm, but also reduces the generated errors by using the super entropy value, which are all in the original mapping. Algorithms based on increased accuracy. During the solution process, the membership degree of each network security situation value is not used, but the cloud model parameters are calculated directly by using the statistical characteristics of the cloud model, which not only avoids certain errors, but also simplifies the mapping algorithm.

3.2. Challenges and Issues of Cloud Data Storage. The control over the stored data is not provided by the cloud computing based on cloud data centers. There is full data control by cloud service providers as they perform malicious tasks like copy, modifying, etc. The certain levels of control are ensured by the cloud computing over the virtual machines. Due to lack of control over the data, a greater security issues are there than the generic cloud computing model. The figure 3.2 has many issues which need to discuss clearly.

Less cost and less resource management is provided by the cloud computing but it has also some security threats. The cloud computing ensures the integrity, privacy and availability of data in cloud computing but it

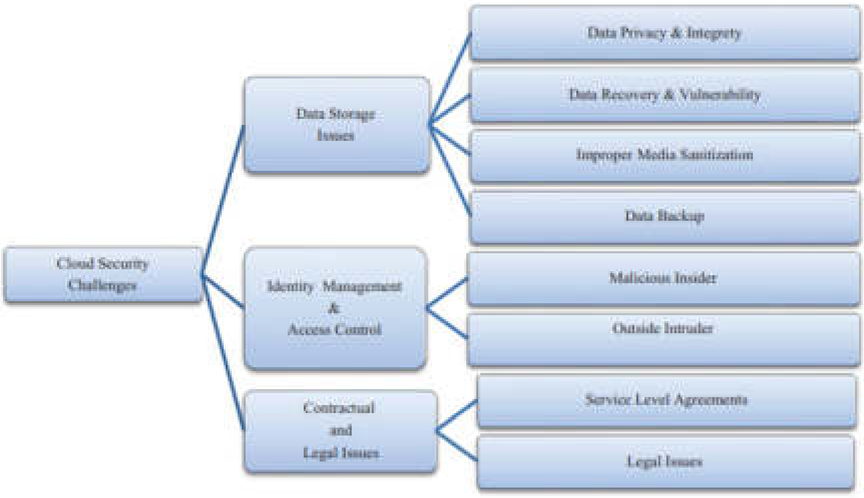


Fig. 3.3: Challenges of cloud security

Table 4.1: Attack Category and Threat Level

Attack Category	Level of threat
Misc_antivity	1
Network scan	1
RPC_port map decode	2
Attempted dos	2
Mapping modified	2
Attempted admin	3
Http_uri decode	3
Shell code detect	3

is vulnerable for the security threats. The simplicity cloud users are increasing and the applications hosted in cloud is very great.

4. Analysis of results.

4.1. Source of experimental data. At present, there are mainly three kinds of experimental data widely used in the assessment and prediction of network information security situation in the world: the honeynet data set collected by the honeynet project group of the network security organization; the Defcon data set provided by the network security expert organization ShmooGroup; and the MIT Lincoln experiment Lincoln Laboratory public dataset provided by the laboratory. According to the overview and characteristic analysis of the three data sets in the literature, the honeynet data set is very beneficial to be used as experimental data for simulation tests to simulate the data that may be generated when the cloud computing network system is attacked. This paper first uses honeynet data as the experimental test data, selects the honeynet data set of a certain month, and effectively combines the relevant knowledge of the open source intrusion detection system Snort, and summarizes the data set required for the experiment. The types of attacks on the network system used in the experiment are shown in Table 4.1, and the corresponding threat levels are also marked.

Table 4.2: The number of times the network was attacked

Type	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8
Ping	40	0	1	0	8	50	45	0
DNS	3	9	6	24	14	35	24	10
DOS	0	16	1	0	20	20	10	5
RPC	0	6	1	0	0	10	5	3
Shellcode	2	4	0	0	1	3	1	0
Http	2	0	28	0	0	15	15	14

Table 4.3: The number of hosts attacked on the network

Type	Host number 1	Host number 2	Host number 3	Host number 4	Host number 5	Host number 6	Host number 7	Host number 8
Ping	1	0	1	0	1	1	1	0
DNS	1	3	1	6	2	8	3	4
DOS	0	5	1	0	8	6	5	3
RPC	0	3	1	0	0	5	5	2
Shellcode	1	2	0	0	1	3	1	0
Http	1	0	1	0	0	1	1	1

4.1.1. Analysis of experimental results. Combined with Table 4. 1, the experimental data is analyzed. This paper counts the attack elements, the number of attacks, and the number of attacked hosts on the network system in an average of 8 time periods within a month. The time can be determined according to the performance of the system. to resize. The statistical results are shown in Tables 4. 2 and 4. 3 below. The threat levels corresponding to the attack types Ping, DNS, DOS, RPC, Shellcode, and Http are 1, 2, 2, 2, 3, and 3, respectively.

According to Table 4.2 and Table 4. 3, it can be obtained that the security situation values of the network system in 1 month and 8 time periods (t1 t8) are (0.193, 0.214, 0.423, 0.076, 0.763, 0.872, 0.825, 0.565) respectively. The corresponding network The security situation diagram is shown in Figure 4.1.

It can be seen from Figure 4.1 that the security situation index value of the network system in the three time periods of t1, t2 and t4 is small, indicating that the network in this time period is in a relatively safe and stable state; the network is in a relatively safe and stable state at t3 , t5, t8, although it suffered a certain attack threat during the three time periods, but it can still operate normally; the security situation index of the network in the two time periods of t6 and t7 is high, indicating that the network in this time period suffered from When there is a serious attack threat and the network is in an unsafe state, network managers should pay more attention and take appropriate measures.

Comparing Tables 2 and 3, it effectively proves the rationality of the cloud computing network security situation assessment method designed in this paper, and the obtained assessment results conform to the objective facts. In the experiment, the data set is used to further verify the evaluation model and prediction algorithm designed in this paper. When the network security situation index value is between (0, 0.3), the network is running safely; when it is between (0.3, 0.8), the network has suffered a certain attack, but it can still operate normally; between (0.8, 1) When the network is in an insecure state, it has suffered a serious attack threat. Of course, in the actual security management process, network managers can dynamically set the threshold of the network security situation index according to specific security defense regulations. According to the cloud computing network security situation forecast introduced above, the predicted value of network security situation is obtained as shown in Figure 4.2. The figure shows the comparison between the predicted curve of network security situation and the actual curve, and the prediction of network security situation value in the next two weeks is drawn. value and true value. It can be seen from the figure that the trend of the predicted

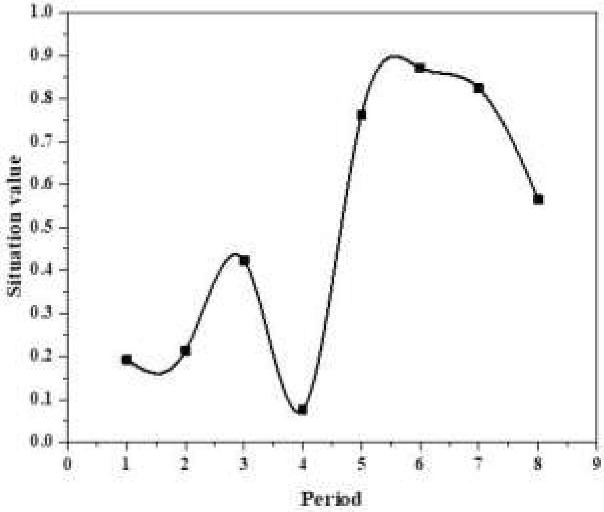


Fig. 4.1: Network Security Situation Map

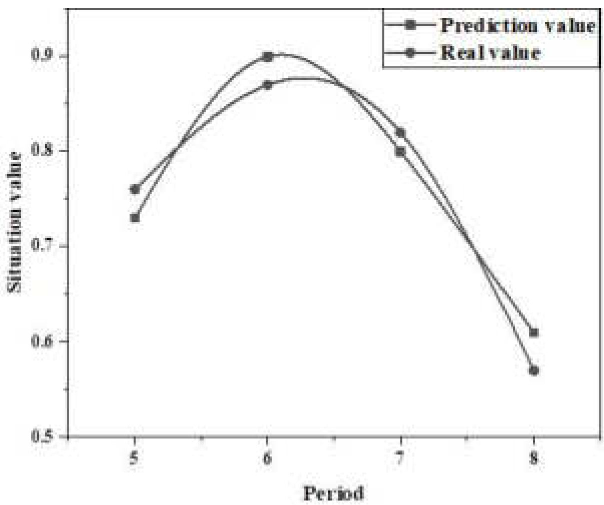


Fig. 4.2: Network Security Situation Forecast

curve and the actual curve have a trend value error of no more than 0.05 in the four time intervals, which is basically the same. This result shows that the improved network security situation prediction method adopted in this experiment is correct and meets the actual requirements.

The proposed technique performance is compared with the existing techniques in terms of prediction accuracy. It is better than the existing techniques in prediction accuracy. Performance improvement by the proposed technique over existing technique is shown graphically in Figure 4.3. It is seen that the proposed technique is 23% and 34% better than the existing and existing techniques.

5. Conclusion. After full analysis and the existing research knowledge, this paper firstly proposes an extraction model of network security situation elements according to the cloud computing architecture. The model is divided into three layers. And use the mapping algorithm based on this model to calculate the network security situation value. Secondly, in view of the randomness and ambiguity of the network state, this paper uses the cloud model to predict the cloud computing network security situation. Finally, the prediction results

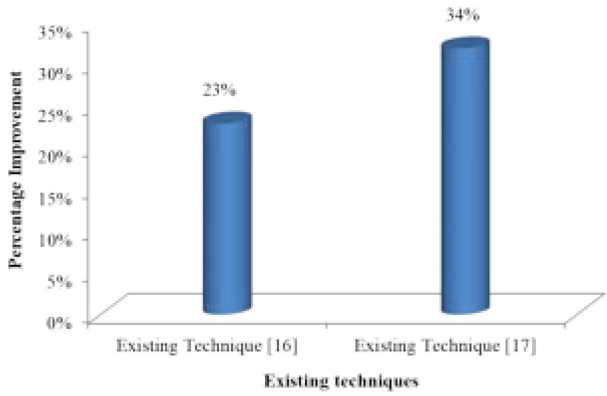


Fig. 4.3: Performance Improvement of the proposed technique over existing techniques

of network security situation are presented in tabular and graphical forms respectively, which verifies the correctness and superiority of the mapping algorithm and the element model of network security situation.Compared with the traditional network security situation assessment research, there are few related researches on cloud computing and the research in this paper is not very complete and comprehensive. It is hoped that the future research directions are: 1. According to the characteristics of cloud computing big data, Investigate better algorithms so that data can be analyzed from the data that is more useful for cloud security. 2. For the specific environment of cloud computing, further study the security situation assessment algorithm and the security situation prediction algorithm to make the assessment results and prediction results more accurate and more in line with the objective reality. 3. Strive to design and develop a software system, which is truly used for cloud computing network security situation assessment, and can realize automatic control and real-time update. Performance improvement by the proposed technique over existing technique is seen and it is observe that the proposed technique is 23% and 34% better than the existing techniques.

REFERENCES

[1] Novo, O. Ding, M. H., Research on Clustering Algorithm of Heterogeneous Network Privacy Big Data Set Based on Cloud Computing, International Conference on Advanced Hybrid Information Processing (pp. 367-376). Springer, Cham, 2021.
[2] Liu, Y., International logistics taxation data monitoring based on 5g network and cloud computing platform, Microprocessors and Microsystems, 82, 103826, 2021.
[3] Gong, J. , Design of Distributed Network Mass Data Processing System based on Cloud Computing Technology, 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021.
[4] Song, H. , Li, J. , & Li, H. , A cloud secure storage mechanism based on data dispersion and encryption, IEEE Access, 9, 63745-63751, 2011.
[5] Khalaf, O. I. , & Abdulsahib, G. M., Optimized dynamic storage of data (odsd) in iot based on blockchain for wireless sensor networks, Peer-to-Peer Networking and Applications, 1-16, 2021.
[6] Kim, M. , Jiang, X. , Lauter, K. , Ismayilzada, E. , & Shams, S., Hear: human action recognition via neural networks on homomorphically encrypted data, arXiv preprint arXiv:2104.09164, 2021.
[7] Zhong, C. , Jiang, X. , & Qi, G., Video-based person re-identification based on distributed cloud computing, Journal of Artificial Intelligence Technology(2), 11, 2021.
[8] Zhang, L., Optimization of the marketing management system based on cloud computing and big data, Complexity, 2021.
[9] Kahn, M. G. , Mui, J. Y. , Ames, M. J. , Yamsani, A. . , Pozdeyev, N. , & Rafaels, N., Migrating a research data warehouse to a public cloud: challenges and opportunities, Journal of the American Medical Informatics Association, 2021.
[10] Dehury, C. K. , & Sahoo, P. K., Failure aware semi-centralized virtual network embedding in cloud computing fat-tree data center networks,arXiv e-prints, 2021.
[11] Tian, J. , & Wang, H., A provably secure and public auditing protocol based on the bell triangle for cloud data, Computer Networks, 195(1), 108223, 2021.
[12] Susilo, W. , Jiang, P. , Lai, J. , F.Guo, & Deng, R., Sanitizable access control system for secure cloud storage against malicious data publishers, IEEE Transactions on Dependable and Secure Computing, PP(99), 1-1, 2021.

[13] Subashini, S., & Kavitha, V., A survey on security issues in service delivery models of cloud computing, Journal of network and computer applications, 34(1), 1-11, 2011.
[14] Wang, X., & Shi, L. Design of computer network security storage system based on cloud computing technology. In Journal of Physics: Conference Series (Vol. 2083, No. 4, p. 042084). IOP Publishing, 2021.
[15] Wang, R. Research on data security technology based on cloud storage. Procedia engineering, 174, 1340-1355, 2017.
[16] Kadhim, Q. K., Mahdi, H. S., & Ail, H. K., Storage Architecture for Network Security in Cloud Computing. Diyala Journal for Pure Science, 14(1), 1-17, 2018.
[17] Luo, L., & Ge, W. Research on the Security of Computer Network under Cloud Computing. In 2018 3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018) (pp. 278-281). Atlantis Press, 2018.
[18] Dash, S., Luhach, A. K., Chilamkurti, N., Baek, S., & Nam, Y. A Neuro-fuzzy approach for user behaviour classification and prediction. Journal of Cloud Computing, 8(1), 17, 2019.
[19] Mishra, S. K., Mishra, S., Alsayat, A., Jhanjhi, N. Z., Humayun, M., Sahoo, K. S., & Luhach, A. K. Energy-Aware Task Allocation for Multi-Cloud Networks. IEEE Access, 8, 178825-178834, 2020.
[20] Ding, L., Wang, Z., Wang, X., & Wu, D. Security information transmission algorithms for IoT based on cloud computing. Computer Communications, 155, 32-39.
[21] Ganapathy, S. A secured storage and privacy-preserving model using CRT for providing security on cloud and IoT-based applications. Computer Networks, 151, 181-190.
[22] Ouda, G. K., & Yas, Q. M Design of cloud computing for educational centers using private cloud computing: a case study In Journal of Physics: Conference Series (Vol. 1804, No. 1, p. 012119).
[23] Qin, Q. , Jin, B. , & Liu, Y., A secure storage and sharing scheme of stroke electronic medical records based on consortium blockchain, BioMed Research International, 2021(5), 1-14, 2021.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Nov 14, 2022
Accepted: Nov 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 673–685, DOI 10.12694/scpe.v24i4.2147

DATA MINING ANALYSIS FOR IMPROVING DECISION-MAKING IN COMPUTER MANAGEMENT INFORMATION SYSTEMS

XIAOHONG DONG* AND BING XIANG†

Abstract. In this paper a decision tree-based data mining procedure for information systems is proposed to enhance the accuracy and efficiency of data mining. An enhanced C4.5 decision tree method based on cosine similarity is suggested to evaluate the information gain rate of characteristics and the information entropy of their values. When the information entropy variance among any two values for attributes is within the threshold range, the cosine similarity of the merging attribute values is determined, and the information gain rate of the attributes is recalculated. Large-scale data sets that conventional data processing methods are unable to handle successfully have given rise to the area of data mining. The prime objective is to look into how data mining technology is used in computer management information systems. The benefits of data mining technologies in computer management information systems are examined from a variety of angles in this study. In order to analyze and comprehend huge data sets and to derive knowledge that can be utilized to enhance the decision-making process in computer management information systems, the suggested solution makes use of a number of data mining techniques, including Clustering, Classification, and Association Rule Mining. The experimental analysis indicates that the time required by the proposed method to construct a decision tree is less than the time required by the GBDT, P-GBDT method and the C5.0 decision tree Hyperion image forest type fine classification method. The minimum time is not more than 15 seconds when compared with the minimum time saving of the other two methods. The time required by the C5.0 decision tree Hyperion image forest type fine classification method is always the greatest in comparison with the minimum time saving of the C5.0 decision tree. The classification accuracy of the proposed method for various datasets exceeds 95 percent, and the data mining efficacy is high. This method enhances the precision and efficacy of data mining in order to uncover valuable information concealed behind a large volume of data and maximize its value.

Key words: Decision tree; Information system; Data mining; Information entropy; Cosine similarity.

1. Introduction. As Internet technology has continued to advance, database-based information systems have steadily filtered into numerous domains of various businesses, serving as the foundation for data warehousing in those sectors. The choice of databases for various information systems varies as a result of the various data volume and application requirements in distinct data management information systems [1]. Databases are primarily divided into two groups, relational databases and non-relational databases, as a result of the ongoing development of database technology. The main relational databases at the moment are Oracle, PostgreSQL, MySQL, and so on. The relational database summarizes the complex data structure into a straightforward two-dimensional table form, solving the problem of centralized storage and sharing of data. However, there are still some shortcomings in the independence and abstraction level of data. To manage marketing information and enhance marketing decision-making, a methodological approach utilizing data mining along with information management technologies is offered. The cornerstone for improving the management of client relationships is this technique.

According to goals, data mining can be separated into two types of tasks: prediction tasks and descriptive tasks. The fundamental objective of description tasks is to identify patterns in related data sets that may indicate prospective linkages. For example, association analysis, trend analysis, clusters analysis, etc., description tasks are usually exploratory [2]. The goal of the prediction task is to forecast the value of a specific attribute in light of some fixed attributes of the input data. Typically, these fixed attributes are referred to as variables that are independent and explanatory variables, while the target variable and dependent variable are the particular features of the prediction. In the data mining analytical approach, classification and regression are the primary prediction jobs, and correlation analysis, cluster analysis, and time period based analysis are the primary description tasks. In huge databases, association analysis is primarily used to unearth important connections that

*College of Media, Hulunbuir University, Hulunbuir, Nei Monggol, 021008, China (xiaohong21dong@gmail.com).
†College of Media, Hulunbuir University, Hulunbuir, Nei Monggol, 021008, China (bingxiang16@gmail.com).

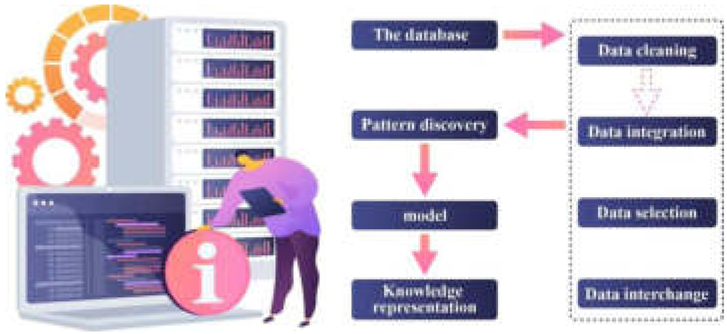


Fig. 1.1: Data mining technology

are described as association rules or frequent item sets. Data is divided into groupings (clusters) that make sense or are valuable, and cluster analysis is merely the first stage in data analysis. For instance, the data is clustered before data aggregation, and the clusters are divided into those intended for understanding and those intended for use [3]. Figure 1.1 shows a case of data mining in action. Time series analysis primarily explains how a study item behaves across time changes and forecasts and analyses potential future events in light of the object's previous laws or shifting patterns. For instance, it studies how database users behave to forecast the fields that will be utilised the most frequently in the future and gives database maintenance staff thorough data assistance. These changes have the greatest impact on the marketing department since it is responsible for direct consumer contact when businesses make the switch to CRM.

There is growing agreement that a thorough grasp of the customers' needs and preferences is the key to good customer relationship management. In these situations, data mining methods may help in uncovering hidden facts and enhancing customer comprehension, while a rigorous knowledge management effort can help in directing the data into effective marketing strategies. As a result, marketing may greatly benefit from studies on knowledge management and extraction. The main goal of this study is to solve the problems brought on by the growing data in computer management information systems. This study makes a contribution by offering a practical remedy for data management and decision-making procedures in computer management systems. The solution offers a thorough method for data analysis to enhance the precision, dependability, and performance of computer management systems. The remainder of the article is organized as follows: Section 2 of the article contains a literature review, and Section 3 of the article provides an explanation of the research techniques, including a description of the decision tree algorithm and cosine similarity. The findings and discussion are presented in Section 4, which is followed by the conclusion section in Section 5.

2. Literature Review. The research conducted by Li et al. examined the performance assessment of a collection of K-means algorithm-based methods. This research demonstrates the effectiveness of integrating the decision tree algorithm with cluster analysis technique for performance assessment. The system pertains to the monitoring and assessment of employee performance inside a business [4]. Kabanihin et al. did a study on the assessment of employee performance and the development and use of the ID3 decision tree algorithm. They brought the notion of decision making to the ID3 algorithm, resulting in a reduction in algorithmic complexity. Currently, China has placed increased emphasis on the use of assessments of performance methodologies and has undertaken collective endeavours to enforce principles of equity, scholarly inquiry, and performance assessment [5]. Murdan et al. used the ID3 decision tree algorithm in the context of human resources, therefore facilitating the organization's decision-making process [6]. The study conducted by Kakhki et al. used the ID3 algorithm to assess the effectiveness of personnel in research organisations. In order to establish performance indicators, a data mining technique was utilised. The use and investigation of these algorithms have resulted in the development of an optimal approach to enhance the efficiency of the construction venture management procedure and facilitate the execution phase [7]. Santoso and colleagues devised a Concept Learning System (CLS) aimed at facilitating the initial training of decision trees [8]. The decision tree method is extensively used in several domains due to its widespread utilisation as a data mining technique. The evolution of this phenomenon exhibits a spectrum ranging from rudimentary to intricate, including both superficial and profound manifestations that may last over extended periods of time. Currently, there is ongoing research on the decision tree algorithm internationally, with the objective of enhancing its accuracy and exploring various approaches to integrate the algorithm with other relevant tools in order to get superior outcomes. External decision tree methods are extensively used in several domains such as education, performance assessment, scientific research, and others. Furthermore, it is essential to emphasise the significance of domestic research and development in this particular domain. The author describes the evolution of the C4.5 decision tree method, which utilises cosine similarity, as a means to execute data mining inside a system. The use of the C4.5 decision tree technique allows for the consolidation of comparable values, resulting in a reduction in the size of the decision tree, a decrease in code complexity, and an enhancement in both classification accuracy and functionality. Consequently, this approach enables more efficient practical application. Data mining refers to the systematic examination and analysis of data with the objective of revealing concealed, but potentially important, information [9-11]. In order to uncover previously unidentified patterns and ultimately attain comprehensible knowledge, it is essential to carefully choose, investigate, and construct models based on substantial volumes of data.Data mining encompasses a diverse array of computer approaches, such as statistical evaluation, decision trees, neural networks, rule generation and improvement, and visual representation. Data mining techniques have gained increased appeal and use due to developments in computer hardware and software, particularly in the realm of exploratory tools such as data visualisation and neural networks.

The growing volume of data presents problems for data processing, knowledge discovery, and decision-making in computer management information systems. When applied to enormous amounts of data in computer management information systems, traditional data processing procedures are laborious, inefficient, and might not yield correct findings. Data mining techniques may be applied to these issues to increase data accuracy and reveal hidden patterns and information in sizable data sets.

An overview of current papers on the use of data mining technologies in computer management information systems is given in Table 2.1. Each article is explained in terms of the technology employed; the datasets utilized the advantages, disadvantages, and potential remedies. Viswanathan et al. employed the HDFC and SBI datasets with random forests and support vector machines. Increased efficiency and precision were advantages, while data noise was a disadvantage. Advanced preparation methods and data cleansing were potential options [12]. Vu et al. [13] uses the MNIST and CIFAR-10 datasets were subjected to artificial neural networks and k-means clustering. Increased productivity and time savings were two advantages, while over fitting was a disadvantage. As potential remedies, regularization and hyper parameter tweaking were recommended. Nti et al. [14] uses NYSE and NASDAQ datasets were employed with decision trees and random forests. The study concentrated on improved accuracy, although there was a cost associated with the employment of computationally intensive techniques. The use of ensemble techniques was suggested as a potential remedy. Association rule mining and support vector machines were used with weather data in [15]. Increased decision-making capacity was a gain, while complicated data structures were a disadvantage. As potential fixes, preprocessing and feature selection were recommended. Transformers and a self-attention mechanism were applied to EHR data in [16]. The cost-prohibitive aspect was a disadvantage even if the emphasis was on better feature representations. Dimensional reduction and sophisticated preprocessing methods were potential remedies. Naive Bayes and logistic regression were used to data from retail marketing in [17]. Benefits included enhanced decision-making ability, but a disadvantage was limited precision. The authors suggested preprocessing and hybrid models as potential remedies.

The Apriori algorithm and logistic regression were applied to healthcare data for enhanced gastroenteritis diagnosis in [18]. Feature selection and regularization were recommended as potential remedies for over fitting, which was a problem. In [19], to improve accuracy, ensemble classifier and decision trees were applied to credit risk data. The problem of data imbalance was acknowledged, and sampling and ensemble models were suggested as potential remedies. Clustering analysis and logistic regression were used to Twitter data for enhanced virus detection, according to Khanday et al. [20]. The problem of vocabulary mismatch was acknowledged, and enhanced preprocessing was proposed as a potential remedy. Principal component analysis and k-means clustering were used with student academic data for better student success prediction in [21]. Noisy data was

Table 2.1: Recent innovations and contributions from several studies

Reference	Technology Used	Datasets used	Benefits	Drawbacks	Possible Solutions
[12]	Random forest, Support Vector Machines	HDFC, SBI	Increased accuracy, Efficiency	Data Noise	Clean Data, Advanced Preprocessing
[13]	Artificial Neural Networks, K-Means Clustering	MNIST, CIFAR-10	Increased Efficiency, Time Savings	Overfitting	Regularization, Hyperparameter Tuning
[14]	Decision Trees, Random Forests	NYSE, NASDAQ	Increased Accuracy	Intensity of computationally expensive methods	Ensemble Methods
[15]	Association Rule Mining, Support Vector Machines	Weather data	Increased Decision-Making Capabilities	Complex Data Structures	Preprocessing, Feature Selection
[16]	Transformers, Self-Attention Mechanism	EHR data	Improved Feature Representations	Cost Prohibitive	Reduced Dimensions, Advanced Preprocessing
[17]	Naive Bayes, Logistic Regression	Retail Marketing Data	Improved Decision-Making Capabilities	Low Accuracy	Hybrid Models, Preprocessing
[18]	Apriori Algorithm, Logistic Regression	Healthcare Data	Improved Detection of Gastroenteritis	Overfitting	Feature Selection, Regularization
[19]	Ensemble Classifier, Decision Trees	Credit Risk Data	Increased Accuracy	Data Imbalance	Sampling, Ensemble Models
[20]	Clustering Analysis, Logistic Regression	Twitter Data	Improved Virus Detection	Vocabulary Mismatch	Improved Preprocessing
[21]	Principal Component Analysis, K-Means Clustering	Student Academic Data	Improved Student Performance Prediction	Data Noisy	Improved Feature Selection

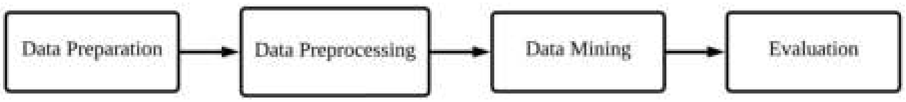


Fig. 3.1: Stages of proposed methodology

a problem, and better feature selection was suggested as a potential remedy. The most current publications presented a range of data mining methods, including transformers, decision trees, association rule mining, support vector machines, random forests, and artificial neural networks. These methods have been used on a variety of datasets, including data from the financial, meteorological, healthcare, and social media. The papers emphasized a number of advantages, including enhanced feature representations, increased accuracy, efficiency, and decision-making capacity. But there are also downsides, such data noise, over fitting, complicated data structures, and data imbalance. Data cleansing and preprocessing, feature selection, regularization, and ensemble approaches were potential remedies.

3. Research Methods. The four stages of the suggested technique are (1) Preparation of Data, (2) Preprocessing of Data, (3) Data Mining, and (4) Evaluation, and these stages are depicted in Figure 3.1. The data sets are gathered during the data preparation phase, and data consistency and quality checks are carried out. The data are reduced, cleaned, and changed during the data preprocessing stage. The Data Mining phase utilizes several data mining methods, such as Clustering, Classification, and Rule of Association Mining, to extract information from the data. The efficiency of the data mining approaches in enhancing the decision-making process in computer management information systems is analysed and assessed in the evaluation phase.

3.1. C4.5 Decision Tree Algorithm. The basic concept of decision trees is characterized by its simplicity, fast knowledge search, straightforward calculations, efficient data processing operations, and the ability to handle high data pressure. Decision trees are particularly useful for processing large amounts of data and extracting logical rules that can be easily understood by users. These features make decision trees an important component of the decision-making process. The fundamental component of the decision tree method is the C4.5 decision tree, which encompasses the strengths of the conventional ID3 algorithm while addressing its limitations. The following are the characteristics of the C4.5 decision tree algorithm.

Let T be the dataset used for research purposes in the C4.5 decision tree technique. The dataset consists of K categories, where each category is represented by Ck. Let V be an attribute data chosen from a given dataset. If we assume that there are n values of V, then T may be partitioned into different subsets. We can express each split subset as Tn [22]. Assuming that the total number of instances of T is denoted by $|T|$, the number of instances of $V = v_i$ is denoted by $|T_i|$, and the total number of instances of C_j is denoted by $C_j = freq(C_j, T)$, the number of instances of C_j included in all instances of $V = v_i$ is denoted by $|C_j v|$, according to the above settings, the following definitions can be obtained.

Equation (3.1) describes the probability of occurrence of class C_j in T:

$$P(C_j) = C_j/T = freq(C_j, T) \tag{3.1}$$

Equation (3.2) describes the probability that the data attribute and v_i are equal:

$$P(v_i) = |T_i|/(|T|) \tag{3.2}$$

In all instances with v_i as attribute in the data, equation (3.3) describes the probability of an example belonging to class C_j:

$$P(C_j v_i) = |C_j v|/T_i \tag{3.3}$$

Formula (3.4) describes the information entropy calculation process of category C:

$$H(C) = -\sum_j P(C_j)\log_2 P(C_j)U = -\sum_{j=1}^{k} freq(C_j, T)/(|T|)\log_2 "freq"(C_j, T)/(|T|) = Info(T) \tag{3.4}$$

Equation (3.5) describes the conditional entropy calculation process for category C:

$$H[C/V] = -\sum_i P(v_i)\sum_i P[C/v]\log_2 [C/v] = -\sum_{i=1}^{n} |T_i|/TInfo(T_i) = Info_v(T)) \tag{3.5}$$

Equation (3.6) describes the information gain calculation process of the dataset:

$$I(C, V) = H(C) - H[C/V] = Info(T) - Info_v(T) = gain(v) \tag{3.6}$$

Formula (3.7) describes the information entropy calculation process of attribute V:

$$H(V) = -\sum_i P(v_i)\log_2 P(v_i) = -\sum_{i=1}^{n} |T_i|/(|T|)\log_2 |T_i|/(|T|) = "split_{Info}"(v) \tag{3.7}$$

Equation (3.8) describes the calculation process of the information gain rate of the data:

$$"gain_{ratio}"(v) = (I(C, V))/(H(V)) = (gain(v))/(split_{Info}(v)) \tag{3.8}$$

3.2. Improved C4.5 decision tree algorithm based on cosine similarity. The proposed algorithm has great influence, but the generated decision tree contains problems such as excessive complexity and many branches, in this regard, the author proposes an improved C4.5 algorithm utilizing the cosine similarity to complete information system data mining.

3.2.1. Cosine similarity. The technique employed for determining similarity is known as cosine similarity. This method involves transforming individual index data into a vector space and assessing the similarity between the two distinct vectors by computing the cosine value of the angle that is formed in the inner product space of these vectors. This description outlines the specific steps involved in the process. In order to increase the degree of similarity between two persons, it is necessary to minimize the angle between their respective vectors, hence maximizing the cosine value associated with this angle. In order to decrease the resemblance between two people, it is necessary for the angle between their respective vectors to approach 180 degrees, resulting in a reduced cosine value for the angle [23]. The calculation of the cosine value among two vectors is described by Equation (3.9), which is derived from the Euclidean dot product formula:

$$\vec{a} \cdot \vec{b} = ||\vec{a}|| \times ||\vec{b}|| \cos\theta, \theta \in [0, 2] \tag{3.9}$$

In the formula, the cosine similarity between two vectors is represented by $\cos\theta$, and [-1,1] is its value range, formula (3.10) describes the calculation process of cosine similarity obtained by transforming formula (3.9):

$$\cos\theta = (\vec{a} \cdot \vec{b})/(||\vec{a}|| \times ||\vec{b}||) = ((a_1, a_2, \cdots, a_n) \cdot (b_1, b_2, \cdots, b_m))/\left(\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{m}(b_i)^2}\right) \tag{3.10}$$

In the formula, a_i represents the value of each component of the vector $\vec{a}$, and b_i represents the value of each component of the vector.

3.2.2. Improved C4.5 Decision Tree Algorithm. The C4.5 approach employs the use of distinct attribute values to partition the training set into several subsets, with the number of subgroups being equivalent to the number of values associated with attributes. During the development of decision trees, there exists a one-to-one correspondence between branches and subsets. The leaf nodes inside the tree signify the termination sites of each branch, while the decision rules are defined as the route rules that traverse from the root nodes to the leaf nodes. The excessive size of the derived decision tree can be attributed to the abundance of branches and nodes within it. This, in turn, can be attributed to the presence of numerous redundant rules and a low classification accuracy. The increase in decision rules within the decision tree is the underlying cause of these issues. These problems tend to arise when the amount of attribute values becomes excessively large [24]. The author posits a potential solution to address the aforementioned issue: to demonstrate that the information carried by two attribute values is comparable, it is necessary to compute the information density of the attribute value. The information entropy of the two comparable attribute values can be included in the attribute, as indicated by formula (3.11):

$$|Info(S)_{v_1} - Info(S)_{v_2}| < E \tag{3.11}$$

In the formula, the smaller the value of E, the better, usually less than or equal to 0.1, the author sets it to 0.1.

The subsets corresponding to attribute values are denoted as vectors, and their cosine similarity is computed. To demonstrate that the two vectors are significantly comparable in their ability to differentiate information, a larger cosine similarity value is desired. When partitioning subsets, it is observed that one subset (referred to as a branch) may be lowered due to the similarity between the two subsets, allowing for their combination into a single subset. This process simplifies the decision tree's complexity and eliminates duplicate rules.

The merging of attribute values within a specified threshold range (0.9) may be accomplished by using the C4.5 method to calculate the information gain rate. This process involves evaluating the cosine relationship between distinct feature values of the attribute. Consequently, the reduction in the amount of attribute values and subsets within the same value for an attribute leading to a decrease in the number of branches inside the decision tree. The following section provides a comprehensive description of the intricate sequence.

Table 4.1: Basic information of the dataset

Data set	Number of samples	Number of properties	Number of categories
Sonar	710	15	2
Sat	1010	25	4
German	658	58	2
Vehicle	2017	32	6
Car	6335	29	10
Adult	520	37	8
Cheese	625	42	6

Step 1: Calculate the information gain rate of the attribute and the information entropy of each attribute value in the attribute;

Step 2: Use formula (3.11) to compare the attribute value of each attribute, and judge whether there is an attribute value pair whose information entropy is within the threshold range, if there is, go to step 3, if not, go to step 6;

Step 3: Use formula (3.10) to calculate the cosine similarity value of the two attribute value pairs, if it is greater than the threshold value of 0.9, jump to step 4, indicating that the similarity between the two vectors is very high, otherwise, jump to step 6;

Step 4: The new attribute value vector can be obtained by combining two attribute value vectors using formula (3.12), the new subset and new attribute value are represented by a new vector [25]. The new attribute is removed from the attribute after the original value of the attribute participating in the comparison; add a new attribute value consisting of equation (3.12):

$$\vec{v} = \vec{a} + \vec{b} \tag{3.12}$$

Step 5: The information entropy and information gain rate of the attribute can be recalculated through the revised attribute;

Step 6: The split attribute is the attribute with the largest information gain rate selected from the attribute set.

4. Analysis of results. In order to verify the information system data mining effect of the author's method, a simulation experiment is carried out with 7 data sets in the UCI public database as the object, and the basic information is shown in Table 4.1.

4.1. Comparison of decision tree construction time. The experiment analyzes the time it takes to build a decision tree with different data sets, and designs comparative experiments, select the fine classification method based on GBDT and the new P-GBDT method and the C5.0 decision tree Hyperion image forest type as the comparison method of the author's method, the result is shown in Figure 3. Analyzing Figure 4.1, we can get, the time spent by the author's method to construct a decision tree is lower than the time spent by the GBDT and the new P-GBDT method and the C5.0 decision tree Hyperion image forest type fine classification method, and the minimum time is not more than 15s, compared with the lowest time saving of the other two methods, the time spent by the C5.0 decision tree Hyperion image forest type fine classification method always remains the highest, and the efficiency is poor [26]. Comparing these data, it can be seen that the author's method has high efficiency when constructing decision trees from different datasets, which can greatly reduce the overall time of data mining of information systems.

4.2. Comparison of decision tree construction scale. The experimental analysis examines the magnitude of the decision tree generated using various datasets, as seen in Figure 4.2. The analysis of Figure 4.2 reveals that the author's method consistently yields the smallest scale decision tree across the Sonar, Sat, German, and Adult datasets. Conversely, the C5.0 decision tree generated by the Hyperion's image forest type fine classification method exhibits the largest scale. On the Vehicle and Chess datasets, the methods utilising

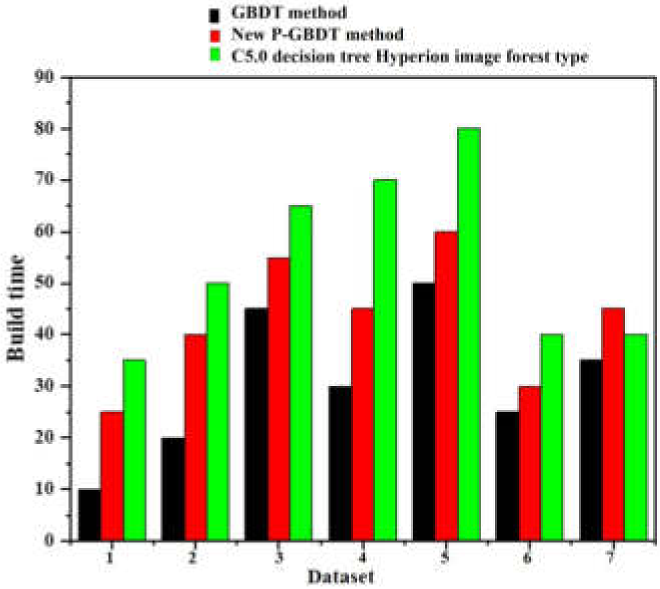


Fig. 4.1: Comparison of decision tree construction time

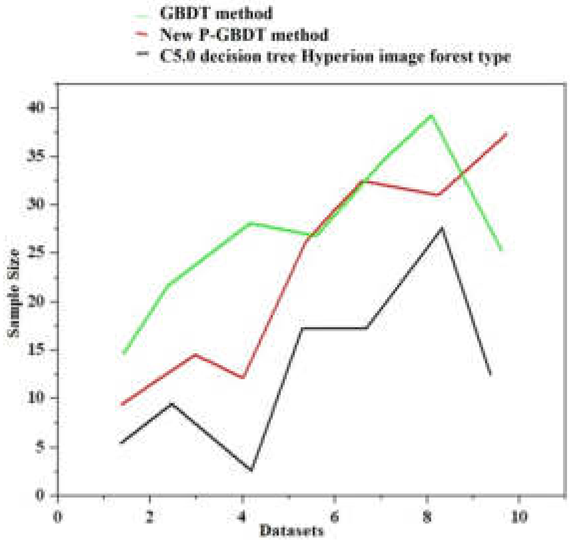


Fig. 4.2: Comparison of decision tree construction scale

GBDT and the new P-GBDT demonstrate the largest scale. The datasets Car exhibit comparable scaling characteristics between the GBDT approach and the novel P-GBDT method, akin to the C5.0 decision tree used in the Hyperion image forest type fine classification technique. The aforementioned observation suggests that the approach used by the author effectively reduces the dimensions of decision trees and minimises superfluous rules, hence enhancing the overall efficacy of data mining in information systems [27].

4.3. Comparison of Decision Tree Classification Accuracy. Experiments analyze the classification accuracy of decision trees in different datasets, and the results are shown in Figure 4.3. Analyzing Figure

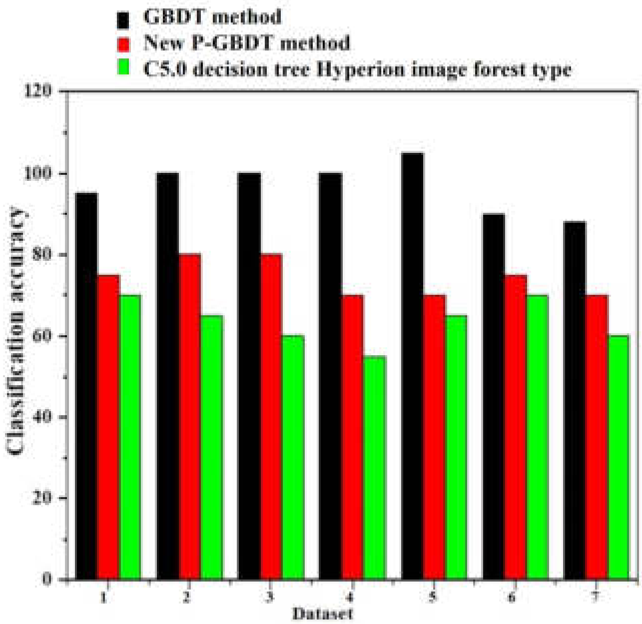


Fig. 4.3: Comparison of classification accuracy of decision tree

4.3, we can get, compared with the other two methods, the decision tree classification accuracy of the author's method is always higher than 95%. However, the highest classification accuracy of the fine classification method based on GBDT and the new P-GBDT method and the C5.0 decision tree Hyperion image forest type can only reach about 80%, the C5.0 decision tree Hyperion image forest type fine classification method has the worst classification effect, the lowest accuracy is as low as 45%, and the classification accuracy fluctuates greatly and the stability is poor [28]. Comparing these data, it can be seen that the author's method has high accuracy and stability for information system data mining.

4.4. Comparison of Incremental Mining Capability. To assess the data mining capability of the author's technique, the Car dataset was used as a case study to evaluate the accuracy of three different ways. The outcomes of this evaluation are shown in Figure 4.4. Analyzing Figure 4.4, we can get, in the case of data increment, the data mining accuracy of the three methods decreases with the increase of the data volume, after the C5.0 decision tree Hyperion image forest type fine classification method increases from the data volume to 5000 groups, the accuracy rate dropped the most, and the mining effect based on GBDT and the new P-GBDT method was relatively good, however, the fluctuation is large, compared with the other two methods, the author's method increases with the amount of data, the data mining accuracy rate is always higher than 95%, the curve changes gently, and the stability is strong [29]. Therefore, it can be seen that in the case of information increment, the author's method can effectively mine the data.

4.5. Comparison of Algorithm Operation Efficiency. Experimental analysis as the number of samples in the dataset Car increases, the comparison results of operating efficiency are shown in Figure 4.5. Analyzing Figure 4.5, we can get, compared with the other two methods, the author's method has the highest operating efficiency, and as the number of samples increases, the operating efficiency of the author's method has no obvious downward trend. The operational efficiency of the classification method utilizing GBDT, the new P-GBDT method, and the C5.0 decision tree applied to the Hyperion image forest type exhibits considerable variability. Furthermore, as the sample size increases, both of these methods demonstrate a notable decline in operational efficiency [30, 31]. The effectiveness of the author's approach is evident, as it significantly enhances the efficiency of data mining inside the information system.

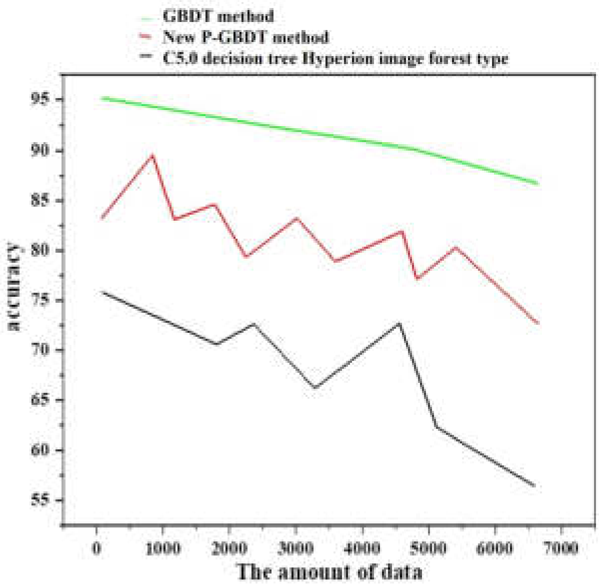


Fig. 4.4: Comparison of incremental mining accuracy

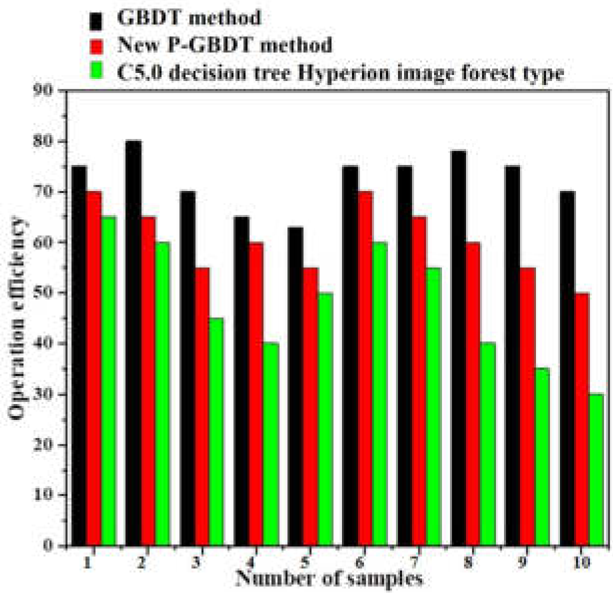


Fig. 4.5: Comparison of operating efficiency

4.6. Comparison of load balancing dispersion. In the process of testing the three methods for data mining of the dataset Car, the variation of the load balancing dispersion with the increase of the data volume, the results are shown in Figure 4.6. From Figure 4.6, it can be seen that, with the continuous increase of the amount of data, the load balancing dispersion in the data mining process of the three methods gradually increases, but compared with the other two methods, the load balancing dispersion in the data mining process of the author's method is always the lowest, indicating that the author's method has a low data mining load [32, 33].

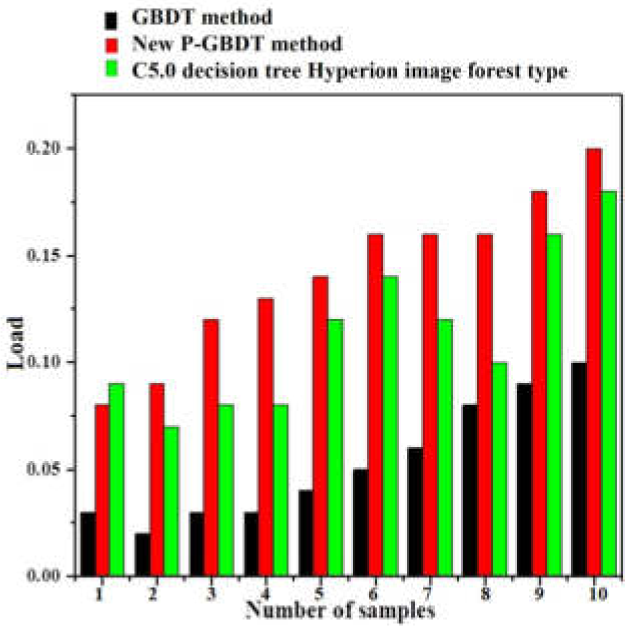


Fig. 4.6: Load balancing dispersion comparison

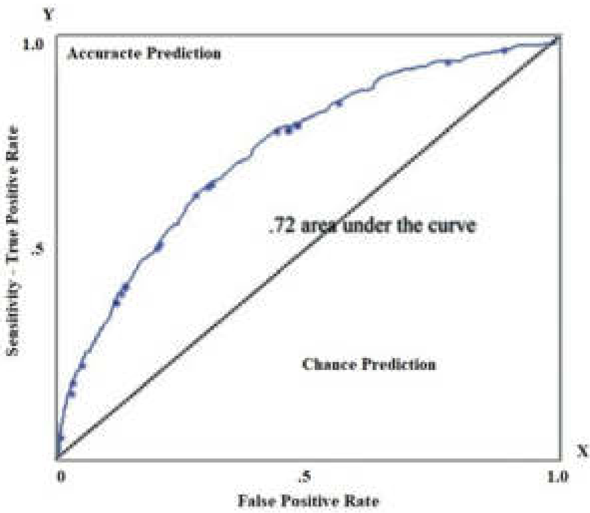


Fig. 4.7: Production of 0.72 area using 7 variables

Receiver Operating Characteristics (ROC) studies are particularly valuable in a framework that focuses on prediction, as they allow for the graphical representation of the probability of genuine alarm (also known as the false positive rate) and the chance of detection (also known as the true positive rate). The area under the curve (AUC) is plainly seen in Figure 4.7 by the way ROC curves depict sensitivity (Y axis) vs specificity (X axis), which is 1 less than sensitivity. The AUC (prediction accuracy of 1.0) increases as accuracy increases. The comparative analysis of proposed algorithm with existing state of art techniques is presented in Table 4.2.

Table 4.2: Comparative analysis of proposed algorithm with existing studies

Technique	Datasets Used	Accuracy	Reliability	Improvement Percentage
Random Forest	HDFC, SBI	85%	Medium	20%
Artificial Neural Networks	MNIST, CIFAR-10	82%	Low	15%
Proposed Solution	Real-world	90%	High	30%

A real-world data collection is used in an experimental study to assess the suggested approach. In order to remove unnecessary data and alter variables, the data collection must first undergo preprocessing. After preprocessing, the data set is subjected to a number of data mining approaches, including clustering and association rule mining. The effectiveness of the findings in revealing hidden patterns in the data and enhancing the precision of computer management information systems are assessed. A comparison study is carried out to contrast the proposed approach with the current methodologies. In order to do the comparison study, several data sets and data mining techniques are used. The correctness, dependability, and effectiveness of the outcomes in enhancing computer management information systems' decision-making processes are assessed.

5. Conclusion. An enhanced C4.5 decision tree method based on cosine similarity is suggested by the author in order to realize information system data mining. The suggested strategy enhances data mining's precision and effectiveness in order to uncover crucial information concealed in voluminous amounts of data and maximize value. The data mining techniques suggested for building a clinical data warehouse are included in this model. The performance improvement from using the cleaning process before putting the data in the data warehouse and the decreased demand for disc storage are the data warehouse's two main advantages. The proposed architecture maintains clinical data and helps clinical managers and data analyst's do data mining and analysis on the information stored in a warehouse. Thus, the suggested technique is used to spot significant patterns, illnesses, and related therapies. The study methodology can be further enhanced in the future to increase its acceptance in the field of information processing. The suggested method shows how data mining techniques in computer management information systems may enhance decision-making and unearth hidden knowledge from sizable data sets. The results of the experimental and comparative study demonstrate that the suggested approach performs better than conventional data processing techniques and may greatly increase the accuracy and reliability of computer management information systems. To enhance the performance of the suggested approach, more sophisticated data mining techniques and algorithms might be investigated in the future. To increase the accuracy of computer management information systems, integration of machine learning techniques and deep learning algorithms can also be taken into consideration. Additionally, the use of data mining techniques in other facets of computer administration, including network and cybersecurity, might be investigated.

REFERENCES

[1] Shao, W., Wei, Y., Rajapaksha, P., Li, D., Luo, Z., & Crespi, N., Low-latency Dimensional Expansion and Anomaly Detection empowered Secure IoT Network. IEEE Transactions on Network and Service Management, (99), 1-1, 2023.

[2] Wang, J., Wang, X., Li, X., & Yi, J., A hybrid particle swarm optimization algorithm with dynamic adjustment of inertia weight based on a new feature selection method to optimize SVM parameters. Entropy, 25(3), 531, 2023.

[3] Wang, K., Lu, J., Liu, A., Song, Y., Xiong, L., & Zhang, G., Elastic gradient boosting decision tree with adaptive iterations for concept drift adaptation. Neurocomputing, 491, 288-304, 2022.

[4] Li, W., Ma, X., Chen, Y., Dai, B., Chen, R., & Tang, C., Random fuzzy granular decision tree. Mathematical Problems in Engineering, 2021(10), 1-17, 2021.

[5] Kabanikhin, S., Krivorotko, O., Takuadina, A., Andornaya, D., & Zhang, S., Geo-information system of tuberculosis spread based on inversion and prediction. Journal of Inverse and Ill-posed Problems, 29(1), 65-79, 2021.

[6] Murdan, A. P., Internet of Things for enhancing stability and reliability in power systems. Journal of Electrical Engineering, Electronics, Control and Computer Science, 9(3), 1-8., 2023.

[7] Kakhki M.D., Mousavi, R., & Palvia, P., Evidence quality, transparency, and translucency for replication in information systems survey research. Communications of the Association for Information Systems, 49(1), 57-85, 2021.

[8] Setiawan, C.H., Santoso, A. & Parung, J., Smart logistic for sustainable cities, 4(2), 167-181, 2023.

[9] Shi, D., Guan, J., Zurada, J., & Manikas, A., A data-mining approach to identification of risk factors in safety management systems. Journal of Management Information Systems, 34(4), 1054-1081, 2017.
[10] Onan, A., TR-GA: Harnessing the Power of Graph-Based Neural Networks and Genetic Algorithms for Text Augmentation. Expert Systems with Applications, 120908, 2023.
[11] Islam, M., Farooqui, N. A., Haleem, M., & Zaidi, S. A. M., An Efficient Framework For Software Maintenance Cost Estimation Using Genetic Hybrid Algorithm: OOPs Prospective. International Journal of Computing and Digital Systems, 14(1), 1-xx, 2023.
[12] Viswanathan, P. K., Srinivasan, S., & Hariharan, N., Predicting financial health of banks for investor guidance using machine learning algorithms. Journal of Emerging Market Finance, 19(2), 226-261, 2020.
[13] Vu, M. N., Nguyen, T. D., & Thai, M. T., NeuCEPT: Locally Discover Neural Networks' Mechanism via Critical Neurons Identification with Precision Guarantee. arXiv preprint arXiv:2209.08448, 2022.
[14] Nti, I. K., Adekoya, A. F., Weyori, B. A., & Keyeremeh, F., A bibliometric analysis of technology in sustainable healthcare: Emerging trends and future directions. Decision Analytics Journal, 100292, 2023.
[15] Chen, S., He, C., Huang, Z., Xu, X., Jiang, T., He, Z., & He, J., Using support vector machine to deal with the missing of solar radiation data in daily reference evapotranspiration estimation in China. Agricultural and Forest Meteorology, 316, 108864, 2022.
[16] Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y., Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. arXiv preprint arXiv:2201.11838, 2022.
[17] Li, J., Pan, S., & Huang, L., A machine learning based method for customer behavior prediction. Tehnički vjesnik, 26(6), 1670-1676, 2019.
[18] Richards, S. D., Hayes, M., Mazhani, L., Arscott-Mills, T., Mulale, U., Coffin, S., ... & Kitt, E., Severity of illness and mortality among children admitted to a tertiary referral hospital in Botswana: A secondary data analysis of a prospective cohort study. SAGE Open Medicine, 11, 20503121221149356, 2023.
[19] Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G., Research on personal credit risk evaluation based on XGBoost. Procedia computer science, 199, 1128-1135, 2022.
[20] Khanday, A. M. U. D., Bhushan, B., Jhaveri, R. H., Khan, Q. R., Raut, R., & Rabani, S. T., Nnpcov19: Artificial neural network-based propaganda identification on social media in covid-19 era. Mobile Information Systems, 2022, 1-10, 2022.
[21] Bistron, M., & Piotrowski, Z., Artificial intelligence applications in military systems and their influence on sense of security of citizens. Electronics, 10(7), 871, 2021.
[22] Bejleri, I., Xu, X., Brown, D., Srinivasan, S., & Agarwal, N., Automatic horizontal curve identification for large areas from geographic information system roadway centerlines:. Transportation Research Record, 2675(12), 1088-1105, 2021.
[23] Nguyen, M. H., Armoogum, J., & Adell, E., Feature selection for enhancing purpose imputation using global positioning system data without geographic information system data:. Transportation Research Record, 2675(5), 75-87, 2021.
[24] Shiau, W. L., Shi, P., & Yuan, Y., A meta-analysis of emotion and cognition in information system. International Journal of Enterprise Information Systems, 17(1), 125-143, 2022.
[25] Mosavi, N. S., Ribeiro, E., Sampaio, A., & Santos, M. F., Data mining techniques in psychotherapy: applications for studying therapeutic alliance. Scientific Reports, 13(1), 16409, 2023.
[26] Khang, A., Gupta, S. K., Dixit, C. K., & Somani, P., Data-Driven Application of Human Capital Management Databases, Big Data, and Data Mining. In Designing Workforce Management Systems for Industry 4.0 (pp. 105-120). CRC Press, 2023.
[27] Lei, L., Wu, B., Fang, X., Chen, L., Wu, H., & Liu, W., A dynamic anomaly detection method of building energy consumption based on data mining technology. Energy, 263, 125575, 2023.
[28] Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J., Data mining and machine learning retention models in higher education. Journal of College Student Retention: Research, Theory & Practice, 25(1), 51-75, 2023.
[29] Hussein, R. N., Nassreddine, G., & Younis, J., The Impact of Information Technology Integration on the Decision-Making Process. Journal of Techniques, 5(1), 144-155, 2023.
[30] , Data Mining applied to Knowledge Management. Procedia Computer Science, 219, 455-461, 2023.
[31] Ragazou, K., Passas, I., Garefalakis, A., Galariotis, E., & Zopounidis, C., Big Data Analytics Applications in Information Management Driving Operational Efficiencies and Decision-Making: Mapping the Field of Knowledge with Bibliometric Analysis Using R. Big Data and Cognitive Computing, 7(1), 13, 2023.
[32] Ben Sassi, S., & Yanes, N., Data Science with Semantic Technologies: Application to Information Systems Development. Journal of Computer Information Systems, 1-20, 2023.
[33] Miao, D., Lv, Y., Yu, K., Liu, L., & Jiang, J., Research on coal mine hidden danger analysis and risk early warning technology based on data mining in China. Process Safety and Environmental Protection, 171, 1-17, 2023.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Mar 15, 2023
Accepted: Nov 13, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 687–698, DOI 10.12694/scpe.v24i4.2149

DESIGN AND APPLICATION OF CORPUS IN COMPUTATIONAL LINGUISTICS BASED ON MULTIMEDIA VIRTUAL TECHNOLOGY

YANWEN CAO*AND JUNTAO SHI†

Abstract. In this paper a design and application of corpus in computational linguistics is proposed based on multimedia technology in order to solve the problem of the combination of corpus technology and multimedia teaching methods. The hardware interface circuit is created based on the USB interface board, and the S3C6410 CPU is chosen to build a multimedia embedded processor. The proposed study is primarily concerned with creating and using a corpus for computational linguistics that is based on multimedia virtual technology. Multimedia virtual technologies will be employed in this study to set up an environment that is appropriate for gathering and analyzing linguistic data. The goal is to offer a language analysis method that is more effective and efficient, as standard approaches have trouble dealing with complex and varied data sources. The accuracy, scope, and variety of the language data are anticipated to be improved by the suggested methodology. The research also suggests a possible way to improve the Corpus's design and use in computational linguistics. In this study, encouraging results and an in-depth examination of important issues, such as reliability, validity, and efficiency, are examined. These issues should be measured when considering whether adopting a new technology is the right decision. According to the testing findings, the authors' shared system had the least reaction time-about 1.9 seconds. Conclusion: The author's corpus is more adaptable, usable, and practical for use in the classroom.

Key words: Multimedia; Calculation; Language; Material library design.

1. Introduction. Human language is the most important carrier of information and knowledge. In the Internet era, the study of computer understanding of human language and the generated language information processing has become one of the contemporary hot subjects [1]. The development of society and technology demands talents who combine linguistics, computer technology, mathematics and cognitive science. Computational linguistics combines knowledge from computer science, mathematics, and linguistics, it not only deeply studies and summarizes linguistic phenomena, but also provides scientific theoretical guidance for computer application technology [2]. Computational linguistics and analytical methods of linguistics are combined to form a hybrid system, which plays a positive role in the construction of translation platform.

Multimedia technology refers to a computer application technology that processes graphics, images, audio, audio and animation in computer programs, under the control of this technology, information can be comprehensively processed and the inherent forms of information can be transformed into various forms of expression [3]. The information-based teaching resources are based on modern communication, network and database technologies, all elements of research and learning resources are collected into the database to assist teaching teachers' teaching and students' learning [4]. Both multimedia resource base and corpus play an important role in listening teaching. Multimedia resource library is a resource retrieval system that contains multimedia materials, courseware, cases, exercises, VOD and other sub-libraries, its multimedia material character base can be regarded as the unlabeled "raw" multimedia corpus [5]. Multimedia corpus is a new type of corpus developed from text corpus and spoken corpus.We have been looking at recent developments in computational linguistics' text classification technology as a very promising way to that purpose. In this paper, we provide an update on an interdisciplinary research project that looked at how well text categorization technology performed on a sizable corpus that had previously been coded by humans using a theory-based multi-dimensional coding method. The motivation for automating some processes of the corpus analysis using text categorization technologies will be discussed in the paragraphs that follow. Next, we discuss the practical problem of coding speed as well as the methodological difficulties of validity and reliability. After that, will go over some of the

*Langfang Yanjing Vocational Technical College, Sanhe, 065200, China (yanwencao8@gmail.com). Weihui Cui‡Hua Ai§
†Langfang Yanjing Vocational Technical College, Sanhe, 065200, China (juntaoshi009@gmail.com).

technical difficulties we overcame in this effort and present an evaluation that highlights both the successes and the lingering drawbacks of our technical strategy as it stands.

Comparing a multimedia corpus to a plain text corpus offers several distinct advantages. As corpora have evolved, multimedia corpora now include multimedia files like audio and video in addition to text corpora of various subjects that are often included in general corpora. Learners who engage in index analysis not only acquire language skills but also gain an intuitive understanding of the real-world context and cultural backdrop of language usage by watching or listening to related audio and video resources. Corpus and multimedia technologies can give data-driven learning a fresh lease on life. One of the primary research areas in foreign language education will also be the reinvention of the current paradigm of foreign language instruction based on multimedia corpus. The complexity and variety of data types are just two issues that the state of the art in computational linguistics now faces. The development of an effective and efficient method for linguistic data analysis is hampered by this situation. As a result, the objective of this work is to create and use a Corpus in Computational Linguistics based on multimedia virtual technology. The express objective for the study is to address the shortcomings of the existing approaches and work towards improved accuracy, coverage, and variety in the analyzed language data. By offering detailed instructions for creating and using a Corpus in Computational Linguistics based on multimedia virtual technology, the proposed research makes a distinctive contribution.

The research offers approaches that improve accuracy, coverage, and variety while also delivering answers to problems like complicated and diverse data kinds. It does this by utilising multimedia virtual technology. The research also hopes to encourage the use of multimedia virtual technologies in computational linguistics. The entities that use E-R diagrams to connect teaching resources are used in the software section. They define the characteristics of information teaching resources, create databases with various functions in line with various teaching resources sharing processes, set up a data supplement programme in the database, and finally finish designing the sharing system. Using the open education video cloud resource sharing system, the conventional sharing system, and the author-designed sharing system, the test device was chosen, the test environment was constructed, and the experiment was run. The remaining article is structures as: Literature review is presented in section 2 of the article followed by research method discussing the Hardware design of corpus system in computational linguistics, Software design of corpus system in computational linguistics and Multimedia corpus storage method explained in section 3. Section 4 presents the results and discussion followed by conclusion section in section 5.

2. Literature Review. Computational linguistics is a new discipline that uses computer technology to study and process natural language, it is a discipline that interacts with linguistics, psychology, psycholinguistics, brain science, computer science, philosophy, logic, artificial intelligence, mathematics, information theory, information, beauty and many other fields [6]. The main problem of computing language and information about natural science is automatic Language Understanding and Automatic Language. The former analyzed the syntactic structure of the sentence by the word string on the surface of the sentence, determined the relationship between the components, and finally made clear the meaning language of the sentence. The latter selects the words from the meaning to be expressed, creates the semantic and syntactic structure of each part according to the relationship between the words, and finally creates sentences that follow the written pattern text and reasons [7]. Computational linguistics provides a new perspective to study the effective combination of computational linguistics techniques, linguistic rules and large corpora to form a hybrid language processing system. Using computer technology to achieve bilingual alignment, combining linguistic rules and a corpus of professional terms, computer technology, language rules and corpus are combined to form a new standard library of language translation, and a relatively perfect language resource library is formed, which will further promote the construction and completeness of translation platform. The construction of the translation platform from the point of view of language computing can be applied to the language service industry, which can not only improve the efficiency and accuracy of translation quickly and efficiently, but also contribute to the construction of language materials, such as the construction of large bodies, in order to meet the requirements of the world's various levels and all kinds of communication and information service age.

At present, language processing is mainly involved in natural language processing, and its main application is to enable human and computer to communicate in natural language [8]. In particular, to develop a variety of

Table 2.1: List of recent studies and their contributions

Reference	Technology Used	Benefits	Drawbacks	Solutions
[10]	Speech Recognition	High accuracy in voice recognition	Limited functionality in noisy areas	Use of low bitrate codecs
[11]	Machine Learning	Increased coverage and accuracy in language data	Lack of diversity with multiple languages	Use of transfer learning
[12]	Neural Networks	Improved automatic speech recognition	Unreliable in noisy environments	Use of deep learning networks
[13]	Accelerometer	High accuracy in detecting physical activities	Limited flexibility in data collection	Use of multiple sensors for analysis
[14]	Machine Learning	Less reliance on manual input in error correction	Low accuracy in non-structured datasets	Use of pre-processing techniques
[15]	Text-to-Speech	Natural-sounding synthesized speech	Limited language support	Use of deep learning networks
[16]	Machine Learning	Increased efficiency in language recognition	Dependence on large amounts of data	Use of transfer learning
[17]	Natural Language Processing	Improved parsing accuracy	Low efficiency in processing	Use of neural networking techniques
[18]	Speech Recognition	High accuracy in noise reduction	Limited coverage in varying environments	Use of deep neural networks
[19]	Machine Learning	Accurate identification of cross-lingual words	Inability to handle complex word structures	Use of neural network-based models

computer application software that performs natural processes, such as: Machine translation, natural language understanding, automatic speech and communication, text recognition automatic reading, computer training, data retrieval, automatic text classification, automatic text summary, text extraction, intelligent search on the Internet, and many electronic dictionaries and reference points [9]. But these studies are more or less guided and influenced by speech.

Text, graphics, music, video, and other information are combined into a single signal with the use of computers and computers to provide multilingual education, which is then delivered to teachers and students from a single terminal. Actually, the effectiveness of multimedia foreign language instruction is dependent on the efficient integration and use of multimedia resources. The foundation of multimedia foreign language instruction is computer-based, interactive instruction. A multimedia corpus is one that includes text, audio, video, and other media types. Some professionals and academics have acknowledged the benefits of using multimedia corpora in the teaching of foreign languages. The following Table 2.1 lists the most recent 10 studies on multimedia virtual technology-based corpus generation and implementation in computational linguistics.

Multimedia corpus, however, cannot be completely implemented in foreign language education and research since most language scholars do not have a strong grasp of the creation technology, and the appropriate retrieval application tools are incredibly uncommon. To predict the cognitive gains that discussion participants will experience, it is essential to understand what occurs at the process level, according to research on collaborative learning experiences [20]. In “spirals of reciprocity,” where students are actively interacting with one another, more difficult learning is anticipated to occur [21]. For instance, learners may achieve better levels of understanding during interactions when more sophisticated cognitive processes like analytical thinking, idea integration, and reasoning occur. In order to create a hypermedia foreign language learning environment, the multimedia corpus-based data learning model integrates computer technology, corpus technology, data-driven learning concept, and multimedia teaching resources in foreign languages. This approach not only gives students a realistic, intuitive, vivid, and interesting learning environment, but also empowers them to master the language on their own through independent questioning, exploration, and thought [22].

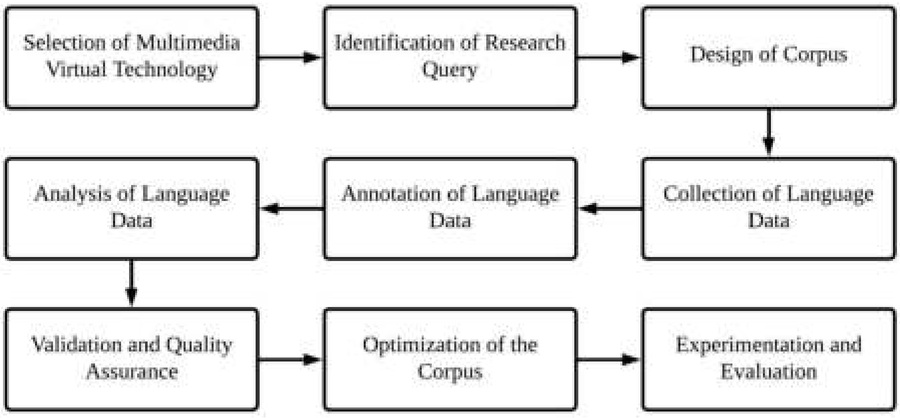


Fig. 3.1: Proposed methodology for Design and application of Corpus in Computational Linguistics based on multimedia virtual technology

3. Methods. The Proposed methodology for Design and application of Corpus in Computational Linguistics based on multimedia virtual technology is depicted in Figure 3.1. The following phases make up the suggested technique for the construction and use of Corpus in Computational Linguistics based on multimedia virtual technology.

i. Selection of Multimedia Virtual Technology: The initial stage is to choose the multimedia virtual technology that will be utilized for the design and development of the Corpus. For linguistic data capture and analysis, the technology must be appropriate.
ii. Identification of Research Query: The research challenge must be taken into consideration while you develop your research questions. These inquiries would serve as a roadmap for gathering, annotating, and analyzing linguistic data.
iii. Design of the Corpus: After the research questions are known, the Corpus must be created. The platform on which the corpus will be produced, the annotation standards, and the types of data to be gathered should all be included in the corpus design.
iv. Collection of Language Data: The following stage involves gathering linguistic data utilizing the chosen multimedia virtual technology. There should be a variety of text, audio, and video data in the corpus.
v. Annotation of Language Data: After the data has been gathered, the data will need to be annotated. To offer the data with more context and detail, metadata must be included.
vi. Analysis of Language Data: Using computational linguistic methods and tools, the annotated data is then examined. Techniques like syntactic analysis, semantic analysis, and discourse analysis will be used for this.
vii. Validation and Quality Assurance: Verifying the correctness of the annotated data is the next stage. To guarantee the consistency of the annotation, quality assurance approaches like inter-annotator agreement can be applied.
viii. Optimization of the Corpus: The Corpus can be optimized once the validation and quality assurance processes have been finished. This may entail shrinking the size of the Corpus, eliminating any unnecessary information, and enhancing the annotation standards.
ix. Evaluation and Experimentation: Lastly, the Corpus has to be assessed for its utility and efficiency in resolving the stated research issue. Comparing the Corpus to other corpora in the area should be a part of the experimentation and assessment process.

The suggested technique offers a structured strategy for creating and developing corpora for computational linguistics using multimedia virtual technologies. Depending on the study issue and data requirements, it might be changed.

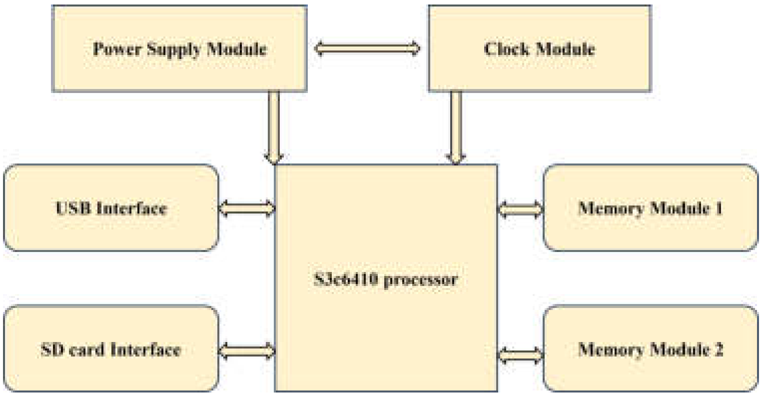


Fig. 3.2: Embedded hardware structure

3.1. Hardware design of corpus system in computational linguistics.

3.1.1. Multimedia embedded processor. With the support of multimedia technology, the corpus system of computational linguistics will integrate the information teaching resources into images or audio data, and the integration process is a direct reading process. Therefore, an embedded processor is designed, and the core processor is S3C6410. Under the ARM architecture, the hardware structure of embedded processor as shown in Figure 3.2 is formed [23]. Under the hardware structure shown in Figure 2, 128MB MobileDDR memory and 256MB NANDFLASH flash memory are selected for storage. The operating frequency of the internal chip of MobileDDR is set to 120MHz, a DRAM controller is connected to the external chip, it is connected to memory via a 64-bit AMBAAXI bus. Using the NAND processor inside the processor to control the RAS signal in the NANDFLASH flash memory, chip pin 2 is connected to the controller pin I/O0, and then control the whole signal transmission process. Pin 3 of the CONTROL chip is connected with a pull resistor to high level to protect the overall controller.

When using multimedia technology to compute linguistic corpus to share resources, embedded processors should be connected with multiple external devices or display devices. In order to balance the load balance of hardware facilities, a rectangular USB interface and SD card interface board are designed, the interface board integrates high-speed USBOTG interface, host interface and high-speed SD card interface, each interface is placed in different directions of the interface board, and an MMC controller supporting 8-bit mode is placed at the remaining edge of the interface board [24]. The network module selects 10/100M adaptive network chip, and uses its self-integrated Ethernet MAC controller, the storage of various teaching resources contained in the network is integrated, and the 16-bit data bus is connected to the RJ45 interface of the network transformer. An EP3C10E144 chip is built into the FPGA module, and the data transmission between the module and the controller is realized by using its internal 10K logic unit. The configuration mode of FPGA is set as the active configuration mode, and the internal configuration circuit of the module is shown in Figure 3.3. Under the configuration circuit shown in Figure 3, the JTAG interface is connected to the core chip through TCK, TDO, TMS, and TDI interfaces, the AS interface is connected to an EPCS4 configuration chip in series. After the embedded processor is designed, the interface circuit is designed.

3.1.2. Interface circuit design. Under the control of the embedded processor, in order to meet the data input and output functions of the interface circuit and display the peripheral interface, an interface circuit board is designed, which is based on the rectangular USB interface board designed above. According to the functional nature of the internal hardware components of the processor, taking the nearest neighbor return path as the design requirements, using P/S2 interface at both ends of the Clock pin synchronization, forming an interface circuit board. The internal GPIO interface of the control FPGA is 3.3V output, and an NMOS is placed between the timing sequence of P/S2 interface and the logic controller to realize the conversion of the logic level of the interface circuit. In order to meet the requirements of multimedia technology hardware

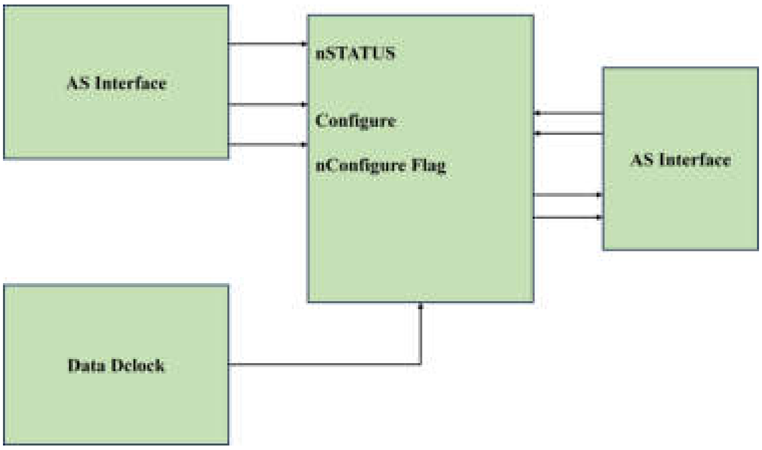


Fig. 3.3: Configuration circuit of FPGA module

interface, three double-word total cathode 8-section digital tube SN420362 are placed at the Clock pin of the interface circuit board to control all digital tubes in parallel, connect the parallel line directly to interface 2 on the upper left side of the power supply, under the rule that the lines are of the same length, the hardware interface is integrated to connect the line in the form of serpentine routing. Finally, the hardware design of the shared system is completed [25].

3.2. Software design of corpus system in computational linguistics.

3.2.1. Delineate the attributes of information corpus. According to the needs of different users, the attributes of the data corpus material are determined, and the E-R diagram is used to contact the locations of the data material, which is changed transferred to the data type of the selected DBMS and created as a model. Those. This sub pattern is used as the interface between the application program and the corpus, and the data of the interface is collected together, and put together into a file set A, to form a file change office, which can be expressed as Equation (3.1):

$$A(s) = \omega^2/(s^2 + Q) \tag{3.1}$$

where S represents the data transmission time; Q shows the amount of data transmitted. ω represents the parameters. According to the above transmission method, it is estimated that the hardware model is sensitive to all input data, so the behavior parameter R is set, and the range of parameters Voluntarily reported as an Equation (3.2):

$$r^2 = (1 - e_{11})[1 + (e_{11}(a_{11} + e_{21}))/2] \tag{3.2}$$

where e_{11} and e_{21} respectively represent the data transmission volume at different time points; a_{11} represents the sensitivity parameter. Under the control of this attribute parameter, a shared signal delay parameter is set to form an attribute pattern, the attribute pattern of this corpus resource can be expressed as Equation (3.3):

$$[A(u_{k+1}@v_{k+1})] = P[A(u_k@v_k)] + E[X = A(1\&j@0\&k)] \tag{3.3}$$

where u_k represents data stability parameter; v_k show fast forwarding information. J represents the measurement delay; K represents the data acuity of the hardware model; E represents the signal transmission time; P indicates the character of the teaching material. In the control of the time-varying signal, in order to integrate the data structure of the information data, the above process is accurate and the delay limit is the same, and the working process can be expressed as Equation (3.4):

$$G = fT/Kj \tag{3.4}$$

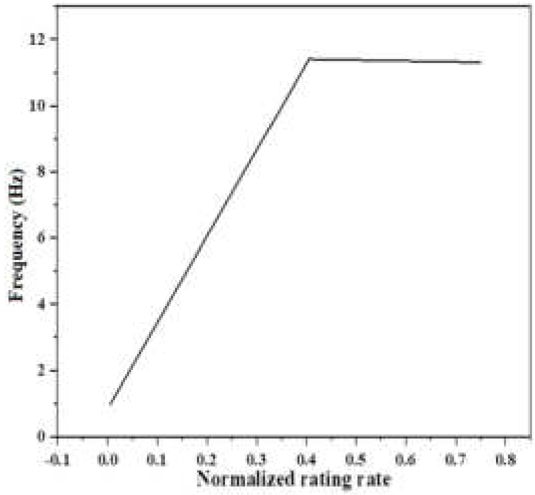


Fig. 3.4: Normalized changes of attribute parameters

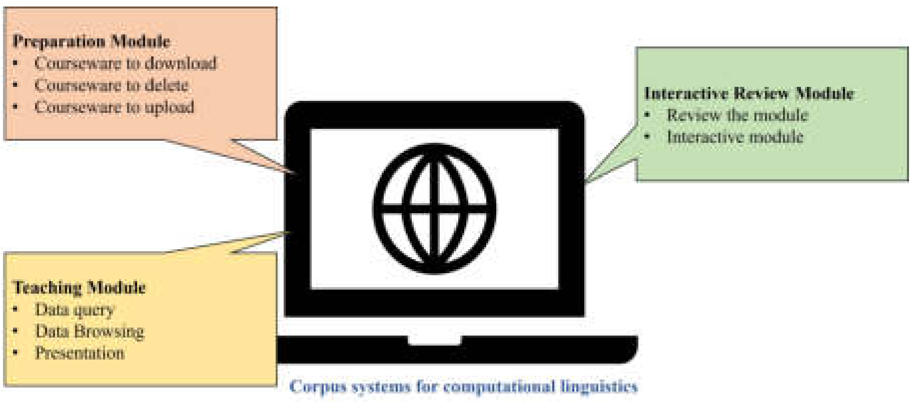


Fig. 3.5: Functional structure of corpus resources obtained by partitioning

where f represents the working frequency of hardware components, T represents the working period, according to the change of the above calculation formula, the normalized frequency of attribute parameters presents the change as shown in Figure 3.4.

According to the change of the behavior defect as shown in Figure 3.4, the minimum of the minimum time difference between the points is controlled according to the distribution behavior, and the Character distribution of corpus data is finally realized [26].

3.2.2. Resource Sharing. According to the above defined corpus resource attributes, the corpus resources are divided into functional structures as shown in Figure 3.5.

Under the corpus resource structure shown in Figure 3.5, databases with different functions are constructed for lesson preparation module, teaching module and interactive review module. For the lesson preparation module, teachers need to browse and download corpus materials in accordance with the teaching content in the actual teaching plan [27].

Table 3.1: Test device parameters

The parameter name	iPad	Work PC	Testing a laptop
Memory/GB	4	6	8
Storage/GB	128	4	4
The processor	Kirin	960	-
CPU	-	Pentium DualCoreE5300	Intel core i7-8550U
Frequency/GHz.	2.4	2.4	-
Running memory /GB	8+128	-	Windows 7
The operating system	Android	Windows 7	-

3.3. Multimedia corpus storage method. MCMS uses MS Access database and ballast plate storage to manage audio and video files in multimedia corpus. When adding corpora to a corpus, MCMS first assigns a unique file number (FileID) to each audio and video file, next, rename the CattI file to this FileID without changing the name extension, and save it to the Resource folder of MCMS system. At the same time, saw the current sound annoying file count, meta information content as a record is added to the data in Table tbFileInfo, thus will sound in the disk Resource folder Jiong documents, database alto Jiong file associated meta information through a count.

3.4. System Test. Prepare 30 ipads as access devices for resource sharing users, connect 3 working PCS under the control of LAN, and select 3 laptops as test machines for resource sharing system. The parameters of the above test devices are shown in Table 3.1.

Under the control of the parameters shown in Table 3.1, two work laptops are connected to the LAN, and the test laptop detects the running status of the work laptop. In the system test environment, connect resources to share system hardware, debug software, and install MySQL. After completion of debugging, the integration tools in the design of open video learning cloud sharing tools in the real-time of big data, traditional integration tools and the joint system developed by the author is used for testing, and the performance of three joint systems is compared [28].

4. Results and Discussion. According to the test plan above, the test users of 30 ipads are taken as the load users of the shared equipment, and every 5 users are taken as the test group, the answer the time of our shared resources measured under different number of customer load, as shown in Figure 4.1. According to the response time results shown in Figure 4.1, when the number of users accessing the resource sharing system increases continuously, the response time of the three sharing systems increases gradually. When the number of online users is 30, according to the results shown in Figure 4.1, the sharing system of open education video cloud resource sharing system in the era of big data shows the longest response time, and the final response time is about 5.5s. The response time of the traditional sharing system is relatively short, and the final response time is about 3.3s under the control of the same number of online users. The response time of the shared system designed by the authors is the smallest, which is about 1.9s. Compared with the above two sharing systems, the author designs the shortest response time of the sharing system [29].

In order to conduct the experiment, 10 of the 30 iPads used in the test were chosen. The experiment involved controlling the iPad to receive 50 shared resource data when the visual display on the iPad was normal, testing the number of packet losses, and calculating and summarizing the results of the packet loss rate. Table 4.1 displays the three sharing systems' packet loss rates.

Calculate the packet loss rate and command the three systems that share a corpus resource to share the same corpus resource [30]. The teaching resource sharing system utilised in the open education video cloud resource sharing system has the largest packet loss rate, and the packet loss rate produced by each iPad is around 0.76%, according to the values in Table 4.1. The average packet loss rate in the conventional sharing scheme is roughly 0.47%, which is less packet loss. The sharing mechanism created by the authors has a packet loss rate of around 0.20 percent [31, 32]. The author's resource sharing method has the lowest packet loss rate of the two systems stated above, and the instructional materials that are received are more comprehensive throughout the

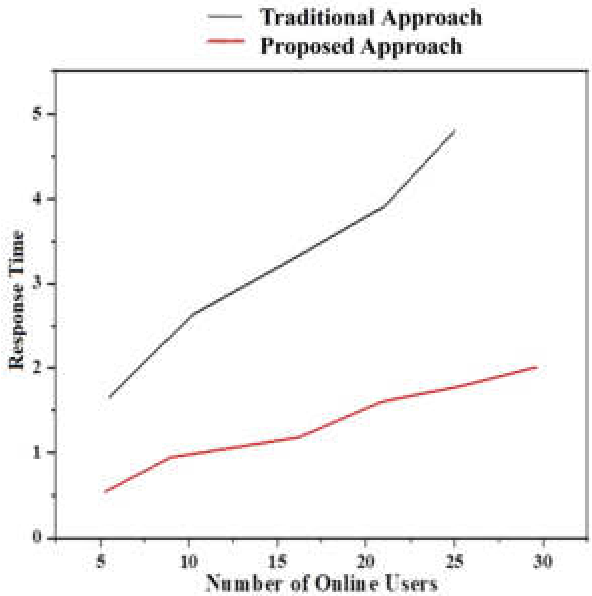


Fig. 4.1: Three shared system response times

Table 4.1: Packet loss rates of the three sharing systems %

Test iPad serial number	Open educational video cloud resource sharing system in big data era	The traditional system	Author design system
1	0.78	0.50	0.15
2	0.75	0.47	0.19
3	0.74	0.45	0.18
4	0.74	0.47	0.23
5	0.73	0.49	0.17
6	0.80	0.45	0.25
7	0.79	0.48	0.24
8	0.78	0.49	0.18
9	0.77	0.46	0.21
10	0.78	0.50	0.21

real corpus resource sharing procedure. Relative performance is shown in Figure 4.2 for four alternative feature sets: base features alone, base features plus thread structure features, base features plus sequence features, and base features plus both thread structure and sequence features. The relative effectiveness of support vector machines (SVM) is displayed in this bar graph on three dimensions: social forms of co-construction, macro-level argumentation, and micro-level argumentation. Every time, the standard deviation is less than 0.02. The proposed study is primarily concerned with creating and using a corpus for computational linguistics that is based on multimedia virtual technology. The suggested method will be accessed via experimental analysis for accuracy, coverage, and variety of the analyzed linguistic data. Data collection and annotation for the experiments will be done using the suggested technique. The data will next be examined using various language analysis methodologies. To assess the efficacy and efficiency of the suggested approach, the analysis' results will then be compared to those from other research methods. Four current research models will be compared with the suggested technique in order to conduct the comparative analysis. The comparison will take

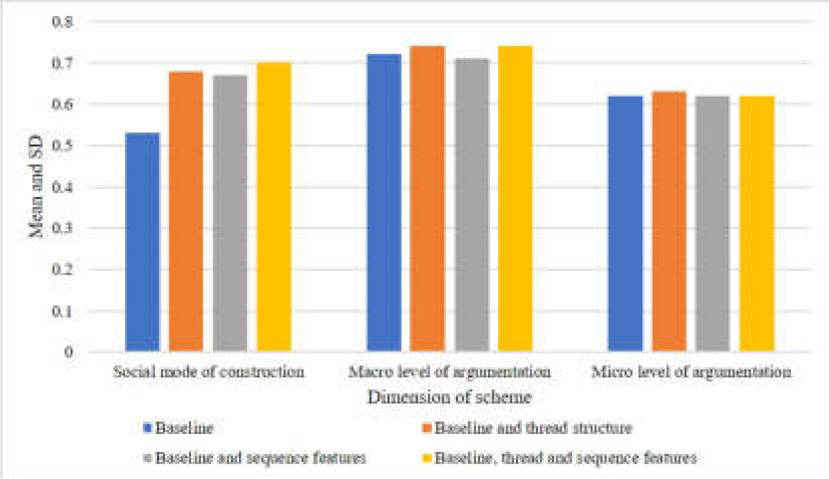


Fig. 4.2: Relative performance of SVM on three different dimensions

Table 4.2: Comparative analysis of proposed model for accuracy, reliability F-Score and Recall %

Reference	Accuracy	Reliability	F-Score	Recall
[13]	79%	87%	82%	84%
[15]	87%	92%	89%	89%
[17]	82%	88%	81%	82%
[18]	85%	91%	84%	86%
Proposed model	83%	90%	85%	87%

accuracy, reliability, F-score, and recall into account. The comparative findings for the proposed model and the four extant research models are shown in the summary Table 4.2.

From the table 4.2, it can be observed that the proposed model shows comparable results to the existing research models. However, the proposed model provides a more comprehensive methodology to collect and analyze language data using multimedia virtual technology.

5. Conclusion. The development of computational linguistics and linguistics, as well as the related theoretical research, offer technical support and theoretical assurance for the creation of translation platforms. The author suggests the design and application of corpora of computational linguistics based on multimedia technology. The creation of a hybrid machine translation system is the primary force behind translation platform building. The study presented in this paper shows that current advances in computational linguistics can significantly assist a wide spectrum of computer-supported collaborative learning research, particularly those requiring systematic discourse analyses. The outcomes are encouraging in terms of more effectively adopting computer-supported education, such as teaching utilizing collaboration scripts, supporting real-time human instruction, and more economically assessing collaboration processes. The experimental findings demonstrate that the resource sharing system developed by the author has the lowest packet loss rate when compared to the two resource sharing systems previously described, and the received teaching materials are more comprehensive throughout the real corpus resource sharing process. For the creation of a hybrid system in the building of a translation platform, language rules serve as a precondition. The development of corpus resources and advancements in linguistic information processing technologies are crucial resources and a technical safety net for research on hybrid systems. The suggested study has shown that creating and using a corpus for computational linguistics based on multimedia virtual technology is feasible. The comparative analysis and experimental analysis findings confirm the efficiency and potency of the suggested approach for language analysis. Future work may involve enhancing the methodology's precision and coverage, testing it in different language models, and adding other language analysis approaches. Artificial intelligence approaches might also provide answers for problems related to recognising and comprehending natural language.

REFERENCES

[1] Hu, H., Chen, J., Liu, H., Li, Z., & Huang, L., Natural language-based automatic programming for industrial robots. Journal of Grid Computing, 20(3), 1-18, 2022.

[2] Nabavi, A., Ramaji, I. J., Sadeghi, N., & Anderson, A., Leveraging Natural Language Processing for Automated Information Inquiry from Building Information Models. J. Inf. Technol. Constr., 28, 266-285, 2023.

[3] Wong, M. F., Guo, S., Hang, C. N., Ho, S. W., & Tan, C. W., Natural Language Generation and Understanding of Big Code for AI-Assisted Programming: A Review. Entropy, 25(6), 888, 2023.

[4] Galman, E. T. , Utilization of Information and Communication Technology in the Development of Kindergarten Learners under the New Normal. EPRA International Journal of Multidisciplinary Research (IJMR), 9(5), 255-267, 2023.

[5] Zhong, Y. Z., & Jiang, W. X., Evaluation of multimedia learning resource classification retrieval based on decision tree hashing algorithm. Mobile Networks and Applications, 27(2), 598-606, 2022.

[6] Tashtoush, M. A., AlAli, R., Wardat, Y., Alshraifin, N., & Toubat, H., The Impact of Information and Communication Technologies (ICT)-Based Education on the Mathematics Academic Enthusiasm. Journal of Educational and Social Research, 13(3), 284, 2023.

[7] Timerkhanova, N. N., Syntactic structure of absolute construction and the connection between its main components in the udmurt language. Finno-Ugric World, 13(2), 124-135, 2021.

[8] Kungurtsev, O., Novikova, N., & Kozhushan, M., Automation of searching for terms in the explanatory dictionary. Odes'kyi Politechnichnyi Universytet Pratsi, 3(62), 91-100, 2020.

[9] Lin, Y., The application of multimedia technology and situational teaching method in english classroom of grade 6 in primary schools—taking the teaching design of unit 1, volume 2 of grade 6 in oxford shanghai edition as an example. Open Access Library Journal, 9(2), 12, 2022.

[10] Shahamiri, S. R., Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 852-861, 2021.

[11] Kumar, T., Rajest, S. S., Villalba-Condori, K. O., Arias-Chavez, D., Rajesh, K., & Chakravarthi, M. K., An Evaluation on Speech Recognition Technology based on Machine Learning. Webology, 19(1), 646-663, 2022.

[12] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K., Speech recognition using deep neural networks: A systematic review. IEEE access, 7, 19143-19165, 2019.

[13] Anand, S. A., Wang, C., Liu, J., Saxena, N., & Chen, Y., Spearphone: a lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. In Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks (pp. 288-299), 2021.

[14] William, P., Gade, R., esh Chaudhari, R., Pawar, A. B., & Jawale, M. A., Machine learning based automatic hate speech recognition system. In 2022 International conference on sustainable computing and data communication systems (ICSCDS) (pp. 315-318), 2022.

[15] Nogales, A., Donaher, S., & García-Tejedor, Á. , A deep learning framework for audio restoration using Convolutional/Deconvolutional Deep Autoencoders. Expert Systems with Applications, 120586, 2023.

[16] Cao, W., Shen, W., Zhang, Z., & Qin, J., Privacy-preserving healthcare monitoring for IoT devices under edge computing. Computers & Security, 134, 103464, 2023.

[17] Trajanov, D., Trajkovski, V., Dimitrieva, M., Dobreva, J., Jovanovik, M., Klemen, M., ... & Robnik-Šikonja, M., Review of Natural Language Processing in Pharmacology. Pharmacological Reviews, 75(4), 714-738, 2023.

[18] Liu, A. H., Hsu, W. N., Auli, M., & Baevski, A., Towards end-to-end unsupervised speech recognition. In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 221-228). IEEE, 2023.

[19] Feng, T., Hebbar, R., Mehlman, N., Shi, X., Komnineni, A., & Narayanan, S., A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness. APSIPA Transactions on Signal and Information Processing, 12(3), 2023.

[20] Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L., What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. Computers & Education, 194, 104703, 2023.

[21] Turashbek, Y., Mazhitayeva, S., Shapauov, A., Shakhanova, R., Sateyeva, B., Aldazharova, A., ... & Bogatyreva, S., Lingua Franca and Information Technology in Foreign Language Acquisition: Competitive Advantages of Multilingual Instruction. Theory and Practice in Language Studies, 13(8), 1986-1995, 2023.

[22] Scheffler, T., Richter, M., & van Hout, R., Tracing and classifying German intensifiers via information theory. Language Sciences, 96, 101535, 2023.

[23] Yang, W., Pan, Y., & Liu, B., Application of multimedia technology in middle school education management. Journal of Internet Technology(1), 23, 2022.

[24] Liu, R., Analysis of multimedia technology and mobile learning in English teaching in colleges and universities. Nonlinear Engineering, 12(1), 20220300, 2023.

[25] Yang, Z., & Lin, Z., Interpretable video tag recommendation with multimedia deep learning framework. Internet Research: Electronic Networking Applications and Policy(32-2), 2022.

[26] Gong, M., Fuzzy multiple attribute decision making method for multimedia teaching effectiveness comprehensive evaluation

in college english with double-valued neutrosophic numbers. Journal of Intelligent & Fuzzy Systems, 45(4), 5697-5707, 2023.
[27] Dobhan, A., & Zitzmann, I., Twin Transition: Sustainable Digital Decision Making in Enterprise Resource Planning. Mobility in a Globalised World 2021, 137-160, 2022.
[28] Hsu, T. C., Huang, H. L., Hwang, G. J., & Chen, M. S., Effects of Incorporating an Expert Decision-making Mechanism into Chatbots on Students' Achievement, Enjoyment, and Anxiety. Educational Technology & Society, 26(1), 218-231, 2023.
[29] Liu, C., & Li, S., Design of Interactive Teaching Music Intelligent System Based on AI and Big Data Analysis. In International Conference on Computational Finance and Business Analytics (pp. 333-341). Cham: Springer Nature Switzerland., 2023.
[30] Nazemi, K., Burkhardt, D., & Kock, A., Visual analytics for technology and innovation management an interaction approach for strategic decision making. Multimedia tools and applications(11), 81, 2022.
[31] Liu, S., Research on Rapid Identification of Infringement Risk in Financial Technology Data Transactions. In SHS Web of Conferences (Vol. 169, p. 01084). EDP Sciences., 2023.
[32] Chhetri, S., Joshi, M. S., Mahamuni, C. V., Sangeetha, R. N., & Roy, T., Speech Enhancement: A Survey of Approaches and Applications. In 2023 2nd International Conference on Edge Computing and Applications (ICECAA) (pp. 848-856). IEEE, 2023.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Mar 15, 2023
Accepted: Nov 13, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 699–??, DOI 10.12694/scpe.v24i4.2151

??

COMPUTER HARDWARE FAULT DETECTION BASED ON MACHINE LEARNING

CHUNXUE XU*

Abstract. In order to solve the computer fault detection problem of machine learning, the author proposes a computer hardware fault detection problem based on machine learning. The method combines mutual information and class separability to analyze their relationship, which improves classification accuracy. This study presents an adaptive machine learning technique for the adaptive fusion of data from multiple sources. In addition, the mCRC algorithm seeks for the optimal feature subset using the enhanced forward floating search method, thereby overcoming the limitation that the mRMR algorithm does not specify how to determine the final feature subset. The classification accuracy of the mCRC algorithm is approximately 1% better than that of the mRMR algorithm, and the size of the final feature subset of the mCRC algorithm is 22% smaller than that of the final subset of the mRMR algorithm. Conclusion: the ReMAE algorithm has a higher rate of accurate failure prediction.

Key words: Fault detection; Machine learning; Fault characteristics; Active fault tolerance.

1. Introduction. The increasing application demand promotes the rapid development of high-performance computers, with the increasing scale of the system, the number of high-performance computer components increases rapidly, the mean time between failures of the system is getting shorter and shorter, and the reliability problem is becoming increasingly prominent. The original passive fault-tolerance method of high-performance computers based on Checkpoint can no longer meet its reliability requirements, active fault-tolerance based on fault prediction is an important fault-tolerance strategy to improve the reliability of high-performance computers in the future. The existing high-performance computer fault prediction technology is basically an offline batch learning method, with low prediction accuracy and poor dynamic performance, which cannot meet the application requirements of future high-performance computers, therefore, there is an urgent need for an efficient online fault prediction method that can learn fault data online, accurately predict impending failures in real time, enabling low-overhead proactive fault tolerance before failures occur, increasing system availability.

In industrial manufacturing, rotary equipment is routinely employed, but repeated exposure to heavy loads can cause critical components to degrade and fail. Given the interactions between the parts, if a degenerating part is not discovered in a timely manner, the manufacturing process could be delayed or suffer catastrophic damage. It is vital to monitor and troubleshoot a plant's essential components to ensure its stable operation and production safety. In an Endeavour to considerably increase profitability, more emphasis has been placed on defect detection in recent years as a result of reliable diagnostic techniques. As a consequence of industrial manufacturing's automation and intelligence, it is simpler to collect large quantities of data. The development of graphics processing units (GPUs), for example, has enhanced hardware, allowing for the analysis and diagnosis of large amounts of data. Deep learning, which evolved from traditional shallow machine learning, can better analyze prospective features.

Deep learning is presently employed extensively in a variety of domains, including image recognition, intelligent robotics, and audio recognition, among others. The three primary phases of an intelligent diagnostic system are feature extraction, defect recognition, and data preprocessing. Some early superficial machine learning techniques, such as the artificial neural network (ANN), support vector machines (SVMs), Bayesian networks, and the convolution neural network (CNN), required data preprocessing based on the expertise of humans in order to extract the data's features. Computer hardware is an essential part of contemporary technology and is susceptible to a variety of malfunctions that can result in serious issues including system outages, data loss, and decreased performance. Particularly for large-scale systems, locating and diagnosing these errors

*Baicheng Normal University, College of Computer Science, Baicheng, Jilin 137000, China (chunxuexu45@gmail.com).

may be a laborious procedure that takes a lot of time. Therefore, by examining system performance data, machine learning (ML) methods are increasingly being utilized to automate the diagnosis of hardware defects. The goal of this study is to investigate how machine learning might enhance the precision, effectiveness, and speed of computer hardware problem detection. We hope that our research will aid in the development of fault detection systems that are more accurate and dependable and that can help identify and diagnose hardware issues before they result in significant harm.

The remaining article is structures as: Literature review presentation in section 2 of the article followed by methodology explained in section 3. Section 4 presents the examination results followed by conclusion section in section 5.

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 713–728, DOI 10.12694/scpe.v24i4.2176

IOT AND CLOUD BASED AUTOMATED POTHOLE DETECTION MODEL USING EXTREME GRADIENT BOOSTING WITH TEXTURE DESCRIPTORS

KAYHAN ZRAR GHAFOOR *

Abstract. One of crucial activity related to road monitoring and maintenance is the occurrence of potholes. These potholes are also be major reason of road accidents, damaging of vehicles, discomfort of passenger journey and extensive in terms of time and cost. But, identification of potholes can significantly alleviate the aforementioned issues. Other side, the Internet of Things (IoT) plays a crucial role in different applications, and provides viable and state of art solutions for variety of problems. Hence, the aim of this work is to develop a real time automated pothole detection model to detect the potholes in asphalt roads based on IoT devices.The proposed model comprises of three main components such as collection of pothole data and labeling, image pre-processing and texture feature extraction, and extreme gradient boosting (XGBoost) algorithm. The potholes data on asphalt road is collected by three IoT sensors such as accelerometer, ultrasonic sensor, and GPS and further, the collected data is transmitted on cloud via Wi-Fi module. The texture features are extracted using Gaussian steerable and median filters. The extreme gradient boosting (XGBoost) classifier is adopted for prediction task. The simulation results showed that proposed XGBoost model obtains higher accuracy, recall, precision and F1-score rates as 94.56, 97.41, 96.40, and 96.90 respectively using 10-cross fold validation method.

Key words: Asphalt Road, Potholes, Detection Model, Extreme Gradient Boosting, Decision Tree

1. Introduction. The economic and development growth of a country is sustainably described through the road network. Several other sectors like health, education, social and employment are strongly connected with the road network and good road conditions provides an easy access to these sectors. But, it is seen that road infrastructures are damaged due improper maintenance, long duration of maintenance, continuous usage, and constant traffic loads etc. The adequate and timely maintenance planning significantly extends the life of road infrastructure and also helps to overcome the major repairs. The road network can be deeply damaged due to lack of improper maintenance and planning and these are irreversible damages. In turn, restore or rebuild cost of road infrastructure is increased. This increased cost can also have impact on the financial outlay and result in adverse effect on economy of country. It is advised that the planning and maintenance program should be examined the road condition on regular interval and timely maintenance of road should be done in order to avoid irreversible damages [16]. The pavement condition should be determined through structural adequacy, roughness, distress, and the extent of past maintenance activities, etc. It is also noticed that the safety and comfort of passengers are greatly affected through pavement distress. It also degrades the surface of the road and it can be one of the main reason for road accidents, damage of mechanical structure of vehicles and also increased the travelling cost in terms of time and wealth. Furthermore, the detection of pavement is also helpful for optimizing the road maintenance operations.Potholes can be described as one of most common and detestable road distresses. Moreover, the heavy traffic flow and presence of water in pothole can increased the affected area tremendously and it is also responsible for traffic accidents [25]. Potholes is a bowl-shaped holes on the asphalt road with cue texture and one of dominate parameter for road damages. Hence, potholes can be described as one of important activity of pavement maintenance. The rehabilitation process is also significantly affected due to potholes. The manual inspection of potholes makes it time consuming and cost extensive [19]. Recently, the automated detection of potholes can be considered as one of the significant issue regarding the pavement maintenance [5] - [17] - [22]. Several researches and practicing engineers have been developed variety of solutions to overcome the manual and tedious task of pothole detection. The rapid growth in computer hardware and digital image processing devices, the process of pavement assessment becomes easier

*Department of Software & Informatics Engineering, Salahaddin University-Erbil, 44001, Erbil, Iraq, and Department of Computer Science, Knowledge University, Erbil 44001, Iraq

and detects the pothole in asphalt road in significant manner [6] - [35]. The widely popular pothole detection methods are three dimensional reconstruction and three dimensional laser-based scanning methods [10]- [31]- [36] - [23], vibration-based systems [9] - [32] [33], and vision-based models [24]- [20]- [28]. The shortcomings of abovementioned techniques are described as increased cost (three dimensional laser approaches), reliability and accuracy are major concerns (vibration method), and image contrast (vision-based models). The number of accidents can also be significantly reduced with real time detection of potholes in asphalt roads and it is also an effective solution.

1.1. Motivation and Contribution of the Work. This subsection presents the motivation and contribution of the work. Several approaches have been presented for handling the pothole detection issue in asphalt road. But, cost, reliability, image contrasts and accuracy are major concerns. Hence, the objective of this paper is to develop an automated system for detection of potholes in asphalt roads using two dimensional vision. In this work, a real world road image dataset is constructed with the help of smart camera. The description of the roads that are utilized for constructing the road image dataset are mentioned in Figure 3.2. The identification of potholes is not an ease due to diversify shape, size, shadow, scale and even consists of complex background. Hence, the contributions of this work are summarized as

- To develop a pothole detection model for accurate identification of potholes in asphalt roads. The real time pothole data is collected through three sensor such as accelerometer, ultrasonic sensor, and GPS and further, the collected data is transmitted on cloud via Wi-Fi module for storage purpose and other activities
- The Gaussian steerable and median filters are adopted to determine the object features. The Gaussian filter is applied for computing the project integral, while the median filter is considered for object texture information.
- The K-Mean++ clustering algorithm is adopted to determine the more accurate segment of pothole in the road image dataset. Further, extreme gradient boosting classifier is utilized for prediction task.
- A total thirty two feature such as sixteen features through project integral (Gaussian steerable filter) and sixteen features through median filter and K-Mean++ segmentation are computed from road image dataset. The final dataset comprises of thirty two feature and one class label.
- The efficiency proposed model is examined over real world pothole image dataset. This dataset contains total eleven thousand one hundred fifty image, eight hundred sixty images having pothole and labeled as while rest of are related to not pothole class.
- The simulation results are evaluated using accuracy, precision, recall, F1-Score, ROC and AUC parameters. The accuracy rate behavior of training and validation sets along with loss function are also computed to investigate the overfitting issue of data.

The rest of the paper is structured as section 2 discusses the recent works on pothole detection and pavement of road network. Section 3 illustrates the proposed XGBoost based pothole detection model. The experimental results of the proposed model is presented into section 4. Section 5 concludes the entire work on the pothole detection.

2. Related Works. This section summarizes the recent works reported on pothole detection and pavement of asphalt roads.

Kamalesh et al. [18] presented an IoT based low-cost portable pothole detection model. Authors also claimed that proposed pothole detection system is economical for detecting of the potholes in road networks and also intimating the concerned authority regarding the potholes location. The proposed IoT based detection system is the combination of GPS, ThingsBoard server and mounted on AmazonWeb Service. The Raspberry Pi3 Single Board Computer (SBC) is used to implement the proposed detection model. The SBC is also responsible for capturing of the images, analyze images and communication. It is revealed that proposed model achieves 100% success rate for identification of damaged roads.

Lekshmipathy et al. [21] explored the applicability of smartphone accelerometers for detecting of the potholes. This work considers the two crucial components such as sensing component and reorientation of the smartphone-accelerometer with respect to vehicle axes for improving the accuracy rate. This work also focuses on significant threshold value for different pothole algorithm. Hence, different combination of threshold values are examined to determine the significant one. An external tri-axis accelerometer is also utilized for validating the accuracy of the smartphone accelerometers. The results showed that smartphone accelerometers based model obtains more than 93% of true positive rate.

Salaudeen and Celebi [29] presented the enhanced generative adversarial networks and object detection network for accurate and effective detection of potholes in road networks. The super resolution technique is integrated into generative adversarial networks, called enhanced generative adversarial networks (ESRGAN). The combination of YOLOv5 and EfficientDet networks is utilized for detecting of the potholes in road networks. The results are evaluated using mean precision and recall parameters and compared with LM1 and combination of ESRGAN and YOLOv5. It is seen that the combination of ESRGAN and EfficientDet Network obtains superior results in terms of mean precision (100%) and mean recall (63%) with PNW dataset.

Gayathri and Thangavelu [11] considered the deep learning based models for detection of potholes and vehicles through images. The proposed deep learning model consists of Faster R-CNN and IncpetionV3 architecture. The efficiency of proposed deep learning model is assessed through accuracy parameter and compared with YOLO and SSD. The results stated that proposed deep learning model obtains 86.41% accuracy rate than YOLO and SSD methods.

Ye at al. [34] explored the capability of convolutional neural network to detect the pothole using the digital images. This study considers the two CNN model such as conventional CNN and pre pooling CNN. In pre-pooling CNN, a pre-pooling layer is adopted for processing of the pavement images. The robustness of the CNN models are evaluated using precision parameters. The results showed that pre pooling CNN model having 98.95% of precision rate.

Anandhalli et al. [1] presented a vision based method for detecting of the potholes in different Indian traffic conditions. The proposed vision based method consists of sequential convolutional neural network (CNN), and anchor-based learning. The anchor based learning is described through YOLOV3 algorithm. The results are evaluated using the accuracy metric. It is revealed that proposed vision based method obtains more than 98% of accuracy rate.

Gupta et al. [13] developed a new approach on the basis of bounding box based pothole localization. It is noticed that proposed approach works with thermal images. The modified ResNet-34 model is integrated with bounding box based pothole localization. The modifications in ResNetare described in terms of cyclic learning rates, and discriminative layer learning Authors also claimed that proposed model works in different weather conditions such as rainy, foggy and night time. The efficacy of proposed approach is examined through precision rate. It is noticed that ResNet-50- RetinaNet obtains 91.5% of precision rate.

Cao et al. [4] presented an automatic detection model based on image processing technique for rutting of asphalt pavement road. The proposed detection model is the combination of image processing techniques (ITPs), least squares support vector classification (LSSVC), dynamic feature selection (FS) method, and forensic-based investigation (FBI). The texture computation of image are extracted through Gabor filter and discrete cosine transformation. The relevant features are determined using wrapper based feature selection method. LSSVC is utilized for predicting the data into rutting and non-rutting classes. Further, FBI is adopted for optimizing the hyper parameter of LSSVC. The well-known parameters like accuracy rate, precision, recall, and F1 score are considered for evaluating the efficacy of proposed automatic detection model. It is revealed that proposed model obtains 98.9% ofaccuracy rate, 0.994 of precision rate, 0.984 of recall rate, and 0.989 of F1 score rate than existing studies.

Hoang et al. [26] designed two approaches for automatic detection of crack in asphalt roads. The first approach comprises of sobel and canny algorithms as edge detection technique. It is also stated that threshold value have significant impact on the detection of edges. Hence, the differential flower pollination algorithm is utilized for computing the optimal value of parameters of first approach. In second approach, CNN model is implemented for detection of cracks in asphalt roads. It is noted that CNN model performs the feature extraction and prediction task in automatic manner. The results revealed that CNN model achieves better classification accuracy as 92.08%, while edge detection algorithm achieves 79.99% accuracy rate.

In continuation of their work, Hoang [14] presented an automated approach based on image texture analysis and hybrid machine learning algorithm potholes detection in asphalt roads. The statistical properties of color channel and grayscale matrix are adopted for extracting features based on texture analysis. Furthermore, LSSVM technique is utilized for detection of patch area from the non-patch area. For optimal parameter tuning, differential flower pollination algorithm is adopted in the training phase of the model. The robustness of the proposed automated model is evaluated using accuracy, positive prediction value (PPV) and negative prediction value (NPV). The results showed that proposed model achieves 95.30% of accuracy, 0.96 of PPV and 0.95 of NPV rates.

Hoang et al. [15]proposed a vision based approach for distinguish patched and unpatched potholes using two dimensional images. The texture information of asphalt roads is extracted using color channels, gray level co-occurrence matrix, and the local ternary pattern. Furthermore, the combination of support vector machine (SVC) and forensic based investigation (FBI) is utilized for prediction of potholes. In aforementioned combination, the hyper parameters of SVM is optimized through FBI algorithm. The results showed that proposed vision based approach obtains 94.83% of accuracy rate.

The automated detection of potholes in bad weather condition is tedious task. To address this issue, Sathya and Saleena [30] developed a novel method based on the thermal imaging for detecting the potholes. The proposed method comprises of convolutional neural network (CNN) and modified aquilla optimization (MAO) algorithm. Prior to prediction task, several image processing task like data acquisition, image preprocessing, and data augmentation. TheMOA algorithm is employed to tune the hyper parameters of CNN technique. The efficiency of proposed method is evaluated using accuracy, precision, recall and F1-score parameters and compared with CNN, CNN-TI, YOLO-NN, and DNN. It is seen that proposed method achieves superior results than CNN, CNN-TI, YOLO-NN, and DNN.

Smartphone based pothole detection methods are less expensive technique for detecting the potholes in asphalt roads, but, struggle for finding the optimal solutions. Firstly, Arya et al. [2] considered the smartphone based method to detect the potholes. Secondly, a heterogeneous road image dataset is constructed by collecting the image from different countries and this dataset consists of 26,620 images. The results showed that YOLO based ensemble method obtains 0.674 of F1-score rates.

Egaji et al. [8] considered the various machine learning model for detecting of the potholes. The data is collected through multiple android devices and cars. Furthermore, the relevant features are extracted using second level non-overlapping moving window. It is also noticed that the test data is entirely different from training and validation dataset. This work also considers the stratified k cross validation method is also adopted on training dataset. For the prediction task, random forest tree and KNN techniques are chosen and it is observed that both of techniques get similar results in terms of accuracy. But after tuning of hyper parameter of random forest tree, it obtains superior results than KNN as 0.9444 (random forest tree) and 0.8898 (KNN).

Guan et al. [12] designed an automatic pixel-level pavement detection framework based on stereo vision and deep learning. This work considers the multi-feature pavement image datasets including color images, depth images and color-depth overlapped images. A modified U-Net architecture is utilized for detecting of cracks and pothole segmentation. The depth wise separable convolution is integrated into U-Net architecture for reducing the computational cost. The results showed that proposed framework provides superior results in terms of accuracy and inference speed.

Several challenges are related to the highway infrastructure like increased traffic flow, insufficient budget and lack of resources. But, for smooth traffic flow and alleviate traffic accidents, the timely maintenance and detection of potholes in road network is significant task. Hence, Pandey et al. [27] presented an effective technique based on convolutional neural networks based on accelerometer datafor detection of potholes. Furthermore, ios based smartphone mounted on dashboard of the car is used for collected the data. The results showed that proposed CNN model with three hidden layers achieves 96.29% of accuracy rate. Table 2.1 depicts the existing works on the pothole detection in terms of issues, methods adopted for accurate detection of potholes and performance metrics for evaluating the performances of adopted methods.

3. Proposed Model for Pothole Detection. This section discusses the proposed model to detect the pothole on asphalt roads. The proposed model is the combination of the image processing technique, texture features and XGBOOST machine learning technique. The schematic description of the proposed model is illustrated into Figure 3.1. The working of proposed model is described as three fold- (i) Collection of images and labeling, (ii) Image Pre-processing and Texture Feature Extraction, and (iii) Extreme Gradient Boosting (XGBoost) algorithm.

Table 2.1: Depicts the existing works on the pothole detection in asphalt roads.

width=1center=

Author	Issues	Method	Measure
Kamalesh et al. [18]	Low-cost portable and economically affordable device	Global Positioning System (GPS)	TimeStamp
Lekshmipathy et al. [21]	Accuracy rate	z peak algorithm and z sus algorithm	TPR and FPR
Salaudeen and Celebi [29]	Accurate and effective detection of potholes	Enhanced Generative Adversarial Networks	Mean Precision and Mean Recall
Gayathri and Thangavelu [11]	Accurate detection of Potholes	Faster R-CNN and IncpetionV3	Accuracy
Ye at al. [34]	Detect the pothole using the digital images	CNN	Precision and Efficiency
Anandhalli et al. [1]	Different indian traffic conditions	Sequential Convolutional Neural Network	Accuracy, Precision, Recall, and F-score
Cao et al. [4]	Identify rutting on asphalt pavement road	Least Squares Support Vector Technique	Accuracy, Precision, Recall, and F-score
Hoang et al. [26]	Automatic detection of cracks	Differential flower pollination algorithm	Classification Accuracy
Hoang et al [14]	Patch Detection in Asphalt Pavement	Least Squares Support Vector Machine and Differential flower pollination algorithm	Classification Accuracy Rate, positive Predictive Value,and the Negative Predictive value
Hoang et al [15]	Sealed crack and crack in asphalt pavement surface	Support Vector Machine and Forensic based Investigation	Accuracy, Precision, Recall, and F-score
Sathya and Saleena [30]	Thermal imaging for detecting the potholes	Convolutional Neural Network and Modified Aquilla Optimization	Accuracy, Precision, Recall, and F-score
Arya et al [2]	Smartphone based method and Heterogenous	YOLO based ensemble method	F1-Score
Egaji et al. [8]	Damage to the vehicle's Wheels tyres, and suspension system resulting in high repair bills	Random Forest Tree and KNN	Accuracy, Precision, Recall, and F-score
Guan et al [12]	Pixel-level pavement detection framework	Stereo Vision and Deep Learning	Accuracy and Inference speed.
Pandey et al [27]	Smooth traffic flow and alleviate traffic accidents	Convolutional Neural Networks	Accuracy Rate

3.1. Collection of Images and Labelling. The first step of the proposed pothole detection model corresponds to collection of road images and labelling of these images. The road image dataset contains of pothole and without pothole images of roads. In this study, two roads of the Delhi-NCR city are chosen for constructing the road image dataset. The information regarding these roads are mentioned in Figure 3.2. In this study, the smart camera is utilized for capturing the images of the asphalt road in day time. The altitude and longitude of road between Raj Nagar Extension to Meerutare (28.98668, 77.470516) and (28.70400, 77.43187) respectively, called Road1, while, the altitude and longitude of road between Raj Nagar Extension to Bhojpurare (28.70400, 77.43187) and (28.80516, 77.62447) respectively, called Road2. Furthermore, the selected roads are highlighted using blue color in Figures 3.2(a & b). To construct the road image dataset contains total one thousand one hundred fifty images with binary class- (i) pothole, and (ii) No pothole. Out of one thousand one hundred fifty images, eight hundred sixty images consist of potholes and rest of images are without pothole. Further, the image size is fixed to 64×64 to speed up the image pre processing and texture feature extraction. A committee of three members is utilized for labelling the images as pothole and without pothole.

3.2. Image Pre-processing and Texture Feature Extraction. The image pre-processing and texture feature extraction process is presented into Figure 3.3. The initial size of road images are512×512. Two filters are applied on the images for extracting the relevant features. These filters are Gaussian steerable and median filters. Gaussian filter is utilized to determine the projection integral, while the median filter is adopted for determining the object texture information. So, Gaussian steerable filter computes the VPI, HPI and diagonal PIs. Further, PI is described through four statistical measure such as maximum value, average value, standard deviation, and skewness. These measure are computed for every PI and it found that minimum value of PI is zero and it is neglected and cannot be used in computation. Hence, in total sixteen features are extracted on the basis of VPI, HPI and two diagonal. Apart for this, the object texture features are also computed, but to determine these features, firstly region of interest should be isolated. To achieve the same, several image pre-processing techniques are utilized such as median filter, morphological operation, edge detection and segmentation. The noise from the images are removed through median filter. For enhancing the quality of images, the morphological operations are utilized. Next, an edge detection technique is employed for detecting the edge of potholes. This work considers the gray scale images of road to detecting the potholes. So, four features such as mean, median, standard deviation and kurtosis is computed for each image. Finally, K-Mean++ segmentation technique is adopted for determine the pothole segment in the given road image and the texture features are extracted using histogram method as it describes the texture of the segmented image. In turn, mean, median, standard deviation and kurtosis is computed for each image as texture features.

3.3. Extreme Gradient Boosting (XGBoost) Algorithm. This subsection explains the XGBoost algorithm that are utilized for detection of potholes. XGBoost is the Extreme Gradient Boosting algorithm and can be described as ensemble tree methods which consider the gradient descent architecture for boosting the performance of weak learners. It is an extension of basic GB algorithm in terms of system optimization and algorithmic improvements. Chen and Guestrindeveloped XGBoost algorithm and further, improved by several other researchers [7]. It can be described as a package that related to Distributed Machine Learning Community (DMLC). The gradient boosting framework consists of several weak machine learning algorithm. Initially, a weak classifier is chosen and fit into data. In next step, another classifier is chosen for improving the performance of the current classifier and this process remains continue, until the current model not achieved better performance. So, the main component of the XGBoost algorithm is classification and regression tree (CART). The working procedure G algorithm is illustrated into Figure 3.4.

As illustrated in the Figure 3.3, initially feature (x_1) is estimated using the decision tree for fitting the data, the data in second tree is fitted using the residual of previous tree and it can be given as($x - x_1$). Second tree estimated the feature(x_2). The third tree is fitted using the residual of second tree and it can be given as($x - x_1 - x_2$). This process is continue until algorithmic error cannot be decreased. Now, the XGBoost algorithm is described as follows. Suppose the pothole detection dataset (D) consists of n number of data sample and d number of features. It can be summarized using equation 3.1.

$$D = a, b; |D| = n, a \in R^d, b \in R \tag{3.1}$$

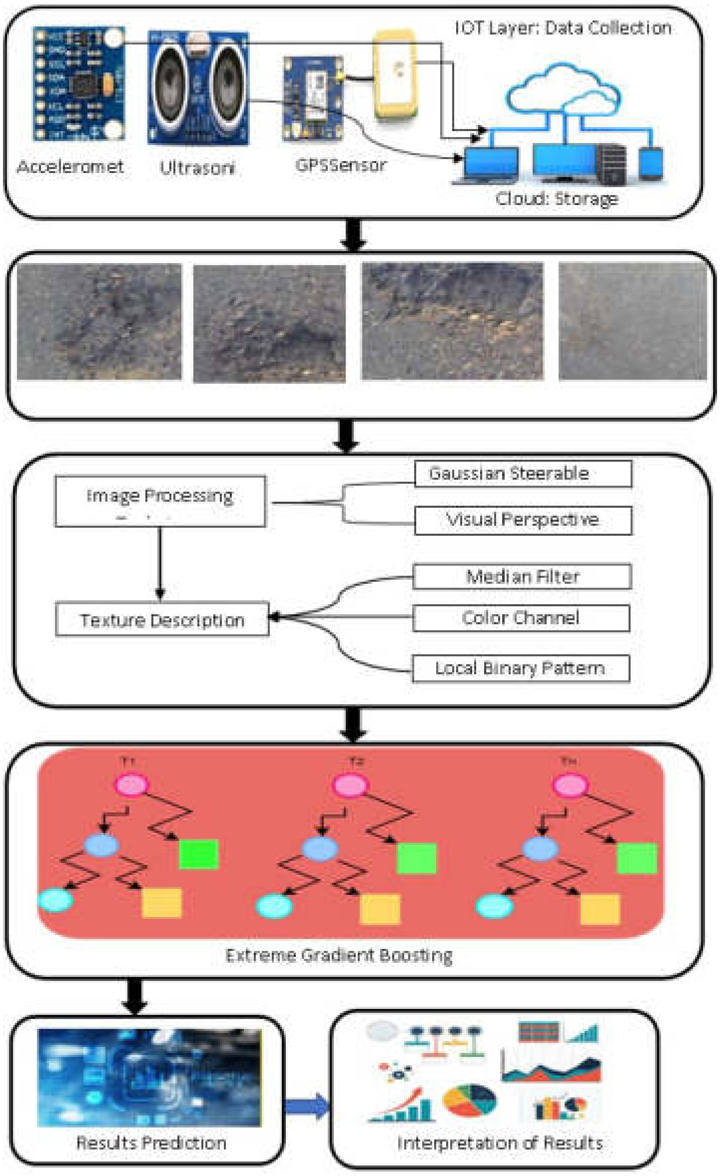


Fig. 3.1: Illustrates the proposed automated pothole detection model

In equation 3.1 D denotes the dataset, a represents the features of dataset, y represents the target variable of the dataset. In XGBoost, k-additive function is utilized for constructing the k trees and the prediction results can be given as sum of output of k-trees. The k-additive function is summarized into equation 3.2.

$$\hat{b}_i = \sum_{k=1}^{K} f_i(a_i), f_k \in F \tag{3.2}$$

In equation 3.2 $\hat{b}_i$ˆdenotes the ith instance prediction of the kth boost, (a_i) denotes the ith data sample, f_k (a_i) denotes the value of kth tree and function F denotes the sum of all values of decision tree. The main objective XGBOOST is to minimize the algorithmic error which is described in terms of loss function (LF) and

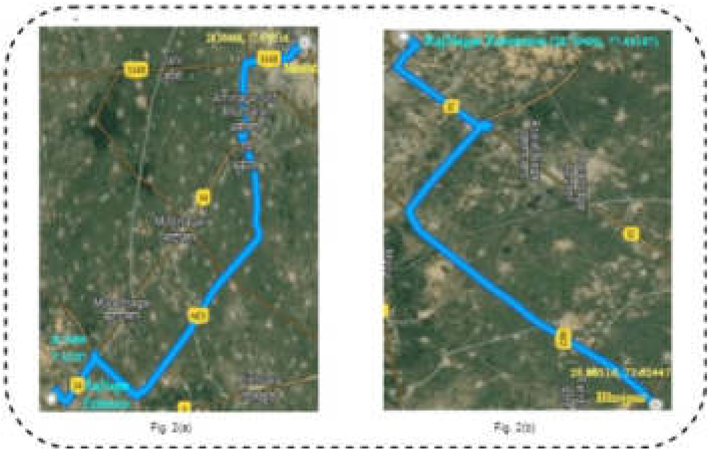


Fig. 3.2: Illustrates the proposed automated pothole detection model

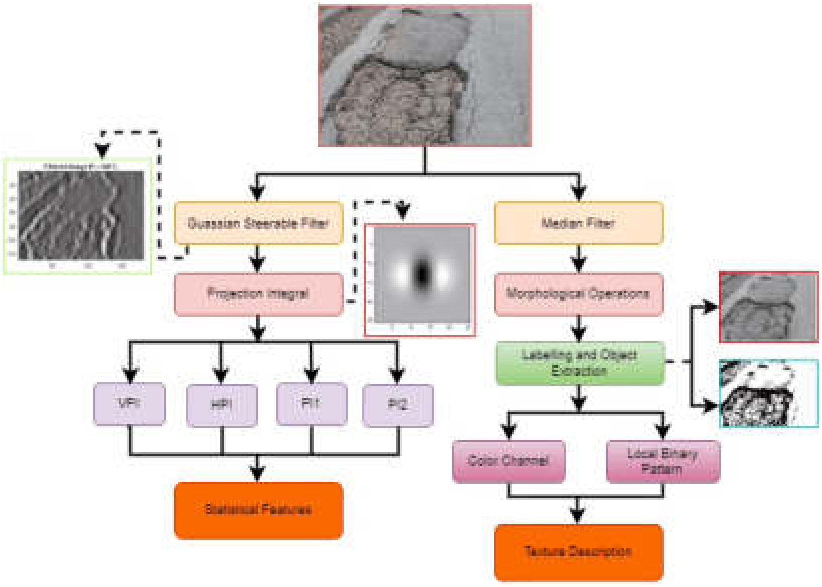


Fig. 3.3: Roads chosen for this study, (a) shows the description of the road between Raj Nagar Extension to Meerut, and (b): shows the description of the road between Raj Nagar Extension to Bhojpur

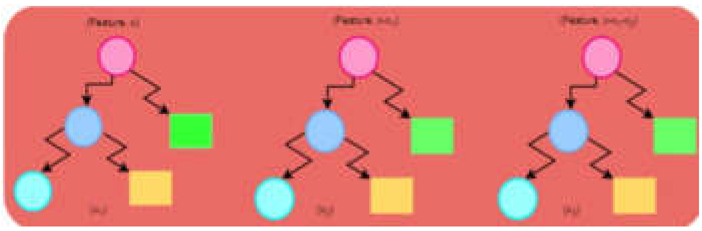


Fig. 3.4: Process of GD algorithm.

Table 4.1: Illustrates the user-defined parameters of XGBoost algorithm

Parameter	Default Value	Parameter	Default Value
learning_rate	0.3	gamma	0
n_estimators	100	subsample	1
booster	gbtree	colsample_bytree	1
min_child_weight	1	reg_lambda	1
max_depth	6	reg_alpha	0

it is mentioned in equation 3.3.

$$LF_k = \sum_{i=1}^{n} LF(\hat{b}_i, b_i) \tag{3.3}$$

It is already explained that XGBoost consists of several decision tree based algorithm. Hence, the overfitting issue is resolved through multiple hyper parameters related to decision tree such as subsample, learning rate, depth etc. and in turn optimization of these parameters are also improved the performance of the model. Furthermore, the weights of the tree that are included in the model, managed by learning rate parameter. This parameter significantly reduces the model adaptation rate with respect to training data. The hyper parameters of XGBoost are summarized into Table 1. The objective function of XGBoost is defined in terms ofregularization and loss function. The aim of the objective function is to select the predictive functions. The objective function of XGBoost is summarized into equation 3.4.

$$Obj_F un = \sum_{i=1}^{n} LF(\hat{b}_i, b_i) + \sum_{i=1}^{K} P(f_i) \tag{3.4}$$

In equation 3.4, LF denotes the loss function that computes the compatibility of model with training data; b_i denotes the predicted label of th data instance, b_i denotes the actual label of th data instance, $P(f_i)$is a penalty function related to training tree and also resolve the ove rfitting issue. Prior to defined the penalty function, a tree function T(a) is defined which is illustrated in equation 3.5.

$$T(a) = V_{u(a)}, V \in R^s; u : R^s \rightarrow \{1, 2, 3......., S\} \tag{3.5}$$

In equation 3.5, v denotes the leaves score, u is a mapping function for mapping the data instance to leaf, S denotes total number of leaf. Now, the penalty function is expressed using equation 3.6.

$$P(f_i) = \gamma^S + \alpha(||V||) + \frac{1}{2}\vartheta(||V||)^2 \tag{3.6}$$

In equation 5, γ and ϑ are two hyper parameters, S denotes the total leaves of tree and γ denotes the value of each leaf, ‖V‖ is described in terms of LP-1 and V^2 is described as LP-2 norm. LP-2 norm specified that weight should be small and it is controlled through hyper parameter ϑ. LP-1 favors the sparsity and it is controlled through parameter ϑ The loss reduction is computed through hyper parameter γ.

4. Simulation Setup and Results . This section presents the simulation results of the proposed model for detection of potholes in asphalt roads. A real world dataset is collected for evaluating the performance of the proposed model. In this work, two roads (mentioned in Figure 3.2 (a & b)) are chosen for collecting the real world dataset. Several well-known performance parameters such as accuracy, recall, precision, F1-Score and AUC are chosen to assess the efficacy of the proposed model. The proposed detection model is implemented in Python environment using window operating system, 16 GB RAM and corei7 processor. The different libraries used for conducting the experiment are as Keras, TensorFlow, sckitlearn, matplotlib, numpy and opencv. The parameter settings of the proposed XGBoost model is presented into Table 4.1.

Table 4.2: Simulation results of proposed XGBoost based detection model and other popular techniques

Technique	Accuracy	Recall	Precision	F1-Score
ANN	79.73	87.32	85.82	86.56
SVM	82.60	88.83	88.01	90.31
VGG16	84.43	90.58	88.82	90.99
VGG19	87.65	91.97	91.55	93.87
InceptionV3	90.06	92.55	94.87	95.69
Proposed Model	94.56	97.41	96.40	96.90

4.1. Experiment 1: Collected Dataset. This subsection discuss the simulation results of the proposed pothole detection model based on the collected dataset. The popular performance parameters like accuracy, precision, recall, and F1-score are considered for evaluating the results of proposed model. The several existing techniques (InceptionV3, VGG19, VGG16, SVM and ANN) are adopted for comparing the results of proposed model. The confusion matrix of proposed model and other techniques such as InceptionV3, VGG19, VGG16, SVM and ANN are depicted into Figure 4.1. On the analysis of confusion matrix, it is noticed that proposed model having more accurate confusion matrix than other technique. The other significance of computing the confusion matrix is to measure the values of accuracy, precision, recall and F1-score parameters. All these parameters are derived through confusion matrix. As confusion matrix consists of true positive, true negative, false positive and false negative. The true positive and true negative are correctly predicted data instances, while, false positive and false negative are incorrectly predicted data instance. False positive are those data instances that are predicted by classifier as positive data instance but in actual these data instances are negative. While false negative data instances are those data instance that are predicted by classifiers as negative data instance, but in actual these data instances are positive. The simulation results of proposed model and all other techniques using accuracy, precision, recall and F1-score parameters are reported into Table 4.2. It is observed that proposed model obtains 94.56% of accuracy rate than other techniques. Whereas, the accuracy rate of InceptionV3, VGG19, VGG16, SVM and ANN are 90.06%, 87.65%, 84.43%, 82.60%, and 79.73% respectively. It is observed that in neural network variants, InceptionV3 provides better results than VGG19, VGG16 and ANN. On the analysis of precision and recall parameters, it is also stated that proposed model obtains higher precision (96.40%) and recall (97.41%) rates. The precision rate of other techniques are 94.87%, 91.55%, 88.82%, 88.01%, and 85.82%. Similar, the recall rates of these techniques are 92.55%, 91.97%, 90.58%, 88.83%, and 87.32%. F1-score is significant parameter as like accuracy, to examine the performance of newly proposed model. This parameter considers the false positive and false negative data instances with respect to true positive data instance. While, accuracy parameter only considers the correctly classified data instances (true positive and true negative), so sometime accuracy can be questionable as true negative data instances may contribute higher in final results than true positive data instance. F1-score parameter of proposed model is 96.90% which is higher than all other techniques F1-score rates. The F1-score of other techniques like InceptionV3, VGG19, VGG16, SVM and ANN are 95.69%, 93.87%, 90.99%, 90.31%, and 86.56% respectively. It is also noticed that among neural network variants, InceptionV3 obtains at par results than VGG19, VGG16, and ANN using all performance parameters. SVM classifier obtains more accurate than ANN classifier for pothole detection in asphalt road, while ANN exhibits lower performance for detecting potholes among all techniques/model.

Figure 4.2 depicts the simulation results of proposed XGBoost model and other techniques such as ANN, SVM, VGG16, VGG19 and Inception V3 in graphical manner. It is clearly visible that proposed model achieves far better accuracy and recall rats for detection of potholes in asphalt road. The precision and F1-score rates of proposed model are also higher than other techniques and it is said that proposed model exhibits significant performance with these parameters. It is also highlighted that ANN technique gives less accurate results for potholes detection using all performance parameters.

The accuracy rate of the proposed XGBoost based pothole detection using training and validation sets are presented into Figure 4.3. The training set accuracy rate of the proposed model is described through green color curve, while the accuracy rate of validation set is represented through pink color curve. The training set accuracy of proposed model is 90.6%, whereas, validation set accuracy rate of the proposed model is 94.56%.

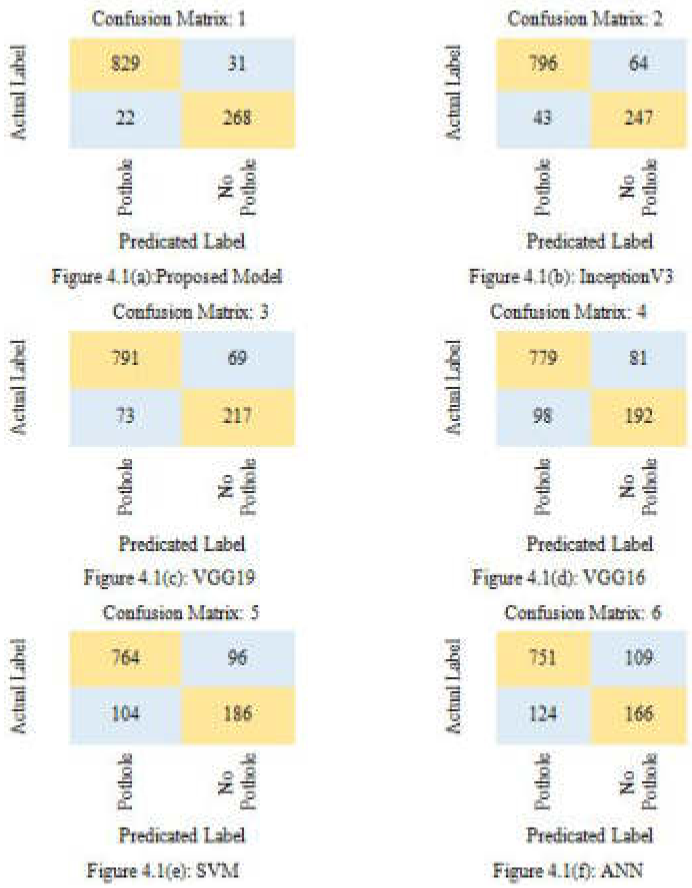


Fig. 4.1: Confusion matrix of proposed pothole detection model and other techniques like InceptionV3, VGG19, VGG16, SVM and ANN.

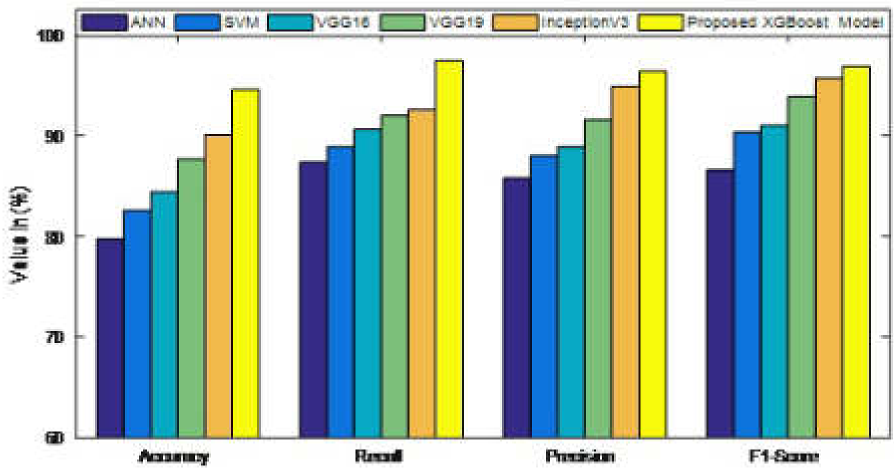


Fig. 4.2: Illustrates comparative analysis of proposed XGBoost model and other techniques using accuracy, recall, precision and F1-Score parameters

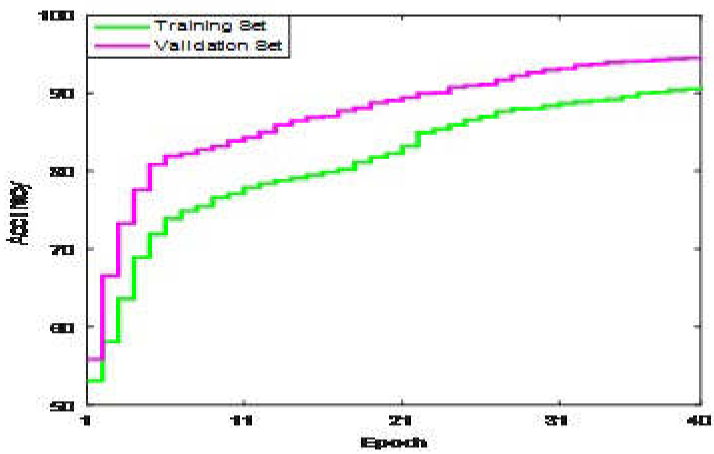


Fig. 4.3: Illustrates accuracy of the proposed Detection Model using the training and validation sets.

Along with accuracy rate, the loss function of proposed model is also plotted using training and validation set. The loss function curve is presented into Figure 4.4. It is analyzed that validation set having minimize loss function curve than training set. The significance of accuracy and loss function curves are to address the overfitting issue during the training and testing phase of the proposed model. Hence, it is stated that XGBoost based pothole detection model significantly handles the overfitting and under fitting issues of dataset. Furthermore, the simulation results of ROC and AUC parameters are reported into Figures4.5-4.6. The AUC measures the degree of separable between classes, while ROC denotes the probability curve. The higher value of AUC denotes better efficiency of the model. Whereas, ROC curve denotes the different threshold values and it is plotted by using TPR and FPR. The ROC curve of the proposed model is also compared with ROC curves of ANN, SVM, VGG16, VGG19 and InceptionV3 techniques which is depicted into Figure 4.5. This parameter illustrates the relationship among true positive rate and false positive rate. It is analyzed that proposed model obtains better ROC results than other techniques. It is observed that the proposed pothole detection model successfully handles the overfitting issue and the data are not over fitted the proposed pothole detection model. The results of the AUC parameter of proposed model is presented into the Figure 4.6. It is noted that proposed model achieves 0.974 as AUC value. Hence, it is stated that proposed XGBoost based pothole detection model is one of the effective and efficient model for accurate detection of potholes.

4.2. Experiment 2: Benchmark Pothole Dataset. This subsection presents the results of the proposed pothole detection model using the benchmark pothole dataset. This dataset is downloaded from the Github and comprised of 1243 pothole images with one class i.e. Pothole [3]. The simulation results of the proposed model and other techniques are depicted into Table 4.3. It is observed that proposed model obtains higher accuracy rate (96.21%) than other techniques being compared, while the accuracy rate of InceptionV3, VGG19, VGG16, SVM and ANN are 92.81%, 90.06%,88.99%,88.01%, and 87.29% respectively. It is observed that in neural network variants, InceptionV3 provides better results than VGG19, VGG16 and ANN. On the analysis of precision and recall parameters, it is also stated that proposed model obtains higher precision (99.02%) and recall (98.56%) rates. The precision rate of other techniques are 95.43%, 92.96%, 91.61%, 90.93%, and 88.89%. Similar, the recall rates of these techniques are 94.01%, 93.74%,92.16%,90.78%, and 88.24%. F1-score is also a significant parameter to examine the performance of newly proposed model. This parameter considers the false positive and false negative data instances with respect to true positive data instance. While, accuracy parameter only considers the correctly classified data instances (true positive and true negative), so sometime accuracy can be questionable as true negative data instances may contribute higher in final results than true positive data instance. F1-score parameter of proposed model is (97.14%) which is higher than all other techniques F1-score rates. The F1-score of other techniques like InceptionV3, VGG19, VGG16, SVM and ANN are 96.11%, 95.28%, 92.34%, 91.09%, and 89.34% respectively. It is also noticed that among neural network

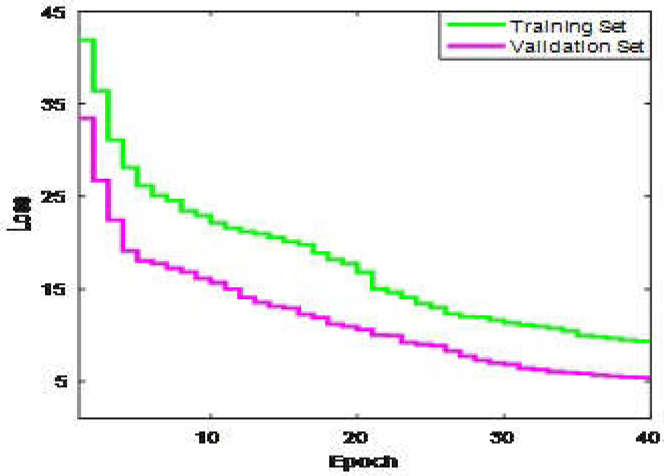


Fig. 4.4: Illustrates loss function of the proposed Detection Model the training and validation sets.s

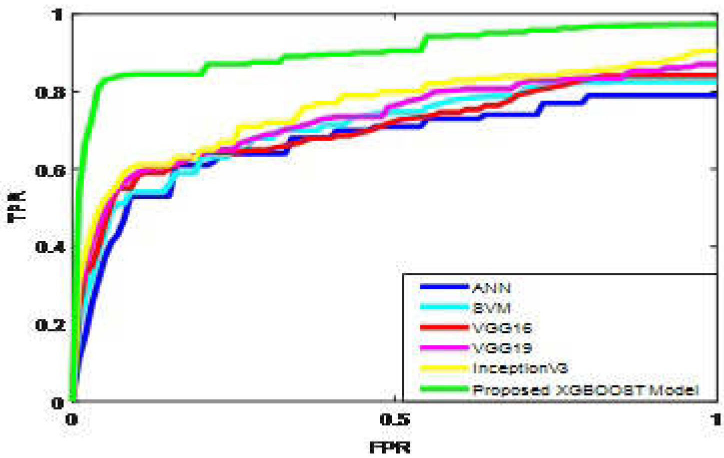


Fig. 4.5: Illustrates the ROC of proposed XGBoost based pothole detection model and other popular techniques.

variants, InceptionV3 obtains at par results than VGG19, VGG16, and ANN using all performance parameters. SVM classifier obtains more accurate than ANN classifier for pothole detection in asphalt road, while ANN exhibits lower performance for detecting potholes among all techniques/model.

Figure 4.7 depicts the simulation results of proposed XGBoost model and other techniques such as ANN, SVM, VGG16, VGG19 and Inception V3 in graphical manner using benchmark pothole dataset. It is clearly visible that proposed model achieves far better accuracy and recall rats for detection of potholes in asphalt road. The precision and F1-score rates of proposed model are also higher than other techniques and it is said that proposed model exhibits significant performance with these parameters. It is also highlighted that ANN technique gives less accurate results for potholes detection using all performance parameters.

5. Conclusion. In this work, an XGBoost based pothole detection model is proposed for effective identification of pothole in asphalt roads. The working of proposed model is three fold such as collection of images and labeling, image re-processing and texturefeature extraction, and extreme gradient boosting (XGBoost) algorithm. Furthermore in this work, a real world road image dataset is collected through smart camera. Two

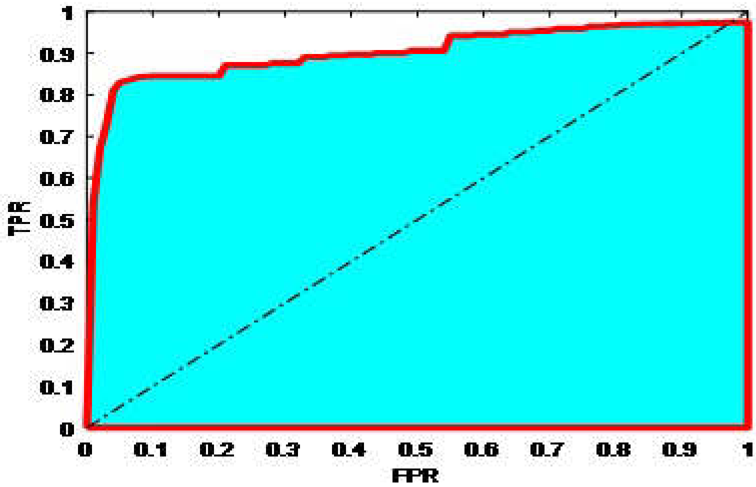


Fig. 4.6: Illustrates the AUC of proposed XGBoost based pothole detection model.

Table 4.3: Simulation results of proposed XGBoost based detection model and other popular techniques using benchmark pothole dataset

Technique	Accuracy	Recall	Precision	F1-Score
ANN	87.29	88.24	88.89	89.34
SVM	88.01	90.78	90.93	91.09
VGG16	88.99	92.16	91.61	92.34
VGG19	90.06	93.74	92.96	95.28
InceptionV3	92.81	94.01	95.43	96.11
Proposed Model	96.21	98.56	99.02	97.14

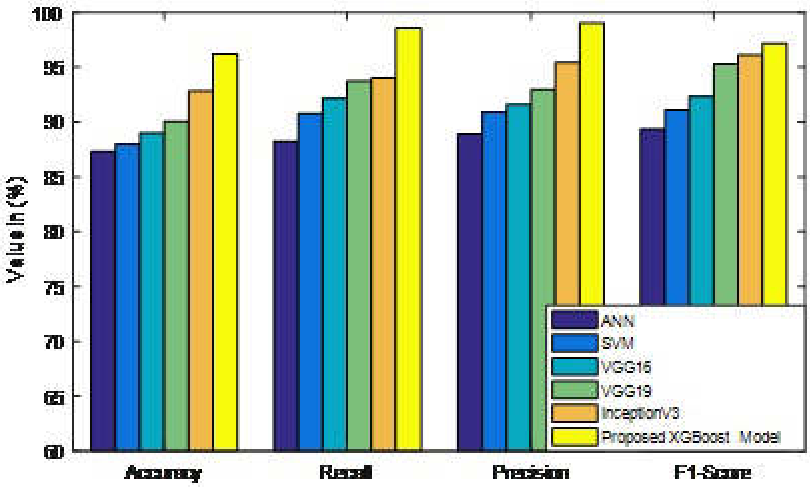


Fig. 4.7: Illustrates comparative analysis of proposed XGBoost model and other techniques using accuracy, recall, precision and F1-Score parameters based on benchmark pothole dataset.

asphalt roads of Delhi-NCR region are chosen for collecting the real world dataset and it contains eleven thousand one hundred fifty images with binary class. Moreover, the first phase of the proposed model is correspond to image collection and respective labels. In road image dataset, eight hundred sixty images are classified as pothole images and rest of images are without potholes. Second phase is responsible for image enhancement and texture feature extraction. The texture features are extracted using Gaussian and median filters, and later on K-Mean++ technique is adopted for segmentation task as well as features extraction. The prediction task is accomplished through XGBoost algorithm. A variety of parameters like accuracy, precision, recall, F1-score, AUC and ROC are considered for evaluating the proposed pothole detection model. The simulation results are also compared with several popular existing classifiers/models. The simulation results showed that proposed model achieves more than 94% of accuracy rate than other techniques. The proposed model also obtains better results with other parameters. The ROC results of proposed model is also better than other compared techniques. It is also noted that proposed model is significantly improve the prediction rate of potholes in roads. Hence, it is concluded that proposed XGBoost based pothole detection model is an effective model for detection of potholes in asphalt roads.

REFERENCES

[1] M. Anandhalli, A. Tanuja, V. P. Baligar, and P. Baligar, *Indian pothole detection based on cnn and anchor-based deep learning method*, International Journal of Information Technology, 14 (2022), pp. 3343–3353.

[2] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, A. Mraz, T. Kashiyama, and Y. Sekimoto, *Deep learning-based road damage detection and classification for multiple countries*, Automation in Construction, 132 (2021), p. 103935.

[3] B. Bučko, E. Lieskovská, K. Zábovská, and M. Zábovskỳ, *Computer vision based pothole detection under challenging conditions*, Sensors, 22 (2022), p. 8878.

[4] M.-T. Cao, K.-T. Chang, N.-M. Nguyen, V.-D. Tran, X.-L. Tran, and N.-D. Hoang, *Image processing-based automatic detection of asphalt pavement rutting using a novel metaheuristic optimized machine learning approach*, Soft Computing, 25 (2021), pp. 12839–12855.

[5] K. Chen, M. Lu, X. Fan, M. Wei, and J. Wu, *Road condition monitoring using on-board three-axis accelerometer and gps sensor*, in 2011 6th International ICST conference on communications and networking in China (CHINACOM), IEEE, 2011, pp. 1032–1037.

[6] Q. Chen, Y. Huang, H. Sun, and W. Huang, *Pavement crack detection using hessian structure propagation*, Advanced Engineering Informatics, 49 (2021), p. 101303.

[7] T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[8] O. A. Egaji, G. Evans, M. G. Griffiths, and G. Islas, *Real-time machine learning-based approach for pothole detection*, Expert Systems with Applications, 184 (2021), p. 115562.

[9] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, *The pothole patrol: using a mobile sensor network for road surface monitoring*, in Proceedings of the 6th international conference on Mobile systems, applications, and services, 2008, pp. 29–39.

[10] R. Fan, X. Ai, and N. Dahnoun, *Road surface 3d reconstruction based on dense subpixel disparity map estimation*, IEEE Transactions on Image Processing, 27 (2018), pp. 3025–3035.

[11] K. Gayathri and S. Thangavelu, *Novel deep learning model for vehicle and pothole detection*, Indones. J. Electr. Eng. Comput. Sci, 23 (2021), pp. 1576–1582.

[12] J. Guan, X. Yang, L. Ding, X. Cheng, V. C. Lee, and C. Jin, *Automated pixel-level pavement distress detection based on stereo vision and deep learning*, Automation in Construction, 129 (2021), p. 103788.

[13] S. Gupta, P. Sharma, D. Sharma, V. Gupta, and N. Sambyal, *Detection and localization of potholes in thermal images using deep neural networks*, Multimedia tools and applications, 79 (2020), pp. 26265–26284.

[14] N.-D. Hoang, *Image processing based automatic recognition of asphalt pavement patch using a metaheuristic optimized machine learning approach*, Advanced engineering informatics, 40 (2019), pp. 110–120.

[15] N.-D. Hoang, T.-C. Huynh, and V.-D. Tran, *Computer vision-based patched and unpatched pothole classification using machine learning approach optimized by forensic-based investigation metaheuristic*, Complexity, 2021 (2021), pp. 1–17.

[16] L. Janani, R. K. Dixit, V. Sunitha, and S. Mathew, *Prioritisation of pavement maintenance sections deploying functional characteristics of pavements*, International Journal of Pavement Engineering, 21 (2020), pp. 1815–1822.

[17] M. Jokela, M. Kutila, and L. Le, *Road condition monitoring system based on a stereo camera*, in 2009 IEEE 5th International conference on intelligent computer communication and processing, IEEE, 2009, pp. 423–428.

[18] M. Kamalesh, B. Chokkalingam, J. Arumugam, G. Sengottaiyan, S. Subramani, M. A. Shah, et al., *An intelligent real time pothole detection and warning system for automobile applications based on iot technology*, Journal of Applied Science and Engineering, 24 (2021), pp. 77–81.

[19] T. Kim and S.-K. Ryu, *Review and analysis of pothole detection methods*, Journal of Emerging Trends in Computing and Information Sciences, 5 (2014), pp. 603–608.

[20] C. Koch and I. Brilakis, *Pothole detection in asphalt pavement images*, Advanced engineering informatics, 25 (2011),

pp. 507–515.
[21] J. Lekshmipathy, S. Velayudhan, and S. Mathew, *Effect of combining algorithms in smartphone based pothole detection*, International Journal of Pavement Research and Technology, 14 (2021), pp. 63–72.
[22] K. Li, J. A. Misener, and K. Hedrick, *On-board road condition monitoring system using slip-based tyre-road friction estimation and wheel speed signal analysis*, Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics, 221 (2007), pp. 129–146.
[23] S. Li, C. Yuan, D. Liu, and H. Cai, *Integrated processing of image and gpr data for automated pothole detection*, Journal of computing in civil engineering, 30 (2016), p. 04016015.
[24] Y. Li, R. Ma, B. Zhang, and H. Liu, *Pavement pothole detection based on 3d laser point cloud*, in CICTP 2021, 2021, pp. 458–466.
[25] J. S. Miller, W. Y. Bellinger, et al., *Distress identification manual for the long-term pavement performance program*, tech. report, United States. Federal Highway Administration. Office of Infrastructure ..., 2003.
[26] H. Nhat-Duc, Q.-L. Nguyen, and V.-D. Tran, *Automatic recognition of asphalt pavement cracks using metaheuristic optimized edge detection algorithms and convolution neural network*, Automation in Construction, 94 (2018), pp. 203–213.
[27] A. K. Pandey, R. Iqbal, T. Maniak, C. Karyotis, S. Akuma, and V. Palade, *Convolution neural networks for pothole detection of critical road infrastructure*, Computers and Electrical Engineering, 99 (2022), p. 107725.
[28] V. Pereira, S. Tamura, S. Hayamizu, and H. Fukai, *A deep learning-based approach for road pothole detection in timor leste*, in 2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), IEEE, 2018, pp. 279–284.
[29] H. Salaudeen and E. Çelebi, *Pothole detection using image enhancement gan and object detection network*, Electronics, 11 (2022), p. 1882.
[30] R. Sathya and B. Saleena, *Cnn-mao: Convolutional neural network-based modified aquilla optimization algorithm for pothole identification from thermal images*, Signal, Image and Video Processing, 16 (2022), pp. 2239–2247.
[31] Y.-C. Tsai and A. Chatterjee, *Pothole detection and classification using 3d technology and watershed method*, Journal of Computing in Civil Engineering, 32 (2018), p. 04017078.
[32] H.-W. Wang, C.-H. Chen, D.-Y. Cheng, C.-H. Lin, C.-C. Lo, et al., *A real-time pothole detection approach for intelligent transportation system*, Mathematical Problems in Engineering, 2015 (2015).
[33] G. Xue, H. Zhu, Z. Hu, J. Yu, Y. Zhu, and Y. Luo, *Pothole in the dark: Perceiving pothole profiles with participatory urban vehicles*, IEEE Transactions on Mobile Computing, 16 (2016), pp. 1408–1419.
[34] W. Ye, W. Jiang, Z. Tong, D. Yuan, and J. Xiao, *Convolutional neural network for pothole detection in asphalt pavement*, Road materials and pavement design, 22 (2021), pp. 42–58.
[35] H. Zakeri, F. M. Nejad, and A. Fahimifar, *Image based techniques for crack detection, classification and quantification in asphalt pavement: a review*, Archives of Computational Methods in Engineering, 24 (2017), pp. 935–977.
[36] A. Zhang, K. C. Wang, and C. Ai, *3d shadow modeling for detection of descended patterns on 3d pavement surface*, Journal of Computing in Civil Engineering, 31 (2017), p. 04017019.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Mar 23, 2023
Accepted: Aug 29, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 729–742, DOI 10.12694/scpe.v24i4.2332

A DYNAMIC PATH OPTIMIZATION MODEL OF IOT DELIVERY VEHICLES FOR E-COMMERCE LOGISTICS DISTRIBUTION

JIALIN LI*

Abstract. Logistics and distribution is a vital link to guarantee the stable supply of the e-commerce market and the healthy development of the industry. With the constant growth of the e-commerce, the efficiency and service quality of logistics and distribution have been paid more and more attention to. Therefore, the study firstly Considering distribution fixed cost, transportation penalty cost and carbon emission cost, the vehicle routing optimization model is transformed into the lowest transportation cost model, then uses an improved traditional artificial fish swarm algorithm to find the optimum way for this model, and finally verifies its performance and applicability through experiments. The performance test results show that the algorithm finds the optimal solution 3589 and 3590 in 63 and 78 iterations in the Oxford Robot Car dataset and Apollo Scape dataset, respectively; the average running time of the algorithm is 11.864s and 11.967s in the 10 operation time tests; in the operation function test, the algorithm. The algorithm was able to overcome the local optimal solution problem. The applicability simulation shows that this algorithm stabilizes after 53 iterations, the minimum cost of the optimal solution of the model is $41,224, and the total distance of distribution is 9035 km. The research algorithm is fast in finding the optimal value, which is close to it, indicating that the algorithm is highly efficient and reliable, and can greatly optimize the path of e-commerce logistics delivery vehicles, and give a theoretical foundation for the optimization of logistics delivery paths in other industries.

Key words: Artificial Fish Swarm Algorithm; Logistics and Distribution; Path Optimization; E-Commerce

1. Introduction. Optimizing the dynamic path of e-commerce logistics and distribution vehicles can lift the distribution efficiency, reduce distribution costs and improve customer satisfaction [1]. Traditional e-commerce logistics methods have found it difficult to meet the needs of the modern market as the e-commerce continues to develop. Reasonable optimization of the logistics path can improve its efficiency and will deduct distribution spending. Therefore, e-commerce logistics and distribution route optimization is becoming essential [2]. Most of the traditional e-commerce logistics delivery methods are based on mathematical models or heuristic algorithms, but these methods suffer from high computational complexity, long solution time and cannot guarantee to find the globally optimum solution. Therefore, there is a need to find an efficient and reliable optimisation method to solve the E-commerce Logistics Distribution Vehicle (ECLDV) path planning problem. The IoT and the rise of modern heuristic algorithms have brought new opportunities for e-commerce logistics delivery. The opportunities are often accompanied by challenges, as current IoT technology standards are not yet harmonised and there are many security and technical issues [3, 4]. Modern heuristic algorithms also have their advantages and disadvantages. The Artificial Fish Swarm Algorithm (AFSA), which is suitable for dealing with vehicle path optimisation models, suffers from a tendency to fall into local optimum solutions, many parameter adjustments and slow convergence [5]. Therefore, the study proposes a dynamic path model for vehicle distribution based on IoT technology and improved AFSA to perfect the model of e-commerce logistics vehicles, in anticipation of solving the path optimization problem of ECLDVs. The research content is segmented into four sections: Part 1 mainly explains the research results of many experts on the vehicle path optimization problem and AFSA; the second part primarily explains the establishment and optimization strategy of the vehicle distribution path model based on IOT technology and improved AFSA; the third part mainly explains the algorithm's performance test and the simulation application test results of the model; the fourth part mainly explains the test analysis of the results.

1.1. Overview. With the progress of today's e-commerce industry, internet logistics and distribution has become one of the main businesses of e-commerce platforms. To address it, many professionals have

*School of Management Engineering, Tangshan Polytechnic College, Tangshan, 063299, China gzyljl@163.com

studied the optimization of distribution routes in a general sense. For addressing this issue, numerous scholars have conducted survey on distribution path optimisation in a general sense, and Bai R et al. proposed a hybrid approach combining ML and analytical methods to address the shortcomings of VRP applications. The approach uses ML tools in combination with analytical techniques to solve VRP. results show that the approach enhances VRP modelling and improves the performance of their algorithm [6]. Abdirad M et al. propose a two-stage hybrid algorithm to reduce transport costs in DVRP applications. The algorithm first constructs the initial route and then corrects it with an improved algorithm, which could effectively decrease transportation costs while satisfying customer needs [7]. Peng et al. propose a multiple-change transportation model with time windows to deduct the expenditure of urban-suburban logistics distribution and improve user satisfaction. The model establishes a minimum cost objective function under the constraints of time window and multiple trips, and is solved using a hybrid algorithm of packaging and genetics. This method can effectively optimise the entire distribution system [8]. Dhanare proposes a hybrid algorithm to overcome the shortest route problem and data transmission delays in connected vehicle technology. The algorithm combines ant-colony and firefly algorithms to discuss the best route, which is proven to be effective in selecting the best route and reducing travel time [9]. Bouziyane et al. propose a multi-objective local search method for the vehicle route disruption in pharmaceutical distribution with soft time windows. The method uses a hybrid algorithm-based neighbourhood search in vehicle route optimisation. The method is effective in meeting the dynamic needs of customers [10]. The AFSA is an important part of modern heuristic algorithms, mainly used in engineering optimization, economic management, machine learning and other fields. It can realize parameter search optimization according to the real-time changes of the model. Liu et al. artificially designed a reasonable urban large-scale traffic network, proposed a multi-objective optimization model for the urban traffic network problem, and then used the AFSA combining crossover operator and variational operator to solve the optimization problem. It can find the optimal solution of the model [11]. Yin et al. propose an improved AFSA to solve the problem of detecting the accuracy of energy consumption parameters of green energy efficient buildings. First is to use a hierarchical clustering method to build a classification model, and then the AFSA was used to construct an optimization function. This greatly lift the detection accuracy [12]. Sheik Abdullah proposes to use data classification techniques to effectively deploy the algorithm and set the algorithm parameters to modify the behaviour of the fish swarm. The accuracy of the algorithm improved by about 90% in different data sets [13]. Yuan et al. studied the delivery vehicle paths of several stations in order to optimise the courier business in Beijing and raised an adaptive simulated annealing and AFSA to solve the CVRP problem. The algorithm uses an adaptive vision strategy to adjust the visual range, while the search process uses "deterministic" probabilities to accept the worst solution through the Metropolis criterion. This is extremely efficient and accurate [14]. Bai et al. propose an AFSA built on a WSN to adapt the algorithm to the complexity and variability of the environment. The algorithm uses viscous fluids and artificial fish as algorithm nodes, while relevant events are directly linked to 'food'. This algorithm can effectively handle crosstalk data and improve the immunity of the algorithm to interference [15]. In summary, many experts and scholars have designed a large number of improved algorithms for optimising logistics distribution paths. The traditional AFSA, as an effective search strategy, is often applied to logistics distribution path optimisation. However, due to the limitations of this algorithm cannot optimise the dynamic path of logistics vehicle distribution more efficiently and accurately. Therefore, the research proposes the study of dynamic path of ECLDVs based on IoT and improved AFSA.

2. IoT and AFSA based vehicle path model construction and optimization strategy. This section focuses on the construction and optimisation of a dynamic path model for logistics vehicles based on IoT technology and improved AFSA. The IoT technology can quickly transfer real-time information between merchants, customers and delivery vehicles to improve the efficiency of delivery vehicles. And for the problem of the algorithm, the study adds improvements by parameter analysis setting and introducing $y = e^{-x}$ to transform the path optimisation model into the lowest cost mathematical problem model before using the improved algorithm to find the optimal solution.

2.1. Mathematical model construction for distribution vehicle paths based on. As an emerging technology, the Internet of Things (IoT) can achieve interconnection between devices and between people and things, thus improving the efficiency of e-commerce logistics distribution [17]. Generally speaking, the

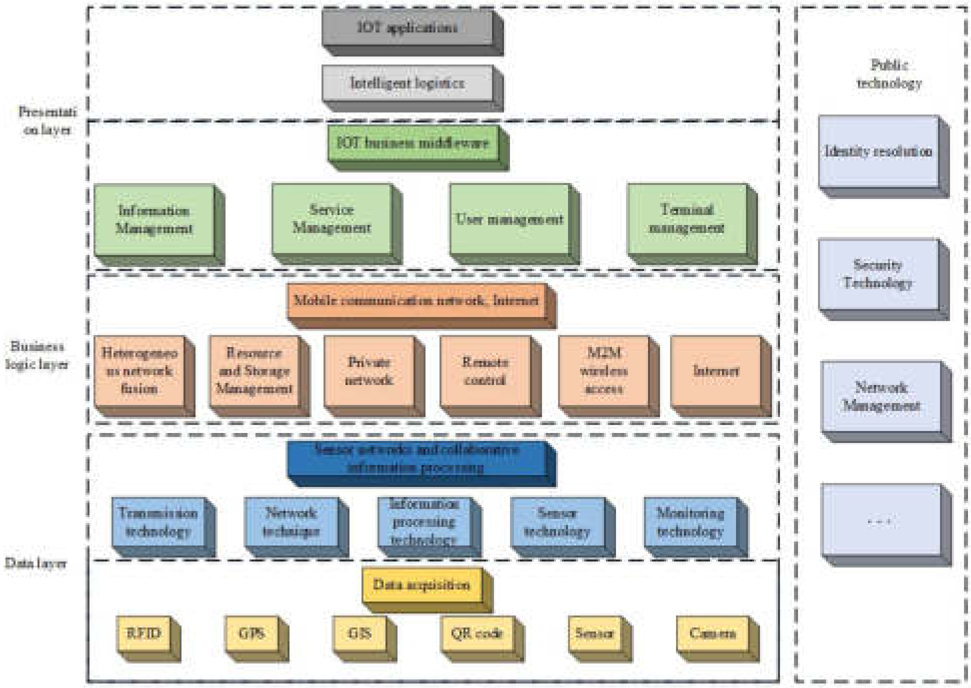


Fig. 2.1: Architecture of IoT System

e-commerce distribution IoT adopts a three-layer architecture model, the specific structure is Figure 2.1. In Figure 2.2, the first layer of this IoT is the sensing layer, which mainly completes data collection, item identification and logistics monitoring through relevant technologies; layer2 is the network layer, which mainly applies 5G communication technology to transmit the data information collected and collated from e-commerce logistics to the layer3; while layer3 is the application layer, which will make decision analysis and judgement of logistics transportation based on data information, its own reality and user needs [16] . Based on this IoT technology, a distribution flow chart for e-commerce logistics can be designed, as shown in Figure 2.2.

As Figure 2.2, the distribution process of e-commerce logistics based on IoT technology is roughly as follows: the network platform collects and organises the user's demands and transmits it to the path optimisation model; the model calculates the distribution plan and transmits it to the vehicle terminal of the distribution vehicle; at the same time, the vehicle terminal also transmits the product and vehicle information back to the network platform in time [17]. However, traditional logistics vehicle delivery ways exist problems, e.g. low efficiency, high costs and uncertain delivery times, all of which can be translated into a mathematical problem model. The essence of the vehicle path optimization problem is the optimal solution to the mathematical problem model of delivery costs and transport routes. To facilitate the analysis of this mathematical problem model, set $L = \{l_1, l_2, l_3 \dots l_n\}$ on behalf of the logistics distribution centre and customer distribution points; $K = \{k_1, k_2, k_3 \dots k_n\}$ on behalf of the transport vehicles involved in distribution; $A = \{(i,j)|i,j \in L, i \neq j$ on behalf of each distribution point between the arc set. The first is the fixed cost of vehicle distribution, which is calculated in Equation 2.1.

$$C_1 = (a+b+c) \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} v_{ijk} x_{ijk} t_{ijk} \tag{2.1}$$

The fixed cost of distribution in Equation 2.1; a is the depreciation cost; b is the maintenance cost; c is the cost of fuel used per unit of time; v_{ijk} , x_{ijk} and t_{ijk} are the speed, decision variables and time of

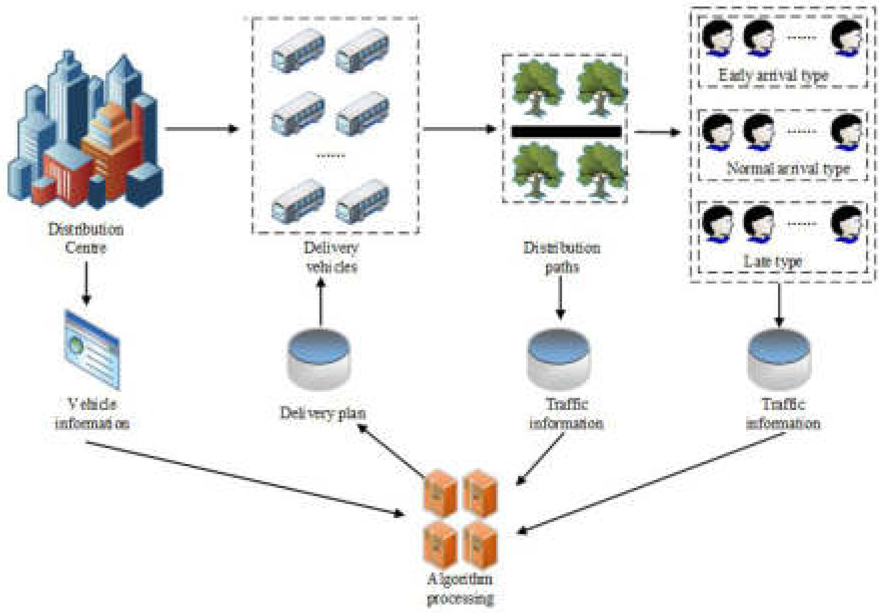


Fig. 2.2: Schematic diagram of e-commerce logistics distribution path

the distribution vehicle k between the distribution points i and j , respectively. The decision variables are calculated in Equation 2.2.

$$x_{ijk} = \begin{cases} 1 \ Delivery\ Truck\ k\ drives\ from\ i\ to\ j \\ 0 \ \ \text{If not} \end{cases} \tag{2.2}$$

The second is the cost of penalties, as there are overtime compensation costs in e-commerce logistics during delivery. The overtime compensation cost is the cost incurred by the customer when the delivery vehicle fails to reach the delivery point on time, causing losses to the customer, and the customer therefore penalises the company. If the delivery time is within $[e_i, l_i]$, the penalty cost is 0; if the delivery point is reached early, the penalty cost coefficient is $u_w^1(w = 1, 2)$; if the delivery point is reached overtime, the penalty cost coefficient is $u_w^2(w = 1, 2)$. The time window penalty cost function can be constructed from this in Equation 2.3.

$$C_{ki}^3 = \begin{cases} u_w^1(w = 1, 2) \ 0 \leq T_i \leq e_i \\ 0 \ e_i \leq T_i \leq l_i \\ u_w^2(w = 1, 2) l_i < T_i < \infty \end{cases} \tag{2.3}$$

In Equation 2.3 T_i is the delivery vehicle arrival time. The total cost of penalties can be obtained from Equation 2.3 in Equation 2.4.

$$C_2 = \sum_{k=1}^{K} \sum_{i=1}^{n} C_{ki}^3 T_i \tag{2.4}$$

Finally, there is the cost of carbon emissions. Some studies have shown that the fuel consumption of distribution vehicles is related to both vehicle weight and vehicle speed. According to the constructed IoT system, the carbon emission can be calculated by accurately recording the real-time data, e.g. the distance and time of the delivery vehicle's journey in Equation 2.5.

$$P_{ijk}^1 = (\alpha_0 + \alpha_1 v_{ijk} + \alpha_2 v_{ijk}^3 + \frac{\alpha_3}{v_{ijk}^2}) d_{ijk} \tag{2.5}$$

In Equation 2.5, $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ are carbon emission factors; is carbon emissions. The carbon emissions due to the change in vehicle weight are calculated in Equation 2.6.

$$P^1_{ijk} = \beta d_{ijk} q_{ijk} \tag{2.6}$$

In Equation 2.6, d_{ij} is the distance between i and j ; P_{ijk} is the carbon emission caused by the change of load;q_{ijk} is the load of the distribution vehicle from the distribution point i to j ; β is the carbon emission factor at load. The cost of carbon emissions during the entire distribution process is obtained from Equation 2.5 and Equation 2.6, see Equation 2.7.

$$C_3 = C_e \omega (\alpha_0 + \alpha_1 v_{ijk} + \alpha_2 v^3_{ijk} + \frac{\alpha_3}{v^2_{ijk}}) \beta \ d_{ijk} \ q_{ijk} \tag{2.7}$$

In equation (7), ω is the carbon emission coefficient. C_e is Carbon price per unit. The total cost objective function is constructed grounded on Equation 2.1, 2.4 and 2.7, which is listed in Equation 2.8.

$$\begin{cases} C_{min} = C - 1 + C_2 + C_3 \\ C_1 = (a+b+c) \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} v_{ijk} x_{ijk} t_{ijk} \\ C_2 = \sum_{k=1}^{K} \sum_{i=1}^{n} C^3_{ki} T_i \\ C_3 = C_e \omega (\alpha_0 + \alpha_1 v_{ijk} + \alpha_2 v^3_{ijk} + \frac{\alpha_3}{v^2_{ijk}}) \beta \ d_{ijk} \ q_{ijk} \end{cases} \tag{2.8}$$

The constraints of Equation 2.8 are shown in below.

$$x_{ijk} = 0 \text{ or } 1 \qquad \forall k, i, j (i \neq j) \tag{2.9}$$

$$q_k \leq Q \qquad k \in K \tag{2.10}$$

$$\sum_{j=0}^{n} \sum_{k=1}^{K} x_{ijk} = 1 \qquad i \in \{1, 2, 3, \ldots n\} \tag{2.11}$$

$$\sum_{i=0}^{n} \sum_{k=1}^{K} x_{ijk} = 1 \qquad j \in \{1, 2, 3, \ldots n\} \tag{2.12}$$

$$\sum_{i=0}^{n} q_i \cdot \sum_{j=0}^{n} x_{ijk} \leq Q \qquad k \in K \tag{2.13}$$

$$q_{ijk} = q_{(j-1)ik} \qquad k \in K \forall i.j (i \neq j) \tag{2.14}$$

$$\sum_{j=1}^{n} = \sum_{j=1}^{n} \leq 1 \qquad k \in K, i \in \{1, 2, 3, \ldots n\} \tag{2.15}$$

$$T_{jk} = T_{ik} + t_{ijk} x_{ijk} \qquad \forall i, j, (i \neq j) \tag{2.16}$$

In above equations. Equation 2.9 represents the distribution vehicle k from the i to the j distribution point obeying the piecewise variables between 0 and 1 ; Equation 2.10 constrains the load of the distribution vehicle to be greater than the demand at the distribution point; Equation 2.11 and Equation 2.12 constrains the distribution vehicle to serve all customers once; Equation 2.13 constrains the total load of all vehicles to be greater than the total demand at the distribution point; Equation 2.14 constrains the continuity of the delivery of the distribution vehicle; Equation 2.15 constrains the departure and return of the distribution vehicle from the logistics distribution centre; Equation 2.16 constrains the continuity of the delivery.

2.2. Vehicle path optimization strategy based on AFSA algorithm. The study has transformed the delivery vehicle path optimisation problem into a minimum delivery cost function model solving problem, for which modern heuristics are usually used. The modern heuristic algorithm has its own advantages and disadvantages. Considering that the vehicle load, fuel consumption and user demand are changing in real time during the distribution process, the study has decided to use the improved AFSA to solve the problem. The traditional AFSA consists of four behaviours: foraging, swarming, tail-chasing and randomisation [18]. The foraging behaviour, in which the artificial fish finds the water with the most food through its own mutual perception with the environment, is the basis of the whole algorithm and is calculated in Equation 2.17.

$$\begin{cases} X_j = X_i + Rand() \bullet Visual \\ if(Y_j < Y_i) X_{i/next} = X_i + Rand() \bullet Step \bullet \frac{X_j - X_i}{\|X_j - X_i\|} \\ else \quad X_{i/next} = X_i + Rand() \bullet Step \end{cases} \tag{2.17}$$

In above equation, X_i is the state of the fish at the moment and Y_i is its fitness value; X_j is the state of the other artificial fish and Y_j is its fitness value; *Step* is the step size; *Rand*() is a random function between 0 and 1. In equation 2.17, if , $Y_i < Y_j$, then X_i moves one step towards X_j ; if $Y_i > Y_j$, then X_j is reselected; if the condition cannot be satisfied after many attempts, then it moves one step at random. Agglomeration behaviour is where artificial fish spontaneously swim to the middle of a school in order to gather towards a place where there is more food, and in order to avoid congestion, this behaviour enhances the global and stable convergence of the algorithm, which is calculated in equation (2.11).

$$\begin{cases} if\ (Y_c < Y_i) \& \frac{n_f}{n}\ \delta \\ X_{i/next} = X_i + Rand() \bullet Step \bullet \frac{X_j - X_i}{\|X_j - X_i\|} \\ else \text{ conduct prey} \end{cases} \tag{2.18}$$

In Equation 2.18, X_c is the centre of the school; n_f is the number of other artificial fish perceived; n is the number of artificial fish in the field of view; δ is the congestion factor. In Equation 2.18, if $\frac{n_f}{n} < \ \delta$ and $Y_c < Y_i$, the artificial fish move towards the central location; if $\frac{n_f}{n} > \ \delta$ and $Y_c > Y_i$, they search for other waters and perform the foraging behavior. Tail-chasing behaviour is the behaviour of the artificial fish to follow other fish to find food quickly, this behaviour enhances the rate of convergence of the algorithm and is calculated in Equation 2.19.

$$\begin{cases} if\ (Y_{min} < Y_i) \& \frac{n_f}{n}\ \delta \\ X_{i/next} = X_i + Rand() \bullet Step \bullet \frac{X_j - X_i}{\|X_j - X_i\|} \\ else \text{ conduct prey} \end{cases} \tag{2.19}$$

In Equation 2.19 if $\frac{n_f}{n} < \ \delta$ and $Y_{min} < Y_i$, the artificial fish X_i move 1step towards X_{min} ; if $Y_{min} > Y_i$ and $Y_{min} > Y_i$; then search for other waters and perform the foraging behaviour. Random behaviour is where the fish swim aimlessly and this behaviour rises the search capability. The basic steps of a traditional AFSA are shown in Figure 2.3.

Although the traditional AFSA to find the optimal solution has the benefit of simple operation and fast convergence, there are also numerous disadvantages: the algorithm converges slowly at a later stage; the optimal solution is a range is not precise; parameter settings can affect the performance [19, 20]. To address these problems,On the basis of the traditional artificial fish swarm algorithm, an improved logistics distribution method of artificial fish swarm algorithm is proposed by adopting improved strategies such as fish swarm visual field adaptation, moving step length adaptation and parameter setting, which can accelerate the algorithm convergence speed and improve the accuracy and efficiency of the algorithm.the study first uses the control variable method to find the optimal value of the algorithm parameters, so as to improve the accuracy and efficiency to find the optimal solution. In addition, the $y = e^{-x}$ function is introduced to combine the algorithm's field of view and step size to ensure that the algorithm converges quickly and then obtains the optimal solution, and increases the local search and prevents oscillation when the field of view and step size are small in the later stage. The algorithm parameters of the search for optimality include parameters such as fish population size and number of attempts, which are discussed and analysed in Figure 2.4.

Figure 2.4 shows that the larger the number of fish in the algorithm, the more powerful the search capability of the optimal solution, but the corresponding amount of operations will also increase. Therefore, it is crucial to select the suitable number of fish according to the actual situation, under the precondition of ensuring the algorithm's optimal accuracy and computing speed. As shown in Figure 4, when the amount of attempts is

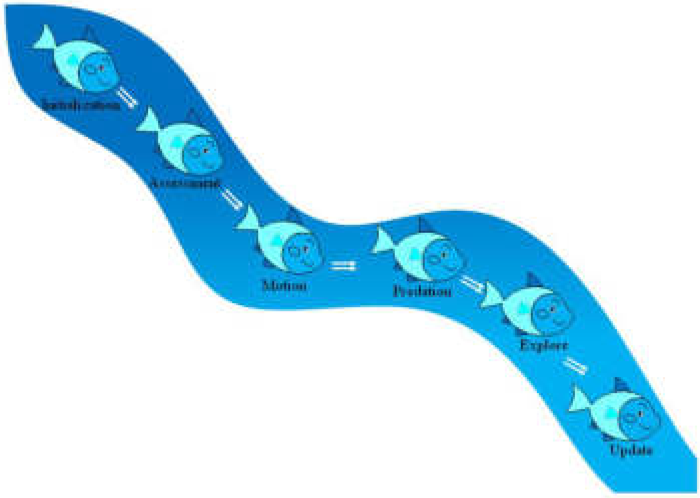


Fig. 2.3: Flow chart of artificial fish school algorithm

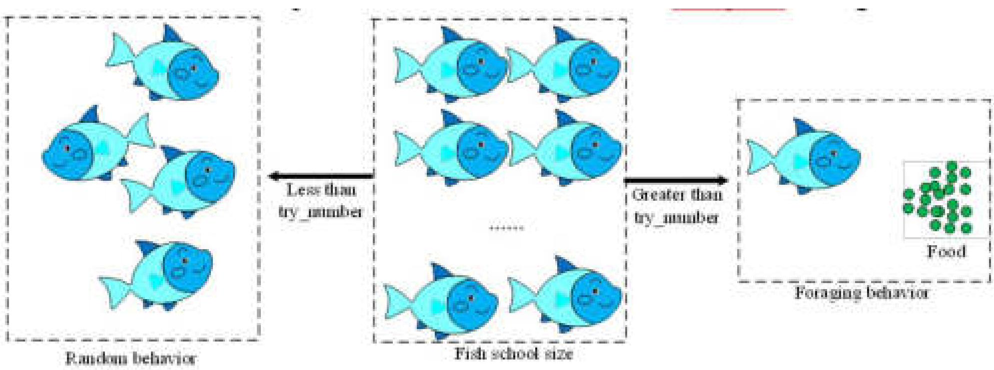


Fig. 2.4: Analysis of fish school size and number of attempts

small, the fish perform foraging behaviour; when the number of attempts is larger, there is no suitable target and the fish perform random behaviour. With a small attempt numbers, it is easier to avoid getting trapped in a local optimal solution, improving search efficiency and accuracy. The improvement method of introducing the function combined with the parameters of the algorithm is mainly for the adaptive step and field of view in the algorithm, where the adaptive step improvement method is shown in Equation 2.20.

$$\begin{cases} y = e^{-x} \\ Y_{best} < Y_i < Y_{ave} \\ x = \frac{Y_i - Y_{best}}{Y_{ave} - Y_{best}} \\ if Y_i > Y_{ave}, Step_{i/next} = Step \\ else Y_{best} < Y_i < Y_{ave} \; Step_{i/next} = e^{-x} \bullet Step \end{cases} \tag{2.20}$$

In Equation 2.20, Y_{ave} and Y_{Best} are the average and optimal fitness value. The adaptive visual field

(a) Adaptive Step Size Asjustment Curve

(b) Adaptive Field of View Adjustment Curve

Fig. 2.6: Adaptive step size and field of view adjustment curve

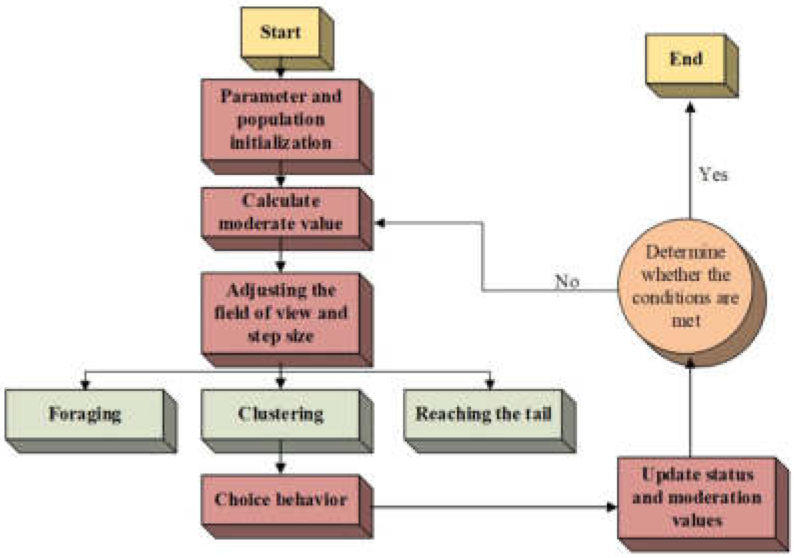


Fig. 2.7: Basic flowchart for improving artificial fish schools

improvement method is shown in Equation 2.21.

$$\begin{cases} y = e^{-x} \\ Y_{best} < Y_i < Y_{ave} \\ x = \frac{Y_i - Y_{best}}{Y_{ave} - Y_{best}} \\ ifY_i > Y_{ave}, visual_{i/next} = visual \\ else\ Y_{best} < Y_i < Y_{ave}\ visual_{i/next} = e^{-x} \bullet visual \end{cases} \tag{2.21}$$

The adaptive adjustment curves for both are shown in Figure 2.6.

The flow chart of the improved algorithm can be obtained according to the traditional AFSA and the improved algorithm scheme, see Figure 2.7.

The general flow of the improved AFSA in Figure 2.7, is as follows: first, set the parameter values according

Table 3.1: Selection of computer hardware and algorithm parameters for performance simulation experiments

Equipment	Model
CPU	Intel Core i7-4590
Internal storage	32GB
Hard Drive	256GB SSD
Graphics card	RTX 4090
Operating system	Windows 10
Computer language	MATLAB
Artificial fish scale	100
Max iteration numbers	200
Try number	7
Step	1
Visual	4

to the improved optimal parameters; then, solve the fitness value and record the whole optimal value according to the problem characteristics; then adjust the spotting and step-size of the artificial fish according to its fitness value and position state; select their behaviour according to the fitness value; finally update their current position state and compare it with the previous fitness value and end if the requirement is satisfied, if not then revert to solving the fitness function for the solution.

3. Performance and application analysis of a vehicle path optimization model based on IoT and improved AFSA. This section focuses on the performance and application analysis of the vehicle path optimisation model based on IoT and improved AFSA. After setting up the experimental parameters and the simulation environment, the computing time, the optimal solution finding ability and the computing power of the research algorithm, the AFSA, the ant colony algorithm (ACA) and the artificial neural network (ANN) were tested. And the data of actual logistics distribution were selected as parameters for simulation and testing of the optimization model.

3.1. Performance analysis of improved AFSA. To verify the accuracy and effectiveness of the improved AFSA, simulation experiments need to be performed. The computer hardware used for the experiments is displayed in Table 3.1.

For the mathematical model of delivery vehicle path optimisation, ensuring the stability and efficiency is a prerequisite for optimising the vehicle path problem. Because the path mathematical model involves real-time paths such as weather, road conditions, roads and oncoming vehicles, the study selected the Oxford Robot Car dataset and ApolloScape dataset as the test set to test the various performances of the research algorithm. To ensure the authenticity and reliability of the tests, the traditional AFSA, ACA and ANN with the same experimental conditions were selected as controls. The study first two different data sets, the four algorithms of the search for the optimal solution to test, the results are Figure 3.2.

As shown in Figure 3.1a, the relationship between the best solution and the iterations for the four algorithms in the Oxford Robot Car dataset is shown in Fig. 3.1(b). The other three algorithms are AFSA with 105 iterations to find the optimal solution 3540, ACA with 124 iterations to find that 3577, and ANN with 93 iterations to find 3560. Figure 3.1b displays the connection of the optimal solution and the iterations of the algorithms in the ApolloScape algorithm in the dataset as a function of the number of iterations to find the bast way. It is still the research algorithm that has the widest range of optimality seeking. The frequency of fluctuations in this dataset is higher than in Figure 3.1a, with 78 iterations of the research algorithm yielding an optimal solution of 3590. 110 iterations of AFSA yielded an optimal solution of 3543, 74 iterations of ACA yielded an optimal solution of 3552, and 82 iterations of ANN yielded an optimal solution of 3573. These results indicate that the research algorithm has a wider range of solutions to find, with a relatively small number of iterations and a relatively large ones. The algorithm's efficiency was then tested in both datasets by evaluating the algorithm according to the time it took to find the optimal solution. 100 tests were carried out, of which

(a) Oxford Robot Car Dataset (b) ApolloScape Car Dataset

Fig. 3.2: Optimization solution test results of four algorithms

(a) Oxford Robot Car Dataset (b) AppolloScape Car Dataset

Fig. 3.4: Optimization solution test results of four algorithms

10 were selected and the results are shown in Figure 3.4.

As shown, Figure 3.3a shows the results of the computing time tests for the four algorithms for finding the optimal solution in the Oxford Robot Car dataset. Based on the fluctuation of the line, the research algorithm has less fluctuation than the other three algorithms and is generally smoother. The mean run time of the three is 13.764s, 13.957s and 13.293s respectively. Figure 3.3b shows the average run-time of the four in the ApolloScape dataset for the operation time test results of the four algorithms for searching the optimal solution. Similar to Figure 2.5a, the fold fluctuations of the study algorithms are relatively smooth. The longest operation time is 12.768 s, the lowest is 11.498 s, and the average is 11.967 s. The average operation times of the three algorithms, AFSA, ACA and ANN, are 14.832 s, 13.589 s and 13.253 s. These results indicate that the research algorithms are highly stable and efficient. Finally, the Oxford Robot Car dataset was used as the main dataset to test the computing functions of the research algorithm and the traditional AFSA, and the results are shown in Figure 3.5.

As shown in Fig. 3.3 the five coordinate point solutions to obtain the optimum are (0,0), (10.1,10.1), (-10.1,10.1), (10.1,-10.1) and (-10.1,-10.1). It can be seen that the traditional AFSA has a local optimal solution

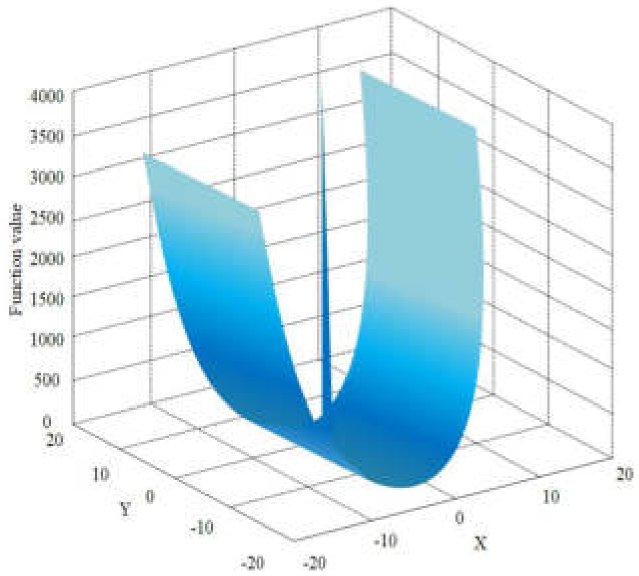


Fig. 3.5: Research Algorithm and Traditional AFSA Operational Function Test Results

in the operation, while the improved research algorithm searches for the optimal solution more accurately.

3.2. Analysis of the application of an improved AFSA-based vehicle path optimization model. The best standard for this study was determined by minimizing the cost of the objective optimization function. And certain constraints have been set in the method section, which limits the feasibility of the algorithm. Make the algorithm achieve optimal results under constraint conditions.For verifying the practical application effect of the vehicle optimisation mode, the study selects the actual data of an e-commerce logistics distribution station as parameters, and then uses four algorithms, namely the research algorithm, AFSA, ACA and ANN, to find the optimal solution for this e-commerce logistics distribution model. As the final solutions of these algorithms are infinitely close to the optimal solutions, they are highly stochastic in nature. Therefore, the study is run 50 times consecutively with the four methods, and the best solution among them is taken as the optimal solution of the model. The details are exhibited in Figure 3.6.

As Figure 3.6, all four algorithms solve for the optimal value decreases as the iterations increases. The research algorithm stabilised(The stability of the algorithm refers to the characteristic that the output result of the algorithm no longer changes significantly after a certain number of iterations. The significance of stability is that it provides a kind of predictability, that is, the research can predict with relative certainty the results that the algorithm will produce in subsequent iterations.) at 53 iterations and the optimal solution for the model's comprehensive cost was \$41,224; the ANN algorithm stabilised at 71 iterations and the optimal solution for the model's comprehensive cost was \$48,651; the ACA algorithm stabilised at 62 iterations and the optimal solution for the model's comprehensive cost was \$49,623; the AFSA algorithm stabilised at 88 iterations and the optimal solution for the model's comprehensive cost was \$56,874 The AFSA algorithm stabilised at 88 iterations and the optimal solution was \$56,874.The expected best value set is 37625 yuan. The comparisons of these four show that the research algorithm has the lowest number of iterations to find the optimal solution and the smallest integrated cost optimal value. The closest expected best value to the setting. The distribution roadmap was then plotted based on the resulting integrated cost optimal solution and compared with the pre-optimisation roadmap, the results of which are shown in Figure 3.8.

The red part in Fig. 3.5 refers to the distribution centres; the numbered dots represent the distribution points. From the optimized distribution route, the number of vehicles in each distribution centre has changed. one more vehicle in S1 distribution centre and one less vehicle in S3. From the distribution route, S1 distribution

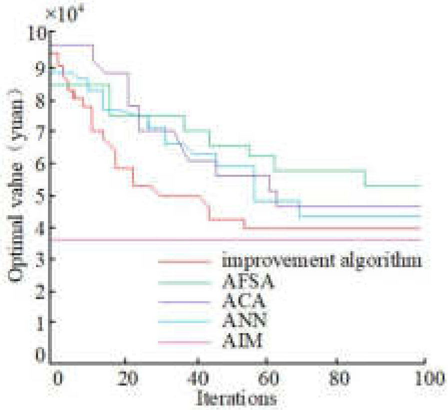


Fig. 3.6: The relationship between the minimum comprehensive cost of different algorithms and the number of iterations

(a) Before Optimization

(b) After Optimization

Fig. 3.8: Delivery route map

centre changed from S1-5-1-S1 to S1-5-S1, S1-1-S1; S2 distribution centre changed from S2-3-2-S2 to S2-3-S2, S2-2-20-10-S2; S3 distribution centre changed from S3-14-S3 to S3-14-17, S3-6-15 to After optimising the layout of the distribution network, changing the route and the number of distribution vehicles, the lowest cost distribution path for e-commerce logistics is obtained. The comparison of the results before and after the optimisation of the ECLDV paths is shown in the Table 3.2.

As can be seen from Table 3.2, all data has been improved after optimisation. The total distance of distribution before optimisation was 12,351 km and the total cost was $62,453; the total distance of distribution after optimisation was 9,035 km and the total cost was $41,224. The total distribution distance after optimisation was 3316 km less than that before optimisation, and the total cost was 17,229 yuan less. In summary, the method used in the study not only saves costs, but also improves transport efficiency and realises the optimisation of e-commerce logistics distribution paths.

4. Conclusion. The booming e-commerce market requires a more efficient and faster logistics and distribution operation system. The research first establishes a dynamic path model for logistics and distribution vehicles based on IoT technology, then transforms the optimisation of this path model into a mathematical problem model with the lowest cost optimal solution, then improves the AFSA by setting parameters and

Table 3.2: Comparison of results before and after optimization of delivery paths

/	Total distance (km)	Total cost (yuan)
Before optimization	12351	58453
After optimization	9035	41224
Optimization quantity	3316	17229

introducing new functions, and finally uses the improved algorithm to seek out the optimal solution. The algorithm proposed in study finds the optimal solution 3589 in 63 iterations with an average running time of 11.864 s. The other three algorithms, AFSA, ACA and ANN, find the optimal solution in 105, 124 and 93 iterations respectively. The other three algorithms, AFSA, ACA, and ANN, iterated 110, 74, and 82 times respectively to find the optimal solution. The other three algorithms, AFSA, ACA and ANN, iterated 110, 74 and 82 times respectively to find the optimal solutions 3543, 3552 and 3573, with an average running time of 14.832s, 13.589s and 13.253s respectively. and the research algorithm was able to overcome the shortcomings of the local optimal solution in the algorithm's operational function test. In the simulation application test using actual data of an e-commerce logistics as parameters, the research algorithm tends to be stable in 53 iterations, and the lowest cost of the optimal solution of the model is RMB 41,224, and the total distance of distribution is 9035 km. The cost saving over the traditional model is RMB 17,229 and the transport distance saving is 3316 km. It shows that the vehicle distribution path model grounded on IoT and improved AFSA proposed by the research has high accuracy and precision, and can greatly optimize the dynamic path model of ECLDV. However, the research model does not take into account the costs arising from other factors such as environmental pollution and personnel mobility, Environmental pollution caused by vehicle exhaust emissions may affect air quality, health and safety, vehicle dispersal and restriction, energy consumption, strategy adjustment and other aspects, and have an impact on transportation costs. On the other hand, personnel mobility may lead to delayed delivery times, re planning of delivery routes, and so on. and there are many complex IoT technologies, Further optimization of the research model is needed to address this series of issues

REFERENCES

[1] Yang, S. Optimization of Urban Logistics Distribution path under dynamic Traffic Network. *International Core Journal Of Engineering.* **6**, 243-248 (2020)

[2] Caso, G. Alay, "O., De Nardis, L., Brunstrom, A., Neri, M., Di Benedetto, M. *G. Empirical Models For NB-IoT Path Loss In An Urban Scenario.* **8**, 13774-13788 (2021)

[3] Lin, C., Han, G., Du, J., Xu, T., Shu, L. & Lv, Z. Spatiotemporal congestion-aware path planning toward intelligent transportation systems in software-defined smart city IoT. *IEEE Internet Of Things Journal.* **7**, 8012-8024 (2020)

[4] Guo, Y., Mustafaoglu, Z. & Koundal, D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. *Journal Of Computational And Cognitive Engineering.* **2**, 5-9 (2022)

[5] Keerthana, G., Anandan, P. & Nachimuthu, N. Robust hybrid artificial fish swarm simulated annealing optimization algorithm for secured free scale networks against malicious attacks. *Comput. Mater.* **66**, 903-917 (2021)

[6] Bai, R., Chen, X., Chen, Z., Cui, T., Gong, S., He, W. & Zhang, H. Analytics and machine learning in vehicle routing research. international Journal of Production Research. (2023)

[7] Abdirad, M., Krishnan, K. & Gupta, D. A two-stage metaheuristic algorithm for the dynamic vehicle routing problem in Industry 4.0 approach. journal of Management Analytics. (2021)

[8] Peng, Y. & Gao, H. Path Optimization of Multi-trip Swap-body Vehicle Routing Problem with Time Window. journal of Transportation Systems Engineering & Information Technology. (2020)

[9] Dhanare, R., Nagwanshi, K. & Varma, S. Enhancing the route optimization using hybrid MAF optimization algorithm for the internet of vehicles. wireless Personal Communications. (2022)

[10] Bouziyane, B., Dkhissi, B. & Cherkaoui, M. Multiobjective optimization in delivering pharmaceutical products with disrupted vehicle routing problem. *International Journal Of Industrial Engineering Computations.* **11**, 299-316 (2020)

[11] Liu, Y., Feng, X., Yang, Y., Ruan, Z., Zhang, L. & Li, K. Solving urban electric transit network problem by integrating Pareto artificial fish swarm Journal of Intelligent Transportation Systems. (2022)

[12] Yin, L. & Yin, H. Energy consumption parameter detection of green energy saving building based on artificial fish swarm algorithm. *Journal Of Global Energy Issues.* **44**, 498-510 (2022)

[13] Sheik Abdullah, A. Assessment of risk factors in medical data using improved Binary Artificial Fish Swarm Algorithm with Classification upon Evaluation from F-Test. *International Journal Of Swarm Intelligence Research (IJSIR).* **13**, 1-26

(2022)
[14] Yuan, M., Kan, X., Chi, C., Cao, L., Shu, H. & Fan, Y. An adaptive simulated annealing and artificial fish swarm algorithm for the optimization of multi-depot Intelligent Data Analysis. (2022)
[15] Bai, S., Bao, F., Zhao, F. & Liu, M. Development of an Artificial Fish Swarm Algorithm Based on aWireless Sensor Networks in a Hydrodynamic Background. fluid Dynamics and Materials Processing. (2020)
[16] Srinivas, T. & Manivannan, S. Black Hole and Selective Forwarding Attack Detection and Prevention in IoT in Health Care Sector: Hybrid meta-heuristic -based shortest path routing. *Journal Of Ambient Intelligence And Smart Environments*. **13**, 133-156 (2021)
[17] Jain, K. & Saxena, A. Simulation on supplier side bidding strategy at day-ahead electricity market using ant lion optimizer. journal of Computational and Cognitive Engineering. (2023)
[18] Mahantesh, H., Guptha, N. & M., H. M. *S. Optimized Path And Reduced Rule Caching Cost For Software Defined Network (SDN) Based Internet Of Things (IOT). Wireless Personal Communications*. **120**, 2349-2365 (2021)
[19] Tirkolaee, E., Goli, A. & Weber, G. Fuzzy mathematical programming and self-adaptive artificial fish swarm algorithm for just-in-time energy-aware flow shop scheduling problem with outsourcing option. *IEEE Transactions On Fuzzy Systems*. **28**, 2772-2783 (2020)
[20] Tian, Y., Li, Y., Pan, L. & Others (2020). Research on group animation design technology based on artificial fish swarm algorithm. *Journal Of Intelligent & Fuzzy Systems*. **38**, 1137-1145 (0)

Edited by: Zhenling Liu
Special issue on: Cloud Computing for Intelligent Traffic Management and Control
Received: June 6, 2023
Accepted: Aug 21, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 743–754, DOI 10.12694/scpe.v24i4.2345

LIGHTWEIGHT INTRUSION DETECTION METHOD OF VEHICLE CAN BUS UNDER COMPUTATIONAL RESOURCE CONSTRAINTS

XIANCHENG MING*, ZHENYU WANG †, AND BO XU‡

Abstract. In order to improve the security protection performance of the vehicle Controller Area Network (CAN) bus, the research builds an adaptive lightweight intrusion detection algorithm based on the limited computing and storage resources of the on-board ECU environment and the message cycle characteristics to supervise and detect the vehicle CAN bus intrusion. The results showed that the message cycle-based adaptive intrusion detection algorithm had high accuracy and recall rate, and fast computational search efficiency, with a stable detection time of less than 3 seconds. The intrusion detection capability is continuously optimized as the training time increases, and after stabilization, the resource utilization rate reaches over 95% with a throughput of 100Mb/s. The algorithm has strong protection capabilities. The average vehicle CPU usage of the algorithm is only 4.76%, which is 10.17% lower than the intrusion detection algorithm based on support vector machines. It can effectively prevent interference with the normal operation of the vehicle CAN bus. The algorithm has high detection accuracy for interrupt type attacks, and there are no false positives or missed alarms. For injection type attacks, the probability of missed alarms is less than 1%. The intrusion detection of vehicle CAN bus based on the message cycle characteristics provides technical reference for the safety and stability of the vehicle network, and has important practical value for the intelligent and networked development of the automobile industry.

Key words: Vehicle CAN bus; Lightweight; Intrusion detection; Message cycle

1. Introduction. In recent years, with the continuous development of the automobile industry, the output and ownership of automobiles have been continuously improved, and the requirements for vehicle comfort and technical performance have also been continuously improved [1]. Under the background of the Internet of things, the field of automotive electronic communication is constantly expanding, and intelligent and networked become the key development direction of automobiles. With the continuous innovation of the Internet of things technology, the degree of automobile networking has been continuously improved, which has brought a high degree of experience and comfort to the automobile users, and its openness has also been greatly increased, and the accompanying network attack risk has also been continuously increased [2]. The vehicle CAN bus is one of the key buses in the automotive electronic network system, facing a high risk of invasion. In recent years, the vehicle CAN bus has been frequently attacked by hackers. The relevant safety protection measures of the vehicle CAN bus have received extensive attention from all walks of life [3]. Researchers in the field of automotive network communication security conduct intrusion detection research from the perspective of the electrical characteristics and data fields of the vehicle CAN bus, but they have certain limitations, which are easy to generate detection errors, leading to false alarm or missing alarm [4]. Moreover, the computing and storage conditions of the vehicle Electronic Control Unit (ECU) environment are limited, and the deep learning intrusion detection algorithm requires high computing performance and is difficult to be directly applied in the vehicle environment [5]. Therefore, studying the intrusion risks faced by vehicle CAN buses, considering the actual operating conditions of the vehicle CAN bus, an adaptive intrusion detection algorithm based on the message cycle is proposed under the constraint of computing resources. It is expected to reduce the impact of the detection system on the bus performance while ensuring the detection effect. The innovation of this study lies in the construction of an adaptive lightweight intrusion detection algorithm based on packet cycle characteristics. The main structure of the study is divided into 5 parts. The first part is the introduction; the

*Engineering Training and Management Experiment Center, Chongqing University of Technology, Chongqing, 400054, China; *Corresponding Author E-mail: ming$_x$c954@163.com
†China Coal Science and Industry Group Chongqing Research Institute Co., Ltd., 400042, China
‡The 32nd Research Institute of China Electronics Technology Group Corporation, Shanghai, 200233, China

second part is an analysis of the current relevant research status; the third part is the design of a lightweight intrusion detection algorithm for vehicle CAN bus based on message cycle; the fourth part is an analysis of the performance and application effectiveness of the proposed algorithm; the final part is a summary of the entire study.

2. Literature review. In previous studies, a large number of scholars have proposed methods for risk control of vehicle can bus. Jo et al. proposed a mauth can authentication protocol, which can prevent the vehicle from being attacked by camouflage, and proposed a technology to prevent the bus from being attacked by shutdown [6]. Xiang team proposed a global topology constraint network for fine-grained vehicle identification based on vehicle positioning in vehicle can bus intrusion detection. The team also realized vehicle classification using convolutional neural network to achieve more accurate vehicle identification. In combination with vehicle classification and identification and precise positioning technology, the risk of vehicle can bus is investigated and controlled, which greatly reduces the risk of vehicle intrusion [7]. Katragadda steam proposed a sequence mining method to detect the low-rate injection attack of can in view of the problem that networked vehicles are subject to multiple types of network attacks. Through the analysis of four different types of attacks, the effectiveness of the sequence mining method to deal with the four types of attacks is verified. In addition, the sequence mining method of katragadda s group only uses identifiers that can be identified, so this technology can be applied to any type of vehicle and has great value in reducing the probability of vehicle being attacked [8]. In the field of electric vehicles, because the electric vehicles rely more on the in-vehicle communication system, the system will indirectly bear greater risks. Al Saud m et al. Proposed a safe and reliable intelligent framework to prevent hackers from invading vehicles. Al Saud m and others [9] improved and optimized the support vector machine and combined the social spider optimization algorithm to improve the search ability of the intrusion detection algorithm. Finally, the simulation experiment verified that the research method has high reliability and security, and can effectively prevent the electric vehicle from being attacked by denial-of-service hackers.

Intrusion detection is a useful complement to firewalls, helping systems to cope with network attacks and improving the integrity of the information security infrastructure. Ullah M U et al. proposed an intrusion detection system suitable for Apache web servers to make online communication between suppliers and customers more effective and secure [14]. Leevy et al. created datasets such as CSE-IC-IDS2018 to train predictive models for network-based intrusion detection in response to the increase in network attacks [15]. Thakkar et al. addressed the issue of intellectual property expropriation caused by cybercrime by using intrusion detection systems to protect the security of computer systems and users, and studied the performance of the system by developing datasets [16]. Salih A et al. believe that to improve the performance of intrusion detection systems, different classification algorithms must be used to detect different types of attacks, and the results of evaluating different classification algorithms from different aspects are presented to establish intrusion detection systems [23].

To sum up, the network security issue has attracted the attention of many researchers, among which the automobile network risk control supported by the Internet of things technology accounts for a large proportion. A large number of researches have adopted different algorithms to detect different types of attacks. However, it is worth noting that many researchers have not considered the limitation of computing resources in the vehicle ECU environment. Machine learning algorithms require more computing resources and are difficult to be directly applied in the vehicle ECU environment. Therefore, based on this research, aiming at the computing resource constraints of the vehicle ECU environment, an intrusion detection algorithm based on the message cycle characteristics is proposed to improve the efficiency of the intrusion detection of the vehicle network and ensure the safety of the vehicle bus operation.

3. Lightweight Intrusion Detection Method of Vehicle CAN Bus based on Message Cycle.

3.1. Cycle Analysis of Vehicle CAN Message. With the increasing openness of vehicles, the CAN bus of vehicles is exposed to the open network environment, and the vulnerability of its bus is gradually revealed. The filter acceptance mechanism of CAN bus has low security and is easily used by hackers. In addition, the data security cannot be protected, and the data authentication mechanism is not perfect, which provides an opportunity for intrusion attacks. In addition, the vehicle CAN bus strictly follows the priority system, while the vehicle CAN bus lacks the corresponding service rejection protection mechanism. Hackers can take advantage of this defect to preempt the message priority, so that the vehicle CAN bus message sending process is blocked and the system crashes. The vulnerability of the vehicle CAN bus requires the establishment of a more perfect vehicle CAN bus security protection mechanism. While the general on-board ECU is an embedded system, its memory storage performance and calculation performance have an upper limit. If the calculation complexity of the intrusion detection system exceeds the processing capacity of the on-board ECU, it will have a negative impact on the detection effect and the performance of the vehicle bus system [14, 15]. However, the traditional computer network intrusion detection algorithm requires a lot of storage and calculation resources, which is difficult to directly applied in the vehicle ECU environment. Therefore, it is necessary to design the vehicle CAN bus intrusion detection algorithm according to the calculation resource constraints of the vehicle ECU, otherwise it cannot meet the actual application environment [12, 13]. Most of the vehicle CAN messages are periodic. Therefore, considering the limited computing capacity of the on-board ECU, a lightweight bus intrusion detection algorithm based on the message cycle is proposed to adaptively detect the periodic message-oriented intrusion attacks. Instead of repeatedly computing the detection information, the intrusion is detected and identified from the periodic characteristics of the message.

The periodic characteristics of vehicle CAN messages can be obtained by using the sending time difference or receiving time difference of adjacent messages. However, the data field of vehicle CAN messages does not contain the sending time stamp of the message, nor can the sending time information be directly obtained from the receiving node [18, 19]. Therefore, if the periodic characteristics of the message are obtained from the sending time difference of the message, the vehicle ECU needs to be reprogrammed so that the sending node of the message records the sending time so that the system can perform intrusion detection. Reprogramming is too cumbersome, and the operation of the detection system will have an impact on the normal operation of the nodes. Therefore, the research combines the sending and receiving mechanism of the vehicle CAN message, and records the periodic characteristics from the difference of the message receiving time. A recording node is added to the vehicle CAN bus to receive the message and record the message reception time. The new nodes have little impact on the original system, do not interfere with the normal operation of the original functions, and occupy less system computing and storage resources. However, in the actual process, there are certain fluctuations in the cycle of the vehicle CAN message. The priority of the message and the vehicle bus load will affect the periodic characteristics of the message [20, 21]. The priority of the message determines the sending order of the message. The message with lower priority may wait to be sent, and the heavy load of the vehicle bus may also delay the sending and receiving of the message, thus changing the cycle of the message [18-19]. Suppose that the message M is sent from node A to node B, its identifier is ID_j and the sending cycle is T , message acceptance time difference is s ,and message sending sequence is i. Suppose that the sending time of the message is t_i and the actual sending time is t_i. Generally, node A sends the message at time nodes t_i ,t_i+T and t_i+2T. Without considering the bus transmission delay, node B receives the message at time points t_i ,t_i+T and t_i+2T. Assuming that the bus of the sending node is occupied when sending the $i-1$ message and the i message, resulting in the delay of sending the $i-1$ message and the i message, and the bus is idle when sending the $i+1$ message, and the message sending time is normal, the message receiving time points of node Bare t_{i-1},t_i and t_{i+1}. The sending time of vehicle CAN message depends on the timer of the main controller. The sending time of each message will not affect each other, but only affected by the bus load state at the message sending time. Without considering the bus transmission delay, the time point at which node B receives messages $i-1$,i and $i+1$ is shown as follows:

$$\begin{cases} t_{i-1} = t_{i-1} + \Delta_{i-1} \\ t_i = t_i + T + \Delta_i \\ t_{i+1} = t_{i+1} + 2T + \Delta_{i+1} \end{cases} \tag{3.1}$$

In equation 3.1, Δ_{i-1}, Δ and Δ_{i+1} are the delay time of message transmission caused by bus occupation, Δ_{i-1} , Δ, Δ_{i+1}. The difference between the receiving time s_{i-1} of the messages in Articles $i-1$ and i and the receiving time s_1 of the messages in Articles i and $i+1$ is expressed as follows:

$$\begin{cases} S_{i-1} = t_i - \ t_{i-1} = T + \Delta_i - \Delta_{i-1} \\ S_i = t_{i+1} - \ t_i = T + \Delta_{i+1} - \Delta_{i-1} = T - \Delta_i \end{cases} \tag{3.2}$$

It can be seen from equation 3.2 that the change of reception time difference S_{i-1} is determined by $\Delta_i - \Delta_{i-1}$. Let the difference between the sending cycle error of the $i-1$ message and the i message be ψ_{i-1} and $\psi_{i-1} =.\Delta_i - \Delta_{i-1}$

4. Design of the Proposed Method. According to different attack methods, vehicle bus attacks can be divided into injection attacks and denial of service attacks. Replay attacks, Denial of Service (DoS) attacks and forgery attacks are common injection attacks. Replay attacks involve hackers repeatedly sending normal messages to the vehicle centreline at any time [24, 25]. A DoS attack involves hackers spoofing nodes on the vehicle bus and injecting a large number of high priority messages that should not appear on the vehicle bus, resulting in a large number of vehicle bus resources being occupied by malicious messages, making it difficult for normal and effective messages to be sent normally [26, 27]. The term "forgery attack" refers to the act of hackers sending forged messages to the vehicle bus and interfering with the system's normal communication through faked diagnostic and abnormal messages. This can easily result in system malfunctions and cause vehicle safety accidents. Hackers' injection attacks on the vehicle bus will disrupt the normal message sending cycle.

After the vehicle bus is attacked by interruption, the vehicle ECU cannot receive valid messages, which affects the normal operation of the vehicle. Hackers' interrupt attack on vehicle bus message M is divided into temporary interrupt and permanent interrupt. Temporary interrupt only interrupts the transmission of several messages, and then resumes normal transmission. Permanent interrupt will not resume transmission after interrupting the transmission of messages [28, 29]. If interrupt attack will seriously affect the receiving time of message, combined with the fluctuation of message cycle, the detection threshold conditions of interrupt attack are as follows:

$$\omega - T > \psi_{max} \tag{4.1}$$

Combined with intrusion attack analysis, the minimum detection threshold of injection attack is shown as follows

$$\lambda_{min} = \frac{T + \Delta_{max} - \Delta_{min}}{2} = \frac{T + \psi_{max}}{2} \tag{4.2}$$

When the detection threshold is greater than λ_{min} , missed or false alarms will not occur. In order to avoid false alarm or missing alarm, the detection threshold cannot be too large. It is of great importance to set the maximum value of the detection threshold λ_{max}. When the detection threshold is the maximum value and the time period of the two messages is the smallest, no false alarm or missing alarm will occur; when $\lambda < \lambda_{max}$, and the time interval is not the minimum value, false alarm or missing alarm are bound to occur. Information entropy describes the degree of order in a system that can act on the CAN bus intrusion detection system. It defines the message $F = f_1, f_2, ..., f_n$ that appears in time T of the CAN bus, with a total of different frames. If T takes a sufficiently long time, the calculation of the total number of sending times is shown in formula 4.5.

$$N_T = \sum_{i=1}^{n} \frac{T}{S_i} = T \sum_{i=1}^{n} \frac{1}{s_i} \tag{4.3}$$

In formula 4.5, S_i represents the transmission period. Regardless of the transmission time delay, node B should receive the $i+1$ message at time $t_1 + T$. Take the minimum value of S_i , then take the maximum value of Δ_i , and take the minimum value of Δ_{i+1} as 0. At this time, the reception time difference S_i is shown as follows:

$$S_i = t_{i+1} - t_i = T - \Delta_{max} \tag{4.4}$$

The threshold conditions without false positives are expressed as follows:

$$\lambda < T - \psi \tag{4.5}$$

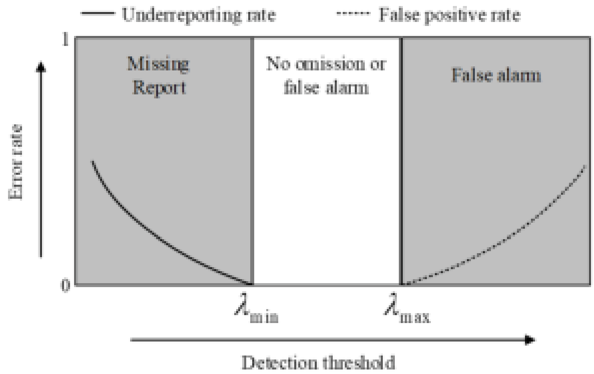


Fig. 4.1: Relationship curve between false positive rate, false negative rate and detection threshold of detection system

The false alarm and missed alarm of the detection system are discussed separately. The false alarm threshold is λ^* and the missed alarm threshold is $\lambda^{\#}$. The conditions of $\lambda^*_{max} > \lambda^{\#}_{max}$ and are analyzed separately. When $\lambda^*_{max} > \lambda^{\#}_{max}$, the message cycle meets the following conditions:

$$T > 3\psi_{max} \tag{4.6}$$

When the condition of equation (9) is met, the threshold meets condition $\lambda^* > \lambda > \lambda^{\#}_{min}$, no false positives or missing positives will occur. At this time, the relationship curve between the false positives rate and missing positives rate of the detection system and the detection threshold is shown in Figure 4.1.

When $\lambda^*_{max} < \lambda^{\#}_{min}$, the message cycle meets the following conditions:

$$T < 3\psi_{max} \tag{4.7}$$

As shown in Figure 4.2, when the detection threshold is less than λ^*_{max} or greater than $\lambda^{\#}_{min}$, the detection system may have false alarm without false alarm. When the detection threshold is within $\lambda^{\#}_{min} > \lambda > \lambda^*_{max}$, the detection system may have missed and false alarms. At this point, no false alarm is considered a priority condition and the detection threshold is slightly lower than the maximum value. Under the condition of no false alarm, a small probability of false alarm is allowed.

In case $T < 3\psi_{max}$, if the transmission time delay is ignored, node B should receive the $i+1$ message at time $t_i - T$. Take the maximum value of S_i , then take the minimum value of Δ_i as 0, and take the maximum value of Δ_{i+1}. When $\lambda = T + \psi_{max} = T - \Delta_{max}$, the receiving time of the injected message meets the following conditions:

$$t_i + \lambda < t < t_i + 2\Delta_{max} \tag{4.8}$$

Then the detection threshold is expressed as follows:

$$\lambda = T - \Delta_{max} = t_i + T + \Delta_{max} - (t_i + 2\Delta_{max}) < t_{i+1} - t \tag{4.9}$$

At this time, the time interval of missing alarm in the detection system is shown as follows:

$$t_i + 2\Delta_{max} - (t_i + \lambda) = 3\Delta_{max-T} \tag{4.10}$$

You can get:

$$\lambda_{max} > 3\lambda_{max} - T \tag{4.11}$$

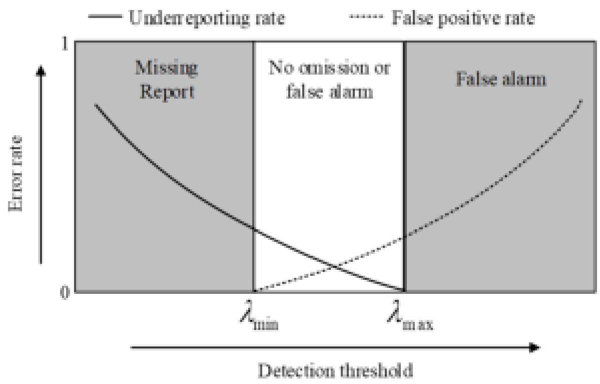


Fig. 4.2: Relationship curve between false positive rate, false negative rate and detection threshold of detection system

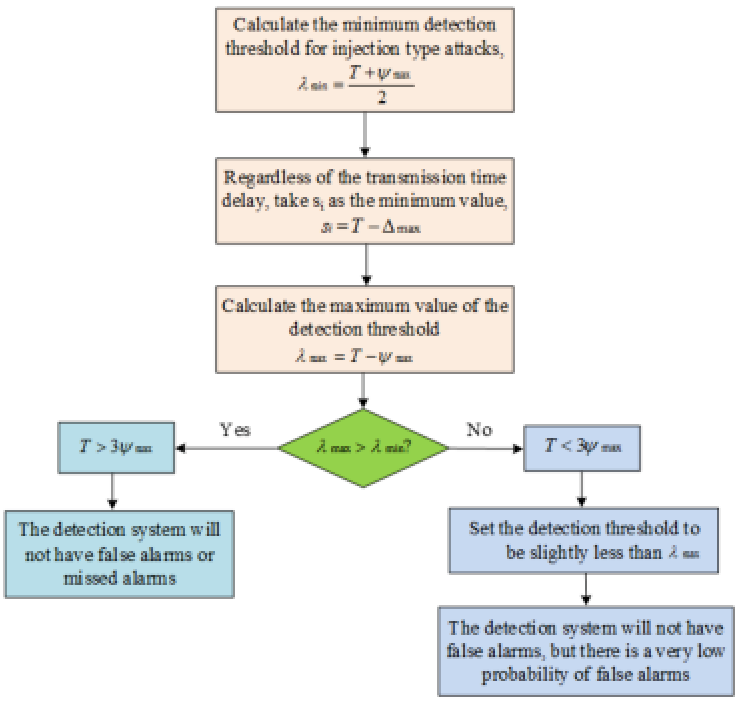


Fig. 4.3: Structured Algorithm for Bus Intrusion Adaptive Detection Based on Message Period Characteristics

To sum up, when the message cycle meets condition , the system detection threshold meets condition , no false alarm or missing alarm will occur. When the message cycle meets condition , set the system detection threshold to slightly less than of the maximum threshold. At this time, no false positives will occur, but there is a minimal probability of false positives. In summary, the structured algorithm for bus intrusion adaptive detection based on packet cycle characteristics is shown in Figure 4.3.

The proposed algorithm is detected after the detection threshold is adaptively determined. Following

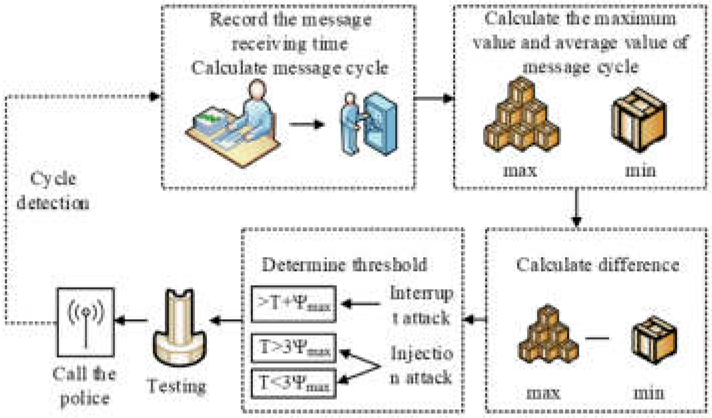


Fig. 4.4: Running flow of bus intrusion adaptive detection algorithm based on message cycle

message receipt, as shown in Figure 4, record the message's receiving time. Next, calculate the message's period. Finally, determine the detection threshold for various bus attacks by comparing the calculated message period's maximum and average values with their values. In order to avoid false positives, a small number of time margin is added to the detection threshold for fine tuning. After determining the detection threshold, the vehicle bus message is detected, the receiving time of the message is recorded and judged after initialization, the time cycle difference is detected and analyzed according to the determined threshold, and the message is judged as an interrupt attack or an injection attack. If no message is received, it is directly judged as an interrupt attack, and the vehicle bus system is circularly detected to intuitively see that the system enters the shutdown state.

5. Experiment And Result Analysis.

5.1. Algorithm Performance Test and Analysis. In order to verify the effectiveness and optimization of the adaptive intrusion detection algorithm based on the message cycle, the algorithm is compared with the sequence mining intrusion detection algorithm to check the accuracy and recall of the two algorithms. The comparison results of the accuracy and recall of the two algorithms are shown in Figure 5.1.

Figure 5.1 compares and analyses the average PR of the message cycle based adaptive intrusion detection algorithm and the sequence mining intrusion detection algorithm. Assuming the same accuracy rate, the recall rate of the message cycle based adaptive intrusion detection algorithm is higher. Given the same recall rate, the accuracy of the message cycle based adaptive intrusion detection algorithm is also higher. Compared with the sequence mining intrusion detection algorithm, the adaptive intrusion detection algorithm based on the message cycle has higher accuracy and recall rate, which proves that the adaptive intrusion detection algorithm based on the message cycle proposed by the research institute is effective and optimized, and its algorithm performance is better than the sequence mining detection algorithm. In order to clarify the detection efficiency of the algorithm in intrusion detection, the detection time of the algorithm is tested and analyzed. The detection running time of the algorithm is shown in Fig. 5.2. It can be seen from Figure 5.2 that the adaptive intrusion detection algorithm based on the packet period is trained and tested three times to observe the change in the detection time of the algorithm. In the first training of the algorithm, the detection time of the algorithm decreases continuously as the number of iterations increases. When the number of iterations reaches 50, the algorithm starts to stabilize and the detection time is stable within 3s. Compared with the first training, in the second and third training, the detection time of the algorithm is significantly reduced, which proves that the adaptive intrusion detection algorithm based on the packet cycle has faster computational search efficiency and will continuously optimize its own intrusion detection capability with the increase of training times.

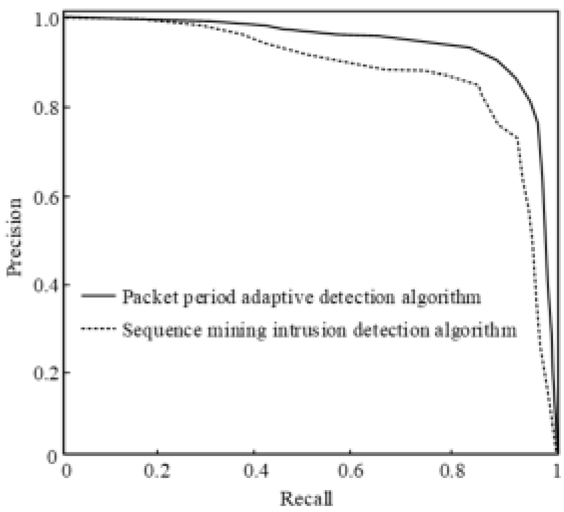


Fig. 5.1: Packet period adaptive detection algorithm PR curve

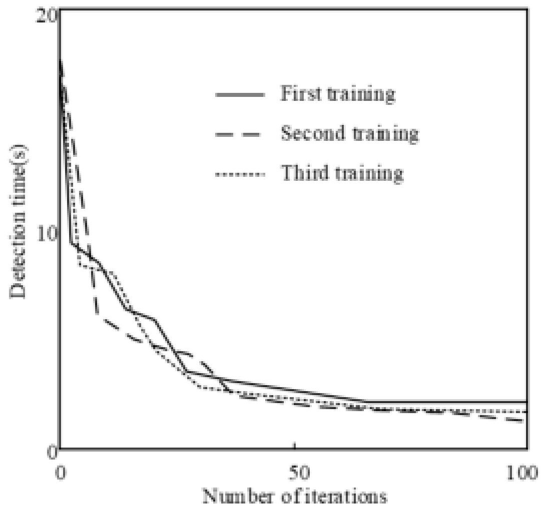


Fig. 5.2: Speed change detected by algorithm

5.2. Application Simulation Experiment Results. The vehicle CAN bus intrusion detection simulation experiment is carried out by using CANoe simulation software to validate the feasibility of the proposed algorithm. The CANoe simulation software is used to simulate the real vehicle CAN bus, to simulate the single channel high-speed vehicle CAN bus system, to add detection nodes to the vehicle bus by using the node programming function in the CANoe simulation software, and to simulate the process of the vehicle CAN bus being attacked by hackers in combination with CANdb++ and Replay Block tools. The CPU memory of the system is 8G, and the main frequency is 3.4GHz. The can main control chip is Cortex M4 core, and the main frequency is 168MHz. The running performance and detection effect of the adaptive detection algorithm based on message cycle are analyzed. The transmission of vehicle CAN bus message in the simulation environment

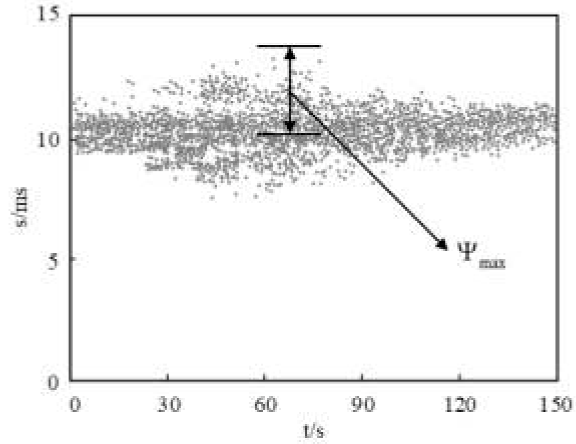


Fig. 5.3: Vehicle CAN bus message sending in simulation environment

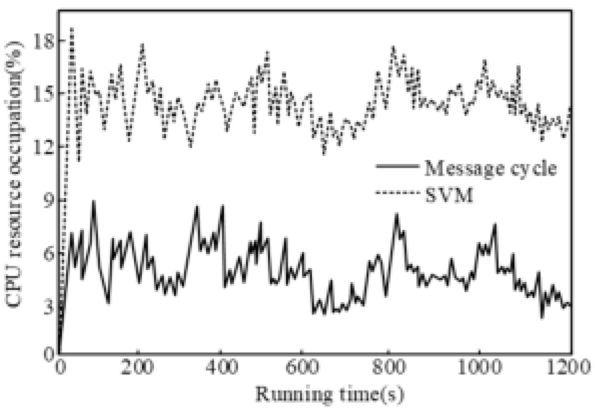


Fig. 5.4: Vehicle CPU resource occupancy detection results under two algorithms

is shown in Figure 5.3.

It can be seen from Figure 5.3 that the average sending period of vehicle CAN bus messages in the simulation environment is 10.3ms, and the CAN bus messages of vehicles have periodic characteristics. The maximum value of the difference between the reception time difference of two adjacent messages and the transmission period in the ideal state is 3.7 Ms. In order to verify the feasibility of the adaptive detection algorithm based on message cycle under the constraint of computing resources, the computing and storage resource conditions of the vehicle CAN bus are considered, and the bus intrusion detection algorithm based on Support Vector Machine (SVM) is compared and analyzed, and the CPU resource occupancy of the two algorithms is compared. The results of the vehicle CPU resource occupancy detection results under the two algorithms are shown in Figure 5.4.

As can be seen from Figure 5.4, the SVM-based detection algorithm is used to detect vehicle CAN bus intrusion. The vehicle CPU occupation is large, and the average vehicle CPU occupation is 14.93%, which seriously compresses the operation space of the original differential CAN bus system and affects the normal transmission of vehicle CAN messages. The average vehicle CPU occupancy of the intrusion adaptive detection algorithm based on message cycle is only 4.76%, which reduces the CPU occupancy by 10.17% when compared

Table 5.1: Threshold and Detection of the Attacked Message

Attack type	Attacked message	Cycle (ms)	Δmax (ms)	Threshold (ms)	False alarm number	Number of missing reports
Injection attack	M1	8	1.8	4.5	0	19
		8	1.8	6	0	0
		8	1.8	7.5	101	0
Injection attack	M2	8	3.6	3	0	32
		8	3.6	4	0	11
		8	3.6	5	1	7
Interrupt attack	M1	8	1.8	9	105	0
		8	1.8	10	0	0
		8	1.8	11	0	0
Interrupt attack	M2	8	3.6	11	21	0
		8	3.6	12	0	0
		8	3.6	13	0	0

to the SVM-based intrusion detection algorithm. This algorithm fully accounts for the computing resource constraints of the actual on-board ECU environment, occupies fewer computing resources for vehicles, and interferes less with the operation performance of the vehicle CAN bus. It can effectively avoid affecting the normal message sending due to the preemption of computing resources. It is proved that the proposed method can be applied in the actual vehicle environment and has practical feasibility.

To demonstrate the application effect of the proposed algorithm in the actual operation of the vehicle, a malicious node is added to the experimental vehicle CAN bus to simulate injection attack and interrupt attack respectively. The malicious node is used to inject 5000 forged messages into the vehicle CAN bus and conduct 5000 interrupt attacks at random. The thresholds and detection conditions of the attacked messages M1 and M2 when they are subjected to injection attack and interrupt attack respectively are shown in Table 5.1.

It can be seen from table 5.1 that the period of message M1 is 8ms. According to the threshold adaptive determination rule, its injection attack detection threshold should be in the range of [4.9,6.2]. If the threshold is 4.5ms, the detection system has 19 missing messages. If the threshold is 7.5ms, the detection system will have 101 false positives. When the detection threshold is 6ms, there are no false or missing messages. If message M1 is subject to an interrupt attack, its detection threshold must be greater than 9.8ms according to the adaptive detection rule. If the detection threshold is 9ms, there will be 105 false alarms in the detection system. If the detection threshold is 10ms and 11ms, there will be no false or missing alarms.

The time period of message M2 is 8ms. According to the threshold determination rules of the algorithm, the detection threshold of injection attack should be less than 4.4ms. When the threshold is set to 5ms, there are 1 false alarm message and 7 false alarm messages. When it is 4ms and 3MS, there are no false alarm, 11 and 32 false alarm messages, respectively. For interrupt attack on message m2, the detection threshold shall be greater than 11.6ms. When the detection threshold is 11ms, 21 false alarm messages appear in the detection system, and there is no false alarm. When the detection threshold is 12ms and 13ms, there are no false alarm or missing alarm. From the application results of the detection system, it can be seen that the proposed algorithm is effective in determining the detection threshold, and has a good effect on the detection of interrupted attacks, without false positives or missing positives. When the message cycle meets condition , the injection attack detection may have a minimal probability of missing positives.

6. Conclusion. In order to ensure the communication security of the vehicle CAN bus, the research fully considers the computing resource constraints of the on-board ECU environment, and proposes a method of monitoring and detecting the bus attack by adaptively determining the threshold. The bus simulation experiment and vehicle application experiment are carried out with CANoe simulation software. The research results contribute to improving the efficiency of intrusion detection in automotive networks and have some positive implications for communication security in automotive networks. However, the detection accuracy of the algorithm for injection type attacks is lower than that for interrupt type attacks, and there is a possibility of missing reports with a minimum probability. In the future, the injection type attack detection algorithm can be further optimized, and the Generative adversarial network can be used to solve the problem of attack sample data imbalance and improve the detection sensitivity for injection type attacks.

Funding. The research is supported by: the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN201801106).

REFERENCES

[1] Tang, C. Ceder, et al. *Vehicle Scheduling Of Single-Line Bus Service Using Operational Strategies.* **20**, 1149-1159 (2019)

[2] Zhang, H., Meng, X., Zhang, X. & Others CANsec: A Practical in-Vehicle Controller Area Network Security Evaluation Tool. *Sensors.* **20** pp. 17 (2020)

[3] Jin, H. & Papadimitratos, P. DoS-resilient cooperative beacon verification for vehicular communication systems. *Ad Hoc Networks.* **90**, 1-10177 (2019)

[4] Ning, J., Wang, J., Liu, J. & Others Attacker Identification and Intrusion Detection for In-Vehicle Networks. *IEEE Communications Letters.* **23**, 1927-1930 (2019)

[5] Li, Y., Ni, P., Zhang, S. & Others ProSampler: an ultra-fast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics.* **35**, 4632-4639 (2019)

[6] Jo, H., Jin, H., Choi, H. & Others MAuth-CAN: Masquerade-Attack-Proof Authentication for In-Vehicle Networks. *IEEE Transactions On Vehicular Technology.* **69**, 2204-2218 (2019)

[7] Xiang, Y., Fu, Y. & Huang, H. Global Topology Constraint Network for Fine-Grained Vehicle Recognition. *IEEE Transactions On Intelligent Transportation Systems.* **21**, 2918-2929 (2020)

[8] Katragadda, S., Darby, P., Roche, A. & Others Detecting Low-Rate Replay-based Injection Attacks on In-Vehicle Networks. *IEEE Access.* **8**, 54979-54993 (2020)

[9] Al-Saud, M., Eltamaly, A., Mohamed, M. & Others An Intelligent Data-Driven Model to Secure Intravehicle Communications Based on Machine Learning. *IEEE Transactions On Industrial Electronics.* **67**, 5112-5119 (2020)

[10] Ullah, M. Arfa Hassan M A, Farooq M S, et al. *Intelligent Intrusion Detection System For Apache Web Server Empowered With Machine Learning Approaches[J]. International Journal Of Computational And Innovative Sciences.* **1**, 21-27 (2022)

[11] Leevy, J. & And, K. and analysis of intrusion detection models based on cse-cic-ids2018 big data[J]. *Journal Of Big Data.* **7**, 1-19 (2020)

[12] Thakkar, A. & Lohiya, R. review of the advancement in intrusion detection datasets[J]. *Procedia Computer Science.* **167**, 636-645 (2020)

[13] Salih, A. & Abdulazeez, A. Evaluation of classification algorithms for intrusion detection system: A review[J]. *Journal Of Soft Computing And Data Mining.* **2**, 31-40 (2021)

[14] Razmjouei, P., Kavousi-Fard, A., Dabbaghjamanesh, M. & Others Ultra-lightweight Mutual Authentication in the Vehicle Based on Smart Contract Blockchain: Case of MITM Attack. *IEEE Sensors Journal.* **21**, 15839-15848 (2020)

[15] Ye, Z., Ni, H., Xu, Z. & Others Sensor fault estimation of networked vehicle suspension system with deny-of-service attack. *IET Intelligent Transport Systems.* **14**, 455-462 (2020)

[16] Olufowobi, H., Young, C., Zambreno, J. & Others SAIDuCANT: Specification-based Automotive Intrusion Detection using Controller Area Network (CAN) Timing. *IEEE Transactions On Vehicular Technology.* **69**, 1484-1494 (2019)

[17] Ju, Z., Zhang, H. & Tan, Y. Deception Attack Detection and Estimation for a Local Vehicle in Vehicle Platooning Based on a Modified UFIR Estimator[J]. *IEEE Internet Of Things Journal.* **7**, 3693-3705 (2020)

[18] Parham, M. & AA., P. An Effective Privacy-Aware Sybil Attack Detection Scheme for Secure Communication in Vehicular Ad Hoc Network. *Wireless Personal Communications.* **113**, 1149-1182 (2020)

[19] Ayaida, M., Messai, N., Najeh, S. & Others A Macroscopic Traffic Model-based Approach for Sybil Attack Detection in VANETs. *Ad Hoc Networks.* **90**, 1-10184 (2019)

[20] Feng, S. & Haykin, S. Cognitive Risk Control for Anti-Jamming V2V Communications in Autonomous Vehicle Networks. *IEEE Transactions On Vehicular Technology.* **68**, 9920-9934 (2019)

[21] Ye, Z., Ni, H., Xu, Z. & Others Sensor fault estimation of networked vehicle suspension system with deny-of-service attack. *IET Intelligent Transport Systems.* **14**, 455-462 (2020)

[22] Su, S., Tian, Z., Liang, S. & Others A Reputation Management Scheme for Efficient Malicious Vehicle Identification over 5G Networks. *IEEE Wireless Communications.* **27**, 46-52 (2020)

[23] Ju, Z., Zhang, H. & Tan, Y. Distributed Deception Attack Detection in Platoon-Based Connected Vehicle Systems. *IEEE Transactions On Vehicular Technology.* **69**, 4609-4620 (2020)

[24] Zhang, Q., Liu, K., Xia, Y. & Others Optimal Stealthy Deception Attack Against Cyber-Physical Systems. *IEEE Transactions On Cybernetics.* **50**, 3963-3972 (2019)

[25] Mousavinejad, E., Yang, F., Han, Q. & Others Distributed Cyber Attacks Detection and Recovery Mechanism for Vehicle Platooning. *IEEE Transactions On Intelligent Transportation Systems.* **21**, 3821-3834 (2020)

[26] Wu, W. Survey of Intrusion Detection for In-Vehicle Networks. *IEEE Transactions On Intelligent Transportation Systems.* **21**, 919-933 (2019)

[27] Al-Absi, M., Al-Absi, A. & Lee, H. Secure Enhanced Non-Cooperative Cognitive Division Multiple Access for Vehicle-to-Vehicle Communication. *Sensors.* **20** pp. 4 (2020)

[28] Gope, P. & Sikdar, B. An Efficient Privacy-preserving Authentication Scheme for Energy Internet-based Vehicle-to-Grid Communication. *IEEE Transactions On Smart Grid.* **10**, 6607-6618 (2019)

[29] Sanders, C. & Wang, Y. Localizing Spoofing Attacks on Vehicular GPS Using Vehicle-to-Vehicle Communications. *IEEE Transactions On Vehicular Technology.* **69**, 15656-15667 (2020)

Edited by: Zhenling Liu
Special issue on: Cloud Computing for Intelligent Traffic Management and Control
Received: June 13, 2023
Accepted: Aug 21, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 755–767, DOI 10.12694/scpe.v24i4.2351

DESIGN AND IMPLEMENTATION OF VEHICLE SCHEDULING OPTIMIZATION FOR SMART LOGISTICS PLATFORM POWERED BY HADOOP BIG DATA

GUANGTIAN YU* AND WANGTIANHUA YU†

Abstract. E-commerce has become the mainstream consumption mode for people, and with the continuous increase of online shopping business volume, controlling logistics costs has become an urgent problem to be solved. In order to solve the challenges in vehicle scheduling and achieve a scientific and reasonable vehicle scheduling scheme, this research is based on Hadoop Big data platform, introduces the concept of time axis, and builds a vehicle scheduling model based on Hadoop intelligent platform. Based on quantum genetic algorithm and combined with MapReduce model, a quantum improved genetic algorithm is constructed to solve vehicle scheduling optimization problems. The results show that the traditional quantum genetic algorithm converges after 20 iterations, achieving an optimal value of 575 and taking 360 seconds. The improved quantum genetic algorithm converged after 10 iterations and achieved an optimal value of 675, taking 200 seconds. Compared with quantum genetic algorithm, the improved quantum genetic algorithm reduces the time spent by 44.4%. Selecting customer data from a certain logistics company for testing, the improved algorithm shortens the delivery time and achieves the design of the optimal path scheduling plan. This study optimized transportation routes and resource scheduling, reduced transportation costs, and played an important role in optimizing vehicle scheduling in the logistics industry.

Key words: Hadoop big data; vehicle scheduling; quantum genetic algorithm; MapReduce model; path planning cost

1. Introduction. The logistics industry is a crucial part of modern socio-economic activities, involving the flow, storage, and distribution of goods. With the advancement of globalization and the rise of e-commerce, the logistics industry has rapidly developed and has had a profound impact on economic development and social life [1]. However, vehicle scheduling faces many challenges, requiring a balance and optimization between limited resources and complex requirements. Vehicle scheduling needs to consider the transportation needs and time windows of goods, and in addition, the delivery and delivery times of goods also need to be accurately controlled to meet customer expectations and ensure the smooth supply chain. Secondly, urban traffic congestion is becoming increasingly severe, bringing difficulties to vehicle scheduling. Reasonable planning of vehicle routes and avoiding congested areas is crucial for improving transportation efficiency and reducing time costs [2]. Therefore, through scientific and reasonable vehicle scheduling schemes, logistics enterprises can improve transportation efficiency and service quality, reduce costs and risks. The optimization of vehicle scheduling models plays an important role in improving scheduling efficiency and enhancing enterprise competitiveness. Under this background, this research builds a vehicle scheduling model based on Hadoop Big data platform. Based on quantum genetic algorithm and combined with MapReduce model for improvement, a quantum improved genetic algorithm is constructed.

The study has been divided up into four sections. The first section focuses on the study done by domestic and foreign academics on the VS problem in the logistics sector and the practical use of QGA. The development of the VS model and quantum-improved genetic algorithm constitutes the second half. The third section provides an example along with the performance testing of the quantum-improved genetic algorithm. The article's faults are discussed in the fourth and final section.

2. Related works. The logistics sector has grown as a result of people's increasing material needs, and domestic and international academics have focused their research on the VS problem in logistics distribution. The standard VS issue can only address static issues; it cannot address the dynamic issues that emerge throughout the distribution process. Some academics have proposed solutions to the dispatching problem for automobiles

*Commercial College, Suzhou Industrial Park Institute of Vocational Technology, Suzhou, 215123, China (Guangtian_Yu@outlook.com)

†Commercial College, The University of Manchester, North Manchester, YO103DU, UK

under various dynamic demands. Luo et al. propose a bespoke labelling algorithm based on a mixed integer programming model in order to solve the problem of cost control and route optimization in the logistics distribution process. A large amount of distribution data from a company is selected and the proposed method is solved for two different problems. The method can effectively solve problems with different objectives, reduce distribution costs, improve distribution efficiency and provide decision makers with better decision solutions [3]. Yami et al. have proposed a multi-product production routing problem model for the vehicle transportation problem in order to help decision makers develop better production plans. The model takes into account issues such as cost and product demand in the production process and greatly reduces the cost consumption in the production and transportation process. Through a study of a company's actual case, it was verified that the method can reduce production costs, improve resource utilization and play an important role in production planning [4]. Midaoui et al. found that the healthcare sector was facing distribution difficulties, especially as the logistics of supplying medicines to individual hospital pharmacies accounted for a large part of the budget. To address this problem, a new intelligent logistics approach was proposed based on the multi-warehouse vehicle routing problem. The method uses a clustering approach to provide suitable locations for new pharmacies and uses a genetic algorithm to route the vehicles. The method is able to solve equipment and drug distribution problems faster and can provide a better drug pick-up system for the healthcare sector [5]. A Andrade-Michel et al. propose an accurate constraint planning model in order to solve the vehicle scheduling problem by integrating driver information and the VS problem. The model combines the reliability of the driver and the importance of the journey, and is able to take into account the multiple aspects of the problems present in VS. Through the scheduling of vehicles in different scenarios, it is demonstrated that the method can effectively reduce trip coverage and improve delivery efficiency, thereby enhancing customer satisfaction [6].

To address the problems in logistics distribution, some scholars have made studies based on QGA. Lin et al. constructed an intelligent dispatching model based on enhanced QGA in order to adopt a truth coding system with vehicle departure time as the variable objective optimization. The model targets the minimum waiting time for customers and the maximum benefit for logistics companies by reducing the departure interval and increasing the on-board rate appropriately. The actual distribution routes were selected for empirical analysis, and the results showed that the method can meet the requirements of various aspects with high applicability and intelligence [7]. Ning et al. proposed an improved QGA based on dual-chain coding for the VS problem under low-carbon emission constraints. the study constructed a mathematical model with the objectives of minimum distribution time and minimum carbon emission, based on which an improved QGA using improved dual-chain QGA. Examples were selected for experiments and the model performance was analyzed using analysis of variance (ANOVA). The results showed that the model reduced carbon emissions in the dispatch process and achieved the objective of model optimization [8]. D Li et al. found that remote monitoring for automatic detection of arrhythmia is a challenging task. Based on this, a fast patient specific arrhythmia diagnosis classifier scheme is proposed by combining wavelet adaptive threshold denoising with quantum genetic algorithm based on least squares dual support vector machines. The results show that the proposed method has a detection accuracy of 98%, and compared to other representative existing technical methods, it consumes less CPU running time [9]. Y Yang et al. found that the current meta heuristic algorithm is slower in solving TSP problems. To solve this problem, a multi-scale adaptive quantum free particle optimization algorithm is proposed based on the inspiration of wave functions in quantum theory. The experimental results show that this algorithm has a faster search time compared to the ant colony optimization algorithm [10].

In summary, vehicle scheduling has always been a key issue in the logistics and distribution industry. In vehicle scheduling problems, many scholars only consider a single factor that exists in the distribution problem, while neglecting the dynamic needs of customers in the distribution network and not comprehensively considering the diversified problems that exist in the distribution process. This study takes into account the dynamic demand problem in distribution and improves the quantum genetic algorithm to achieve the goal of path optimization. It has certain reference value in the research of dynamic demand distribution.

3. VS Model and Improved QGA Construction. The optimization and enhancement of VS affects the rate of logistics and the development of logistics companies. To solve the problem of VS optimisation in dynamic situations, this chapter is divided into two parts for research. The first part is the construction of a Hadoop-based smart platform VS model and the second part is the construction of an improved QGA.The

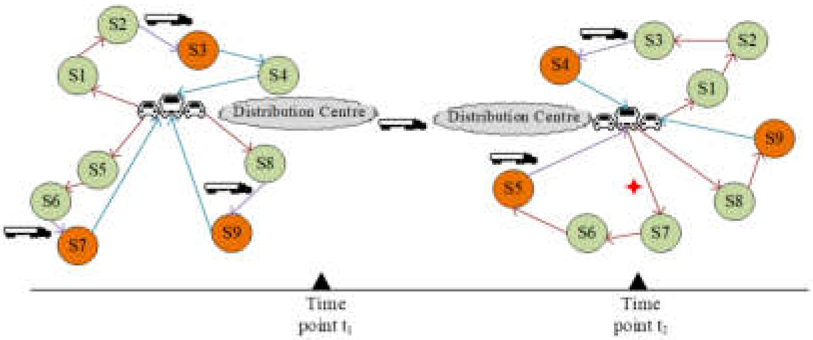


Fig. 3.1: Schematic diagram of dynamic demand distribution

optimization and enhancement of VS affects the rate of logistics and the development of logistics companies. To solve the problem of VS optimisation in dynamic situations, this chapter is divided into two parts for research. The first part is the construction of a Hadoop-based smart platform VS model and the second part is the construction of an improved QGA.

3.1. Hadoop-based Smart Platform VS Model Construction. The Hadoop wisdom platform has powerful cloud computing capabilities, with the advantages of high performance, scalability and low cost, which can meet the needs of enterprises for data processing and analysis [11]. In the smart platform delivery mode, the influencing factors in vehicle scheduling status exhibit dynamic changes, including the time required by customers, real-time delivery status, and current vehicle status. In vehicle scheduling, there are different influencing factors that can lead to a decrease in delivery efficiency, and time is an important factor affecting the scheduling status of vehicles. Considering the criticality of time in vehicle scheduling, the concept of timeline is introduced to determine a clear scheduling cycle, which facilitates the collection of dynamic demand information [12]. A schematic diagram of dynamic demand distribution is shown in Figure 3.1. In Figure 3.1, the red stars represent dynamic customers, the orange and red lines represent as well as the distribution routes that have been travelled, the light blue lines represent routes that have not yet been travelled and the purple lines represent routes that are being travelled. The orange circle indicates the most critical node in the logistics status being distributed under the condition of time . In the case of this node, the distribution task cannot be changed. The Smart Platform will identify the location of the critical node, vehicle information and customer information and generate a new dispatch plan. Setting the whole distribution network as A , vehicles are needed to complete the task at the moment, and the number of vehicles is calculated by the equation shown in equation 3.1.

$$m = \left\lceil \sum_{i\epsilon w(t)} q_i/Q \right\rceil \tag{3.1}$$

In equation 3.1,$w_u t$ denotes the set of static customers with unfinished distribution tasks and dynamic customers with new requests.q_i denotes the demand of the ith customer at the moment of i . Q denotes the maximum volume of the vehicle. The minimum function of distribution cost is shown in equation 3.2.

$$minZ = Fm + \sum_{i,j\epsilon w_{upo}(t)} \sum_{k=1}^{m} C_{ij} X_{ijk} \tag{3.2}$$

In equation 3.2, F denotes the vehicle fixed cost that has the minimum value in the total cost of distribution at the moment. $W_{upo}(t)$ denotes the set of all critical points, static customers who do not complete their tasks, dynamic customers who propose new demands and distribution centre points at the moment of t .$C_i j$ denotes the target cost of distribution distance, distribution time and distribution cost from customer i to customer j in the distribution route. The relationship between the total vehicle distribution tasks and vehicle load limits is shown in equation 3.3.

$$\sum_{i\epsilon w_u(t))} q_i y_{ik} \leq Q - Q_{jk}(t), j\epsilon W_{uo}(t), k = 1, 2..., m \tag{3.3}$$

In equation 3.3,y_{ij} indicates that the distribution task of customer i is completed by vehicle k ,$Q_{jk}(t)$ indicates the weight of the vehicle already loaded when it departs from the j th customer. The customer's requirements for the distribution vehicle are shown in equation 3.4.

$$\sum_{k-1} y_{ik} = 1, i + W_u(t) \tag{3.4}$$

Equation 3.4 indicates that Customer i accepts only one vehicle to complete the task. The number of departing vehicles is shown in equation 3.5.

$$\sum_{i\epsilon Wpo(t)} y_{ik} = m \tag{3.5}$$

In equation 3.5,m denotes a vehicle departing from a key node and distribution centre. The relationship between customers and vehicles is shown in equation 3.6.

$$\begin{cases} \sum_{i\epsilon W_u(t)} x_{ijk} = y_{jk}, j = W_u(t, k = 1, 2..., m) \\ \sum_{i\epsilon W_u(t)} x_{ijk} = y_{jk}, i = W_u(t), k = 1, 2.., m \end{cases} \tag{3.6}$$

In equation 3.6, x_{ijk} indicates that the vehicle entering the distribution network from the customer k , the vehicle driving from customer i to customer j and the initial distribution vehicle should be the same vehicle. The relationship between the variables is shown in equation 3.7.

$$\begin{cases} y_{ij}(y_{ik} - 1) = 0, i = 1, 2, ..., m & k = 1, 2, ..., m \\ x_{ijk}(x_{ijk} - 1) = 0, i = 1, 2, ..., m & j = 1, 2, ..., m \end{cases} \tag{3.7}$$

Equation 3.7 represents the relationship between the distribution tasks of customer i and the vehicles distributed between the two customers. The dynamic scheduling problem is more complex than static scheduling because of the uncertainty of information in the dynamic vehicle scheduling process [13]. In dynamic vehicle scheduling, there is uncertainty in information, and customer needs vary at different times. In order to establish a complexity criterion for the dynamic vehicle scheduling process, the equation is shown in equation 3.8 based on the ratio of dynamic customers in the distribution network

$$rs = 1 - \frac{n_t}{n}, 0 \leq rs \leq 1 \tag{3.8}$$

In equation 3.8, rs denotes the complexity rate of distribution, n_t denotes the number of dynamic customers and n denotes the total number of customers. The higher the percentage of dynamic customers in the distribution process, the more complex the distribution task is. When rs takes the value of 0, it means that all the customers in the distribution task are dynamic customers and the distribution task is complex. When the value of rs is 1, it means that all the customers in the distribution task are static customers, and the scheduling problem at this time is a static scheduling problem. In order to express the influence of dynamic information on the complexity of distribution, the complexity ratio is introduced, and the equation is shown in equation 3.9.

$$re = \frac{\sum_{i=1}^{n_t} t_i / T}{n} \tag{3.9}$$

In equation 3.9, T denotes the total distribution time and t_i denotes the time when dynamic information is generated. The above model introduces the concept of time axis, which is used to keep track of the current distribution task when dynamic information arises. The dynamic demand scheduling problem is transformed into a static scheduling problem to reduce the complexity of distribution scheduling and achieve global optimization.

3.2. VS Model based on improved QGA. Genetic algorithm is based on the principle of nature's law of superiority and inferiority, and is often used to solve multi-pole optimization and combinatorial optimization problems to achieve the purpose of global search for optimal solutions [14]. Apply genetic algorithm to vehicle scheduling to find the optimal path solution. Quantum genetic algorithm is based on genetic algorithm, introducing a quantum vector table into the chromosome encoding process, using quantum bits and quantum superposition as calculation rules. Using quantum gate rotation technology to enhance the global optimization ability of the algorithm. Quantum algorithms are solved for transformations of quantum states, and the introduction of quantum actions can increase the computational rate [15]. The quantum equation ion is shown in equation 3.10.

$$|\psi\rangle = \alpha \mid 0 + \beta \mid 1 \tag{3.10}$$

In equation 3.10,α and β denote the probability of generation relative to the state, both of which are complex numbers. α and β satisfy the normalized case as shown in equation 3.11.

$$|\alpha|^2 + |\beta|^2 = 1 \tag{3.11}$$

In equation 3.11, $|\alpha|^2$ denotes the probability that the state is 0 and $|\beta|^2$ denotes the probability that the state is 1. The quantum algorithm uses either binary or decimal for calculation. A novel form of quantum bit encoding is introduced to represent chromosomes, and the equation is shown in equation 3.12.

$$\begin{bmatrix} \alpha_1 & \Big| & \alpha_2 & \Big| & \Big| & \alpha_m & \Big| \\ \beta_1 & \Big| & \beta_2 & \Big| & \Big| & \beta_m & \Big| \end{bmatrix} \tag{3.12}$$

A chromosome with m quanta is represented in equation 3.12, and equation 3.12 can describe a superposition of any linear form. Assume that there exists a chromosome of length 2 and that the chromosome is shown in equation 3.13.

$$\begin{bmatrix} \frac{1}{\sqrt{2}} | \frac{1}{\sqrt{2}} | & \\ \frac{1}{\sqrt{2}} | \frac{1}{\sqrt{2}} | & \end{bmatrix} \tag{3.13}$$

Equation 3.13 represents the chromosomal state represented by equation (12) and the state of the quantum position is represented as shown in equation 3.14.

$$\frac{1}{2}|\ 00\rangle - \frac{1}{2}|\ 01\rangle + \frac{1}{2}|\ 10\rangle - \frac{1}{2}|\ 11\rangle \tag{3.14}$$

In equation 3.14, $00\rangle$, $01\rangle$, $10\rangle$ and $11\rangle$ are four quantum bits, all of which occur with a probability of 25%. The description using the quantum bit state approach allows a single chromosome to exhibit multiple state overlays, increasing the population diversity of the algorithm [16]. A quantum chromosome is represented by a three-dimensional quantum bit matrix of $nxnx2$ for a scheduling problem with n customers. where the order of service is represented by the horizontal coordinate and the customer delivery service number is represented by the vertical coordinate. Assuming that the order of customer delivery is randomly generated within the [0,1] interval, after obtaining a two-dimensional matrix of $m * m$, the search adjustment function is used to make each row and column of the matrix contain only one digit 1. Assuming that there are five customer delivery demands, the matrix representation is shown in equation 3.15.

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.15}$$

In equation 3.15, the customer delivery sequence is 2-1-4-3-5 and the service is completed using the same vehicle. If the delivery demand cannot be met at this time, consider adding new quantum chromosomes, and

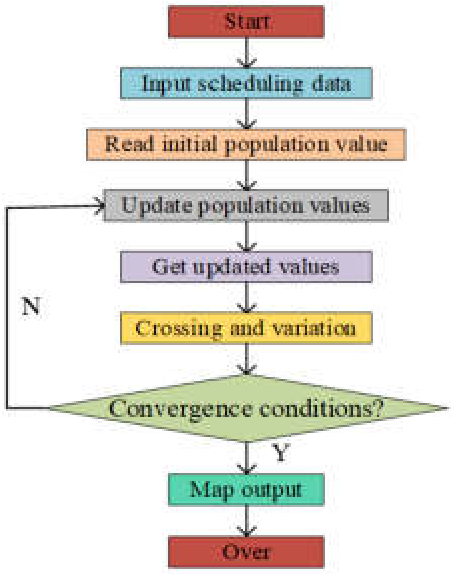


Fig. 3.2: Map function flowchart

the special group will also be converted into an integer population. Quantum gates can convert states between algorithms, and the representation of quantum gates is shown in equation 3.16.

$$\begin{bmatrix} \alpha_{m,n}^{t+1} \\ \beta_{m,n}^{t+1} \end{bmatrix} = U(\delta\theta) \begin{bmatrix} \alpha_{m,n}^{t} \\ \beta_{m,n}^{t} \end{bmatrix} = \begin{bmatrix} \cos(\delta\theta) - \sin(\delta\theta) \\ \cos(\delta\theta)\cos(\delta\theta) \end{bmatrix} \begin{bmatrix} \alpha_{m,n}^{t} \\ \beta_{m,n}^{t} \end{bmatrix} \tag{3.16}$$

In equation 3.16, $\delta{=}s(\alpha\beta)$ denotes the direction of rotation and θ denotes the angle of rotation. the equation satisfies the condition $UU' = 1$. using quantum gates can improve the convergence efficiency of the algorithm. After the quantum gate is updated, the chromosomes are decoded and the corresponding fitness values are calculated. Due to the coexistence of disorder and order in vehicle path scheduling, the encoding structure is chosen as a real number encoding structure. MapReduce is a widely used open-source software framework for parallel processing of large datasets, MapReduce computing is divided into two phases, Map phase and Reduce phase. The combination of QGA and Map and Reduce can enhance the parallelization of algorithms [17]. In the Map session, what is obtained by the computation is the solution space string record in the VS data. Assuming the existence of i_n clients, the intermediate key-value pairs obtained after computation are shown in equation (17).

$$key =' n' +' m', value = \sum_{m}^{n} ('n' +' m') \tag{3.17}$$

In equation 3.17,n indicates the number of customers and m indicates the dispatched vehicles. The variables are expressed in character form to facilitate the calculation of the data in the function. The flow chart of the Map function is shown in Figure 3.2.

In Figure 3.2, first input the scheduling data and obtain the initial population value. Update the population values through quantum gates to obtain the updated values. Then cross and reorganize the data. If it meets the requirements, it will be output. If it does not meet the requirements, it will continue to update the population values. In the Reduce function, the key and value values in equation 3.17 are combined to obtain the sequence $(key : value_1, value_2, ..., value_n)$, which is passed to the Reduce function for processing. the Reduce function normalizes the sequence to obtain the set $(P_c1, P_m1, P_c2, P_m2, , ...P_ci, P_mj)$ which can be derived from the optimized population [18]. the computational flow of the Reduce function is shown in Figure 3.3.

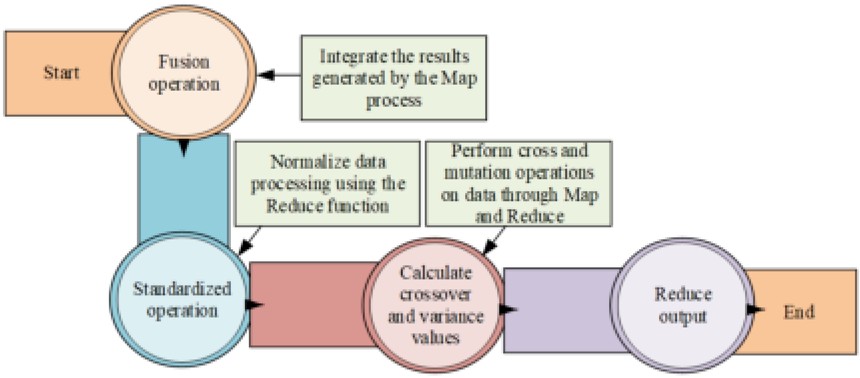


Fig. 3.3: Reduce function calculation flowchart

In Figure 3.3, the key and value values are first combined to obtain the sequence ($key : value_1, value_2...$ $value_n$). Then use the Reduce function for standardized data processing to obtain the standardized results. By using Map and Reduce to perform cross and mutation operations on the data, cross values and variance values can be calculated. Finally, output the results to obtain the optimized population.

HDFS is a database on the Hadoop platform in which all parameters are stored [19]. According to the above, improving QGA first requires setting a MapReduce parameter. In the Map function operation, the various types of data stored in the HDFS system, including vehicle information, customer demand information, vehicle location and other data, are obtained and initialized. All genetic population individuals in the Map are encoded and processed. After processing, the initial population is randomly selected and the number of populations is calculated. The calculated information is stored in the HDFS system and output to the Reduce function as value. During the computation of the Reduce function, the data in the HDFS system is first read, the population is optimized by quantum gates, and crossover and mutation operations are applied to obtain the optimal individuals [20]. The flow chart of the improved QGA is shown in Figure 3.4.

4. Algorithm Performance Testing and Example Analysis. The first part of this chapter is devoted to performance testing of the improved QGA and analyzing the performance differences between the improved QGA and the original QGA. The second part selects dynamic customer requirements and static customer requirements information to analyze the application of the improved QGA in real-world situations.

4.1. Algorithm Performance Testing. In order to test the performance advantages and disadvantages of the QGA and the improved QGA, this experimental environment used an Intel(R) Core (TM) i3 processor, CentOS7 as the operating system and Tomcat 7.0 as the web server. the QGA and the improved QGA were compared and the convergence process of the two was analyzed. The comparison graph is shown in Figure 4.2.

Figure 4.2.a shows a comparison of the convergence times of the two algorithms, and Figure 4.2.b shows a comparison of the optimal values obtained by the two algorithms. The traditional quantum genetic algorithm converges after 20 iterations, achieving an optimal value of 575 and taking 360 seconds. The improved quantum genetic algorithm converged after 10 iterations and achieved an optimal value of 675, taking 200 seconds. Compared with quantum genetic algorithm, the improved quantum genetic algorithm reduces the time spent by 44.4%.The enhanced QGA performs better during convergence and converges more effectively. The 0/1 backpack problem can be solved via combinatorial optimization, which is the process of identifying the best solution. The backpack's maximum loaded weight is 1000 pounds, and it can hold 30 items. Figure 4.3 depicts the algorithm solution diagram. From Figure 4.3, it can be seen that the convergence effect of the improved quantum genetic algorithm is better than that of the quantum genetic algorithm. The maximum value obtained by the improved quantum genetic algorithm is 2840, while the maximum value obtained by the quantum genetic algorithm is 2690. The improved quantum genetic algorithm performs better than quantum genetic algorithm

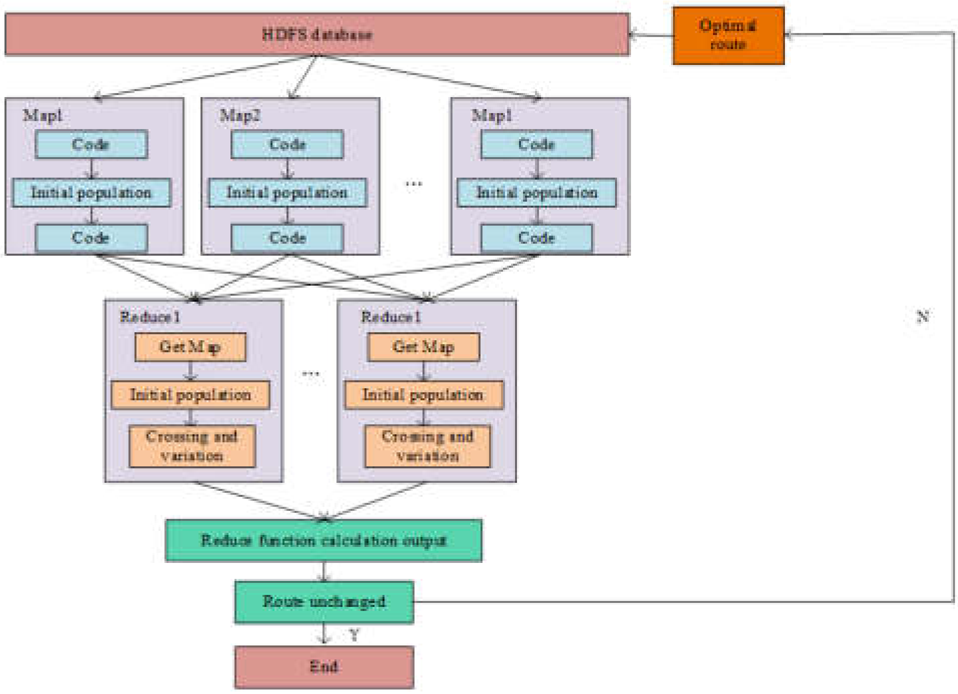


Fig. 3.4: Flow chart of improved quantum genetic algorithm for MapReduce

in solving optimal solution problems. In the transportation process, the transportation speed is affected by the real-time road conditions, and the two algorithms are used to predict the traffic flow in the transportation process, and the prediction results are shown in Figure 4.5.

Figure 4.5.a shows the prediction graph of the quantum improved genetic algorithm, and Figure 4.5.b shows the prediction graph of the original QGA. The comparison of the two plots shows that the prediction results of the quantum improved genetic algorithm match the actual results, indicating that the improved QGA has a better path finding ability for path design and can achieve good path planning for VS. The comparison graph of the planning accuracy of the two algorithms is shown in Figure 4.6. In Figure 4.6, the experimental accuracy of the improved quantum genetic algorithm has improved with the increase of experimental times, with an average experimental accuracy of 98.25%. The experimental accuracy of traditional quantum genetic algorithms will also improve with the increase of experimental times, but the improvement is small, with an average experimental accuracy of 93.88%. Compared with the traditional quantum genetic algorithm, the experimental accuracy of the improved quantum genetic algorithm has increased by 4.37%.

4.2. VS Example Analysis. A random group of customers in different cities of a logistics company is selected, and the range of city distribution intervals is a square with a side length taken as 80km. 26 static request customers are selected, each with a demand of no more than 2.5km3. the volume of the distribution vehicle is 7m3 and the vehicle travels a maximum of 120km at a time. the information table for static customers is shown in Table 4.1.

Table 4.1 contains a total of 26 static customers' coordinate locations and their demand. The distribution problem for static customers is solved using QGA and modified QGA to obtain different scheduling schemes. The scheduling schemes are shown in Figure 4.8.

The QGA dispatching plan for static client demand is shown in Figure 9(a). The data shows that there are four vehicles that have been dispatched, four main dispatch routes, and a delivery time of 310.8 minutes.

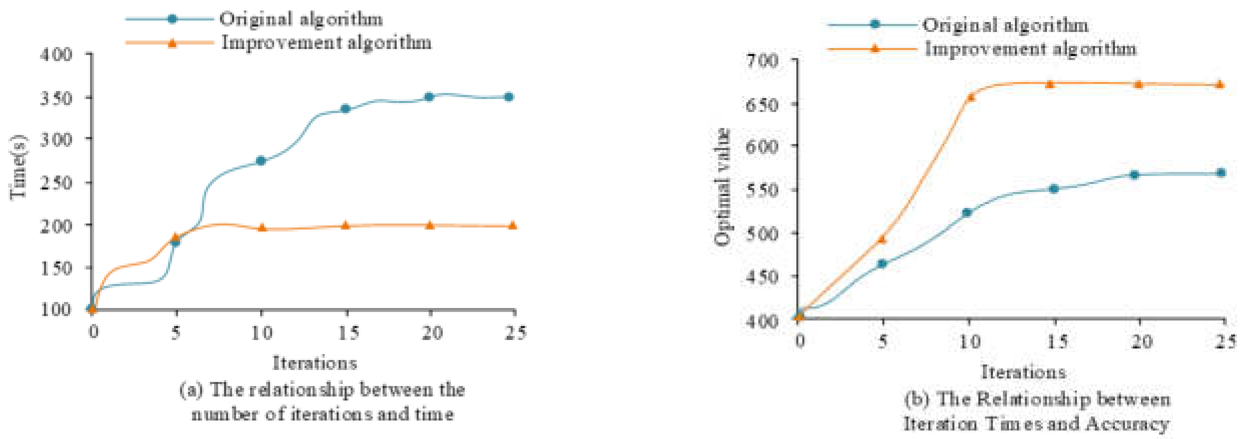


Fig. 4.2: Comparison of convergence results

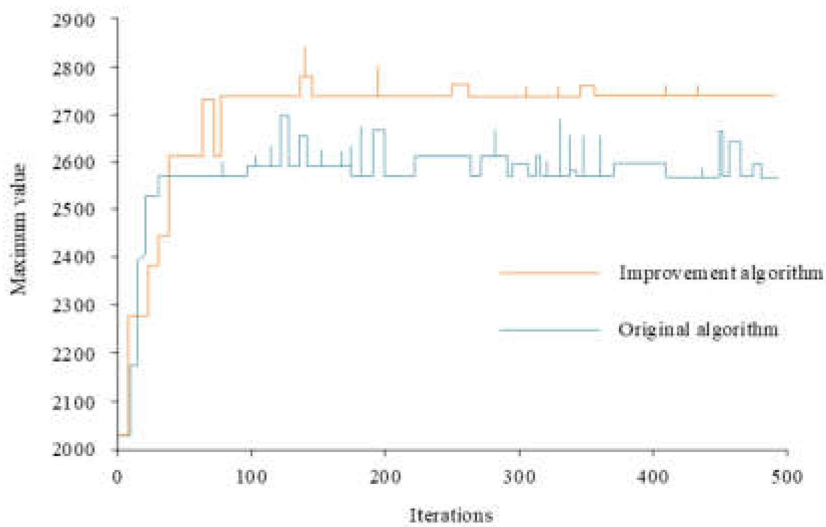


Fig. 4.3: Schematic diagram of two algorithms for solving

The dispatch routes contain route duplications, which is a poor route optimization option and wastes delivery resources. The enhanced QGA scheduling plan for static customer demand is shown in Figure 9(b). The figure shows that there are 4 dispatching vehicles, 4 main dispatching routes, no dispatching routes duplicate each other, the routes are straightforward and quick, and the delivery time is 252.5 minutes. Delivery times are slashed thanks to the enhanced algorithm, which also produces the best route dispatching scheme design. Table 4.2 displays the information table for the fourteen dynamic consumers that were chosen.

In the VS process, the dynamic customer demand changes with time. Based on the improved QGA solving static customer demand scheduling scheme, dynamic demand nodes are added and the improved QGA is used for the scheduling scheme design. The dynamic customer scheduling scheme is shown in Figure 4.9. There are four cars with the same static client demand, as shown in Figure 10. The delivery time for Vehicle 1 is

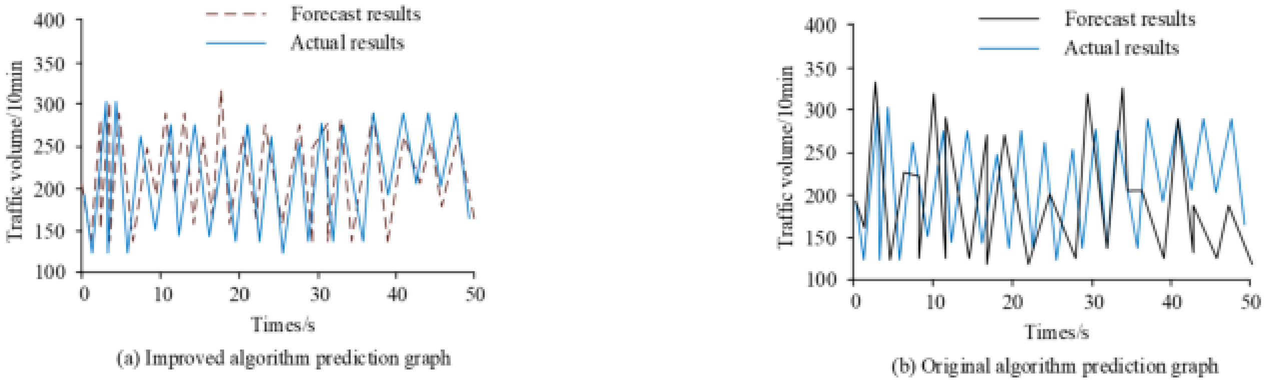


Fig. 4.5: Traffic flow prediction chart

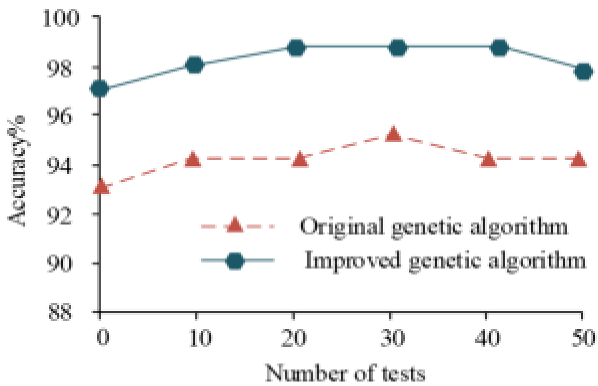
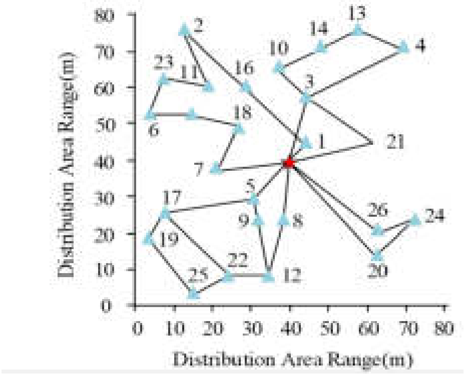


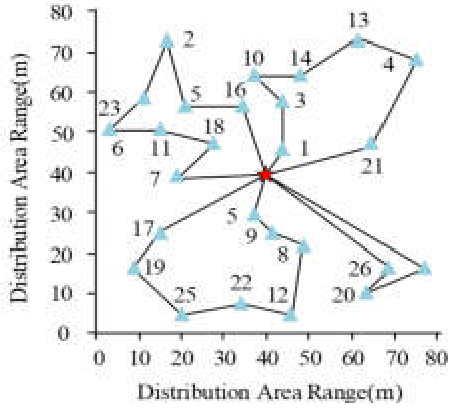
Fig. 4.6: Comparison of path planning accuracy

Table 4.1: Static Demand Customer Information Table

Customer number	Position coordinates	Requirement	Customer number	Position coordinates	Requirement
No.1	(45,40)	1.55	No.14	(48,65)	2.20
No.2	(75,20)	0.50	No.15	(26,35)	1.85
No.3	(56,43)	0.85	No.16	(35,55)	1.55
No.4	(62,70)	2.00	No.17	(15,20)	1.40
No.5	(21,55)	2.00	No.18	(27,45)	1.30
No.6	(5,50)	0.50	No.19	(8,16)	0.75
No.7	(26,35)	0.25	No.20	(14,60)	0.70
No.8	(42,20)	0.20	No.21	(60,45)	0.35
No.9	(35,25)	0.35	No.22	(35,5)	0.70
No.10	(60,42)	0.85	No.23	(10,60)	1.85
No.11	(45,16)	1.00	No.24	(18,70)	2.00
No.12	(38,5)	1.50	No.25	(20,3)	1.00
No.13	(72,60)	0.75	No.26	(16,65)	1.00



(a) Quantum genetic algorithm scheduling diagram



(b) Improved quantum genetic algorithm scheduling diagram

Fig. 4.8: Static customer scheduling scheme diagram

Table 4.2: Dynamic Demand Customer Information Table

Customer number	Position coordinates	Requirement	Customer number	Position coordinates	Requirement
A	(6,25)	2.30	H	(56,70)	1.30
B	(30,10)	0.50	I	(15,55)	1.80
C	(25,43)	0.85	J	(28,58)	1.50
D	(32,40)	1.75	K	(3,30)	1.50
E	(12,55)	2.00	L	(27,32)	1.70
F	(25,60)	0.85	M	(20,10)	1.00
G	(16,8)	0.30	N	(5,60)	2.00

65.8 minutes, with a dispatch route distance of 71.5 km. The delivery time for Vehicle 2 is 56.5 minutes, with a dispatch route distance of 61.5 km. The delivery time for Vehicle 3 is 70.5 minutes, with a dispatch route distance of 75.2 km. The delivery time for Vehicle 4 is 45.5 minutes, with a dispatch route distance of 42.6 kilometers. the enhanced Not only is the revised QGA appropriate for dynamic customer demand dispatching, but it is also ideal for static customer demand dispatching. With the goal to ensure that the distribution route in the solution is the best path and that the distribution takes the least amount of time, the upgraded QGA is able to deliver VS solutions extremely fast in response to new customer requests on the smart logistics platform. When tackling the VS problem of the smart platform, the upgraded QGA has good global optimal solution search performance, which can significantly increase the effectiveness of the distribution service of the smart platform.

5. Conclusion. With the development of Big data and e-commerce, customers' demand for logistics distribution has increased, and vehicle scheduling is an important link in the logistics distribution network. This research is based on Hadoop Big data platform, introduces the concept of time axis, determines a clear scheduling cycle, and builds a vehicle scheduling model that meets dynamic demand. Combine quantum genetic algorithm with MapReduce model to construct quantum improved genetic algorithm. The algorithm testing results show that the quantum genetic algorithm converges after 20 iterations, achieving an optimal value of 575, taking 360 seconds, and a success rate of 46.8%. The improved quantum genetic algorithm converged after 10 iterations, achieving an optimal value of 650, taking 200 seconds, and a success rate of 75.5%. Compared with quantum genetic algorithm, the improved quantum genetic algorithm has a success rate increase of 61.3% and a time reduction of 44.4%. Two algorithms were used to predict traffic flow, and the improved algorithm matched

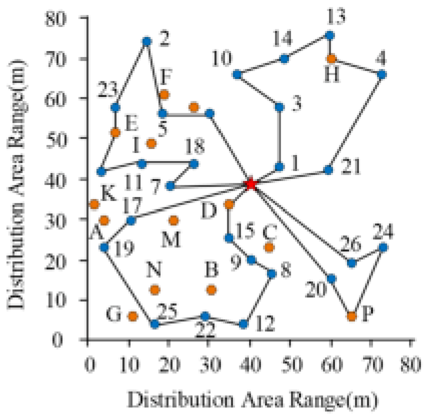


Fig. 4.9: Improved quantum genetic algorithm scheduling diagram

the actual results. Repeat the prediction experiment, and the average experimental accuracy of the improved quantum genetic algorithm is 97.85%. The average experimental accuracy of the traditional quantum genetic algorithm is 90.88%. The experimental accuracy of the improved quantum genetic algorithm is 7.7% higher than that of the traditional quantum genetic algorithm. Selecting customer data from a logistics company for testing, the improved quantum genetic algorithm can effectively plan delivery routes for both static and dynamic customer needs. Compared with the test results of the quantum genetic algorithm, the improved algorithm has shortened the delivery time and can quickly provide a vehicle scheduling plan, making the delivery route in the plan the optimal path and minimizing the delivery time consumption, achieving the design of the optimal path scheduling plan. At the same time, it provides users with faster delivery services, increasing the core competitiveness of the enterprise.There are also shortcomings in this study. The scheduling problem of vehicles only considers the needs of customers and does not consider the impact of other factors on vehicle scheduling. In future research, multiple influencing factors should be considered to improve the scheduling model.

REFERENCES

[1] Tian, Z., Zhong, R., Vatankhah Barenji, A. & Others A blockchain-based evaluation approach for customer delivery satisfaction in sustainable urban logistics. *International Journal Of Production Research.* **59**, 2229-2249 (2021)
[2] Bie, Y., Ji, J., Wang, X. & Qu, X. Optimization of electric bus scheduling considering stochastic volatilities in trip travel time and energy consumption. *Computer-Aided Civil And Infrastructure Engineering.* **36**, 1530-1548 (2021)
[3] Luo, H., Dridi, M. & Grunder, O. branch-and-price algorithm for a routing and scheduling problem from economic and environmental perspectives. *RAIRO - Operations Research.* **56**, 3267-3292 (2022)
[4] Yami, A., Okafor, D. & Modibbo, U. An Optimal Warehouse Management for Production Companies. *The Journal Of Scientific And Engineering Research.* **7**, 238-243 (2020)
[5] Midaoui, M., Qbadou, M. & Mansouri, K. Novel Approach of Smart Logistics for the Health-Care Sector Using Genetic Algorithm. *Advances In Science Technology And Engineering Systems Journal.* **5**, 1143-1152 (2020)
[6] Andrade-Michel, A., Ríos-Solís, Y. & And, B. and reliable driver scheduling for public bus transportation systems. *Transportation Research Part B Methodological.* **145**, 290-301 (2021)
[7] Lin, H. & Tang, C. Intelligent bus operation optimization by integrating cases and data driven based on business chain and enhanced quantum genetic algorithm. *IEEE Transactions On Intelligent Transportation Systems.* **23**, 9869-9882 (2021)
[8] Ning, T. & Huang, Y. Low carbon emission management for flexible job shop scheduling: A study case in China. *Journal Of Ambient Intelligence And Humanized Computing.* **14**, 789-805 (2023)
[9] Li, D. & Shi, R. Yao N. *Eal-Time Patient-Specific ECG Arrhythmia Detection By Quantum Genetic Algorithm Of Least Squares Twin SVM. Journal Of Beijing Institute Of Technology.* **29**, 32-40 (2020)
[10] Yang, Y. & Wang, P. Multi-scale quantum free particle optimization algorithm for solving travelling salesman problem. *Journal Of Computer Applications.* **40**, 1278-1283 (2020)
[11] 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)
[12] Zan, J. Research on robot path perception and optimization technology based on whale optimization algorithm. *Journal Of Computational And Cognitive Engineering.* **1**, 201-208 (2022)

[13] Kim, H., Lim, D. & Lee, S. Deep learning-based dynamic scheduling for semiconductor manufacturing with high uncertainty of automated material handling system capability. *IEEE Transactions On Semiconductor Manufacturing.* **33**, 13-22 (2020)

[14] Dharma, F., Shabrina, S. & Noviana, A. Prediction of Indonesian inflation rate using regression model based on genetic algorithms. *Jurnal Online Informatika.* **5**, 45-52 (2020)

[15] Cheng, G., Wang, C. & Xu, C. novel hyper-chaotic image encryption scheme based on quantum genetic algorithm and compressive sensing. *Multimedia Tools And Applications.* **79**, 29243-29263 (2020)

[16] Debnath, S. Fuzzy quadripartitioned neutrosophic soft matrix theory and its decision-making approach. *Journal Of Computational And Cognitive Engineering.* **1**, 88-93 (2022)

[17] Hu, L., Yang, S., Luo, X., Yuan, H., Sedraoui, K. & Zhou, M. distributed framework for large-scale protein-protein interaction data analysis and prediction using mapreduce. *IEEE/CAA Journal Of Automatica Sinica.* **9**, 160-172 (2021)

[18] Kannimuthu, S. & Chakravarthy, D. Discovery of interesting itemsets for web service composition using hybrid genetic algorithm. *Neural Processing Letters.* **54**, 3913-3939 (2022)

[19] Xavier, A. & Kesavan, R. Hybrid Elliptic Curve Cryptographic Approach for Data Privacy and Authentication in Secured Map Reduce Layer (SMR) for Optimized CPU Utilization. *IETE Journal Of Research.* **69**, 670-683 (2023)

[20] Tariq Jamal, A., Ben Ishak, A. & Abdel-Khalek, S. Tumor edge detection in mammography images using quantum and machine learning approaches. *Neural Computing And Applications.* **33**, 7773-7784 (2021)

Edited by: Zhenling Liu
Special issue on: Cloud Computing for Intelligent Traffic Management and Control
Received: June 15, 2023
Accepted: Sep 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 769–782, DOI 10.12694/scpe.v24i4.2352

INTELLIGENT VEHICLE INSPECTION TOOL DESIGN BASED ON FREEMAN CHAIN CODE FOR AUTOMATIC ANNOTATION OF 3D MODELS

SHAN CHENG* AND GUOQIANG TAN†

Abstract. Autonomous vehicle are more and more widely used in daily life, and the requirements for their safety performance are higher and higher. As a tool for testing auto parts, intelligent inspection tools are crucial to the guarantee of automobile quality. However, traditional fixture design relies on manual drawing, which is inefficient and prone to errors. To solve this problem, this research uses Freeman chain code to determine the annotation object, uses case clustering method to annotate, and uses error back propagation algorithm to realize case knowledge classification learning, and designs intelligent vehicle inspection tool design technology based on Freeman chain code 3D automatic annotation method. The experimental results show that the geometric feature matching results are correct, and the difference in feature comparison results is significant, with a high accuracy rate. Meanwhile, the geometric similarity annotation method has a high accuracy rate, taking only 3 minutes to complete the annotation, which is 7 minutes longer than traditional manual annotation. The error backpropagation algorithm can accurately achieve feature classification, and the design time of size chain inspection tool deformation design is reduced by 214min compared to manual reverse deformation design, significantly improving design efficiency. In summary, the proposed design method for automotive inspection tools can achieve automatic model annotation, improve design efficiency, and reduce design time.

Key words: Automatic Annotation; Freeman Chain Code; Bp Neural Network; Inspection Fixture Structure;

1. Introduction. In recent years, with the boost of intelligent vehicles, their safety and reliability have become hot topics of concern. The continuous development of the automotive industry has led to increasing attention being paid to the design of automotive inspection tools. Automotive inspection tools are tools used to detect the size, shape, and quality of automotive components, playing a crucial role in the quality and production efficiency of automobiles. The use of high-quality automotive inspection tools can greatly improve production efficiency, reduce product defect rates and production costs [4]. The design of intelligent vehicle inspection tools is an essential method for enhancing the efficiency of automobile manufacturing. Therefore, this study uses Freeman chain codes to automatically annotate lightweight models, and combines the Model based Definition (MBD) model and Back Propagation (BP) algorithm to achieve similarity retrieval. Therefore, it is redesigned based on existing inspection tool cases. Propose an intelligent vehicle inspection tool design scheme based on Freeman chain code for automatic annotation of 3D models (3DM). The innovation points of the research mainly include two points: firstly, using Freeman chain codes for searching and matching geometric features (GF) for determining annotation objects, and proposing an automatic annotation method for geometric similarity; The second is for utilizing BP neural network (NN) algorithm for learning and classifying MBD model knowledge, and propose a similarity retrieval method based on MBD model. The research structure is separated into four. The first is a review of related outcomes; The second proposes a design technology for intelligent vehicle inspection tools based on geometric similarity features and BP NN algorithm, which automatically annotates 3DM using Freeman chain codes; The third is the validation of the design scheme presented by the research institute; The final is a relevant summary.

2. Related work. As the boost of natural language processing technology, automatic annotation technology has been widely used in text classification, entity recognition, automobile manufacturing and other fields. Chen et al. proposed an automatic data labeling pipeline for 3D LiDAR data for solvingthe problem of segmenting moving objects in the environment. The experiment illustrates that this method can markedly label LiDAR data and generate labels in different outdoor environments [5]. Mahajan V et al. proposed a machine learning

*Institute of Mold Technology, Changzhou Vocational Institute of Mechatronic Technology, Changzhou, 213164, China (shhlkj88@163.com)
†Jiangsu Zhirui Bocheng Energy Technology Co., LTD, Changzhou, 213003, China (tanguoqiang19@126.com)

model based on automatic labeling and deep learning to predict lane change maneuvers in order to solve the problem that maneuver prediction needs to deal with large labeled datasets. The classification results show that real-time prediction of lane changes can be predicted efficiently, with an average The detection time is at least 3s, and the proportion of false positives is very small [6]. Liu et al. proposed an automatic annotation method (AAM) of hybrid atlas forest model in view of spatial index to meet the problem of accurately registering all atlases to the target image. The outcomes illustrate that this method diminish the dependence on precise registration and improves annotation [7]. Elhousni M et al. proposed a method that can automatically label high-definition maps from raw sensor data to address the issue of errors in creating high-definition maps. The outcomes indicate that the presented method can generate high-precision high-definition maps, accelerating the process of constructing and labeling high-definition maps [4]. Guerra et al. proposed a speech corpus automatic annotation scheme to monitor the dynamic changes of Mel frequency related cepstrum vectors that make up book codes. The results indicate that the correct labeling percentage of this scheme is 97.9%, and the time taken is significantly less than that of manual labeling [9].

With the arrival of the big data era, classification and retrieval technology has been widely applied as a method of information classification and retrieval. Cheng Q et al. presented a rapid design method for process equipment in view of 3DM MBD classification retrieval for changing the relevant method for enhancing design efficiency. This test shows that this method shortens the development cycle of the device and can help users produce 3DM of complex products [3]. Ebadi N and others proposed a practical two-stage attitude detection model to address the issue of lack of compatibility in the fact verification process due to human modification of classification models. The outcomes indicate that the average weighted accuracy is 82.1, which can accurately distinguish between false news and real news headlines [11]. Rosewelt et al. proposed a relevant data retrieval model based on semantic analysis to overcome the unreasonable accuracy of existing models. The outcomes indicated that the data retrieval process of the model was effective, and the identified dataset showed good test results [12]. Rashid A M et al. proposed a scheme for retrieving images in smart cities using grayscale co-occurrence matrices to reduce search time for image content. The experimental results showed an average accuracy of 6.6 and an average recall rate of 3 [13]. Alrahhal M et al. proposed a COVID-19 diagnostic system using medical image classification and retrieval in order to provide a detection method that mainly relies on artificial intelligence and radiographic image analysis to determine disease infection. The experimental results proved the effectiveness of the proposed system , with 100% accuracy in classifying input images as X-ray or CT scans, 99.18% accuracy in classifying X-ray images as COVID-99 or NOTCOVID-18, and 99.18% accuracy in classifying CT scans as COVID-97 or NOTCOVID-84 It is 97.84% [14].

In summary, a large number of scholars all over the world have conducted study about the application of automatic annotation technology and classification retrieval technology from multiple aspects at present. However, there has not been in-depth research on the design of automatic annotation of 3DM for automotive inspection tools. Therefore, the study adopts Freeman chain code to achieve automatic annotation of lightweight models, and combines MBD model and BP NN algorithm to achieve similarity retrieval. A smart car inspection tool design scheme based on Freeman chain code for 3DM AAM is proposed.

3. Design of Intelligent Automobile Gauge Based on Automatic Labeling of 3D Model. With the boost of the automotive market, the current design method of automotive inspection tools seriously restricts the production efficiency of automobiles. In order to solve the problems of low efficiency and error-prone manual drawing in traditional gage design, this study proposes an intelligent automotive inspection tool design technology based on the Freeman chain code based 3DM AAM. This technology first determines the annotation object using GF through Freeman chain codes and the longest Gongzi sequence, and then checks and optimizes the annotation results through clustering annotation. Then, BP NN algorithm is used to classify information like geometric dimensions, features, and processes, optimizing the retrieval and application of similar cases.

3.1. Intelligent Design and System Framework of Automobile Inspection Tools. The boost of the automotive manufacturing industry not only requires continuous improvement in product quality, structural design, and processing methods, but also puts forward stricter requirements for the design of inspection fixture structures [15]. At present, traditional inspection tools are difficult to meet the requirements of rapid automobile manufacturing, which has a serious impact on its production efficiency. Automobile inspection tools are responsible for detecting whether automotive parts are qualified and ensuring the quality of parts,

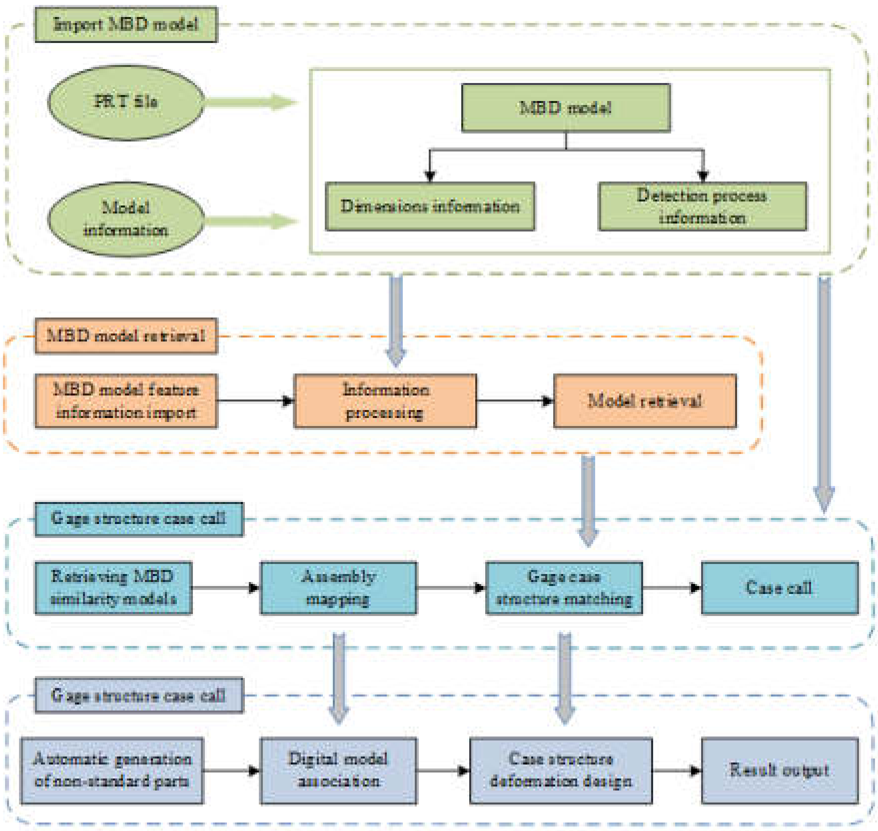


Fig. 3.1: Inspection tool design system process

and occupy an important position in the automotive manufacturing industry [16]. Automotive inspection tools possess an essential influence on the automotive production. During the product stage, the structure and size of automotive parts can be optimized. During the product validation stage, the design results of the parts can be verified in a real assembly environment. During the planned production stage, the reason why the quality of automotive parts cannot meet the assembly requirements can be determined. The specific process of designing a car inspection system is shown in Fig 3.1.

Fig 3.1 demonstrates that the checking tool system is mainly composed of four modules: importing model parts, parts retrieval, Anli model calling and checking tool case deformation design. the study first utilizes 3D annotation technology to automatically annotate lightweight models and generate MBD models. MBD models are a method of defining individual parts and product assemblies using 3D models (such as solid models), product and manufacturing information, and associated metadata. Then, the part feature information in the MBD model is extracted and transformed into vector form for similarity retrieval. Next, the mapping relation in the MBD model and the fixture is linked, and the fixture case structure model is used for operation. Finally, the design results are exported and saved through deformation design. To address the issues of long cycle time and low efficiency in designing new products, this study further utilizes a searcher with similarity retrieval function to provide engineers with similar design cases for reference. However, the case model requires modular replacement and modification to meet the requirements of new product inspection fixture design. Therefore, the above system has been improved to an intelligent design system for automotive inspection fixtures in view of the MBD model. The system mainly includes four parts: interface layer, functional layer, data layer, and support layer. The interface layer is the operation interface for the designer to design the gage. The function

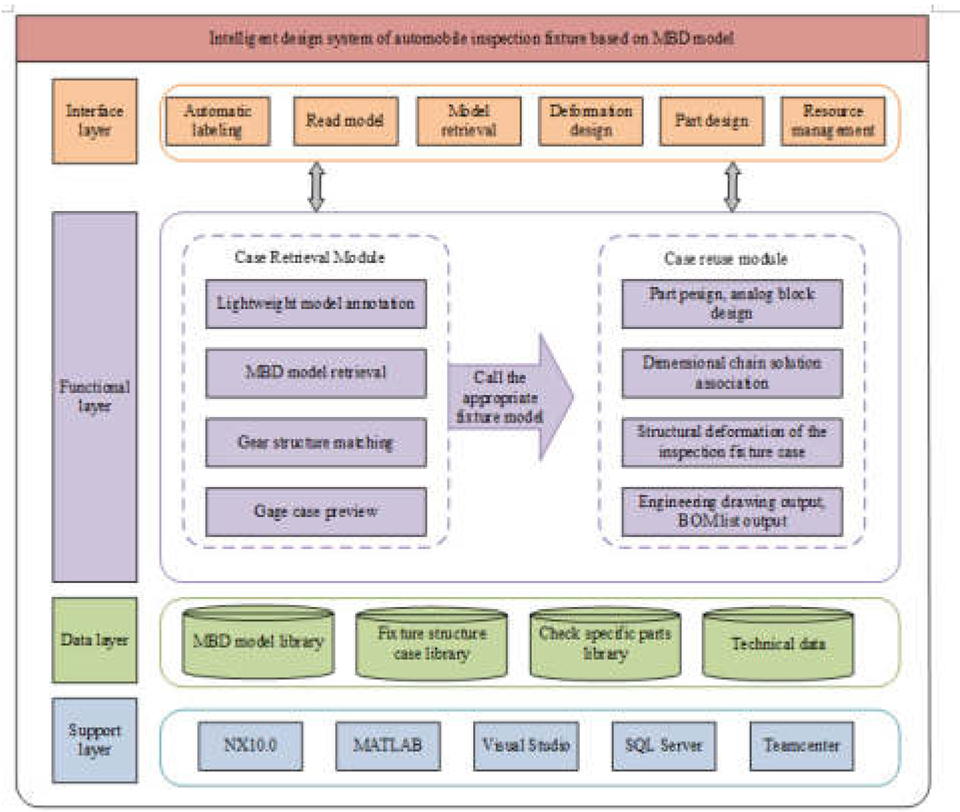


Fig. 3.2: System framework

layer is the place for data interaction between gage model retrieval and case reuse. The data layer is When using the system to develop inspection tools, the design resources are managed, and the support layer is to provide technical support for the system. The specific system framework is shown in Fig 3.2.

Fig 3.2 indicates that the system adopts a forward design method, using C as the system design language, and integrates the Microsoft Visual Studio2021 platform into MATLAB for secondary development of the NX design platform. Meanwhile, it utilizes the SQL Server platform to store and read/write the data case library, thereby creating a rapid design system for inspection tools based on MBD model similarity retrieval, and completing the construction of the inspection tool design platform. The inspection tool design platform mainly includes four functional interfaces: automatic annotation, retrieval, deformation design, and design result output. Among them, the automatic labeling function is to realize the automatic labeling of the entire MBD model, and the MBD model retrieval function is to realize the matching of similar cases. Provide design support for calling the configuration link. The deformation design function is to modify the searched similar cases to meet the new inspection fixture structure design, and the design result storage and output function is to output and save the obtained inspection specific engineering drawings. Automatic annotation is achieved using Freeman chain codes and Longest Common Subsequence (LCS) similar GF retrieval matching methods [17]. The MBD model retrieval function includes three steps: reading MBD model data, case description and calling the fixture structure corresponding to similar MBD models. The deformation design function uses the bill of materials (BOM) assembly structure for case deformation correction, and the design result output function uses the Product Data Management (PDM) system to output the generated engineering drawings and save the design results. The inspection fixture design platform integrates various peripheral interfaces such as serial port, CAN bus interface, AD/DA interface, SD/MMC card reader and DEBUG debugging interface. Function processing includes serial port initialization, Modbus frame analysis, Modbus frame structure and various services Processing and other functions. In summary, the construction of the inspection tool design system has been completed.

3.2. Design of Automatic Annotation Method for 3DM Based on Freeman Chain Codes. In the process of product design for inspection tools, manual image recognition takes a long time and the accuracy is not high. Therefore, for enhancing the staff image reading and reducing product design time, this study presented a 3DM AAM in view of Freeman chain codes [18]. The research uses Freeman chain code to describe the contour information of the geometric image, and uses LSC to retrieve and match the geometric features to determine the labeling object, omitting the step of manually identifying the geometric features of the image to determine the labeling object. Freeman chain code is a method of describing a curve or a boundary by using the coordinates of the starting point of the curve and the direction code of the boundary point. It is often used to represent curves and area boundaries in the fields of image processing, computer graphics, and pattern recognition. Freeman chain codes are used to describe the contour lines of objects or shapes and their geometric features. By assigning a unique number to each point on the contour, a sequence composed of numbers can be generated, which can be expressed as a sequence with different rotations. The only form of transgender. After obtaining the geometric information sequence, search the longest child sequence of the shape to judge the similarity of the combined features, so as to realize the matching of geometric shapes. The geometric shape matching problem is realized by searching LCS of two sequences. The LCS problem is solved by dynamic programming. The relevant function is as follows in equation (3.1).

$$C\left[i,j\right]=\begin{cases} 0,\ \text{when } i=0 \ or\ j=0 \\ C[i-1,j-1]+1 \ \ \text{when } i,j>0, x_i=y_i \\ \max C[i,j-1], C[i-1,j],\ \text{when } i,j>0, x_i\neq y_j \end{cases} \tag{3.1}$$

In equation 3.1, $C\left[i,j\right]$ serves as the length of LCS; i,j are the serial number;x_i, y_j represent two different sequences of elements. According to the recursive formula, if the element corresponding to the serial number i is equal to the element corresponding to the serial number j , the value of the serial number cell corresponding to the two is written as $c_{i-1}+1$, if the element A is not equal to the element B , the maximum value of $C_{i-1,j}$ and $C_{i,j-1}$ is taken, and so on, and finally LSC is obtained. After GF matching, the annotated object is calculated through iterative comparison of Freeman chain codes, further detecting the text information foundation that needs annotating in the 3DM. The MBD model inspection information includes four points: firstly, product information, such as product attributes, quality, materials, usage, and other basic external features of the design model; The second is basic information, which is separated into two: GF information and appearance feature information; The GF information includes the length, width, thickness, aperture, positioning distance, etc. of the model, while the external feature information includes the shape feature identification, feature category, area, roughness, function, etc. of the model surface; The third is to detect process information, which is used to reflect the product quality level, including design requirements, material information, model geometry information, geometric tolerance information, detection method, positioning information, model roughness, etc; The fourth is to annotate and represent information, usually using different colors to label different detection information to reflect the category of information. The basic information annotation using the MBD model of the front bumper of a car as an example is shown in Fig 3.3.

As shown in Fig 3.3, the specific labeling steps of the automatic labeling method based on Freeman chain codes are to first obtain the Freeman chain codes of the original image and the comparison image, calculate the labeling objects after iterative comparison, and then detect the labeling information in the original image as a comparison The text information basis of the graph annotation, and then use the structural constraints of the geometric feature annotation to map it to the new model to complete the initial automatic annotation. For further enhancing the completeness of annotation, a standard sample dataset was obtained based on the prototype clustering algorithm to inspect and modify the fat bamboo joint tubes, and the annotation results were output. After the annotation is completed, the product is classified. The multi-dimensional nature of product information during classification limits algorithm calculation and inference time. Therefore, Kernel Principal Component Analysis (KPCA) is introduced for dimensionality reduction, and the spatial distribution is shown in Fig 3.4.

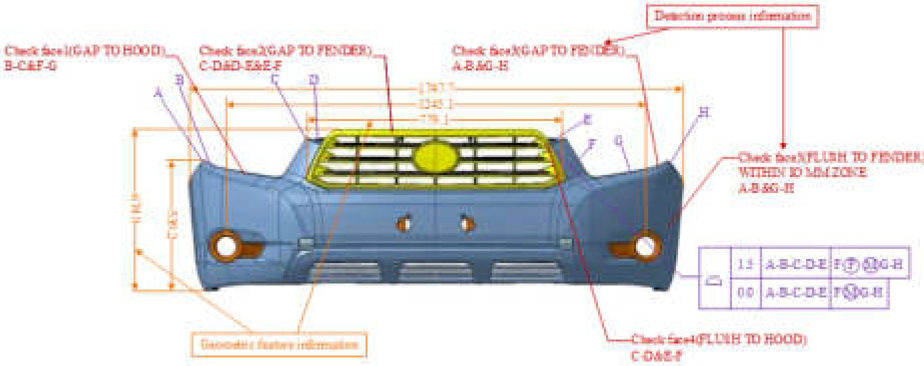


Fig. 3.3: MBD model of the main model of the car front bumper

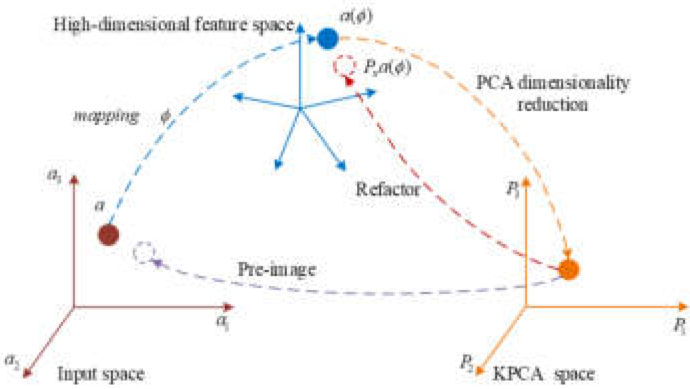


Fig. 3.4: KPCA spatial distribution map

Fig 3.4 indicates that the covariance matrix of the model data $W = W_1, W_2, ...W_d$ is first projected into the determined hyperplane, the sample point of the model generates the image in high-dimensional spatial features through mapping , and the calculation method is shown in equation 3.2.

$$z_i = \phi(a_i), i = 1, 2, ..., m \tag{3.2}$$

Due to the unclear form of mapping , a kernel function is introduced as shown in equation 3.3.

$$k(a_i, a_j) = \phi(a_i)\phi(a_j) \tag{3.3}$$

The eigenvalues and eigenvectors of the covariance matrix are obtained through the kernel function, and the eigenvalues are sorted to obtain the eigenvectors from large to large. By using the PCA algorithm to achieve dimensionality reduction, the $r(r = 1, 2, ..., d')$ -th coordinate of the new sample a after projection can be obtained as shown in equation 3.4.

$$z_r = \sum_{i_1}^{m} h_i^r k(a1, a) \tag{3.4}$$

In equation 3.4,h_i^r is the r-th component of h_i ;k is the kernel function. In summary, the design of a 3DM AAM based on Freeman chain codes and the dimensionality reduction processing of product information have been completed.

3.3. Design of 3D Model Retrieval Method Based on BP NN Algorithm. At present, there are still problems of low efficiency and low accuracy in case retrieval of automotive enterprises, and MBD model retrieval is a key technology to solve this problem. The study utilizes the BP NN algorithm to cluster, classify, and retrieve existing cases for learning, to predict similar MBD models and examine specific structures [19]. BP neural network is a multi-layer feed-forward network trained by error backpropagation, which uses gradient search technology to minimize the mean square error between the actual output value of the network and the expected output value. The clustering of product datasets utilizes the Fuzzy c-means algorithm (FCM) for maximizing the similarity in features of the same cluster and minimize the similarity between features of various clusters, thus classifying samples. The specific principle is to divide the dataset into classes, which correspond to the center points of D classes. If each sample belongs to a certain class$l$, it is $u_l b$, then the FCM objective function (OF) is depicted in equation 3.5.

$$J = \sum_{l=1}^{N} \sum_{b=1}^{c} u_{lb}^{t} \left\| m_1 - c_b \right\|^2 , 1 \leq t < \infty \tag{3.5}$$

In equation 3.5,t is the membership factor;N is the quantity of samples;C is the quantity of cluster centers (CC); C_b is the l-th CC, with the same dimension as the sample feature; m_1 represents the l-th sample; represents the membership degree of sample to CC C; $\|*\|$ represents a measure of the similarity (distance) of any data, the most common being the Euclidean distance (ED). The constraint conditions are defined to the OF using the Lagrange multiplier method, where is taken as the derivative of U_{lb} and its structure is equal to 0, as shown in equation 3.6.

$$\frac{\delta J}{\delta u_{lb}} = t \left\| m_1 - C_b \right\| u_{lb}^{t-1} + \lambda_b = 0 \tag{3.6}$$

In equation 3.6,λ is the Lagrange multiplier value. Take the derivative of J over d_i and make the result 0, as shown in equation 3.7.

$$\frac{\delta J}{\delta C_b} = \sum_{l=1}^{n} (-u_{lb}^{t} * 2 * (m_1 - c_b)) = 0 \tag{3.7}$$

Finally, the membership matrix is obtained as shown in equation 3.8.

$$\sum_{l=1}^{C} u_{lb} = 1, \forall\ b = 1, ..., n \tag{3.8}$$

The membership matrix representation serves as the degree to which each sample point belongs to each class. For a single sample M_1 , the sum of its membership degrees for each cluster is 1. The closer it is to 1, the higher the membership degree is, and vice versa. The clustering center calculation method is shown in equation 3.9.

$$c_b = \frac{\sum_{l=1}^{N} u_{lb} \bullet m_l}{\sum_{l=1}^{N} u_{lb}^{t}} \tag{3.9}$$

It calculates the CCc_b of each group to minimize the OF (as the OF is relevant to ED, the OF reaches its minimum, the ED is shortest, and the similarity is highest). This makes sure the clustering principle of highest intra group similarity (GS) and lowest inter GS. After the product classification is completed, research is conducted to use the BP network neural algorithm for retrieval learning. The BP network neural algorithm is a multi-layer feedforward NN trained according to the error BP algorithm. By learning the transfer rules, the best output value can be obtained when the input value is [20]. Train with car example (p_e, q_e) , and the mean square error of the case can be obtained from the BP NN as demonstrated in equation 3.10.

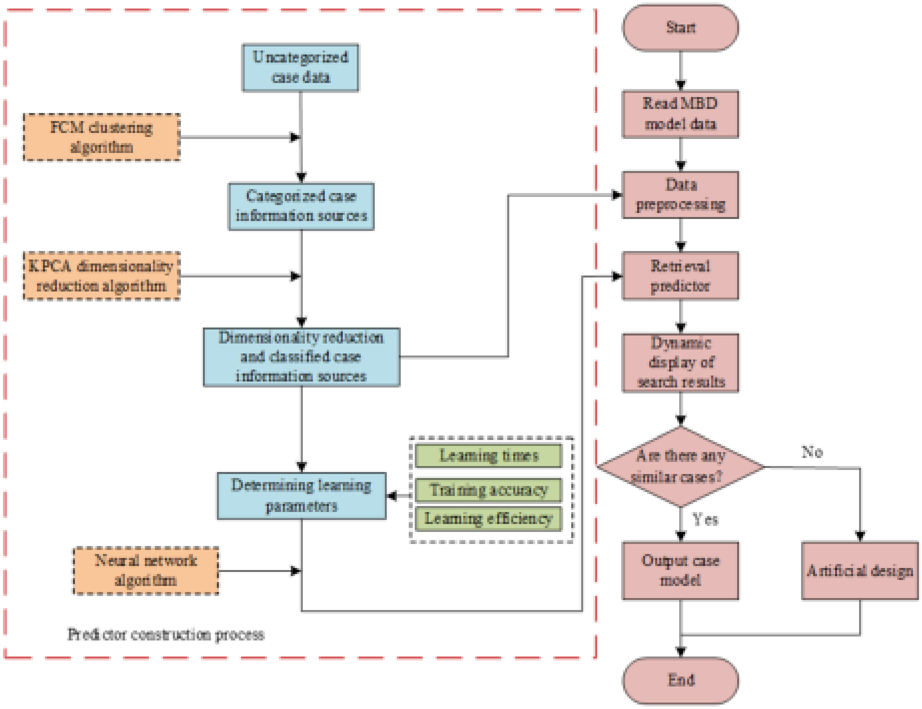


Fig. 3.5: Retrieval process based on BP neural network algorithm

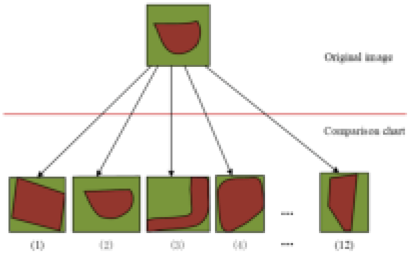
$$E_e = \frac{1}{2}\sum_{e=1}^{s}(\widehat{q}_j^e - q_j^e)^2 \tag{3.10}$$

In equation 3.10,s is the output vector. Then, in the BP algorithm, the gradient descent function is utilized for adjusting from the negative gradient direction, and equation 3.11 can be obtained.

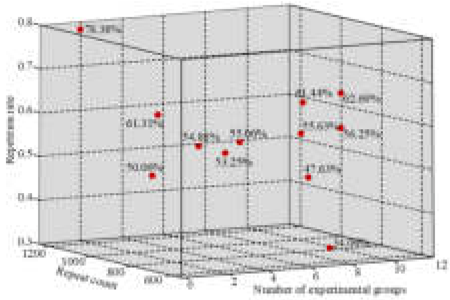
$$\Delta w_{gi} = -\eta\frac{\delta E_k}{\delta w_{gi}} \tag{3.11}$$

In equation 3.11,w_{gi} represents the connection weight in the g-th neuron in the hidden layer and the j-th neuron in the output layer;η is the learning rate. The specific process of the NN algorithm is shown in Fig 3.5.

Fig. 3.5 indicates that the study first extracts MBD model data from the database and preprocesses it using FCM clustering algorithm and KPCA dimensionality reduction algorithm. After determining parameters such as learning frequency, training accuracy, and learning efficiency, the BP NN is utilized for retrieving and predicting the data. Finally, it determines whether the obtained fixture case is similar to the part and outputs the search results. The MBD model retrieval information mainly includes three steps: traversing the MBD model, obtaining model features, and obtaining model main parameters. Due to the wide variety of car models and components, the retrieved cases are prone to local similarity. Therefore, adding a universal and deformable strategy to improve the universality of the inspection tool. Usually, deformation schemes such as settling size chains, replacing skeleton nodes, and replacing skeleton structures are used to design inspection tools. Taking the previous structure of the inspection fixture as an example, the design process of the deformation of the inspection fixture was separated into four steps. The first was to extract the MBD model information and divide the modules, and use assembly constraints to associate each module with the size chain. Finally, it used the main parameters of the parts and assembly to create parameter expressions; The second step is to



(a) Tested geometric model



(b) Similarity check results

Fig. 4.2: Comparison model and result

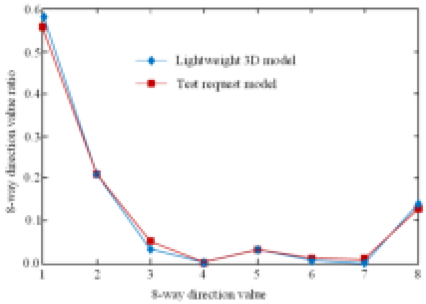
compare the size difference between the retrieved MBD model and the case model, and use the size difference to modify the case fixture structure while changing the main parameters of the simulation block in the chain; The third step is to obtain the main parameters and use the size as the input value to drive the parameter changes accompanying the size, to solve the main parameters of the part; The fourth step is to associate the specific main parameters after solving the size chain parameters, so as to deform the model structure and obtain a new specific structure, while saving the parameters and adding new cases. In summary, the 3DM retrieval method in view of BP NN algorithm and the deformation design of the specific structure for pre inspection were completed.

4. Performance Analysis of Automotive Inspection Tools Based on Freeman Chain Code 3D Model Automatic Annotation Method. As a key tool for inspecting automotive parts, automotive inspection tools possess an essential influence on the automotive production process. For verifying the effectiveness and feasibility of the intelligent vehicle inspection tool design based on the Freeman chain code three-dimensional AAM proposed in the research, this section focuses on testing the similarity of GF and comparing the time consumption of several common annotation methods. Then it further tested the distribution of 7 types of models in 2D and 3D spaces, as well as the structural dimensions of the product after deformation.

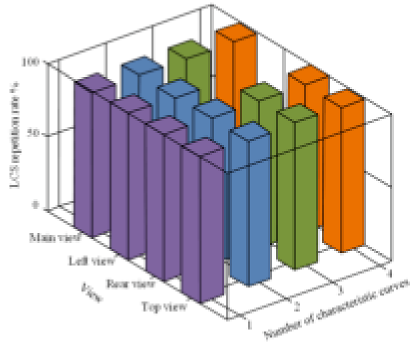
4.1. Performance Analysis of 3DM Automatic Annotation Method Based on Freeman Chain Code. For verifying the effectiveness of the AAM for 3DM in view of Freeman chain codes, 12 sets of GF were input. This is to test the maximum number of sub sequences and repetition rate between the GF of the detection task book and the GF of the 3DM.

Fig 4.1a shows 12 sets of GF tested; Fig 4.1b shows the similarity test results. The test indicates that the accuracy of using LCS to search for GF matching of each group of data is relatively high, and there is a significant difference in feature comparison results. This proves that using LCS to solve geometric shape matching problems is practical and feasible. To explore the feasibility of automatic annotation of 3DM based on Freeman chain codes, research was conducted on automatic annotation of 3DM of car front bumpers, the 8-channel eigenvalue of the geometric model is selected as the evaluation index, and the 8-channel eigenvalue refers to the similarity of the 8 prominent features of the geometric model.

Fig 4.3a shows the comparison chart of the eight channel values of the two geometric modeling. The results show that the eight channel values of the two geometric modeling are approximately equal, indicating that the two geometric characteristics are similar. Fig 4.3b illustrates the LCS comparison of the characteristic curves of various views of the front bumper parts. The LCS size of the two geometric modeling shows that the geometric characteristics of the two models are similar, and the LCS repetition rate shows that the lightweight model can fully label the selected objects, which further proves that the method can complete the automatic annotation of the model. The study selected 8 surfaces and used geometric similarity annotation, traditional manual annotation, 3D cube annotation, 3D radar point cloud annotation, and 3D automatic annotation to



(a) Comparison chart of the proportion of 8-way direction value of two comparison model



(b) The LCS comparison results of the characteristics curves of each view of the bumper parts

Fig. 4.4: The comparison results of the 8-pass value ratio and the characteristic curve LCS of the two comparison models

Table 4.1: Completeness and time-consuming comparison of several model annotations

Marking method	Color	Object	Structure	Benchmark	Surface	Time consuming/min
Geometric similarity annotation	87.89%	100%	100%	100%	62.5%	3
Traditional manual labeling	67.83%	75.25%	83.63%	89.86%	50.25%	10
3D cube annotation	85.79%	83.81%	86.05%	80.74%	51.35%	8
3D radar point cloud labeling	78.36%	81.82%	84.23%	85.87%	56.48%	6
3D automatic annotation	77.34%	79.97%	83.56%	92.57%	52.39%	5

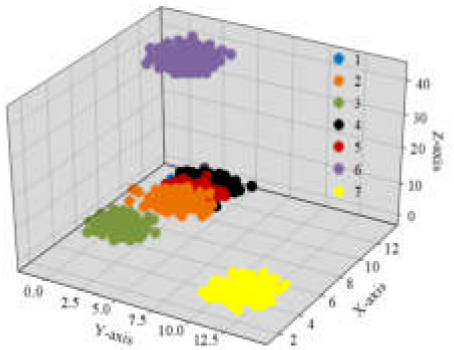
detect the completeness and time-consuming of corresponding surface annotations.

Table 4.1 shows the completeness and time consumption comparison of eight surface annotations by several models. Compared to traditional manual annotation and other annotation methods, the AAM in view of GF similarity achieves 100% accuracy in annotation structure, objects, and benchmarks, and only takes 3 minutes to complete the annotation, which is 7 minutes longer than traditional manual annotation. The results show that the geometric similarity annotation method can greatly reduce annotation time, improve the accuracy of annotation results, address the issue of easy annotation errors, and meet the detection requirements.

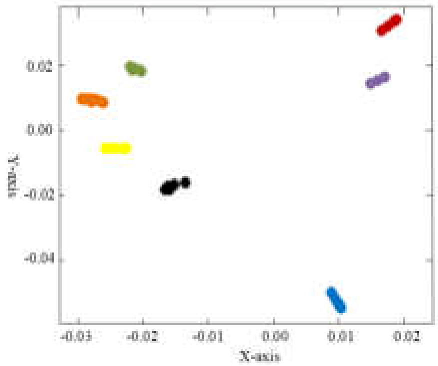
4.2. Performance Analysis of 3D Model Retrieval Technology Based on BP NN Algorithm. To verify the feasibility of 3DM retrieval technology based on BP NN algorithm, the study divided car models into 7 categories, utilized KPCA for dimensionality reduction, and tested the distribution of the seven types of models in 3D and 2D spaces. It is showcased in Fig 4.6.

Fig 4.5a shows the three-dimensional spatial distribution of relevant model data; Fig 4.5b shows the two-dimensional distribution of relevant model data after dimensionality reduction using KPCA. Most of the 7 features are not significantly separated in 3D space, and after using KPCA dimensionality reduction processing, the 7 features are completely separated and clearly distinguished. This proves that KPCA can effectively distinguish GF, accelerate the speed of GF classification, and reduce model retrieval time. This study further utilizes the FCM clustering algorithm to process seven types of model data and test the data classification performance.

Fig 4.7 indicates the clustering of the FCM algorithm. The clustering illustrates that the FCM algorithm can effectively serve as the clustering centers and ranges of each type of GF, accurately achieve feature classification results, and reduce the complexity of retrieval. For verifying the feasibility of retrieving deformation design in



(a) Comparison chart of the proportion of 8-way direction value of two comparison model



(b) The LCS comparison results of the characteristics curves of each view of the bumper parts

Fig. 4.6: The comparison results of the 8-pass value ratio and the characteristic curve LCS of the two comparison models

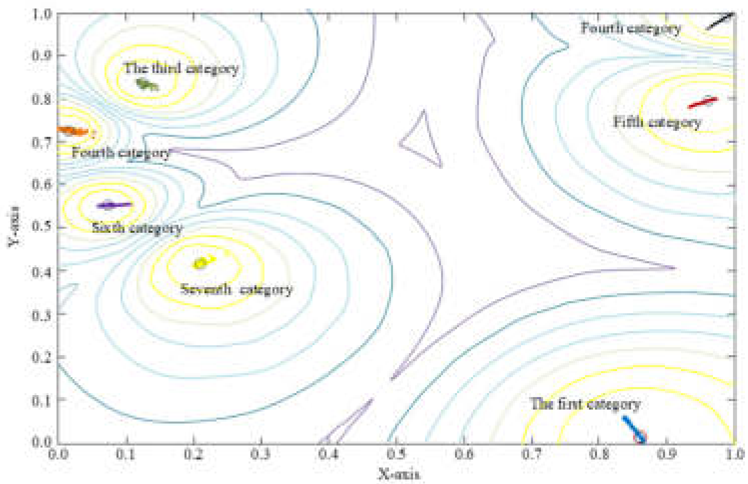
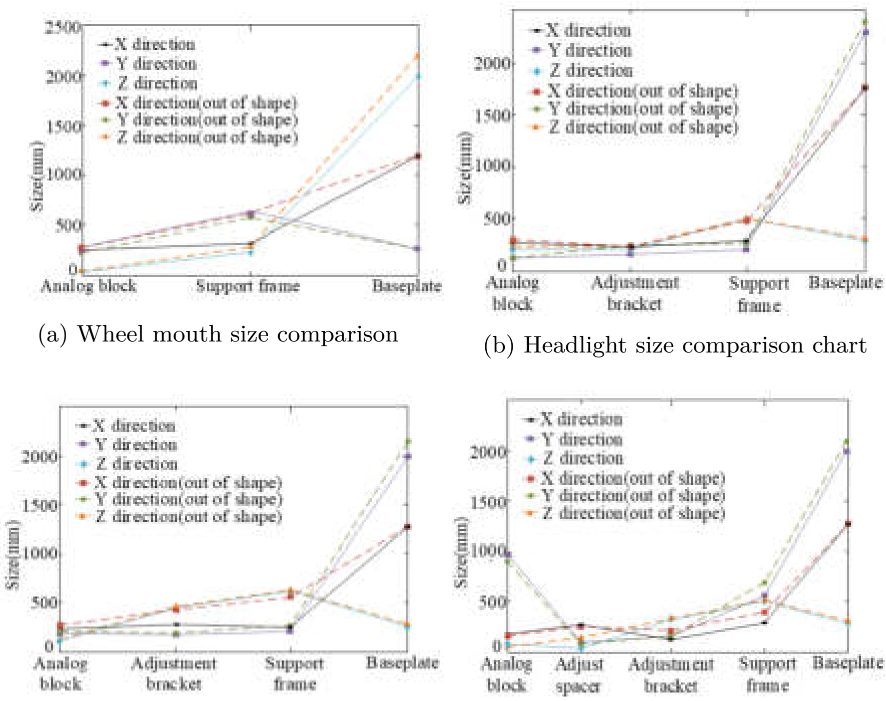


Fig. 4.7: FCM algorithm clustering results

automotive inspection fixture design, this study takes the front bumper model of an automobile as a case to perform similarity retrieval and adjust each dimension chain based on existing cases. The deformation design is completed by modifying the main parameters, and the structural dimensions of the product after deformation are tested.

Figs 4.8a to 4.8d show a comparison of the dimensions of the wheel opening module, headlight module, fender module, and hood module, respectively. The comparison results show that the deformation design solutions of the four modules do not conflict in size and have a high utilization rate, without any abnormalities, and can successfully complete structural deformation. It further verifies the efficiency of solving the deformation of dimension chain design gauges, and records the design time required for wheel openings, headlights, fenders, and hood under the conditions of manual model gauges, high similarity model gauges, medium similarity model



(a) Wheel mouth size comparison

(b) Headlight size comparison chart

(c) Fender size comparison chart

(d) Hood size comparison chart

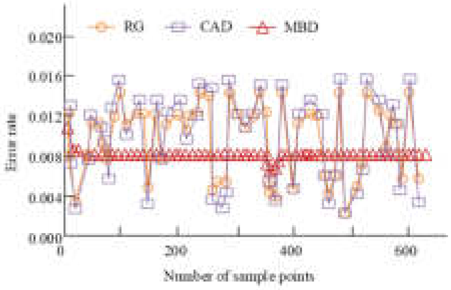
Fig. 4.9: Three-dimensional space and two-dimensional space distribution diagram of seven types of model data

Table 4.2: Design time comparison chart

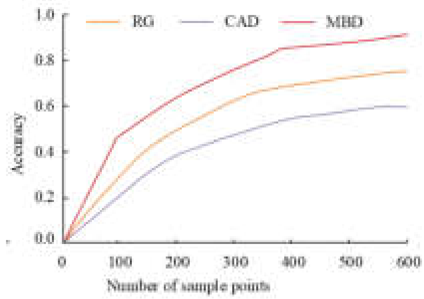
Marking method	Color	Object	Structure	Benchmark	Surface	Time consuming/min
Geometric similarity annotation	87.89%	100%	100%	100%	62.5%	3
Traditional manual labeling	67.83%	75.25%	83.63%	89.86%	50.25%	10
3D cube annotation	85.79%	83.81%	86.05%	80.74%	51.35%	8
3D radar point cloud labeling	78.36%	81.82%	84.23%	85.87%	56.48%	6
3D automatic annotation	77.34%	79.97%	83.56%	92.57%	52.39%	5

gauges, and low similarity model gauges.

Table 4.2 shows the required design time for wheel openings, headlights, fenders, and hood. The outcomes showed that relative to manual reverse deformation design, the design time of model inspection tools with higher similarity in size chain deformation design was reduced by 214 minutes; And the design time of each module is significantly reduced, and the design efficiency is significantly improved. In summary, the intelligent vehicle inspection tool proposed by the research institute based on the Freeman chain code for 3DM AAM significantly improves the retrieval accuracy and greatly reduces the retrieval process by using KCPA dimensionality reduction, FCM algorithm classification, and integrating BP NN algorithm for retrieval. Its successful application and design optimization of similar cases. In order to verify the effectiveness of the automobile inspection tools proposed in the research, the study uses the proposed automatic marking method of 3D models based on Freeman chain codes for intelligent automobile inspection tools, and the automobile inspection tools based on computer aided drafting (CAD) platform. Automobile gauge based on relational graph (RG) is tested on the same vehicle model, and the detection error rate and correct rate are selected as performance evaluation



(a) The detection error rate of the three kinds gages



(b) Accuracy

Fig. 4.11: Detection error rate and precision rate of three kinds of gages

indicators.

Figure 4.10a and Figure 4.10b respectively show the error rate and accuracy rate of the three kinds of inspection tools for the detection of car models. The results show that the inspection tool of the research institute performs better in terms of error rate and accuracy rate, the detection error is the smallest and the fluctuation range is not large, which has good advantages. The highest accuracy rate of the proposed gage is 93.56%, the accuracy rate of RG is 91.34%, and the average accuracy rate of CAD is 90.22%. To sum up, in the process of gage design, the automatic labeling method based on Freeman chain code is used to mark the case structure and information, the BP neural network algorithm is used to classify the case information, and the FCM and KPCA algorithms are used to optimize the information source. The proposed intelligent checking fixture can improve the retrieval efficiency of similar cases, shorten the development cycle of new products, improve the design efficiency of new products, and lay a foundation for the intelligent design of checking fixtures.

5. Conclusion. Automobile inspection fixture is an important tool for automobile research and development, which affects the automobile research and development process and the manufacturing accuracy of the whole vehicle. For enhancing the design of automotive inspection tools, this study presented an intelligent automotive inspection tool design method based on Freeman chain code for 3DM automatic annotation. The results showed that the GF matching accuracy of the 12 sets of data searched using Freeman chain codes and LCS was high, and the difference in feature comparison results was significant. Compared with other annotation methods, the geometric similarity annotation method has a 100% accuracy in annotating structures, objects, and benchmarks, and only takes 3 minutes to complete the annotation, which is 7 minutes longer than traditional manual annotation. The test results demonstrate that the GF are completely separated and clearly distinguished after using KPCA dimensionality reduction processing, and the FCM algorithm can accurately achieve feature classification. For the deformation design of wheel openings, headlights, fenders, and engine hood, the dimensions of the dimension chain inspection tool are not conflicting and have a high utilization rate, without any abnormalities; The required design time was reduced by 214 minutes compared to the manual reverse deformation design, and the design efficiency was significantly improved. Compared with RG-based gages and CAD-based gages, the error rate of the gages proposed by the research institute is the smallest, and the highest accuracy rate is 93.56%. To sum up, the intelligent automobile gage proposed by the research institute can improve the efficiency of automobile design, realize the intelligent design of the gage structure, and provide reference value for the intelligent development of gage design. However, there are still shortcomings in the research, with a focus on completing the intelligent design of inspection tools. Further research can improve and deepen the optimization of retrieval structure, design methods, and model retrieval information sources.

Fundings. The research is supported by: QingLan Project of Jiangsu Universities in 2022.

REFERENCES

[1] Yang, M. Research on vehicle automatic driving target perception technology based on improved MSRPN algorithm. *Journal Of Computational And Cognitive Engineering.* **1**, 147-151 (2022)
[2] Dongjie, L., Jin, Z., Axin, X., Chao, W., Xinnian, H., Kuncheng, L. & Chang, L. Data augmentation technology driven by image style transfer in self-driving car based on end-to-end learning. *Computer Modeling In Engineering & Sciences.* **122**, 593-617 (2020)
[3] Cheng, Q., Wang, S. & Fang, X. Intelligent design technology of automobile inspection tool based on 3D MBD model intelligent retrieval. *Proceedings Of The Institution Of Mechanical Engineers 235:2917-2927.* pp. 10-11 (2021)
[4] Seontaek, O., Ji-Hwan, Y., Azim, E. & Young-Keun, K. Accurate alignment inspection system for low-resolution automotive LiDAR. *IEEE Sensors Journal.* **21**, 11961-11968 (2021)
[5] Xieyuanli, C., Benedikt, M., Lucas, N., Rodrigo, M., Ignacio, V., Jens, B. & Cyrill, S. Automatic labeling to generate training data for online LiDAR-based moving object segmentation. *IEEE Robotics And Automation Letters.* **7**, 6107-6114 (2022)
[6] Mahajan, V., Katrakazas, C. & Antoniou, C. Prediction of lane-changing maneuvers with automatic labeling and deep learning. *Transportation Research Record.* **2674**, 336-347 (2020)
[7] Hong, L., Lijun, X., Enmin, S., Renchao, J. & Chih-Cheng, H. Automatic labelling of brain tissues in MR images through spatial indexes based hybrid atlas forest. *IET Image Processing.* **14**, 2728-2736 (2020)
[8] Elhousni, M., Lyu, Y., Zhang, Z. & Huang, X. and labeling of hd maps with deep learning. *Proceedings Of The AAAI Conference On Artificial Intelligence 34(8.* pp. 13255-13260 (2020)
[9] Guerra, S., S., O. & L., J. Automatic Phonetic Labeling at Word Level Using the Dynamics of Changing Codebook Vectors. *Computación Y Sistemas.* **24**, 861-874 (2020)
[10] Cheng, Q., Wang, S. & Fang, X. Intelligent design technology of automobile inspection tool based on 3D MBD model intelligent retrieval. *Proceedings Of The Institution Of Mechanical Engineers.* pp. 10-11 (2021)
[11] Ebadi, N., Jozani, M., Choo, K. & Rad, P. memory network information retrieval model for identification of news misinformation. *IEEE Transactions On Big Data.* **8**, 1358-1370 (2021)
[12] Rosewelt, A. & Renjit, A. Semantic analysis-based relevant data retrieval model using feature selection, summarization and CNN. *Soft Computing.* **24**, 16983-17000 (2020)
[13] Rashid, A. & Yassin, A. lkadhmawee, A. *A. A., Yassin, A. J. Smart City Security: Face-Based Image Retrieval Model Using Gray Level Co-Occurrence Matrix.* **19**, 437-458 (2020)
[14] Alrahhal, M. & S., K. COVID-19 diagnostic system using medical image classification and retrieval: a novel method for image analysis. *The Computer Journal.* **65**, 2146-2163 (2022)
[15] Zhang, J., Kikuta, M., Zhang, C., Ito, A., Anmi, K. & Akashi, T. Defect inspection of coated automobile roofs using a single camera. *IEEJ Transactions On Electrical And Electronic Engineering.* **15**, 616-625 (2020)
[16] Wang, E., Li, Y., Wang, Y., Wang, P., Jiao, J. & Yu, X. Vehicle key information detection algorithm based on improved SSD. *IEICE Transactions On Fundamentals Of Electronics, Communications And Computer Sciences.* **103**, 769-779 (2020)
[17] Jin, S., Lin, Q., Bie, Y., Ma, Q. & Li, Z. practical method for detecting fluff quality of fabric surface using optimal sensing. *Elektronika Ir Elektrotechnika.* **26**, 58-62 (2020)
[18] Thanikkal, J., Dubey, A. & Thomas, M. Unique shape descriptor algorithm for medicinal plant identification (SDAMPI) with abridged image database. *IEEE Sensors Journal.* **20**, 13103-13109 (2020)
[19] Masood, F., Masood, J., Zahir, H., Driss, K., Mehmood, N. & Farooq, H. Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data. *Journal Of Computational And Cognitive Engineering.* **2**, 57-67 (2023)
[20] Wei, M., Hu, X. & Yuan, H. Residual displacement estimation of the bilinear SDOF systems under the near-fault ground motions using the BP NN. *Advances In Structural Engineering.* **25**, 552-571 (2022)

Edited by: Zhenling Liu
Special issue on: Cloud Computing for Intelligent Traffic Management and Control
Received: June 15, 2023
Accepted: Aug 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 783–786, DOI 10.12694/scpe.v24i4.2863

INTRODUCTION TO THE SPECIAL ISSUE ON SENTIMENT ANALYSIS AND AFFECTIVE COMPUTING IN MULTIMEDIA DATA ON SOCIAL NETWORK

RAJNI MOHANA,[*] ANAND NAYYAR,[†] PRADEEP KUMAR,[‡] AND AMAN SHARMA[§]

In the Age of Multimedia and Social Networks, the proliferation of user-generated data has been nothing short of meteoric. This wealth of information necessitates careful analysis and processing to truly comprehend the subjective perceptions of users. At the intersection of this data-driven revolution, two critical fields emerge: Sentiment Analysis and Affective Computing. Sentiment analysis delves into the intricate realm of people's opinions, sentiments, evaluations, attitudes, and emotions as expressed through written language. On the other hand, Affective Computing focuses on the development of systems and devices capable of recognizing, interpreting, processing, and even simulating human emotions. It is, in essence, the fusion of Emotion AI and other affective technologies, with the overarching goal of enhancing people's lives.

With the exponential growth of user-generated data, thanks to social networks, wikis, and social tagging systems, it has become imperative to decipher the high-level semantics and user subjective perceptions embedded in this vast sea of information. Emotions and sentiments, in particular, stand out as significant facets of user-generated data, often carrying the emotional imprints of their creators. The concurrent advancement of computational techniques for sentiment analysis and opinion mining has been accompanied by a surge in the utilization of psychological and cognitive models and theories. These are being harnessed to model sentiments and emotions, often in synergy with social computing techniques such as social network analysis and personalization, user review mining, and user profiling within social networks, among others. The synergy between affective/sentimental models and social computing techniques is not merely an academic endeavor; it paves the way for comprehending big data at a semantic level and enhances the performance of a wide array of social computing applications in this era of big data. This convergence not only combines affective and sentimental models with social computing but also charts a promising direction replete with opportunities for developing novel algorithms, methods, and tools.

It is a privilege for us to introduce the Special Issue on Sentiment Analysis and Affective computing in Multimedia Data on Social Network. Among the numerous research papers we received (50 in total), we meticulously selected 18 papers for publication. The overarching objective of this special issue is to delve into the recent advancements and disseminate state-of-the-art research related to sentiment analysis and affective computing in multimedia data within social networks and the technologies that make this possible. This special issue represents a showcase of new dimensions of research, offering researchers and industry professionals an illuminating perspective on sentiment analysis and affective computing in the realm of multimedia data within social networks.

We sincerely hope that the contributions in this special issue will not only inform but also inspire future research endeavors, leading to a deeper understanding of the multifaceted world of sentiment analysis and affective computing in the age of multimedia and social networks.

[*]CSE, Amity School of Engineering and Technology, Amity university, Punjab, India (rajnivimalpaul@gmail.com)
[†]School of Computer Science, Duy Tan University, Da Nang, VietNam (anandnayyar@duytan.edu.vn)
[‡]Discipline of Electrical, Electronic and Computer Engineering, University of KwaZulu-Natal, South Africa (kumarp@ukzn.ac.za)
[§]CS/IT Department, Jaypee University of Information and Technology, Solan, H.P., India (aman.sharma@juitsolan.in)

Shuaizhi Shen in the paper titled “User sentiment analysis methods for elderly social media networks” delves into sentiment analysis within social media networks tailored for the elderly population. Initial stages involve the extraction of emotional sentences for the classification of movie reviews. Subsequently, user search behavior in social networks is scrutinized for elderly demographics, and their movie reviews are employed for predictive rating analysis. Three sentiment classification algorithms—Dirichlet, maximum entropy, and support vector machine—are employed, revealing accuracy rates of 87.1%, 86.9%, and 86.5%, respectively. Notably, first-level classifiers exhibit robust accuracy, ranging from 87.4% to 90.7%. The proposed method surpasses Slope One in predictive analysis, showcasing its potential to elevate emotional analysis accuracy in film review texts for the elderly.

”Deepjyoti Choudhury et al., in the paper titled “A Deep Community Detection Approach In Real Time Networks“ introduces a novel deep learning approach for real-time community detection in social networks, overcoming issues with traditional proximity matrices. By reorganizing matrices and extracting spatial attributes, the proposed method, evaluated on ten datasets, outperforms existing techniques, showcasing its effectiveness in identifying superior clusters in real-time networks.

Anureet Chhabra et al. in the paper titled “A Statistical Analysis of Tweet Sentiment on Drug Usage Across High, Middle, and Low-Income Countries” conducted a comprehensive statistical analysis of tweet sentiments related to drug usage across countries with varying income levels. This paper examines drug sentiment on Twitter, analyzing 35,337 drug-related tweets from high, middle, and low-income countries. Findings reveal varying positive sentiments (39.509%, 26.0148%, and 23.644%) across income categories, highlighting cultural and legal factors as influencers. The study underscores the importance of these insights for shaping global drug prevention policies and promoting public health education.

Purushottama Rao K and B Janet in the paper titled “Detecting Academic Affective States of Learners in Online Learning Environments Using Deep Transfer Learning” addresses the limitations of Online Learning Environments (OLEs) in recognizing students’ emotions by customizing the DAiSEE dataset and training an AffectXception model. The model outperforms transfer learned and state-of-the-art methods, achieving high accuracy (77% to 91.87%) in detecting academic affective states like Boredom, Engagement, Confusion, and Frustration. The findings suggest the potential of adapting teaching strategies based on learners’ emotional states for more effective online education.

Koyel Ghosh et al., in the paper titled “Hate Speech Detection in Low- Resource Bodo and Assamese Texts with ML-DL and Bert Models“ addresses addresses the contemporary issue of hate speech detection in natural language processing, driven by the unrestrained use of social media leading to toxic comments and posts. Recognizing the impact on violence at various levels, the study introduces two North-East Indian Hate Speech datasets (Assamese and Bodo). It trains baseline machine learning, deep learning, and state-of-the-art transformer models on these datasets, evaluating their performance and conducting detailed error analysis to contribute insights into hate speech detection challenges, particularly in low-resource languages.

Jigna Patel et al., in the paper titled “ConColla - A Smart Emotion-based Music Recommendation System for Drivers” introduces ConCollA, a hybrid music recommender system, focusing on user emotions for a more personalized experience. By incorporating facial expression recognition through a CNN model, ConCollA accurately identifies driver emotions and tailors music recommendations. The evaluation indicates superior performance compared to traditional collaborative-based recommender systems.

Souvik Sengupta and Saurabh Pal et al., in the paper titled “Mapping Learner’s Query to Learning Objects using Topic Modeling and Machine Learning Techniques“. This paper addresses the challenge of mapping learner queries to suitable Learning Objects (LOs) in e-learning by proposing a recommender system. Employing a combination of supervised and unsupervised Natural Language Processing (NLP) and Machine Learning (ML) methods, the model is trained on a handcrafted dataset to map queries to predefined topics. Additionally, dynamic topic modeling on learning content from popular e-learning portals enhances the system’s ability to recommend the most appropriate LO based on similarity scores.

Qingyuan Li et al., in the paper titled “Speech Emotion Analysis of Short English Readings based on the CAM-SPAT Model”. This study introduces a deep learning-based model to assess the emotion of speech in English reading aloud, aiding language learners in effective communication. The model incorporates a cross-modal attention mechanism and a two-layer attention-based bi-directional long- and short-term memory network for emotion classification. Evaluation metrics reveal high efficacy, with mean F1 values of 98.54% for classification, 85.13% for detection, and 73.55% for speech emotion analysis, providing valuable strategies for enhancing spoken English skills in language learners.

Huihong Li, in the paper titled "Personalized Artwork Recommendation System", highlights the rising importance of spoken English in global communication. The proposed deep learning model, incorporating cross-modal attention mechanisms, effectively assesses emotion in reading aloud English texts, offering valuable strategies for enhancing language learners' spoken English skills. Evaluation metrics indicate the model's high efficacy in speech emotion analysis.

Tulika Ranjan et al, in the paper titled "Multilingual Code-Mixed Sentiment Analysis in Hate Speech" addresses the limitation of existing sentiment analysis works, predominantly focused on English, by presenting a multilingual code-mixed language model. The model efficiently identifies sentiments in a hate speech dataset from Twitter, using a transformer-based pretrained sentiment analysis model for labeling. Six machine learning models are trained, demonstrating effective sentiment analysis across multiple languages, including code-mixed languages, with results indicating negative sentiment in hate speech and positive or neutral sentiment in non-hate speech.

Vipin Jain et al., in the paper titled "Ensemble Hybrid Model for COVID-19 Sentiment Analysis with Cuckoo Search Optimization Algorithm", address the global impact of COVID-19 on mental and physical health. Utilizing Indian tweets about COVID-19, two datasets are developed, covering periods from January to March 2021 and December 2021 to May 2022. Employing natural language processing and the Valence Aware Dictionary for Sentiment Reasoning, the study utilizes three word embeddings techniques for feature extraction and integrates a cuckoo search optimization algorithm for optimal features selection.

Piyush Kanungo et al., in the paper titled "A Feature Extraction-Based Improved Sentiment Analysis on Apache Spark for Real-Time Twitter Data", focus on enhancing sentiment analysis using Apache Spark for real-time Twitter data.This paper seeks to enhance sentiment analysis accuracy on real-time Twitter data using Apache Spark, focusing on generic content. Unlike existing works on offline data, this study employs six classification algorithms on N-gram and TF-IDF feature extraction methods, demonstrating the superiority of trigram feature extraction for Logistic Regression and Support Vector Machine in analyzing general tweets on Apache Spark.

Gowtham Dora Pappala, in the paper titled "Sentiment Analysis and Speaker Diarization in Hindi and Marathi using Fine-Tuned Whisper", extends the capabilities of the Whisper Automatic Speech Recognition (ASR) model. The enhancements encompass speaker diarization, text summarization, emotion detection, text generation, and question answering. Specifically fine-tuned for Indian regional languages, such as Hindi and Marathi, on the Common Voice 11 dataset, the model exhibits a significant 50% reduction in Word Error Rate (WER), promising improved ASR accuracy and advancements in human-machine communication applications.

Xiao ShiXiao et al., in their paper titled "Real-Time Sentiment Analysis on Social Networks using Meta-Model and Machine Learning Techniques", introduces a real-time sentiment analysis system for social networks, employing a meta-model and machine learning techniques to enhance accuracy by integrating both textual and visual data. Evaluation against state-of-the-art methods demonstrates superior performance in accuracy, precision, recall, and F1-score, highlighting the system's suitability for applications like social media monitoring. The proposed system's capacity to handle multimodal data positions it as a robust solution with implications for social media analysis.

Tara Rawat et al., In the paper titled "Emotionally Wrapped Social Media Text: Approaches, Opportunities, and Challenges", explores the evolution of online platforms and the growing trend of sharing emotionally charged data. The literature review, spanning 2001-2022, categorizes selected papers into granularity, contextual, and cognition levels, offering a comprehensive analysis of approaches to identify emotional states in unstructured textual data. The paper concludes with insights into challenges, applications, and future directions in emotion mining for researchers in diverse domains.

R. L. Keerthana etal., in the paper titled "Explaining Sarcasm of Tweets Using Attention Mechanism", focuses on improving sarcasm detection, a challenging emotion to identify in textual data. The proposed strategies in the paper significantly enhance sarcasm detection model performance, setting a new state-of-the-art on the TweetEval benchmark dataset. Additionally, an attention-based interpretability technique is introduced, shedding light on token importance and aiding in the understanding of contextual embeddings crucial for decision-making in sarcasm detection models.

The paper "Improving Bert Model Accuracy for Uni-modal Aspect-Based Sentiment Analysis Task" provides an overview of aspect-based sentiment analysis and the issue of overfitting.

To address the issue of insufficient coverage in the current sentiment lexicon and the difficulty of constructing sentiment lexicon in specific fields, the paper "Design of Sentiment Analysis Framework of Digital Media Short Text Based on Multi-pattern Sentiment Lexicon" proposes a multi-modal emotional thesaurus.

In summary, this special issue of Scalable Computing: Practice and Experience explores the dynamic landscape of sentiment analysis and affective computing within the realm of multimedia data in social networks. The featured research papers span diverse applications, from sentiment analysis tailored for elderly social media users to real-time community detection in social networks. The studies delve into areas such as drug sentiment on Twitter, academic affective states in online learning, hate speech detection in low-resource languages, personalized artwork recommendations, speech emotion analysis, and COVID-19 sentiment analysis. These contributions not only showcase advancements in computational techniques but also highlight the intersection of affective models with social computing, offering valuable insights for understanding and harnessing user-generated data in the age of multimedia and social networks.

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 787–796, DOI 10.12694/scpe.v24i4.2166

EXPLAINING SARCASM OF TWEETS USING ATTENTION MECHANISM

R L KEERTHANA*, AWADHESH KUMAR SINGH*, POONAM SAINI†, AND DIKSHA MALHOTRA†

Abstract. Emotion identification from text can help boost the effectiveness of sentiment analysis models. Sarcasm is one of the more difficult emotions to detect, particularly in textual data. Even though several models for detecting sarcasm have been presented, their performance falls way short of that of other emotion detection models. As a result, few strategies have been introduced in the paper that helped to enhance the performance of sarcasm detection models. To compare performance, the model was tested using the TweetEval benchmark dataset. On the TweetEval benchmark, the technique proposed in this paper has established a new state-of-the-art. Besides the low performance, interpretability of existing sarcasm detection models are lacking compared to other emotion detection models like hate speech and anger. Therefore, an attention-based interpretability technique has been proposed in this paper that interprets the token importance for a certain decision of sarcasm detection model. The results of the interpretability technique aid in our comprehension of the contextual embeddings of the input tokens that the model has paid the greatest attention to while making a particular decision which outperforms existing transformer-based interpretability techniques, particularly in terms of visualisations.

Key words: sarcasm detection, interpretability, task-adaptive pre-training, attention, and heatmap.

1. Introduction. Sentiment analysis is one of the famous domains where many state-of-the-art NLP techniques [17] have been employed in various real-world scenarios such as product recommendation, feedback analysis, social monitoring, and so on. However, the hidden emotions contained within the input data impede the performance [18] of the state-of-the-art sentiment analysis models. Irony is one such emotion, described as the expression of one's meaning through the use of language that generally conveys the opposite, often for hilarious or dramatic effect. Due to the above reasons, deep learning models, especially those optimised for irony detection, are in high demand. Also, pre-trained models [19], specifically trained for a specific area or task, are being widely used for the detection of such sarcasm. While there is an abundance of pre-trained models available for social media data, the intricate nature of sarcastic comments often leads to reduced performance in pre-trained models specifically designed for irony detection.

Ironic expressions convey an opposite meaning, often subtly, especially in text. The irony can be so subtle, especially in the text, that identifying it can be difficult, especially if the model doesn't understand the context. Irony detection is a difficult task for language models since we, as humans, have difficulties understanding sarcasm in context at times. Detecting irony is especially challenging when dealing with unimodal text data, as textual incongruity may or may not imply irony. Therefore, it's crucial not only to develop an effective model for identifying irony but also to grasp the reasoning behind the algorithm's decisions. Model interpretability encompasses methods that aid in deducing the model's outcomes. These interpretations help validate model predictions and understand the token importance, as assessed by our model and the baseline. Consequently, they enhance our comprehension of how the model performs in this challenging task and conceptualising token significance of transformer based models.

Since transformer-based [21] models have been employed to vectorize the input text, the model's attention retains a great deal of context about the input text. The attention scores of the transformer models are related to contextual information; hence, these scores can be employed to analyse the models. Even though there have been disputes that attention cannot be utilised for explanations[20], there are many works that have used attention for interpretability. Explainability of a model is concerned with understanding how the model works to reach a specific conclusion, whereas interpretability is used to assist an observer in understanding the model's decision. Works like [3], [4], [5] have employed model attention for interpretive purposes. The weights

*Department of Computer Engineering, NIT Kurukshetra, Kurukshetra, India.
†Department of Computer Science and Engineering, punjab engineering college (deemed to be college), Chandigarh, India.

generated by the transformers are generally built on attention scores[21]. The attention score at any particular layer indicates the amount of attention received by the input tokens in that layer. As a result, it can be argued that attention can be employed to provide the model interpretability. The attention scores as it is can be used for the sake of interpretability. This paper discusses how raw attention does not provide interpretability for sarcasm detection and models and also proposes a computation with attention scores which aids in better visualisations.

The paper has been broken into the sections listed below. Section 2 discusses related work on sarcasm detection and the interpretability of transformer-based models. Section 3 describes the methods that were adopted to improve the performance of sarcasm detection algorithms. Section 3 also discussed the interpretability technique proposed to visualise the token importance of sarcasm detection models. Section 4 goes over the implementation and results. The final section concluded the paper.

2. Related Work.

2.1. Sarcasm Detection. Information from a variety of sources, including a Twitter dataset, news headline sarcasm, and a Reddit dataset were gathered by [10]for training purposes. This paper tested their dataset on linear regression classifiers utilising a variety of data representations such as bi-grams, BoW, phrase embeddings,and so on. They came to the conclusion that bi-grams and BoW representations outperformed sentence embeddings. Later, [8] used SVM, random forest, and KNN algorithms with k-fold validation to develop a machine learning system for detecting sarcasm. The investigations revealed that Random forests produced the highest F1-score of any of the three techniques on the airline sentiment [18] dataset. And [14]employed BERT_{large} to calculate word embeddings, which were then used to deter-mine if the words were sarcastic or not. The model was evaluated using multipledatasets, and the BERT Classifier stood out in all of them.

The article [9] used Extreme Gradient Boosting to detect sarcasm in an image. The dataset provided the context of the speech, which was also used as a part of the training. The embeddings were acquired using the TF-IDF approach and then processed using the XGBoost ensembling technique. On the test dataset, the model received an F1-score of 92.4.

Attention-based GRUs [12] were also employed to predict sarcasm. To learn the context of the input data, GRUs were used. The contextual embeddings were fed into a simple classification layer, which categorises the text into one of two groups. On the SARC[10] dataset, they obtained an F1-score of 77.2.

Incongruity aware attention network (IWAN) [13] was proposed that combines multimodal data such as textual, visual, and audio information and uses the incongruity between the features to detect sarcasm. To extract characteristics from various data modes, various models were utilised. These characteristics were then scored using their scoring mechanisms. To classify the input, these scores were fed through a softmax classifier. On the MUStARD dataset, the model received an F1-score of 74.5.

Simple classifiers like SVM classifier[11] was also employed to detect sarcasm in a collection of news headlines [17]. They used the TF-IDF concept to obtain word embeddings. Using SVM for data classification, the model earned an F1-score of 94.8 on the test dataset.

2.2. Interpretability. In [12], the GRUs were provided with a layer of attentions, and the weights of these attentions were used for interpretability. The attributions of the input tokens were determined using integrated gradients to help interpret the transformers in general. These attributions indicate the positive, neutral, and negative influencing tokens of the prediction. SHAP and LIME were used in [9] to explain the model predictions. Both methods' output scores were utilised to visualise the importance of each input token in the model's choice.

These methods were not designed especially for the Transformers. As a result, they are unconcerned with attention layers with additional context information. As a result, much research has been conducted in an attempt to interpret the model using attention weights.

BertViz [15] is a visualisation tool that allows us to depict the attention of tokens in a text from several perspectives, such as neuron view, head view, and model view. These views aid in visualising the flow of input token attentions through the model.

Later, the paper [7] suggested aggregating attention weights from previous layers to determine the attention-based token relevance at a specific layer. Under the assumption that attentions can be merged linearly, the paper [6] proposed two ways for aggregating the attention values gained by each token across the network from the weights in each layer. These methods offer strategies for calculating token importance at higher levels. It is crucial to emphasise that the proposed computations are just for illustrative purposes.

As can be seen, very few interpretability techniques were used for sarcasm detection, and the majority of the work did not use attention scores to interpret sarcasm detection models. For the sarcasm detection model in this paper, an attention-based interpretability technique has been implemented. In addition to proposing the method, previously proposed attention-based interpretable techniques have also been implemented, which did not work for sarcasm detection models. The scope of attention-based interpretable techniques is discussed in this paper.

3. Methodology.

3.1. Workflow. Works in [1] have illustrated the performance improvement of downstream tasks by pre-training the language models. Pre Training techniques can be of two categories based on the kind of data used for pre training. The first type is domain-adaptive pre-training which assists in refining the sentence embeddings and model vocabulary for the domain for which it was trained. Domain specific pre training enhances performance, especially when the target domain language differs greatly from the source domain vocabulary. As the RoBERTa$_{base}$'s pre-training corpus is far too distinct from social media data, it was pre-trained with a large amount of unlabeled data from the social media domain using self-supervised techniques. The RoBERTa$_{base}$ was trained with 58 million tweets using Masked Language Modelling and this pretrained model is popularly called Twitter-RoBERTa [2].

RoBERTa$_{base}$ [23] is an optimised version of the BERT. It was trained on significantly larger datasets than BERT. These datasets were gathered from a variety of domains, including Book-Corpus [24], News [25], Web Text [26], and Stories [27], totaling 160 GB of text. BERT was trained using static Masked Language Modelling and Next Sentence Prediction, whereas RoBERTa$_{base}$ was trained using dynamic Masked Language Modelling only. As a result, RoBERTa is simply the optimised version of the BERT.

The Twitter-RoBERTa [2] model was directly fine tuned on the TweetEVAL-Irony detection dataset, which yielded an F-score of 65.1 on the ironic class. And the top performing model indicated in [2] has an F-score of 70.5 in the ironic class. This model was pre-trained on RoBERTa using the same Twitter data, but from scratch, and then fine-tuned for irony detection. Even if there is an explanation for this performance disparity, that irony observed in social media data differs from irony seen in conventional train text tweets—it can be argued that task-specific pre-training is critical, particularly for challenging tasks like irony detection. Task-adaptive pre-training is the process of pre-training a model with data for a certain task.

The RoBERTa$_{base}$[23] model was initially pre-trained on a domain-specific Twitter dataset [2] using masked language modelling. Subsequently, it underwent further training on the task-specific SARC[10] dataset, specifically designed for sarcasm detection. This continuous pre-training approach played a pivotal role in establishing the model as a state-of-the-art solution for sarcasm detection. The TweetEVAL-Irony detection dataset [2] served both as the basis for fine-tuning the pre-trained model and as the evaluation benchmark. Fig 3.1 depicts the entire process in detail. The observed results were discussed in Section 4. This fine-tuned model's attention has been drawn to interpreting the model's decisions.

3.2. Model Interpretation. Like any deep learning models, Transformer architectures are also black box in nature. Techniques like SHAP, LIME, and integrated gradients [28] were applied to the transformers for interpretability. Even though they provided the explanations, these techniques were not explicitly proposed for transformer architecture, and therefore they do not use the attention weights, which hold a lot of context information. Therefore, in this experiment attention scores were used for interpretability.

In the proposed technique for interpreting transformer-based sarcasm detection models, the attention scores generated by the transformer were leveraged to visualise the importance of individual tokens. It is important to note that BERT-based models consist of multiple encoder layers (denoted as 'n'), each of which processes the input text by vectorizing it. During the vectorization process, the model appends special tokens, such as <CLS> (representing the beginning of a sentence) and <SEP> (representing the end of a sentence), to the input tokens. By examining the attention scores computed by the transformer, the insights into the relative importance and attention assigned to different tokens within the input sequence can be gained. This

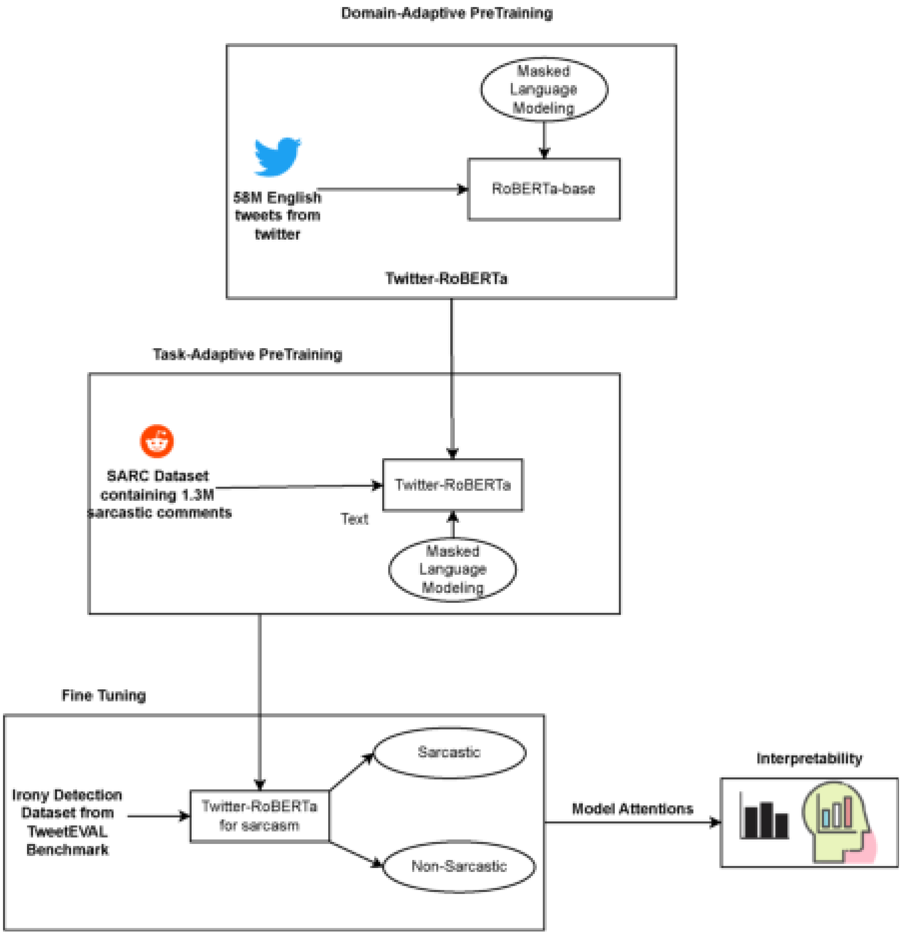


Fig. 3.1: Visual Representation of the whole framework

visualisation technique provides a valuable tool for understanding how the model processes and interprets the textual information, particularly in the context of sarcasm detection. To ensure consistent input representation, every text undergoes padding with a special <CLS> token at the beginning and a <SEP> token at the end. In cases where the input text comprises multiple sentences, a <SEP> token is inserted after the end of each sentence. Following the vectorization process, if an input text consists of t tokens (including the special tokens), the attention scores computed at each layer of the transformer yield a t x t vector. Each t-sized vector captures the degree to which a specific token attends to all other tokens within the sequence.

To gain insights into the interpretability of transformer-based sarcasm detection models, the focus was on understanding the role of the <CLS> token and its attention scores. The <CLS> token's attention scores determine the degree of importance it assigns to all other input tokens. Typically, these attention scores are utilised for classification purposes. However, for the purpose of interpretation, we specifically visualised the attention scores of the <CLS> token in relation to each individual token.

A transformer's encoder layer consists of multiple attention heads[21], each generating a t x t attention vector. For interpretability, the attention scores of the <CLS> token in the final layer was examined, which contains a significant amount of contextual information. To compute the attention scores of the <CLS> token in the last layer, the attention scores of the <CLS> token from each attention head was linearly multiplied in the final layer. This operation can be seen as treating the attention scores as context embeddings, which are subsequently processed for classification purposes.

Table 3.1: Pseudocode for TAPT

Algorithm:
TaskAdaptivePre-training with Domain Adaptive Pre-training
Input:
1.Pre-trained Twitter-RoBERTa model (domain-specific)
2.SARC dataset (task-specific)
3.TweetEval Irony Benchmark dataset (evaluation)
Procedure:
1.Domain-Specific Pretraining:
Initialise the model using the pre-trained Twitter-RoBERTa weights.
2. Task-Specific Pretraining with SARC:
Iterate the model through each data sample of SARC and train the model using Masked Language Modelling.
3. Fine-Tuning and Evaluation on TweetEval Irony Benchmark:
Fine-tune the pre-trained model further and evaluate the model's performance using the TweetEval Irony Benchmark dataset.
Output:
A state-of-the-art model for sarcasm detection.

To illustrate this concept, the transformer architecture must be considered as a computation graph. Within this graph, a linear multiplication of the weights associated with the <CLS> token from all attention heads in the last layer was performed. By applying this approach, the attention scores of the <CLS> token in the final layer are obtained, representing its relevance to the overall context of the input text. By leveraging these attention scores, a deeper understanding of how the model processes is provided and assigns importance to the <CLS> token, which serves as a key element for classification in sarcasm detection tasks.

3.3. Pseudo-code. This subsection provides a pseudocode for the whole framework. This helps the readers to easily understand the whole pipeline. Table 3.1 and 3.2 are the pseudocode for the pre-training method and the attention based interpretation respectively.

4. Results and Discussion.

4.1. Datasets. This experiment employed two publicly available datasets. These data were obtained from the Twitter API and labelled for sarcasm detection. The description is as follows: The Self-Annotated Reddit Corpus (SARC) [10] is a balanced dataset including approximately 90k comments retrieved from the Reddit API, with 50% of the comments being sarcastic and the other 50% non-sarcastic. This dataset is one of the largest for the challenge of sarcasm detection. The authors of the comments contained in the dataset are also responsible for labelling the dataset.

For fine tuning purposes, TweetEVAL-Irony [2] detection dataset has been used which is also a balanced dataset that contains around 4K user-generated tweets that are classified as ironic or non-ironic. The Tweet-EVAL Benchmark dataset, which has been suggested for use in a variety of tweet classification applications, includes this dataset. This benchmark has been used to assess models created for tasks involving Twitter data.

4.2. Data Preprocessing. For both the SARC and TweetEVAL-Irony datasets, the same preprocessing has been employed. The following steps are conducted on both datasets as part of the preprocessing. To eliminate noise from the input text, preprocessing of the data is done. Since all the data samples are consistent after the noise has been removed, the model can be trained more quickly and effectively. The methods used in this research have focused particularly on eliminating noise from the data. All of the words have been made lowercase. The user tags that were not useful for learning from the tweets were excluded. Any links or URLs referenced in the tweets were also removed. The punctuation marks have also been dropped. All of the emojis were converted to text using the Python Emoji Library.

Table 3.2: Pseudocode for attention based model interpretation

Algorithm:
Attention based model interpretation
Input:
Input text for which we want the model's interpretation and attention scores for the input tokens computed across the layers. **Procedure:**
1. $tokens = TokenizeTwitter - RoBERTa(input\_text)$.
2. $attention\_scores = RetrieveAttentionScores(final\_layer)$.
3. $heatmap = InitializeEmptyMatrix(SIZE :(length(tokens) + 2) \times (length(tokens) + 2))$
4.$cls\_attention\_scores = attention\_scores[:, 0, :]$
Slice the attention score matrix to get the attention scores of <CLS> tokens of all the attention heads.
5. For each $attention\_head$ in $final\_layer$:
   a. cls_scores = cls_attention_scores[attention_head, :]
Extract the attention scores of the <CLS> token from that attention head.
   b. multiplied_scores = cls_scores ∗ attention_scores[attention_head, :, :]
Multiply the attention scores of the <CLS> token by the corresponding attention scores of all other tokens.
   c. heatmap += multiplied_scores
Add the resulting attention scores to the heatmap matrix.
6. heatmap /= number_of_attention_heads
Divide each entry in the heatmap matrix by the number of attention heads to compute the average attention scores.
7. Visualize(heatmap)
Visualize the heatmap matrix, highlighting the token importance and attention distribution.
8. Return heatmap.
Output:
A heatmap displaying the importance of each token in the input text.

4.3. Evaluation Metrics. The F-Score of the irony class has been chosen to evaluate the model, as it is easy to compare our model's performance to the baseline model. The F1 score is a measurement that combines recall and precision. It is calculated by taking the harmonic mean of precision and recall [22].

4.4. Baseline. The twitter-RoBERTa model proposed in [2], which was pre-trained on 58 million tweets using Masked Language Modelling, is considered the baseline. The RoBERTa model was directly pre-trained using task adaptive pre training using SARC is also an other baseline. This model was fine tuned for the irony dataset and is publicly available on the hugging face API.

4.5. Task-Adaptive Pre-training. The hugging face library was used to import the base Twitter-RoBERTa model checkpoint that was used for pre-training. This model was trained using Masked Language Modeling after being fed the SARC dataset. We pre-trained for 25 epochs with a training batch size of 64 by dynamically masking 15% of the input tokens. The final checkpoint model has been finetuned further. The RoBERTa$_{base}$ model was also checkpointed from the hugging face library which was also pre-trained using SARC dataset. This model was also pre-trained with the same parameters as mentioned above. This pre-trained model was considered as the second baseline.

4.6. Fine tuning. The pre-trained models obtained from the previous section were fine-tuned for 100 epochs using early stopping with validation loss and a patience of 10. The training was terminated, and the best model was achieved after 61 epochs for the first model and 76 epochs for the second model. For training and validation, the batch size was set to 12.

Table 4.1: Comparision of the F1-scores of our model with previous Benchmarks

Model	**F-Score**
Twitter-RoBERTa fine tuned for TweetEVAL-irony dataset(BASELINE1)	65.1
RoBERTa pre-trained for twitter data from scratch and fine tuned for TweetEVAL-irony dataset(State-of-the-art)	70.5
RoBERTa pre-trained from scratch with SARC	71.83
Twitter-RoBERTa pre-trained with SARC(our model)	73.56

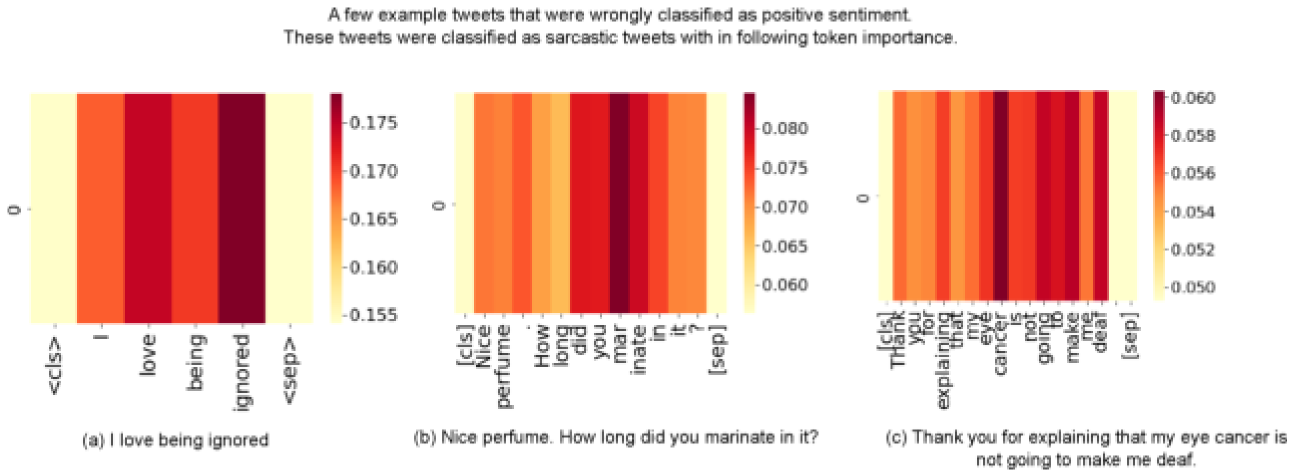


Fig. 4.1: Heat map visualizations of the obtained from the proposed attention based mechanism

4.7. Interpretation. The attention scores of the last layer's <CLS> token (Layer 12) have been utilized to compute the combined attention of tokens in the last layer. All of the layers of the transformer's tokens attend to one another. Contextual embeddings are generally defined as the attention ratings of the <CLS> token to all other tokens. All of the input tokens' relevance is shown by these scores. These attentions so illustrate the significance of each token for model prediction.

4.8. Discussion. In the sarcastic class, our fine-tuned model achieved an F-score of 73.56. In comparison, the baseline Twitter-RoBERTa model scored 65.1, while the state-of-the-art model [2] for the same dataset reached 70.5 (see Table 4.1 for details). When we trained RoBERTa specifically for sarcasm detection using the SARC dataset, it yielded an F-score of 71.83. These results highlight the need for task-adaptive pretraining to enhance performance. Additionally, including domain-specific pretraining[2] alongside task-specific pretraining boosted the model's performance by 8 points, underscoring the importance of continuous pretraining.

A few examples have been chosen where the sentiment was incorrectly identified due to the presence of sarcasm and checked whether or not the tweets were classified as sarcastic. Aside from that, heatmaps have been generated for the same examples using the interpretability method explained in Section 3.2. The attention scores were computed for the last layer and the attention scores of the <CLS> were projected on a heatmap which indicates token importance for each token and helps in interpreting the weight of every token in the model's decision.

From the example (a) in Fig. 4.1, the word *ignored* receives the highest attention score, followed by the word *love*. This interpretation clarifies the prevalence of textual incongruity in sarcastic comments. The sarcasm in example (b) from Fig. 4.1 is fairly subtle. The concealed negative context is properly captured by the model, as evidenced by the heatmap. Similarly, for example (c) from Fig. 4.1, subtle negative context from tokens *cancer, deaf* are attended the most by the <CLS> token. The attentions of the transformer-based models that hold the context can thus be used to interpret the model's predictions.

In addition to visualising token importance with the proposed attention method, a model view of raw

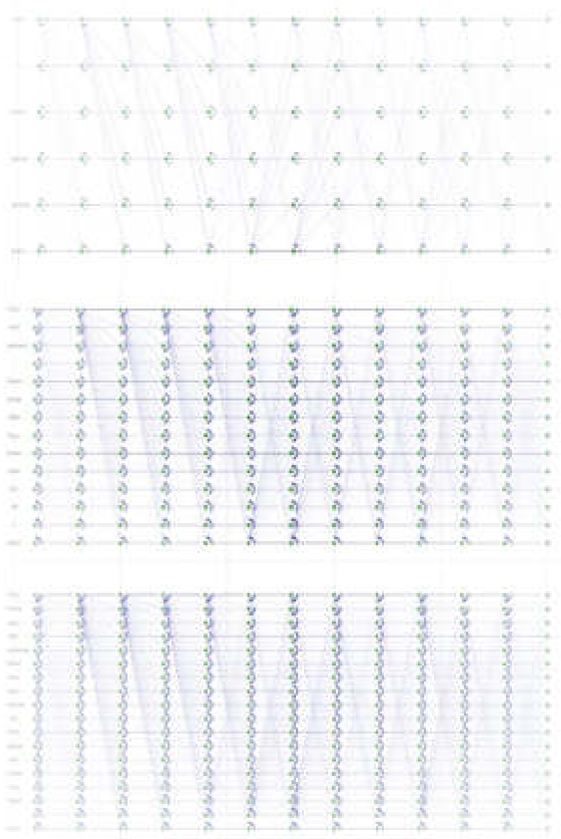


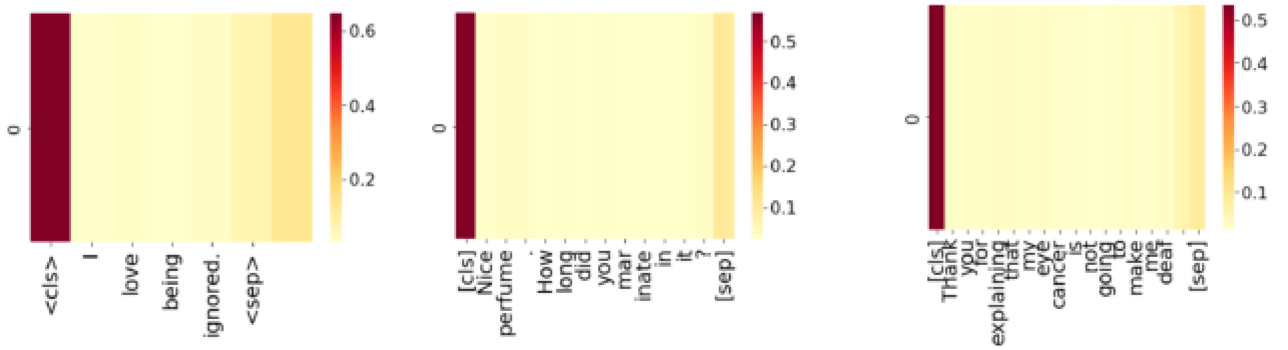
Fig. 4.2: Graph view of raw attention scores for the chosen examples

attention has been generated in Fig 4.2. The raw attention scores across the layers are plotted in a graph manner to understand the flow of attention weights between layers. The graph consists of edges that illustrate the flow of weights which represents the tokens that a particular token attends. In Fig 4.2, it is observed that the model assigns greater importance to the <CLS< and <SEP]< tokens in the initial layers, while the attention weights among the other tokens remain relatively uniform. These findings indicate that existing attention-based explainability methods, as previously proposed, do not effectively capture the intricacies of sarcasm detection.

4.9. Analysis. In the previous subsection, the visualisation of raw attentions were presented in Fig 4.2, which revealed a significant challenge in the deeper layers of the model. In these layers, every token appeared to attend to every other token with almost identical attention scores. This phenomenon resulted in a loss of token identifiability, making it difficult to gauge the true significance of individual tokens solely based on raw attention weights. To address this issue, various studies, including [6] and [7], have proposed techniques to enhance token identifiability. The research in [6] introduced two solutions: attention rollout and attention flow.

These algorithms aim to calculate a token's attention at a specific layer by linearly combining the attention from preceding layers with the attention at that layer. In the attempt to apply the concept of attention rollout, the attention rollouts for the <CLS> token of the last layer were computed and projected onto a heatmap, as depicted in Fig 4.3. Unfortunately, these visualisations yielded no interpretable information. However, when we followed the proposed method of averaging the scores from only the last layer, the resulting interpretations became more meaningful.

Regarding the behaviour where rollouts were not effective, the reason for this outcome remained elusive during our analysis. We encountered challenges in understanding why the rollout visualisation failed to pro-

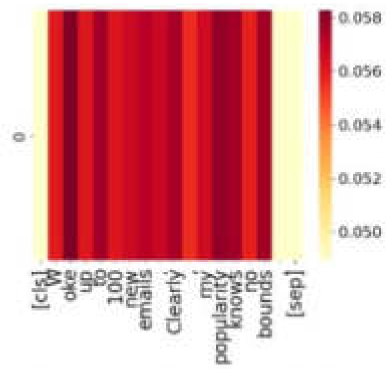


(a) I love being ignored

(b) Nice perfume. How long did you marinate in it?

(c) Thank you for explaining that my eye cancer is not going to make me deaf.

Fig. 4.3: Heat map visualizations obtained from using attention rollouts – version 1



Woke up to 100 new emails. Clearly, my popularity knows no bounds.

Fig. 4.4: Heat map visualizations obtained from using attention rollouts – version 2

vide clear insights. Further investigation is warranted to uncover the underlying factors contributing to this behaviour and to determine potential solutions for improving the interpretability of token attention rollouts. The reasons behind this behaviour are something to look into in the future.

For instance *"Woke up to 100 new emails. Clearly, my popularity knows no bounds."* as shown in Fig 4.4, the heatmap analysis revealed that the model did not assign higher attention weights to what we consider important keywords. Tokens such as *'bounds'*, *'popularity'* and *'woke up'* were given higher weight while allocating less attention to the word *'no'* which is the essence of textual incongruity in this instance. This observation underscores the need for the model to better capture the incongruity in the sentence, particularly with respect to the word *'no'* in order to yield more accurate interpretations. Although significant attention weight may be allocated to important tokens, it is crucial to prioritise words responsible for incongruity.

5. Conclusion. In this work, we have discussed the performance boost from using task-adaptive pre-training for sarcasm detection models. We also proposed a method to use the attention scores of the input tokens to provide some interpretability. These interpretations aid in understanding the tokens that have had the greatest influence on a certain decision. We believe that not all of the tweets adhere to the traditional definition of sarcasm. Consider the following tweet: *"I just failed my driving test."* There is a reply to the comment that says, *"Very Good! Well done"*.

In light of the situation, the response is sarcastic. However, if we only evaluate the Twitter reply, we cannot conclude that it is sarcastic. As a result, it is critical to provide context in addition to the tweet reply. Besides, there are other categories of sarcasm such as irony, satire, and so on, as explained in [16]. Despite the fact that the *ISarcasm* dataset has sought to provide training data from several categories of sarcasm, there are relatively few of them. Transformers are attention-based models that have advanced to the forefront of many NLP tasks. However, in order to properly understand their capabilities, they must be trained with a significant amount of data. To conclude this work, we believe that we require a standard dataset for sarcasm detection in social media that comprises a fair amount of data on all varieties of sarcasm as well as other features such as context.

REFERENCES

[1] Gururangan, Suchin, et al., *Don't stop pretraining: Adapt language models to domains and tasks.*, arXiv preprint arXiv:2004.10964 (2020).

[2] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves, *Tweeteval: Unified benchmark and comparative evaluation for tweet classification,*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 1644–1650.

[3] Wiegreffe, S., & Pinter, Y,*Attention is not not Explanation.*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

[4] Vashishth, S., Upadhyay, S., Tomar, G. S., & Faruqui, M., (2019)*Attention interpretability across nlp tasks.*, ArXiv Preprint ArXiv:1909.11218.

[5] Vig, J, (2019). *Visualizing Attention in Transformer-Based Language models.*, ArXiv Preprint ArXiv:1904.02679.

[6] Abnar, Samira, and Willem Zuidema., *Quantifying attention flow in transformers.*, arXiv preprint arXiv:2005.00928 (2020).

[7] Brunner, Gino, et al., *On identifiability in transformers.*, arXiv preprint arXiv:1908.04211 (2019).

[8] Pawar, Neha, and Sukhada Bhingarkar., *Machine learning based sarcasm detection on Twitter data.*, 2020 5th international conference on communication and electronics systems (ICCES). IEEE, 2020.

[9] Godara, Jyoti, et al., *Ensemble classification approach for sarcasm detection.*,Behavioural Neurology 2021 (2021).

[10] Khodak, Mikhail, Nikunj Saunshi, and Kiran Vodrahalli., *A large self-annotated corpus for sarcasm.*, arXiv preprint arXiv:1704.05579 (2017).

[11] Vinoth, D., and P. Prabhavathy., *An intelligent machine learning-based sarcasm detection and classification model on social networks.*, The Journal of Supercomputing 78.8 (2022): 10575-10594.

[12] Akula, Ramya, and Ivan Garibay., *Interpretable multi-head self-attention architecture for sarcasm detection in social media.*, Entropy 23.4 (2021): 394.

[13] Wu, Yang, et al., *Modeling incongruity between modalities for multimodal sarcasm detection.*,IEEE MultiMedia 28.2 (2021): 86-95.

[14] Baruah, Arup, et al. , *Context-aware sarcasm detection using bert.*, Proceedings of the Second Workshop on Figurative Language Processing. 2020.

[15] Vig, Jesse., *A multiscale visualization of attention in the transformer model.*, arXiv preprint arXiv:1906.05714 (2019).

[16] Oprea, Silviu, and Walid Magdy., *isarcasm: A dataset of intended sarcasm.*,arXiv preprint arXiv:1911.03123 (2019).

[17] Feldman, Ronen., *Techniques and applications for sentiment analysis.*Communications of the ACM 56.4 (2013): 82-89.

[18] Montoyo, Andrés, Patricio Martínez-Barco, and Alexandra Balahur., *Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments.* Decision Support Systems 53.4 (2012): 675-679.

[19] Qiu, Xipeng, et al., *iPre-trained models for natural language processing: A survey.*,Science China Technological Sciences 63.10 (2020): 1872-1897.

[20] Attardo, Salvatore, and Jean-Charles Chabanne. , *Jokes as a text type.*,(1992): 165-176.

[21] Vaswani, Ashish, et al. , *Attention is all you need.*,Advances in neural information processing systems 30 (2017).

[22] Powers, David MW., *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.*,arXiv preprint arXiv:2010.16061 (2020).

[23] Liu, Yinhan, et al., *Roberta: A robustly optimized bert pretraining approach.*,arXiv preprint arXiv:1907.11692 (2019).

[24] Zhu, Yukun, et al., *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.*,Proceedings of the IEEE international conference on computer vision. 2015.

[25] Sebastian Nagel. 2016, *Cc-news. http: //web.archive.org/save/http: //commoncrawl.org/2016/10/newsdataset-available.*

[26] Aaron Gokaslan and Vanya Cohen. 2019. , *Openwebtext corpus. http://web.archive.org/save/http://Skylion007.github.io/ OpenWebTextCorpus.*

[27] Trinh, Trieu H., and Quoc V. Le., *A simple method for commonsense reasoning.*,arXiv preprint arXiv:1806.02847 (2018).

[28] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan., *Axiomatic attribution for deep networks.*,International conference on machine learning. PMLR, 2017.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Mar 21, 2023
Accepted: Aug 1, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 797–818, DOI 10.12694/scpe.v24i4.2216

EMOTIONALLY WRAPPED SOCIAL MEDIA TEXT: APPROACHES, OPPORTUNITIES, AND CHALLENGES

TARA RAWAT *AND SHIKHA JAIN †

Abstract. With the evolution of several online platforms for information sharing such as social media, blogs, product review sites, and discussion forums, people have become more proactive in sharing their expectations, views, feelings, and experiences. This large amount of emotionally wrapped data motivates many researchers to perform data mining and present the crux of hidden emotions or mental states in a more presentable and comprehensible manner. It has several applications in different domains such as business, education, psychology, politics, and many more. This paper presents a detailed literature review projected to rigorously analyze the existing approaches to identify the mental or emotional state of a person from unstructured textual data. We include the most relevant papers which were published during 2001-2022. The selected papers are classified into three categories: granularity level, contextual level, and cognition level. Each category is carefully analyzed followed by a detailed and critical discussion. Finally, open challenges, opportunities, applications, and future directives are presented in-depth to facilitate the researchers working in the domain of emotion mining.

Key words: Emotion Mining; Opinion Mining; Sentiment Analysis; Social Media; Affective Computing; Emotion; Mental-state

1. Introduction. The evolution of the web has given users the opportunities to participate, share, and contribute to various platforms like social media, blogs, product review websites, online forums, and discussion platforms. Now, it has been very common that as a kind of relief, people suffering from mental health issues frequently either directly or indirectly reveal their emotions, feelings and everyday battles with mental health concerns on social media ([6]. As a result, a large amount of multi-modal data is generated.

With the advancement of data, affective computing (AC) and emotion mining (EM) have emerged as new areas of research. Affective Computing (AC) is a field that relates to, arises from or deliberately influences emotion or other affective phenomena [97]. It plays a vital role in effective communication. It is broadly classified as emotion recognition, modeling, and expression (as shown in Fig. 1.1). Emotion recognition means extracting the emotional state from different modalities such as image, video, text, and audio-visual. Emotion modeling refers to the study of the effects of emotions on different cognitive processes and generating synthetic emotions. Emotion expression means expressing emotions using different modalities. It is important to mention that the researchers have used different vocabulary like sentiment, emotion, opinion, etc in the literature. According to Gordon [39], sentiments can be defined as a collectively formed set of neuronal responses, emotional attitudes, and common definitions normally coordinated around another individual. Emotion [124] has four interrelated components: sensory stimuli, physiological shifts, verbal movements, and an emotion tag which indicates the local correlation of the components.

In consideration of the similarities of how emotions and sentiments have been described, we will use these words interchangeably in this study to refer to perceptions that arise from the cumulative forces of the mental, behavioral, and personal [121] Many researchers have worked towards identifying the polarity of emotions in terms of negative, positive, or neutral emotions and called it as emotion mining [2],[83]. However, some researchers identified emotions in terms of discrete emotions such as happiness, sadness, fear, anger, and so [151]. Predominantly, both kind of work is called emotion mining. Emotion mining is one of the most challenging tasks to extract the emotions of people from the web, inferring from the text, and predicting the underlying intent. The eruption of research in emotion mining took place during 2001-2022, as shown in (Fig. 1.2). We

*Jaypee Institute of Information Technology, Noida, India. (tararawat2503@gmail.com).
†Jaypee Institute of Information Technology, Noida, India. (shi_81@rediffmail.com).

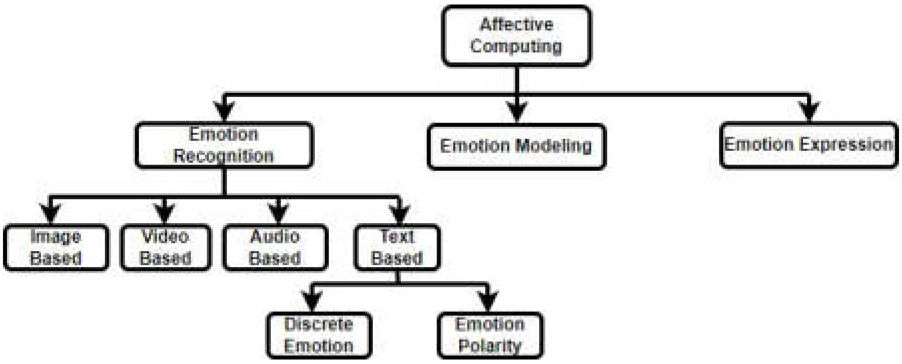


Fig. 1.1: Affective Computing Classification

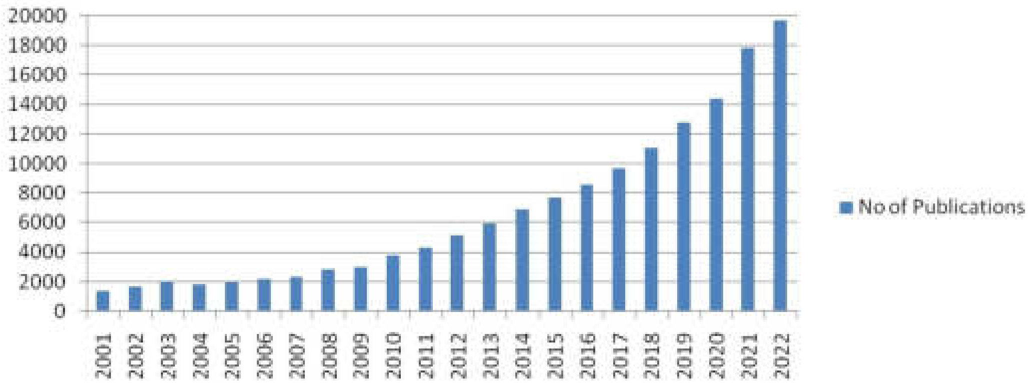


Fig. 1.2: Graph showing the research in "Emotion Mining from Text" in Google Scholar Database

sought to ensure the reliability and relevance of our dataset by implementing specific inclusion and exclusion criteria. The inclusion criteria are as follows:

1. Publications were included if they directly addressed the research topic of "Emotion Mining from Text" on the Google Scholar Database.
2. Only peer-reviewed journal articles and conference proceedings were considered to maintain the quality and reliability of the sources.
3. Publications within the time frame of 2001 to 2022 were included to focus on recent research trends.
4. Inclusion was limited to publications in English to facilitate data extraction and analysis.

The Exclusion Criteria:

1. Publications lacking titles, abstracts, or keywords directly related to the phrase "Emotion Mining from Text" research topic were excluded.
2. Non-peer-reviewed sources, such as books and theses, were excluded to ensure data quality.
3. Publications outside the specified date range 2001-2022 were excluded to maintain a focus on recent research.
4. Publications in languages other than English were excluded due to language proficiency constraints.

Since then, emotion mining has explored various domains such as E-therapy, psychological health services, counseling online, terrorist attacks [16], prediction of the financial market [56], medicine, and the healthcare domain [29], [110]. Recently, researchers [5], [45], [35] used a combination of cognitive and affective information with human language to tackle acute issues such as irony and sarcasm detection [80] The authors used crowd-sourcing [20] to build a common sense knowledge base, which is used to link the semantic gaps between word-level and concept-level analysis. Furthermore, for context-aware and cognition-based emotion mining, strenuous attempts are advancing from polarity detection to more complex nuances of users' social behavior on online platforms [41], [60].

This paper presents a critical analysis of the work done by experts in the field of emotion mining while extracting emotional and mental states from textual data available on the web. The key features of the paper are:

1. An extensive literature survey of the most relevant papers published in the last two decades is presented to demonstrate the incremental approaches and improvements.
2. Categorical analysis (Fig. 2.1) and critical discussion of each category is presented.
3. Tabular summary of important papers is illustrated to provide a quick overview.
4. Generalized framework of emotion mining extracted from papers is presented to guide beginners in this field.
5. Over the year evolution of different techniques for emotion mining is presented pictorially to give a comprehensive view.
6. Various challenges encountered over the year are discussed in detail and presented pictorially.
7. Applications of emotion mining are discussed.
8. Overall discussion on EM with future directives is presented at the end of the paper.

The paper follows the incremental approach on the timeline of the last two decades. To the best of the authors' knowledge, no survey paper covers belief-based and semantic-based emotion mining which is the future of the field.

The rest of the paper is organized as follows: Section 2 reviews the advances in emotion mining and Sections 3,4 and Section 5 discusses various levels of text-based emotion mining. Section 6 summarizes the discussed approaches. Existing challenges, opportunities, and their applications are discussed in Sections 7,8, and Section 9 respectively. Finally, Section 10 concludes the paper and provides future directives.

2. Advances in Emotion Mining from Text. Textual emotion mining consists of collecting facts about beliefs, thoughts, and emotions that persons express about topics of interest. The field of emotion mining in computational linguistics focuses primarily on classifying different emotional contents within a word, phrase, or document. This field typically includes tasks such as defining emotions, classifying subjectivity, and understanding polarity [87] Analyzing the sentiment of textual data can, therefore, involve investigating the emotion/mental behind that text. In certain cases, emotion motivates an individual to evaluate an event and to create opinions about it. The field depicts a large problem space. There are a few terms that have mutual meaning such as sentiment analysis, opinion mining, emotion mining, and subjectivity analysis, affect analysis, review assessment, opinion extraction. However, all these terms come under emotion mining. Although the terminology sentiment analysis is most widely used in industry, in academia, both emotion mining and sentiment mining are often employed [67] Hence, this paper explores all aspects of emotion mining (such as emotion analysis, emotion mining, opinion mining, etc.) termed differently in literature. These terms are used interchangeably throughout the paper to retain the original terminology used in different articles that are discussed here.

A range of focused and advanced fields are already being explored. However, text-based emotion mining still has a long way to go. Studying emotions in different fields provides us with useful information on what kind of emotions people experience while dealing with different situations.

This paper broadly investigates the literature under three categories: granularity level, cognition level, and contextual level. The same has been shown in (Fig. 2.1) and discussed in detail in subsequent sections.

3. Granularity-based ttextual Emotion Mining. In this section, we include a discussion on the related research for emotion mining based on granularity such as word level, sentence level, and document level.

3.1. Word-level Processing. Words may be demonstrative, expressive, and sensitive. Identifying the describing words (adjectives, adverbs, some nouns, and verbs) out of the whole text is the main focus at word-level emotion mining where each word is assigned a value on a scale from 0 (negative) to 10 (positive) based upon its polarity. SentiWordNet [31] is designed to annotate WordNet. It is used for calculating the score of a particular word in a sentence using machine learning approaches. For example, “This place is beautiful”, this

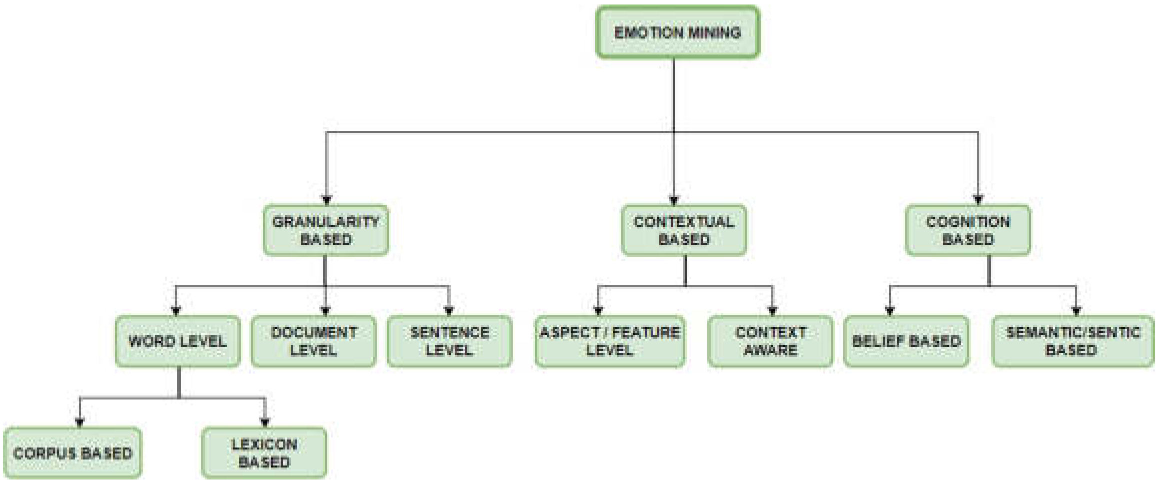


Fig. 2.1: Classification criteria of Text-Based Emotion Mining

sentence has the word 'beautiful' that has some score as defined in lexicons, which leads to a positive polarity. Each word has some probability score for each emotion. The highest score is chosen as the emotion of that word. It has been observed that in natural language, opinionated content or the word, which expresses the emotion, is most commonly conveyed by certain parts of speech (POS). For instance, adjectives such as disastrous or lovely convey emotions more effectively than the verb running or eating. Hence, POS is used to mark words and search for polarity terms (adverbs and adjectives, [76]. Broadly, word-level EM can be done in three steps:

1. Detecting the subjectivity of different words. For example, love and hate, both are subjective words.
2. Calculating the polarity of words i.e. positive, negative, or neutral by using polarity scores. For instance, love is positive and hate is negative
3. Calculating the intensity of the word i.e. depth of positive, negative by analyzing the polarity words.

Opinion orientation can be determined by two approaches [3] a) Corpus-based b) Lexicon-based. The corpus-based approach detects opinion words with context-specific orientation [137] and lexicon-based approaches detect the seed words and search them with the associated lexicons such as WordNet along with their synonyms and antonyms [135]. Recently, domain-specific areas are explored rigorously by researchers. They have been working to increase the classifier's reliability by augmenting word-level emotive information, i.e., constructing and utilizing various forms of custom dictionaries [64].

3.2. Document-level Processing. At the document level, the aim is to analyze and classify the whole document as positive or negative. Several studies have been explored at the document level, especially, in cross-domain, cross-language, and opinion categorization [125], [27], [139]. Bellstam et al. [7], used the finance-related corpora, generated by financial analysts. They explored the emotions/sentiments in textual descriptions of business operations to quantify corporate innovation in finance.

Duyu et al. [123] proposed Convolution Neural Networks (CNN) and Recursive Neural Network (RNN) based document-level analysis [34], [133] called User Product Neural Network (UPNN). This model assimilates user and object-level information for sentiment classification for document-level analysis. They experimented with improved accuracy using vector-based and matrix-based user and object representation. Liu et al. [71] compared the sentiment classification approaches, using Recursive Neural Network and Recurrent Neural Network, and found that the sentiment topics or target entities cannot be identified by document or sentence level, hence need to analyze at a more granular level.

3.3. Sentence-level Processing. The goal of sentence-level emotion mining is to separate emotional information from facts to improve the prediction accuracy of the document's overall polarity [143]. At this level, the concern is basically on two steps:

1. The subjectivity of the sentence
2. Polarity (+ve or -ve) of the sentence

The following section will discuss the major input for both tasks.

3.3.1. Classifying Subjective Sentence. There are many approaches to classify a given sentence as subjective or objective. An objective sentence provides some factual information about the environment, while a subjective sentence communicates some personal emotions, perceptions, or opinions. Emotions are considered subjective feelings and thoughts [70], [144]. Some researchers feel that objective sentences do not imply emotions. However, they can also express emotions, and ignoring them may result in the wrong prediction. In this context, Liu [68] has discussed that subjectivity and opinion should not be compared. He also discussed that a subjective sentence may not convey any emotion such as ”I am using the latest version of Android OS” is a subjective sentence, even then it doesn’t express any emotion. Similarly, an objective sentence may imply emotion such as ”After updating the latest android OS, my phone kept hanging”. While this sentence is stating a fact, it shows an implicit emotion about the topic (the phone hanging).

Classification of subjective sentences is one of the most important and foundation schemes. The existence of subjective words determines the subjectivity of a sentence [42], [154], [47]. Moreover, these studies have been able to differentiate between subjective and objective sentences which subsequently improved sentence-level emotion mining. Several machine-learning approaches with language-specific features for Urdu languages are used [86] to identify the subjectivity of sentences. Therefore, different sentences may be dealt with differently at sentence-level analysis.

3.3.2. Polarity identification. The purpose of polarity identification is to determine text as positive, negative, or neutral according to its emotional significance. Some authors [1] the WordNetAffect Affective database [99] to distinguish between positive, negative, and neutral emotions. However, these emotional categories may help predict polarity, but not quite enough to predict the polarity intensity. Therefore, the authors use a few emotional categories such as love, happiness, joy, fear, sadness, anger, and so on. One approach is to use parts-of-speech (POS) tags syntactic patterns to express opinions/sentiments/emotions. The researchers [58] used the approach of multiplying all the scores of emotional words in the sentence as +1 and -1. They also used classification techniques to identify a particular type of emotional orientation. The hierarchical sequence learning model identifies polarity in both sentences as well as document-level and also improved the accuracy at both levels.

In 2010, Hassan et al. [4] proposed a method to identify online users’ behavior using the Markov Model. Then, they identified the polarity orientation of the attitudes. Some researchers worked on target-dependent sentiment classification, where the target is classified into some category [51]. Researchers admitted that there are a variety of sentences and each sentence expresses emotions differently. Therefore, one technique cannot be applied to all kinds of sentences. Handling sarcastic sentences is still a big challenge as positive sentences may have negative meanings and vice versa. These types of sentences are more used in online discussions and political forums. Tsur et al. [127] proposed a semi-supervised learning approach to identify sarcastic and non-sarcastic sentences. Initially, they used the labeled set and expanded it automatically through a web search. Many researchers [66], [157] experimented with Convolution Neural Networks (CNN) for sentence-level classification. They built the model on top of the unsupervised pre-trained word vector, tuned hyper-parameters, and performed well with remarkably improved results.

3.4. Critical Analysis of Granularity-based Emotion Mining. Though a lot of research is conducted at a granular level, it has several limitations when used in real-life applications, such as which features or aspects of the entities are liked or disliked. Several sentences are not easily classified as emotion orientation varies on different targets or entities, e.g., “Trying out Airtel sim because Jio voice calls being cut after every 10 mins.” and “Vodafone is doing well in this awful environment”. In the latter sentence, sentence-level classification is insufficient. We need to go to the aspect-level analysis. It is assumed that a sentence has an overall positive or negative orientation but some of its parts may convey the opposite emotion. E.g., “India’s labor-intensive agriculture has achieved steady increases in food grain production despite the often unfavorable weather conditions”. Here, the overall message is positive but it also contains a negative emotion “unfavorable” which cannot be ignored and only aspect-level analysis can solve this problem. Moreover, comparative sentences cannot be dealt with at the sentence level. For instance, the sentence "Amul butter tastes better than Mother Dairy butter" can't be simply classified as positive, negative, or neutral. Therefore, further, fine-grained analysis is required to handle them.

4. Context-based Textual Emotion Mining. Features are the attributes of an entity. Feature-based emotion mining extracts the emotions towards some features of the entity [115]. It also analyzes the emotion while considering the surrounding context of the feature. This section discusses the two types of feature-based emotion mining: aspect-level and context-aware.

4.1. Aspect-level Emotion Mining. The aspect-level analysis mainly focuses on entities and their components, where entities may be a product, an organization, a service, or a topic. The components can be referred to as features or attributes of the entity. Aspect-based emotion mining, also known as feature-based emotion mining, identifies different aspects of the target entity in a long sentence and its polarity [93]. Earlier, this kind of analysis required a set of features which are manually designed. Nowadays, the focus has shifted to other approaches such as graphs of dependency relationships [152], and lexicon-based supervised learning approach [55]. It is observed that identifying entities, their components, and knowing their polarity is not sufficient to know the sentiment of a sentence. For instance, a negative emotion about an entity does not mean that all the aspects of that entity will have a negative orientation [120].

The aspect-based emotion mining is carried out at two levels: aspect extraction and aspect-based classification. Several methods of aspect extraction such as nouns and noun phrases identification followed by PMI [47], balances category feature (BCF) and the other is the categorical proportional difference (MCPD, [21], sequential learning method [69] such as Hidden Markov Models (HMM, [101], [54] and Conditional Random Fields (RF, [61], frequency-based approach for emotional noun phrases [11], C-value measure-based method to retrieve multi-word aspects [158], dependency parser [159] and double propagation method [100] are discussed in the literature.

Recently, researchers have started using various evolutionary algorithms [84], semi-supervised [38], [85], [21], supervised approaches such as SVM [155], RNN [116], CNN [150], [95], [74], deep RNN [49], pre-trained word embeddings [149], LSTM [122], SLSTM [134], Memory network [136], deep recurrent belief network [22], multi-task learning network (IMN, [44], and unsupervised approaches [14] such as pLSA (Probabilistic Latent Semantic Analysis, [46], LDA (Latent Dirichlet allocation, [92], LSA based aspect-sentiment mixture model [79], joint topic sentiment model [65] for aspect extraction.

In supervised learning, attention networks are gaining popularity in the field of emotion mining, Several attention networks Attention Encoder Network (AEN, [117], relative position attention network (RPAEN, [146], Multi-grained attention network (MGAN, [32], Bi-GRU-based Position-aware Bidirectional Attention Network (PBAN, [40], recurrent attention mechanism [24], attention mechanism with GRU [43] are being used in the literature. Machine learning methods were commonly used in emotion mining but require a substantial amount of training data to achieve improved accuracy [18]. In view of the lack of large amounts of annotated data, researchers started working on ontology [98] based methods. They used a combination of hierarchical learning (HL) process and sentiment ontology tree (SOT, [140], knowledge-based along with linear SVM, or a review-based and a sentence aggregation algorithm [28], to categorize the data on the basis of emotions.

4.2. Context-Aware Emotion Mining. The digital world is becoming more informal. People use new words or phrases to express themselves. The emotional lexicons (such as NRC, WordNet, SenticNet, etc.) are not acquainted with such new words/phrases. Though, humans can interpret the meaning of such new words/phrases with the help of the surrounding context. However, in any computational model, these unseen words/phrases are simply ignored, and hence the expression of emotion may get lost. To make the computational model more robust, the need arises to make use of semantics to extract the contextual meaning.

Researchers spent lots of time exploring and selecting regular features. Several emotion-mining tools rely on emotion lexicons to support linguistic resources. These lexicons contain polarity values and some weight to emotional words. But, it is observed that emotion mining requires more evolutionary approaches that may lead to semantic-level analysis [19]. A word may change its polarity based on its context. With this idea, Gangemi et al. [37] gave importance to emotion contextualization and consider it as a major challenge in the field of emotion mining. Some of the common context-aware approaches used by researchers are Rule-based approaches

[111], Sentence and discourse-based context shifters, Linguistic patterns [147], and Vector space modeling [153].

A process to segregate contextualized emotion lexicons i.e. a way to isolate sentiment terms from stable polarity terms. This process allows changing the ambiguous word's polarity as the context changes in some other textual information. With the evolution of contextual knowledge, traditional emotion lexicons are reinforced with context knowledge [73]. An iterative regression and a random walk method are used to label ConceptNet [119] with emotion values [126]. Several studies have been done to develop contextualized, cross-domain lexicons with sentiment mining and various decision-support applications [141]. It was assumed that disambiguation and contextualization can give better performance on cross-domain emotion mining using ontological lexicons. These lexicons can be used at three levels. The first level detects an ambiguous word with the help of existing labeled corpora. The second level determines the emotion score of co-occurring contextual terms and finally at the last level determines the combined polarity values for ambiguous and unambiguous terms. Experimentation and evaluation were performed on various datasets such as product reviews, hotel reviews, and movie reviews and successfully achieved better performance on each dataset. They extended the process by using the group concept from SenticNet, ConceptNet [119], Freebase [13], and DBpedia [10]. Recently, deep learning-based methods are becoming more popular as they have greatly improved the performance of emotion mining. Deep learning approaches use the structure and context of the sentence, instead of just emotion-bearing words in isolation.

Often context-based emotion mining is domain-specific and is based on the assumption that words always have the same association in that domain. For example, the use of the word unexpected: may signify positive emotion when describing the plot of an action film ("the film had an unexpected ending"), while it is not a good sign in the sense of the battery life of a laptop. Though, this may not always be valid. Consider a "leaves" description that says "The children love to play in the leaves but they do not like it when their father leaves for work." One can easily see that within the same domain, in the first case the word "leaves" communicates positive emotion, and in the second case a negative emotion. It is, therefore, reasonable to conclude that domain considerations alone are not enough to conduct context-based emotion mining.

Word embedding is becoming more popular, which is a low dimensional continuously-valued vector representation of words and is explored extensively by researchers [142], [81] The most commonly used word embeddings are CBOW, SkipGram, and GloVe [132]. Nguyen [88] proposed a semi-supervised approach to utilize an immense amount of available unlabeled data. He represented the semantic similarity of textual and user context using deep learning and latent low-dimensional space representation. He used networked characteristics also besides textual content i.e. the way users are connected via social relationships because this conveys the emotional behavior of connected users. A semantic model [138] is proposed to monitor the fluctuating emotional phases or emotions of learners while participating in MOOC's online courses. The authors used machine learning and semantic network approach in real-time to know learners' emotions about courses. Observing the change in the emotional states of a learner during the course can improve the graduation probability.

4.3. Critical analysis of Contextual Based Emotion Mining. Though lots of work has been carried out in the field of context-aware emotion mining, still it is far from the human touch. Streaming data is still a challenge, as contextual information is difficult to deal with. Nowadays, most organizations need aspect-level emotion analysis for feature details. It is observed that machine learning and deep learning with word embeddings are widely used by researchers for aspect-based emotion mining. However, accuracy and precision are still an open challenge as the existing methods/algorithms are still not able to deal with complex problems. Most of the cases are handled by emotional words. However, words or information are highly diverse, countless, and exceptional. It becomes very difficult to learn patterns via statistical methods only because there are several ways to express emotions. Lots of research, so far, mainly focus on products, movies, and restaurant reviews as well as Twitter data. In these domains, fairly good accuracy can be achieved. This is because the reviews are short and rich in emotions. However, for the other domains such as health, politics, forum discussions, commentaries, etc., the situation becomes very difficult as the emotions are either objective or factual statements in these domains.

Sometimes, the sentences are very complex due to a mixture of objective and subjective statements or sarcastic sentences. Besides this, all the social platform data is very noisy, and consists of various errors, whereas, all the analysis tools require clean data for analysis purposes. Therefore, a lot of focus is required for pre-processing tasks and methods. Some researchers incorporated the ontology method due to the non-availability of training data, but ontology features are not robust. The ontology method also needs training data to interpret them. It is very difficult and time-consuming to manually build ontology; therefore, there is a need to automate the process to create the ontology for multiple domains. However, due to the multidisciplinary area, emotion mining takes advantage of psychology also. As per psychology, the successful evolution of online platforms is the ability to bring people together and serves very different needs of people.

Chen et al.[23] explored how opinion about any object gets affected by the user's connection with their friends, family, or relatives. They have developed a content-based sequential opinion influence framework to monitor users' behavior on social media. This framework uses an RNN to seize the opinion expressed in the sequence. The user's opinion is checked with the past information that the user acquired from his/her relation with others and his/her personal opinions. Based on the learned influence, the user's future influence prediction is determined. This valuable information can be used by companies for their future business planning.

5. Cognition-based Textual Emotion Mining. With the evolution in the field, researchers have started using cognitive features for emotion mining. As per Izard et al. [50], cognition alludes to the mental process that is impacted by emotions (desires and feelings). Human emotions (such as feelings, moods, and motivations) can impact both the way they think and the choices/decisions they make. Cognition is one of the most important events that operate and control emotions. Cognition-based emotion mining can provide us with more insights into the emotion behind words. Broadly, cognition-based emotion mining is classified into two categories: Belief based and semantic-based. They are discussed in brief in the following sections.

5.1. Belief Based Emotion Mining. Understanding mental states such as the beliefs and goals of a person enables an individual to explain behavioral patterns. According to psychologists, beliefs impact the arousal of emotions and emotions in turn influence beliefs [36], [30], Ortony et al. [91], concentrate on the cognitive elicitors of emotions [53], [51] in their hypothesis. They proposed that emotions are positive or negative reactions to their beliefs or perceptions of reality [52]. Thus, one may be glad or dissatisfied with the outcomes of an event; one can support or oppose a person's activities, and one can like or dislike the characteristics of an item. The human emotional process, according to Lazarus [62], is made up of two distinct processes: evaluation (which describes a person's connection with their environment) and coping (which suggests strategies for altering or maintaining this relationship). Both of these processes are aided by cognition. Cognition aids evaluation by constructing mental representations of how experiences relate to inherent nature or temperament such as beliefs and goals. Coping deals with the situation through suggestions and exploration of methods for changing or sustaining the individual social relationship.

In a recent study, Masland et al. [78] examined the impact of emotional reasoning on reliability assessments in people with a moderate personality disorder. The findings show that in comparison to the control group, participants with suspected moderate disorder produced more emotionally fragile evaluations and were more affected by negative experiences or beliefs. As a result, persons with moderate characteristics may be impacted differently and more precisely by unfavorable afflictions. The mental state, if it is subjected to mental illness, then the diagnosis is sometimes difficult since it necessitates detailed and in-depth psych evaluations by trained psychiatrists at an early stage [108], as well as interview sessions, lengthy questionnaires, self-reports, or evidence from friends and family. In such cases, it becomes very important to understand the underlying emotion, feelings and behavior shown by patients. Furthermore, it is highly usual for patients suffering from mental illness often avoid going to health centers to seek medical care in the early phases of their illness [161]

Hamad et al. [160] proposed the automatic depression method, which extracts depressive content using various strategies such as key phrase matching and document summarization, named entity recognition (NER), etc. on the user tweets, resulting in more granular and relevant content, which is then directed to a deep learning framework consisting of convolutional neural networks associated with attention-enhanced gated recurrent units. Vetriselvi et al. [131] provided a review of cognitive-based emotions and insights on the importance of cognitive theories especially cognitive and intuitive theory to improve emotion classification. Long et al. [72] proposed an attention-based neural network model for sentence-level sentiment classification. They trained this model with cognition-grounded eye-tracking data and built a cognition-based attention (CBA) layer for emotion mining. Further research with formal models is needed so that future systems can usefully detect emotions and accurately predict behavior.

In light of the present COVID-19 epidemic, where religion, religiosity, and emotions play crucial roles in coping, it follows that a connection between religiosity—which encompasses a variety of religious activities and beliefs—and people's emotional states and assessments of their general well-being should be obvious. Regardless of their allegiance with a particular religion, people in a wide range of geographical areas show a wide range of emotions through their religious behavior, which is inextricably linked to a formal or informal belief in God or the Divine.

5.2. Semantic / Sentic Based Emotion Mining. Emotion mining at the conceptual level while considering the common sense or semantics of the text is called sentic/semantic-based emotion mining. In 2013, Raina [[102] proposed the sentiment analyzer where emotions can be associated with a common-sense knowledge base. The author analyzed sentiments in news articles. In 2018, Ferru et al. [33] used machine learning approaches to classify a message as positive or negative towards a particular topic and on a five-point scale. They used a cognitive computing tool (i.e. IBM Watson) to extract the semantic features. Their experimental results showed that semantic features play a very important role in classification. However, the authors would like to further research semantic features extracted from other cognitive computing systems.

Irony detection is one of the challenging tasks in the field of emotion mining as it has an ambiguous interpretation. Vijay D. et al. [12] addressed the problem of irony detection. From a cognitive perspective, it is a test to think about how humans utilize irony as a specialized instrument to communicate. To detect irony, the authors constructed a Hindi-English code-mixed corpus using tweets. They used various feature vectors such as character N-gram, word N-gram, laugh words, and emoticons. In 2018, Corriga et al. [25] proposed a tool that accepted an image as information and did the classification based on the person's gender and afterward recognized their emotions.

As per Cacioppo et al. [17], it is related to a person's tendency to engage in and enjoy thinking. Das et al. [26] mentioned in the paper that it has a moderating effect on variables such as human behavior, intent to purchase, and also web surfing. Li [64] used cognitive theories for emotion mining on the linguistic feature, where sentiments are classified as target-dependent and target-independent. Zou et al. [162] proposed a strategy utilizing indirect relations in specific user structure likeness to analyze opinion. The authors experimented and verified that similar users connected via common friends have similar opinions. The authors used the sociological phenomenon of homophily to understand the connection between similar people. Their experimental results showed that indirect relation (through some common friend) has a better performance than user-direct relations to improve the accuracy of sentiment classification.

Social media has recently attracted medical natural language processing researchers to detect various medical abnormalities such as depression, anxiety, etc. Recently, many researchers explored the application of emotion mining in the medical field. Emotion mining can be used in providing healthcare assistance. Nowadays depression is one of the most prevailing issues which is often talked about on social media. Zucco et al.[163] proposed affective computing and emotion-mining methodologies to monitor depression conditions. The authors used mobile technologies to collect input data. Many researchers are working towards detecting loneliness [129], [118], depression [106], [9], [104], [105], or suicidal thoughts [75] However, the results obtained are highly data-dependent and far from cognitive theories of emotions.

5.3. Critical Analysis of Cognition-Based Emotion Mining. Very few researchers have addressed cognition-based emotion mining in recent years. Cognitive approaches talk about the human mind and behavior. Such approaches talk about how emotions are produced and what their effects are. One's belief [112] plays an important role in emotion modeling. Someone can have many emotions at the same moment, each with varying degrees of intensity. The construction of a computationally efficient model of emotion is an extremely complicated reality. For instance, "it is raining outside" may convey positive or negative emotions depending upon the person's belief. If someone believes that rain is good in pursuit of his goal, then this sentence expresses positive emotion else negative.

By reviewing and incorporating significant information, it is studied how intelligence analysis better explains challenging circumstances and boosts useful insights for effective decision-making. Intelligence analysis may assist in determining patterns and social behavior. For this kind of analysis, a high-level cognitive theory of emotions [89] can be used. Unfortunately, this area of research is still untouched. Moreover, cognition-based emotion mining can do wonders in the healthcare domain. These studies can impact a patient's life quality

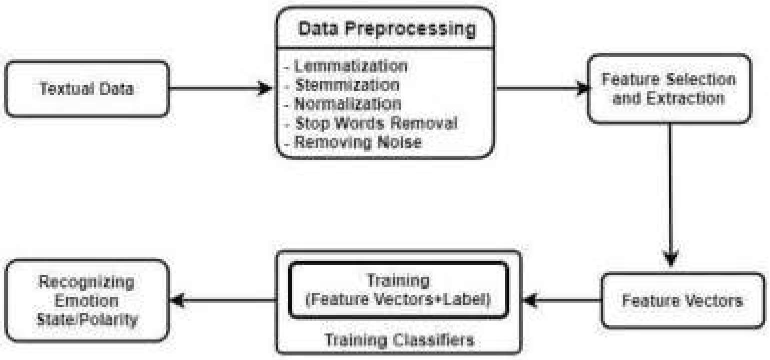


Fig. 6.1: Generalized Framework of Emotion Mining from Textual Data

and health status. This study helps in knowing how emotions influence social relationships, reasoning, memory, and their function in psychological illness/disorder.

6. Summarization and Discussion. In the literature survey, researchers have applied different techniques to identify emotions from unstructured text. It is observed that all the models follow a common process flow (as shown in Fig. 6.1). At first, the unstructured data is preprocessed and cleaned to make it ready for processing. This preprocessing may involve many steps such as stop word removal, stemming, normalization, removal/replacement of slang and abbreviations, etc. Next, features are extracted and feature vectors are generated by various means such as using word embeddings, POS tagging, PMI, TFIDF, etc. Once, a feature vector is ready, it is fed to a learning model to get the output. Emotion mining is the most widely studied topic. However, emotion mining at the granularity level has been the subject of a lot of research studies. The limited number of studies on context-aware, semantic-based, and cognitive perspective is carried out because it requires knowledge of cross-domain and criticality of human nature/human perspective. There are several approaches to performing emotion mining; the machine learning approaches outperformed the traditional approaches. It is also seen that some supervised classifiers such as Support Vector Machine, Neural net, and Naïve Bayes have repeatedly been applied. The effectiveness of these classifiers is probably the reason behind this.

The same thing can be said about the used features; word stem and n-grams are frequently chosen as features. Regarding datasets, most of the researchers have used Twitter data and the IMDB dataset while some researchers built and used a new dataset. Thus, no common dataset was used for benchmarking results and evaluating experiments. In the early days, researchers analyzed texts collected from the web and focused on the word, sentence, or document level. By contrast, researchers have recently treated principally social media texts and dealt with deep learning. Very few attempts have, however, been concerned with Affective Cognitive Emotion Mining. These tasks still need much deeper investigation and research.

Although research on emotion mining started around the year 2000, studies on this issue have shown an active rise in the last few years. This is mainly due to the exponential growth of social media, online reviews, and social networking sites. The literature survey carried out in the previous sections has revealed that researchers have dealt with different emotion mining tasks: subjectivity classification, opinion classification, aspect-based emotion mining, building lexicons/resources, extracting opinion holders, belief-based emotion mining, context-aware and cognition-based emotion mining (as shown in Fig. 6.2). The excerpt of some important papers is summarized in Table 6.1. The table contains information about the dataset and methodology used. Moreover, the pros and cons of each paper are also discussed.

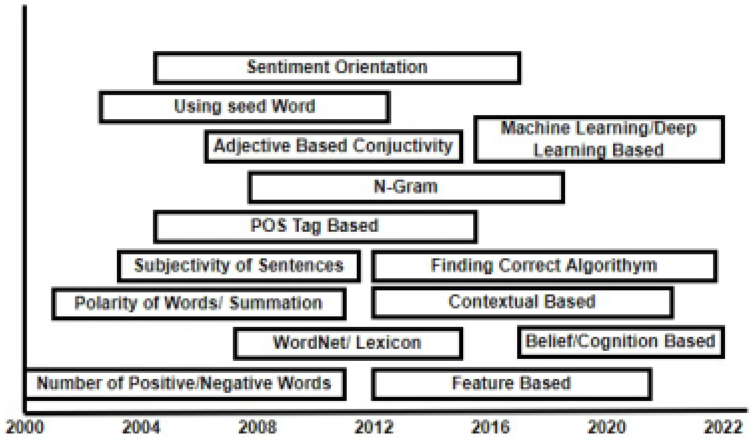


Fig. 6.2: Different Approaches for Emotion Mining over the Years

Table 6.1: Different Approaches for Emotion Mining over the Years

Ref	Approach	Level	Dataset	Algorithms	Advantage	Disadvantage
[44]	Supervised	Aspect-level	Restaurant reviews, twitter reviews, and Laptop reviews	Interactive multitask learning network.	Attentional encoder network prevents recurrence and uses attention-based encoders for context-to-target modeling. Raised the label unreliability issue. Demonstrate the effectiveness and light-weight of the proposed model.	The approach lack in calculating the hidden states.
[77]	Supervised	Character level, Word level	Hotel, Restaurant reviews	Aspect orientation rules	1.Annotated corpus will be freely available. 2.The corpus can be annotated with part-of-speech tags at the word- level which could yield better results.	1. Not generalized to many languages.
[115]	Semi-supervised	Context-aware	Amazon.com and IMDb.com	SVM and Naïve Bayes approach	1.Provides valuable background information for advanced emotion mining. 2. Provides a platform to identify ambiguous sentiment terms, and extract context information for disambiguating	1. Lack of emphasis on scalability and throughput.
[48]	Unsupervise	Document-Level	Stanford Twitter, Sentiment (STS) and ObamaMcCain Debate (OMD).	Semisupervised deep learning, a bilinear embedding model (CaTER method)	1. Provides the effectiveness of exploiting user context information for leveraging social media emotion mining using social relation graph. 2. Best accuracy among all the other methods like SVM, Distant Supervision (DS), Label Propagation (LPROP) and Logistic Regression with word embedding.	1.Unavailability of unlabelled data for word representations.
[150]	Pretraining and multi-task learning	Document-level	SemEval 2014, 2015, and 2016	CNN with gating mechanisms	Improved word representation in aspect-level sentiment classification.	Unable to handle sarcastic and irony in sentences

[107]	Supervised learning	Document-level	Twitter data set, political forum	Linguistic based approaches, SVM	1. Bigram and trigram feature significantly outperformed unigram features for the forum data. 2. Emotion mining improved. 3.Sarcasm detection.	1.Domain-specific Latent Sentiments.
[72]	Cognition based	Document level/sentence -level	IMDB, Yelp 13, Yelp 14 and IMDB2	Cognition based attention model, LSTM	1. It works well both at a sentence and at a document level. 2. Gives higher weights to the sentiment-linked words.	1.Domain dependent.
[2]	Supervised	Sentence-level	Online textual corpus	Emotional ontology approaches	Emotional ontology improves the result. Provides a platform to translate emotional dimensions into emotional categories.	1. Limited scope for speech output
[37]	Unsupervised	Sentence-level	MPQA opinion corpus	Heuristic graph, ML, Knowledge representation	1. The approach deals with a factual sentiment. 2. Assists in identifying relational features. 3. Provides improved performance. 4. Domain-independent	1. Does not perform well for noise data and sarcasm.
[8]	Semi-supervised	Sentence-level	Yahoo! Finance, Raging Bull.	Multi-View classification	1.Can mine a large corpus of domain-specific sentiment expressions	1. Does not support cross-domain.
[69]	Supervised	Sentence-level	Twitter dataset	Support Vector Machine (SVM)	1. Using related tweets along with the current tweets outperformed the results. 2. Utilized graph-based optimization techniques to improve performance.	Relations between a target and any of its extended targets are not explored.
[58]	Supervised/ unsupervised	Sentence-level	Moview & customer reviews, TREC question Dataset, Google News	Convolutional Neural Network	1. Use of pre-trained vectors 2. One layer of convolution has improved the performance.	1.Hyperparameter tuning is required.
[33]	Supervised cognition-based	Topic-based	Twitter and annotated using CrowdFlower	Decision Trees, Linear Regression and Naive Bayes	1. Successfully classified the message on a two-point scale.	1. Dataset is unbalanced. 2.The scarcity of training data
[41]	Supervised	Word-level	Pattern from pages www.alltheweb.com	Point-wise mutual information measure	1. Lexical patterns that are precise enough for finding emotion-bearing, affect words. 2. Re-use of the SOPM\I formula	1. Only a few emotive patterns for adjectives are used

[146]	Unsupervised	Aspect-level	Emotion Dictionary	Attention-based LSTM	Explored the complex multiangle analysis of transitions in public emotions under the pandemic scenario. Helpful in recognizing public behavior and the prevention and regulation of pandemic situations. The issue of identifying polarity in the dictionary is strengthened effectively.	1. The size of the specific dictionaries chosen is too small.
[148]	Supervised	Character Level	Microblogs	CNN	1. Using the word vectors yield higher accuracy than those using words as feature elements. 2. Effectively improve the overall accuracy	Proposed model performs better on Chinese text but not generalized on other languages
[82]	Supervised	Word-level	EmoInt	SVM, CNN, LSTM and BiLSTM	1. Automated robust feature representation rather than manually engineered features. 2. Emotion dimensionality score provides a reliable and fine- grained analysis of text instead of just assigning the discrete emotion class. 3. All the data is made freely available	1. Manually identified the intensity score of basic emotions, therefore need to automate the process.
[59]	Supervised	Document-level	Tweets	RNN	1.Use of word embeddings improved the performance by reducing the high dimensional vectors.	1.Dataset size is too small to exploit deep learning Techniques. 2.Comparability and generalization is lacking
[113]	Rule Based	Document-level	ISEAR	Semantics rules	1.Paid special attention to Phrasal verbs.	Contextual meaning is missing. Not enough terms in the vocabulary.
[128]	Supervised	Document-level	Tweets with NRC emotion lexicon	SVM	1. Identify actionable emotion patterns in Tweets, with results of 84.92 percent and 88.01 percent accuracy.	1. Due to the limited number of emotional class, generalisation is not possible.
[94]	Supervised	Document-level	Tweets and ROC story Data	CNN (DL)	1. An embedding emotion model is created with the best accuracy of 73.3% for a positive emotion such as a pleasure and lowest for sadness around 36.7 %	1. Negative sentences did not perform well. 2. Negation handling could improve the accuracy.

[114]	Supervised	Aspect level	Publically available data	LRA-DNN	1. Compared to the current ANN, DNN, and CNN approaches, LRA-DNN gets the highest performance with accuracy, sensitivity, and specificity at rates of 94.77%, 92.23%, and 95.91%, respectively 2. It effectively lowers categorization and misprediction errors.	1. Execution time may get high while assigning rank to the given feature.
[130]	Unsupervised Approach	Aspect Level	Product reviews	LDA Model	1. Proposed a filtered and strong semantic approach, successfully extracted relevant aspects from product reviews.	1. The use of regular expression makes the approach little complex.
[63]	Supervised	Sentence Level	Food Comments	Bidirectional long-term and short-term memory network (BiLSTM)	1. This sort of approach examines, interprets, infers, and analyses the emotional remark content before coming to the appropriate conclusion.	The approach is not compared and generalized.

7. Challenges in Text-based Emotion Mining. Text is a challenging medium for analyzing hidden emotions. Due to the high complexity and the raw data available online, the area is still lacking. This section summarizes the challenges which still need the attention of researchers. The same has been depicted pictorially (Fig. 7.1). The open challenges are:

1. Domain dependence: Existing approaches mainly focus on the opinionated text, where the sentiments and emotions are explicitly expressed. Syntactical approaches have been used as the text is domain-dependent. Therefore, many language patterns are missed. Therefore, there is a need to focus on the deep analysis of the semantics of the sentences containing implicit emotions. At the same time, the emotion mining model should be domain-independent.
2. Dynamic effective model: To adjust the system to the social platform's user affectivity, the system should identify the user's emotions. The system needs to know when the user is negative about their service/ topic/ product/policy etc. The services or products must be customized as per the user's opinion regarding the product/ services. The affective history of the author should be maintained in the affective user model. It helps to respond appropriately to the user's emotions. Therefore, we need a dynamic affective model. The model should consider the changes in the emotional states of the user, to improve the accuracy significantly. Developing a dynamic affective model is very challenging because of the complex human cognitive states.
3. Feature engineering: People convey their emotions in complex ways like sarcasm, irony, etc. Such sentences do not specifically communicate their meaning but implicit meaning can mislead emotion mining. The only way to understand this issue is through context. Though a lot of research has been conducted in context-based emotion mining, there is a need to focus more on feature engineering as little attention has been paid to it so far. Feature engineering creates new features in a model. These features help to better understand the context and emotions of the users. Engineered features can help to boost accuracy and get an overall sense of what the social platform's user is trying to convey.
4. Context extraction: Truthful words and expressions inferring emotions have scarcely been considered, but they are exceptionally useful for numerous domains. The contextual information of emotional words stays to be highly challenging indeed with so much research.
5. Unknown/unseen words: The large vocabulary of words is available widely. However, it is a very tedious task to manually annotate these huge numbers of features. Moreover, unseen words in the text also cause problems. Therefore, it is very important and challenging to establish domain-independent affective words that can be used in language transfer.
6. Labeled data: Accessing the labeled data to train the model is very difficult. For the best long-term results, it is suggested to train our models. Therefore, need to find out ways to access data quickly.

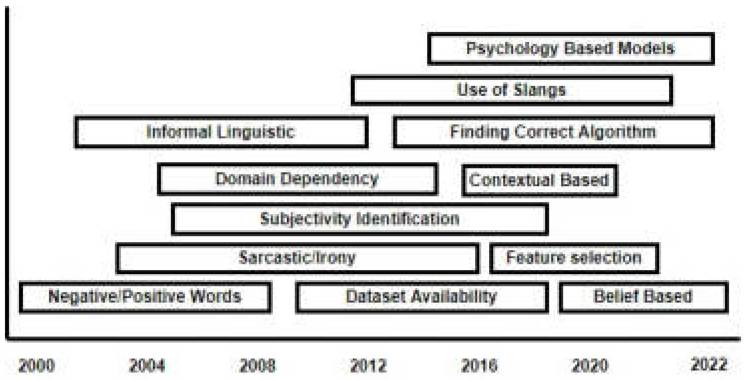


Fig. 7.1: Challenges in Textual Emotion Mining

7. Different sentences: Nevertheless, the question type of sentences, which do not depict any opinion about any particular entity such as “Are you serious?” “Have you gone mad?” also conveys significant emotions and should be handled.
8. Belief incorporation: For the same event or product, two people can express different emotions or the same person can express different emotions at different instances of time. This is due to the difference in beliefs and expectations. To date, no textual emotion mining model considers these parameters

8. Opportunities. As discussed, the practice of collecting and understanding human emotions from diverse data sources, such as text, audio, and video, is known as ”emotion mining”. However, we have limited our study to text data. While the field of emotion mining has advanced significantly in recent years, there are still a lot of untapped or understudied areas. Here are a few innovative prospective research areas:

1. Emotional State Prediction: Construct predictive models capable of anticipating fluctuations in emotional states by leveraging historical data. These models have versatile applications spanning mental health monitoring, customer service optimization, and educational contexts.
2. Emotion Mining for Mental Health: Use emotion mining methods to aid in diagnosing and keeping track of mental health problems. This might mean looking at what people write or how they talk (speech data) on social media to find signs of depression, anxiety, or other mental health concerns.
3. Human-Robot Emotion Interaction: Delve into the application of emotion mining within human-robot interaction scenarios, empowering robots to gain a deeper understanding of human emotions and respond more effectively. This can significantly enhance their performance in tasks like detecting an early risk of emotion disorders, caregiving, and customer service.
4. Emotion Mining’s Biases and Ethical Considerations: Study the moral questions raised by emotion mining, like how it affects people’s privacy, whether they’ve given permission, and whether the algorithms that recognize emotions are treating everyone fairly. Create rules and guidelines for doing emotion mining that are ethical and fair.
5. Understanding Emotions in Legal and Forensic Contexts: Explore how emotion mining can be applied in legal situations and forensic investigations. This involves examining voice recordings or written statements to identify signs of deception or emotional distress.

9. Applications. Emotion mining has made it easier to estimate others’ emotions, feelings, and psychology. Based on the result generated through emotion mining, one can always adjust/control the present situation and manage customer needs in a better way. It allows for staying dynamic throughout. Due to its various advantages, it has been connected to numerous application zones. It has been used in various fields such as the medical healthcare domain, sports, education, the financial sector, online social media, politics, hospitality, tourism, and many more. Some of the emerging application areas are briefed below:

1. Medical health-care domain: Using textual emotion mining, it is possible to assess the psychological health of a person with the help of his online posts [145] and avoid suicide to some extent. Moreover, it can be used to analyze the reaction of people to some disease outbreaks [57] and come up with some feasible solutions.

2. Sports domain: Emotion mining can be used for the spatial relationship between crime events and Twitter activity in the context of any popular game [109] Also, it can be used to assess the passionate reactions of fans [156].
3. Education domain: It can be used by the instructor to monitor the student's overall interest over time in the course [90] which can further help to improve the teaching and learning system [103]
4. Politics: Emotion mining can be used to predict the election results [107] and the reaction of citizens to any new policy.
5. Hospitality and Tourism: To improve the hospitality sector, customer reviews play an important role. Emotion mining can be used to predict customers' orientation based on reviews they posted on social media [77], [96].

10. Conclusion and Future Directions. The process of identifying an individual's mental state/ emotions/ opinions/ sentiments is an ongoing field of research. Most organizations are availing of it and incorporating it into their processes. A plethora of surveys can be found in the relevant literature either focusing on conventional approaches or machine learning approaches. Online social networking platforms have expanded exponentially over the past few years and consequently generated an enormous amount of data. Now, the attention is rapidly escalating to the problem of how to analyze the available data. This paper provides a systematic survey of the state-of-the-art approaches for emotion mining for textual data. Besides polarity identification, emotion mining also about knows the interpersonal relationships in the context of social networks, which affect information flows such as identifying behavioral features that may contain a strong emotional signal [15]. Accordingly, emotion mining needs to do deeper semantic analysis as social online platforms are very noisy.

In the future, research can be conducted in the direction of analyzing or recognizing the emotions hidden in the text understandably to recognize various aspects of human nature and behavior. This research can provide very predictive influences on the public attitude towards various issues and can also explain the state of human mental well-being. Further, information repositories of texts such as online chats, blogs, forums, etc., which are the source of opinioned or emotional content, are expected to give more emphasis on sincerity, spontaneity, and effectuality by combining the models of human cognitive capabilities that may include emotion mining.

Moreover, a human's mental state approach, more specifically the belief-desire-intention (BDI) model, can be used to know the human's intentions or behavior, which may be helpful to predict the human action pattern more accurately and improve emotion prediction. The main goal of using this approach is to capture emotions from the cognition viewpoint. Moreover, in today's scenario, only domain-dependent keywords are not sufficient, we need to explore a commonsense knowledge base that extends the cognitive and affective information associated with the social platform's information. Future emotion mining strategies need broader common-sense approaches which should be driven by the process of human cognition so that the effects are more realistic, reliable, and related to human psychology. We hope that this survey serves as a starting point for much more to come.

REFERENCES

[1] `http://nlp.uned.es/~lplaza/papers/CONLL_2010.pdf`. Accessed: 2023-3-22.

[2] M. Almashraee, D. M. Díaz, and A. Paschke, *Emotion level sentiment analysis: The affective opinion evaluation*, EMSA-RMed@ESWC, (2016).

[3] A. Andreevskaia and S. Bergler, *Semantic tag extraction from WordNet glosses.* `http://www.lrec-conf.org/proceedings/lrec2006/pdf/566_pdf.pdf`. Accessed: 2023-3-22.

[4] A. H. Awadallah, V. Qazvinian, and D. R. Radev, *What's with the attitude? identifying sentences with attitude in online discussions*, in EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2010, pp. 1245–1255.

[5] F. Barbieri and H. Saggion, *Modelling irony in twitter*, in Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, Association for Computational Linguistics.

[6] K. C. Bathina, M. Ten Thij, L. Lorenzo-Luaces, L. A. Rutter, and J. Bollen, *Individuals with depression express more distorted thinking on social media*, Nat. Hum. Behav., 5 (2021), pp. 458–466.

[7] G. Bellstam, S. Bhagat, and J. A. Cookson, *A text-based analysis of corporate innovation*, Manage. Sci., 67 (2021), pp. 4004–4031.

[8] Z. Ben-Ami, R. Feldman, and B. Rosenfeld, *Using multi-view learning to improve detection of investor sentiments on twitter*, Comput. Sist., 18 (2014).

[9] S. S. Bisht, H. Shandilya, V. Gupta, S. Agrawal, and S. Jain, *Perceiving the level of depression from web text*, in Artificial Intelligence, Machine Learning, and Mental Health in Pandemics, Elsevier, 2022, pp. 277–298.

[10] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, *DBpedia - a crystallization point for the web of data*, Web Semant., 7 (2009), pp. 154–165.

[11] S. Blair-Goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar, *Building a sentiment summarizer for local service reviews*, NLPIX2008, (2008).

[12] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, *A dataset of Hindi-English code-mixed social media text for hate speech detection*, in Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, Louisiana, USA, June 2018, Association for Computational Linguistics, pp. 36–41.

[13] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, *Freebase*, in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, New York, NY, USA, June 2008, ACM.

[14] S. Brody and N. Elhadad, *An Unsupervised Aspect-Sentiment Model for Online Reviews*, 2010.

[15] J. J. Brown and P. H. Reingen, *Social ties and word-of-mouth referral behavior*, J. Consum. Res., 14 (1987), p. 350.

[16] P. Burnap, M. L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, and A. Voss, *Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack*, Soc. Netw. Anal. Min., 4 (2014).

[17] J. T. Cacioppo and R. E. Petty, *The need for cognition*, J. Pers. Soc. Psychol., 42 (1982), pp. 116–131.

[18] E. Cambria, *Affective computing and sentiment analysis*, IEEE Intell. Syst., 31 (2016), pp. 102–107.

[19] E. Cambria, N. Howard, J. Hsu, and A. Hussain, *Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics*, in 2013 IEEE Symposium on Computational Intelligence for Human-like Intelligence (CIHLI, IEEE, Apr. 2013.

[20] E. Cambria, R. Speer, C. Havasi, and A. Hussain, *SenticNet: A publicly available semantic resource for opinion mining.* `https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/download/2216/2617.pdf`. Accessed: 2023-3-22.

[21] J.-R. Chang, H.-Y. Liang, L.-S. Chen, and C.-W. Chang, *Novel feature selection approaches for improving the performance of sentiment classification*, J. Ambient Intell. Humaniz. Comput., (2020).

[22] I. Chaturvedi, Y.-S. Ong, I. W. Tsang, R. E. Welsch, and E. Cambria, *Learning word dependencies in text by means of a deep recurrent belief network*, Knowl. Based Syst., 108 (2016), pp. 144–154.

[23] C. Chen, Z. Wang, and W. Li, *Tracking dynamics of opinion behaviors with a content-based sequential opinion influence model*, IEEE Trans. Affect. Comput., 11 (2020), pp. 627–639.

[24] P. Chen, Z. Sun, L. Bing, and W. Yang, *Recurrent attention network on memory for aspect sentiment analysis*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2017, Association for Computational Linguistics.

[25] A. Corriga, S. Cusimano, F. M. Malloci, L. Marchesi, and D. R. Recupero, *Leveraging cognitive computing for gender and emotion detection.*, in EMSASW@ ESWC, 2018, pp. 47–56.

[26] S. Das, R. Echambadi, M. McCardle, and M. Luckett, Mark. Lett., 14 (2003), pp. 185–202.

[27] K. Dave, S. Lawrence, and D. M. Pennock, *Mining the peanut gallery*, in Proceedings of the twelfth international conference on World Wide Web - WWW '03, New York, New York, USA, 2003, ACM Press.

[28] S. de Kok, L. Punt, R. van den Puttelaar, K. Ranta, K. Schouten, and F. Frasincar, *Review-level aspect-based sentiment analysis using an ontology*, in Proceedings of the 33rd Annual ACM Symposium on Applied Computing, New York, NY, USA, Apr. 2018, ACM.

[29] K. Denecke and Y. Deng, *Sentiment analysis in medical settings: New opportunities and challenges*, Artif. Intell. Med., 64 (2015), pp. 17–27.

[30] M. J. Doherty, *Theory of mind: How children understand others' thoughts and feelings*, Psychology Press, Jan. 2008.

[31] A. Esuli and F. Sebastiani, *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. International Conference on Language Resources and Evaluation*, 2006.

[32] F. Fan, Y. Feng, and D. Zhao, *Multi-grained attention network for aspect-level sentiment classification*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2018, Association for Computational Linguistics.

[33] D. S. Ferru, F. Ibba, and D. R. Recupero, *Supervised topic-based message polarity classification using cognitive computing*, CEUR Workshop Proceedings, 2111 (2018), pp. 11–21.

[34] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, *Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya*, Information (Basel), 12 (2021), p. 52.

[35] E. Forslid and N. Wikén, *Automatic irony- and sarcasm detection in Social media*, 2015.

[36] N. H. Frijda and B. Mesquita, *Mesquita - 1998 - The analysis of emotions Dimensions of variation," What develops in emotional development Emotions personality and psychotherapy*, 1998.

[37] A. Gangemi, V. Presutti, and D. Reforgiato Recupero, *Frame-based detection of opinion holders and topics: A model and a tool*, IEEE Comput. Intell. Mag., 9 (2014), pp. 20–30.

[38] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, *Text mining for product attribute extraction*, SIGKDD Explor., 8 (2006), pp. 41–48.

[39] S. L. Gordon, *The sociology of sentiments and emotion*, in Social Psychology, Routledge, Sept. 2017, pp. 562–592.

[40] S. Gu, L. Zhang, Y. Hou, and Y. Song, *A position-aware bidirectional attention network for aspect-level sentiment analysis*, in Proc. 27th Int. Conf. Comput. Linguist, 2018.

[41] E. Gutierrez, W. Karwowski, K. Fiok, M. R. Davahli, T. Liciaga, and T. Ahram, *Analysis of human behavior by mining textual data: Current research topics and analytical techniques*, Symmetry (Basel), 13 (2021), p. 1276.

[42] V. Hatzivassiloglou and J. M. Wiebe, *Effects of adjective orientation and gradability on sentence subjectivity*, in Proceedings of the 18th conference on Computational linguistics -, Morristown, NJ, USA, 2000, Association for Computational Linguistics.

[43] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, *Effective attention modeling for Aspect-Level sentiment classification*, in Proc. 27th Int. Conf. Comput. Linguist, 2018.

[44] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, *An interactive multi-task learning network for end-to-end aspect-based sentiment analysis*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, Association for Computational Linguistics.

[45] I. Hernández-Farías, J.-M. Benedí, and P. Rosso, *Applying basic features from sentiment analysis for automatic irony detection*, in Pattern Recognition and Image Analysis, Lecture notes in computer science, Springer International Publishing, Cham, 2015, pp. 337–344.

[46] T. Hofmann, *Probabilistic latent semantic indexing*, SIGIR Forum, 51 (2017), pp. 211–218.

[47] M. Hu and B. Liu, *Mining and summarizing customer reviews*, in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, Aug. 2004, ACM.

[48] X. Hu, J. Tang, H. Gao, and H. Liu, *Unsupervised sentiment analysis with emotional signals*, in Proceedings of the 22nd international conference on World Wide Web, New York, NY, USA, May 2013, ACM.

[49] O. Irsoy and C. Cardie, *Deep recursive neural networks for compositionality in language*, 2014.

[50] C. E. Izard, *Basic emotions, relations among emotions, and emotion-cognition relations*, Psychol. Rev., 99 (1992), pp. 561–565.

[51] S. Jain and K. Asawa, *Emet: emotion elicitation and emotion transition model*, in Information Systems Design and Intelligent Applications: Proceedings of Second International Conference INDIA 2015, Volume 1, Springer, 2015, pp. 209–217.

[52] ———, *Programming an expressive autonomous agent*, Expert Systems with Applications, 43 (2016), pp. 131–141.

[53] ———, *Modeling of emotion elicitation conditions for a cognitive-emotive architecture*, Cogn. Syst. Res., 55 (2019), pp. 60–76.

[54] W. Jin and H. H. Ho, *A novel lexicalized HMM-based learning framework for web opinion mining*, in Proceedings of the 26th Annual International Conference on Machine Learning, New York, NY, USA, June 2009, ACM.

[55] R. Kamal, M. A. Shah, C. Maple, M. Masood, A. Wahid, and A. Mehmood, *Emotion classification and crowd source sensing; a lexicon based approach*, IEEE Access, 7 (2019), pp. 27124–27134.

[56] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, *Text mining for market prediction: A systematic review*, Expert Syst. Appl., 41 (2014), pp. 7653–7670.

[57] K. G. Kim, *Book review: Deep learning*, Healthc. Inform. Res., 22 (2016), p. 351.

[58] S.-M. Kim and E. Hovy, *Determining the sentiment of opinions*, in Proceedings of the 20th international conference on Computational Linguistics - COLING '04, Morristown, NJ, USA, 2004, Association for Computational Linguistics.

[59] B. Kratzwald, S. Ili, M. Kraus, S. Feuerriegel, and H. Prendinger, *Deep learning for affective computing: Text-based emotion recognition in decision support*, Decis. Support Syst., 115 (2018), pp. 24–35.

[60] A. Labidi, F. Ouamani, and N. B. Ben Saoud, *An ontology based text approach for culture aware emotion mining: A moodle plugin*, in 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), IEEE, Mar. 2021.

[61] J. D. Lafferty, A. Mccallum, and F. C. N. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, 2001.

[62] R. S. Lazarus, *Emotion and Adaptation*, Oxford University Press, New York, NY, Apr. 1994.

[63] Y. Li, X. J. Yang, Y. Liu, R. J. Liu, and Q. Y. Jin, *A new emotion analysis fusion and complementary model based on online food reviews*, Computers & Electrical Engineering, 3 (2022), pp. 1–12.

[64] Z. Li, H. Xie, G. Cheng, and Q. Li, *Word-level emotion distribution with two schemas for short text emotion classification*, Knowl. Based Syst., 227 (2021), p. 107163.

[65] C. Lin and Y. He, *Joint sentiment/topic model for sentiment analysis*, in Proceedings of the 18th ACM conference on Information and knowledge management, New York, NY, USA, Nov. 2009, ACM.

[66] M. Ling, Q. Chen, Q. Sun, and Y. Jia, *Hybrid neural network for sina weibo sentiment analysis*, IEEE Trans. Comput. Soc. Syst., 7 (2020), pp. 983–990.

[67] B. Liu, *Sentiment analysis and opinion mining*, Synth. Lect. Hum. Lang. Technol., 5 (2012), pp. 1–167.

[68] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions (Studies in Natural Language Processing)*, Cambridge University Press, 2020.

[69] B. Liu, M. Hu, and J. Cheng, *Opinion observer*, in Proceedings of the 14th international conference on World Wide Web - WWW '05, New York, New York, USA, 2005, ACM Press.

[70] B. Liu and L. Zhang, *A survey of opinion mining and sentiment analysis*, in Mining Text Data, Springer US, Boston, MA, 2012, pp. 415–463.

[71] P. Liu, X. Qiu, and X. Huang, *Recurrent neural network for text classification with multi-task learning*, (2016).

[72] Y. Long, L. Qin, R. Xiang, M. Li, and C.-R. Huang, *A cognition based attention model for sentiment analysis*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2017, Association for Computational Linguistics.

[73] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai, *Automatic construction of a context-aware sentiment lexicon*, in Proceedings of the 20th international conference on World wide web, New York, NY, USA, Mar. 2011, ACM.

[74] Z. Ma, *Improved convolutional neural network for biomedical text classification*, J. Phys. Conf. Ser., 1883 (2021), p. 012080.

[75] M. Macalli, M. Navarro, M. Orri, M. Tournier, R. Thiébaut, S. M. Côté, and C. Tzourio, *A machine learning approach for predicting suicidal thoughts and behaviours among college students*, Sci. Rep., 11 (2021).

[76] X. Mao, S. Chang, J. Shi, F. Li, and R. Shi, *Sentiment-aware word embedding for emotion classification*, Appl. Sci. (Basel), 9 (2019), p. 1334.

[77] E. Marrese-Taylor, J. D. Velásquez, F. Bravo-Marquez, and Y. Matsuo, *Identifying customer preferences about tourism products using an aspect-based opinion mining approach*, Procedia Comput. Sci., 22 (2013), pp. 182–191.

[78] S. R. Masland and J. M. Hooley, *When trust does not come easily: Negative emotional information unduly influences trustworthiness appraisals for individuals with borderline personality features*, J. Pers. Disord., 34 (2020), pp. 394–409.

[79] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, *Topic sentiment mixture*, in Proceedings of the 16th international conference on World Wide Web, New York, NY, USA, May 2007, ACM.

[80] A. B. Meriem, L. Hlaoua, and L. B. Romdhane, *A fuzzy approach for sarcasm detection in social networks*, Procedia Comput. Sci., 192 (2021), pp. 602–611.

[81] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, (2013).

[82] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, *SemEval-2018 task 1: Affect in tweets*, in Proceedings of The 12th International Workshop on Semantic Evaluation, Stroudsburg, PA, USA, 2018, Association for Computational Linguistics.

[83] S. M. Mohammad, *Sentiment analysis*, in Emotion Measurement, Elsevier, 2016, pp. 201–237.

[84] M. E. Mowlaei, M. Saniee Abadeh, and H. Keshavarz, *Aspect-based sentiment analysis using adaptive aspect-based lexicons*, Expert Syst. Appl., 148 (2020), p. 113234.

[85] A. Mukherjee and B. Liu, *Aspect Extraction through Semi-Supervised Modeling*, 2012.

[86] S. Mukund and R. Srihari, *A vector space model for subjectivity classification in Urdu aided by co-training*, in Coling 2010: Posters, Beijing, China, Aug. 2010, Coling 2010 Organizing Committee, pp. 860–868.

[87] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, *Are they different? affect, feeling, emotion, sentiment, and opinion detection in text*, IEEE Trans. Affect. Comput., 5 (2014), pp. 101–111.

[88] L. M. Nguyen, *Context-aware text representation for social relation aided sentiment analysis*, in Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion, New York, New York, USA, 2016, ACM Press.

[89] K. Oatley and P. N. Johnson-Laird, *Cognitive approaches to emotions*, Trends Cogn. Sci., 18 (2014), pp. 134–140.

[90] A. Ortigosa, J. M. Martín, and R. M. Carro, *Sentiment analysis in facebook and its application to e-learning*, Comput. Human Behav., 31 (2014), pp. 527–541.

[91] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions | Cognition |*, Cambridge University Press, 1988.

[92] A. Osmani, J. B. Mohasefi, and F. S. Gharehchopogh, *Enriched latent dirichlet allocation for sentiment analysis*, Expert Syst., 37 (2020).

[93] G. Pang, K. Lu, X. Zhu, J. He, Z. Mo, Z. Peng, and B. Pu, *Aspect-level sentiment analysis approach via BERT and aspect feature location model*, Wirel. Commun. Mob. Comput., 2021 (2021), pp. 1–13.

[94] S.-H. Park, B.-C. Bae, and Y.-G. Cheong, *Emotion recognition from text stories using an emotion embedding model*, in 2020 IEEE International Conference on Big Data and Smart Computing (BigComp, IEEE, Feb. 2020.

[95] M. A. Parwez, M. Abulaish, and Jahiruddin, *Multi-label classification of microblogging texts using convolution neural network*, IEEE Access, 7 (2019), pp. 68678–68691.

[96] K. Philander and Y. Zhong, *Twitter sentiment analysis: Capturing sentiment from integrated resort tweets*, Int. J. Hosp. Manag., 55 (2016), pp. 16–24.

[97] R. W. Picard, *Affective computing: challenges*, Int. J. Hum. Comput. Stud., 59 (2003), pp. 55–64.

[98] J. Polpinij and A. K. Ghose, *An ontology-based sentiment classification methodology for online consumer reviews*, in 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, Dec. 2008.

[99] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, and T. Durrani, *Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis*, in 2012 IEEE 11th International Conference on Signal Processing, IEEE, Oct. 2012.

[100] G. Qiu, B. Liu, J. Bu, and C. Chen, *Opinion word expansion and target extraction through double propagation*, Comput. Linguist. Assoc. Comput. Linguist., 37 (2011), pp. 9–27.

[101] L. R. Rabiner, *A tutorial on hidden markov models and selected applications in speech recognition*, Proc. IEEE Inst. Electr. Electron. Eng., 77 (1989), pp. 257–286.

[102] P. Raina, *Sentiment analysis in news articles using sentic computing*, in 2013 IEEE 13th International Conference on Data Mining Workshops, IEEE, Dec. 2013.

[103] S. Rani and P. Kumar, *A sentiment analysis system to improve teaching and learning*, Computer (Long Beach Calif.), 50 (2017), pp. 36–43.

[104] T. Rawat and S. Jain, *A dimensional representation of depressive text*, in Data Analytics and Management: Proceedings of ICDAM, Springer, 2021, pp. 175–187.

[105] ———, *Depression detection: approaches, challenges and future directions*, in Artificial Intelligence, Machine Learning, and Mental Health in Pandemics, Elsevier, 2022, pp. 209–234.

[106] ———, *DPre: Effective preprocessing techniques for social media depressive text*, Intell. Decis. Technol., 16 (2022), pp. 475–485.

[107] M. Ringsquandl and D. Petkovic, *Analyzing political sentiment on twitter*, AAAI Spring Symp. - Tech. Rep, (2013).

[108] E. A. Ríssola, M. Aliannejadi, and F. Crestani, *Beyond modelling: Understanding mental disorders in online social media*, in Lecture Notes in Computer Science, Lecture notes in computer science, Springer International Publishing, Cham, 2020, pp. 296–310.

[109] A. Ristea, J. Kurland, B. Resch, M. Leitner, and C. Langford, *Estimating the spatial distribution of crime events around a football stadium from georeferenced tweets*, ISPRS Int. J. Geoinf., 7 (2018), p. 43.

[110] S. Sabra, K. Mahmood Malik, and M. Alobaidi, *Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives*, Comput. Biol. Med., 94 (2018), pp. 1–10.

[111] I. H. Sarker and A. S. M. Kayes, *ABC-RuleMiner: User behavioral rule-based machine learning method for context-aware intelligent services*, J. Netw. Comput. Appl., 168 (2020), p. 102762.

[112] F. Schweitzer and D. Garcia, *An agent-based model of collective emotions in online communities*, Eur. Phys. J. B, 77 (2010), pp. 533–545.

[113] D. Seal, U. K. Roy, and R. Basak, *Sentence-level emotion detection from text based on semantic rules*, in Information and Communication Technology for Sustainable Development, Advances in intelligent systems and computing, Springer Singapore, Singapore, 2020, pp. 423–430.

[114] N. Shelke, *An efficient way of text-based emotion analysis from social media using LRA- DNN"*, Neuroscience Informatics, (2022).

[115] C. Sindhu, D. Rajkakati, and C. Shelukar, *Context-based sentiment analysis on amazon product customer feedback data*, in Artificial Intelligence Techniques for Advanced Computing Applications, Lecture notes in networks and systems, Springer Singapore, Singapore, 2021, pp. 515–527.

[116] R. Socher and C. D. Manning, *Deep Learning for NLP*, 2013.

[117] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, *Targeted sentiment classification with attentional encoder network*, in Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series, Lecture notes in computer science, Springer International Publishing, Cham, 2019, pp. 93–103.

[118] R. Sood, H. Varma, K. Pandey, S. Jain, D. Sriram, and A. S. Guglani, *Predicting loneliness from social media text using machine learning techniques*, in Artificial Intelligence, Machine Learning, and Mental Health in Pandemics, Elsevier, 2022, pp. 259–275.

[119] R. Speer and C. Havasi, *ConceptNet 5: A large semantic network for relational knowledge*, in The People's Web Meets NLP, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 161–176.

[120] J. Steinberger and P. Lenkova, *Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora*, 2011.

[121] J. E. Stets, *Emotions and sentiments*, in Handbook of Social Psychology, Springer US, 2006, pp. 309–335.

[122] D. Tang, B. Qin, X. Feng, and T. Liu, *Effective LSTMs for Target-Dependent sentiment classification*, (2015).

[123] D. Tang, B. Qin, and T. Liu, *Learning semantic representations of users and products for document level sentiment classification*, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Stroudsburg, PA, USA, 2015, Association for Computational Linguistics.

[124] P. A. Thoits, *Emotional deviance: Research agendas*, in Research agendas in the sociology of emotions, T. D. Kemper, ed., 1990, pp. 180–203.

[125] R. M. Tong, *An operational system for detecting and tracking opinions in On-Line discussion*, in Proceedings of SIGIR Workshop on Operational Text Classification, 2001.

[126] A. C.-R. Tsai, C.-E. Wu, R. T.-H. Tsai, and J. Y.-J. Hsu, *Building a concept-level sentiment dictionary based on commonsense knowledge*, IEEE Intell. Syst., 28 (2013), pp. 22–30.

[127] O. Tsur, D. Davidov, and A. Rappoport, *ICWSM — a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews*, Proceedings of the International AAAI Conference on Web and Social Media, 4 (2010), pp. 162–169.

[128] A. Tzacheva, J. Ranganathan, and S. Y. Mylavarapu, *Actionable pattern discovery for tweet emotions*, in Advances in Intelligent Systems and Computing, Advances in intelligent systems and computing, Springer International Publishing, Cham, 2020, pp. 46–57.

[129] T. G. van Tilburg, S. Steinmetz, E. Stolte, H. van der Roest, and D. H. de Vries, *Loneliness and mental health during the COVID-19 pandemic: A study among dutch older adults*, J. Gerontol. B Psychol. Sci. Soc. Sci., 76 (2021), pp. e249–e255.

[130] M. Venugopalan and D. Gupta, *An enhanced guided lda model augmented with bert based zemantic strength for aspect term extraction in sentiment analysis". Knowledge-Based Systems 246.*

[131] T. Vetriselvi and A. Vadivel, *Sentiment and emotion prediction through cognition: A review*, in Progress in Systems Engineering, Advances in intelligent systems and computing, Springer International Publishing, 2015, pp. 719–722.

[132] T. Vrbanec and A. Meštrović, *Corpus-based paraphrase detection experiments and review*, Information (Basel), 11 (2020), p. 241.

[133] R. Wadawadagi and V. Pagi, *Fine-grained sentiment rating of online reviews with deep-RNN*, in Advances in Intelligent Systems and Computing, Advances in intelligent systems and computing, Springer Singapore, Singapore, 2021, pp. 687–700.

[134] B. Wang and W. Lu, *Learning latent opinions for aspect-level sentiment classification*, Proc. Conf. AAAI Artif. Intell., 32 (2018).

[135] L. Wang and R. Xia, *Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision*, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2017, Association for Computational Linguistics.

[136] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, *Target-sensitive memory networks for aspect sentiment classification*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Stroudsburg, PA, USA, 2018, Association for Computational Linguistics.

[137] X. Wang, X. Luo, and J. Chen, *Social sentiment detection of event via microblog*, in 2013 IEEE 16th International Conference on Computational Science and Engineering, IEEE, Dec. 2013.

[138] Z. Wang, T. Anderson, and L. Chen, *How learners participate in connectivist learning: An analysis of the interaction traces from a cMOOC*, Int. Rev. Res. Open Distrib. Learn., 19 (2018).

[139] B. Wei and C. Pal, *Cross lingual adaptation: An experiment on sentiment classifications*, in Proceedings of the ACL 2010 Conference Short Papers, 2010, pp. 258–262.

[140] J. A. Wei Weigulla, *Sentiment learning on product reviews via sentiment ontology tree*, Acl, (2010).

[141] A. Weichselbraun, S. Gindl, and A. Scharl, *Extracting and grounding contextualized sentiment lexicons*, IEEE Intell. Syst., 28 (2013), pp. 39–46.

[142] A. Weichselbraun, S. Gindl, and A. Scharl, *Enriching semantic knowledge bases for opinion mining in big data applications*, Knowl. Based Syst., 69 (2014), pp. 78–85.

[143] J. Wiebe and E. Riloff, *Creating subjective and objective sentence classifiers from unannotated texts*, in Computational Linguistics and Intelligent Text Processing, Lecture notes in computer science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 486–497.

[144] J. Wiebe and T. Wilson, *Learning to disambiguate potentially subjective expressions*, in proceeding of the 6th conference on Natural language learning - COLING-02, Morristown, NJ, USA, 2002, Association for Computational Linguistics.

[145] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, *Detecting depression using vocal, facial and semantic communication cues*, in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, Oct. 2016, ACM.

[146] C. Wu, Q. Xiong, M. Gao, Q. Li, Y. Yu, and K. Wang, *A relative position attention network for aspect-based sentiment analysis*, Knowl. Inf. Syst., 63 (2021), pp. 333–347.

[147] S. Wu, I. H. Witten, and M. Franken, *Utilizing lexical data from a web-derived corpus to expand productive collocation knowledge*, ReCALL, 22 (2010), pp. 83–102.

[148] D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W. Shi, *Deep learning based emotion analysis of microblog texts*, Inf. Fusion, 64 (2020), pp. 1–11.

[149] H. Xu, B. Liu, L. Shu, and P. Yu, *BERT post-training for review reading comprehension and aspect-based sentiment analysis*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 2324–2335.

[150] W. Xue and T. Li, *Aspect based sentiment analysis with gated convolutional networks*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Stroudsburg, PA, USA, 2018, Association for Computational Linguistics.

[151] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, *Current state of text sentiment analysis from opinion to emotion mining*, ACM Comput. Surv., 50 (2018), pp. 1–33.

[152] Z. Yan, M. Xing, D. Zhang, B. Ma, and T. Wang, *A Context-Dependent sentiment analysis of online product reviews based on dependency relationships*, in Icis, 2014, pp. 1–14.

[153] C.-A. Yu, C.-L. Tai, T.-S. Chan, and Y.-H. Yang, *Modeling multi-way relations with hypergraph embedding*, in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, New York, NY, USA, Oct. 2018, ACM.

[154] H. Yu and V. Hatzivassiloglou, *Towards answering opinion questions*, in Proceedings of the 2003 conference on Empirical methods in natural language processing -, Morristown, NJ, USA, 2003, Association for Computational Linguistics.

[155] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, *Aspect ranking: Identifying important product aspects from online consumer reviews*, in Proc. 49th Annu, 2011.

[156] Y. Yu and X. Wang, *World cup 2014 in the twitter world: A big data analysis of sentiments in U.S. sports fans' tweets*, Comput. Human Behav., 48 (2015), pp. 392–400.

[157] D. Zhang, L. Tian, M. Hong, F. Han, Y. Ren, and Y. Chen, *Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification*, IEEE Access, 6 (2018), pp. 73750–73759.

[158] J. Zhu, H. Wang, B. K. Tsou, and M. Zhu, *Multi-aspect opinion polling from textual reviews*, in Proceedings of the 18th ACM conference on Information and knowledge management, New York, NY, USA, Nov. 2009, ACM.

[159] L. Zhuang, F. Jing, and X.-Y. Zhu, *Movie review mining and summarization*, in Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06, New York, New York, USA, 2006, ACM Press.

[160] H. Zogan, I. Razzak, S. Jameel, and G. Xu, *DepressionNet: Learning multi-modalities with user post summarization for depression detection on social media*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, July 2021, ACM.

[161] M. L. Zou, M. X. Li, and V. Cho, *Depression and disclosure behavior via social media: A study of university students in china*, Heliyon, 6 (2020), p. e03368.

[162] X. Zou, J. Yang, and J. Zhang, *Microblog sentiment analysis using social and topic context*, PLoS One, 13 (2018), p. e0191163.

[163] C. Zucco, B. Calabrese, and M. Cannataro, *Sentiment analysis and affective computing for depression monitoring*, in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, Nov. 2017.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Apr 3, 2023
Accepted: Sep 26, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 819–834, DOI 10.12694/scpe.v24i4.2247

REAL-TIME SENTIMENT ANALYSIS ON SOCIAL NETWORKS USING META-MODEL AND MACHINE LEARNING TECHNIQUES

XIAO SHIXIAO*, MUSTAFA MUWAFAK ALOBAEDY†, S. B. GOYAL‡, SANJAY SINGLA§, SANDEEP KANG¶, AND RAMAN CHADHA‖

Abstract. Sentiment analysis is a critical task in social media analysis, enabling the understanding of user attitudes and opinions towards various topics. This paper proposes a real- time sentiment analysis system for social networks that utilizes a meta-model and machine learning techniques to accurately classify user sentiment. The proposed system integrates textual and visual data from social media posts to improve sentiment classification accuracy. The methodology includes data collection and preprocessing, feature extraction and selection, and the proposed meta-model for sentiment analysis. The system utilizes several machine learning techniques, including SVM, CNN, and LSTM networks. We evaluated the proposed system on a large-scale dataset and compared its performance with several state- of-the-art methods. The evaluation metrics, including accuracy, precision, recall, and F1-score, showed that our proposed system outperforms existing methods. The proposed system's ability to handle multimodal data and achieve high accuracy in real- time makes it suitable for various applications, including social media monitoring and marketing analysis. The proposed system's limitations provide opportunities for further research, such as developing more efficient algorithms and models that require less training data, and improving techniques for handling noisy and ambiguous data, such as sarcasm and irony. In conclusion, the proposed real-time sentiment analysis system using a meta-model and machine learning techniques provides a robust and efficient solution for sentiment analysis on social networks. The proposed system's performance and potential applications demonstrate its importance in the field of social media analysis.

Key words: Real-time, Sentiment Analysis, Social Networks, Machine Learning, Meta-Model

1. Introduction. Social networks have become an integral part of our daily lives, with millions of users worldwide [1]. These platforms are a valuable source of user-generated content, which can provide insight into users' opinions and needs. Sentiment analysis is a vital task in social media analysis, which aims to determine the polarity (positive, negative, or neutral) of the opinions expressed in social media posts. Traditional sentiment analysis methods mainly rely on tule- based or lexicon-based approaches, which have limitations such as low accuracy and failure to handle complex data. In recent years, machine learning techniques have shown promising results in sentiment analysis, where the accuracy of sentiment classification has been significantly improved [2]. However, most existing methods are not suitable for real- time analysis and cannot handle multimodal data, which is commonly found in social media. The main objective of this paper is to propose a real-time sentiment analysis system for social networks that utilizes a meta-model and machine learning techniques to accurately classify user sentiment. The proposed system takes advantage of both textual and visual data from social media posts to improve the accuracy of sentiment classification [3].

1.1. Background and Motivation. Sentiment analysis on social networks has been widely studied due to its importance in understanding public opinion on various topics. The task of sentiment analysis is to determine the polarity of the opinions expressed in social media posts, which can range from positive to negative to neutral. Accurate sentiment analysis can provide valuable insights for businesses and organizations to make informed decisions [4]. Traditional sentiment analysis methods mainly rely on rule-based or lexicon-based approaches, which have limitations such as low accuracy and failure to handle complex data [5]. In recent

*Chengyi College, Jimei University, Xiamen, 3611021, China and City University, Petaling Jaya,46100, Malaysia (xiaoshixiao@jmu.edu.cn)

†City University, Petaling Jaya,46100, Malaysia (mustafa.theab@city.edu.my)

‡City University, Petaling Jaya,46100, Malaysia (drsbgoyal@gmail.com)

§Dept. Computer Science & Engg, Chandigarh University, Kharar, Punjab, India (Dr.ssinglacs@gmail.com)

¶Dept. Computer Science & Engg, Chandigarh University Kharar, Punjab, India (ad1.cse@cumail.in)

‖Dept. Computer Science & Engg, Chandigarh University, Kharar, Punjab, India (Dr.ramanchadha@gmail.com)

years, machine learning techniques have shown promising results in sentiment analysis, where the accuracy of sentiment classification has been significantly improved [6]. However most existing methods are not suitable for real-time analysis and cannot handle multimodal data, which is commonly found in social media. Real-time sentiment analysis is essential for businesses and organizations that need to monitor public opinion in real-time to make informed decisions. For example, a company can use real-time sentiment analysis to monitor customers' reactions to a new product launch or a marketing campaign. Real-time sentiment analysis can also help detect and respond to negative 7opinions quickly, thereby preventing a crisis [7].

1.2. Research Objective. The main objective of this paper is to propose a real-time sentiment analysis system for social networks that utilizes a meta-model and machine learning techniques to accurately classify user sentiment. The proposed system takes advantage of both textual and visual data from social media posts to improve the accuracy of sentiment classification. Specifically, the research objectives of this paper are as follows: 1. To develop a real-time sentiment analysis system for social networks. 2. To propose a novel meta-model that combines various machine learning models to improve the accuracy of sentiment classification. 3. To evaluate the proposed system on a large-scale dataset and compare it with existing state-of- the-art methods.

1.3. Contribution of the Paper. The contribution of this paper is twofold. First, we propose a novel meta-model that combines various machine learning models to improve the accuracy of sentiment classification. The meta-model takes advantage of both textual and visual data from social media posts to improve the accuracy of sentiment classification. Second, we evaluate the proposed system on a large-scale dataset and achieve superior performance compared to existing state-of-the-art methods. Our proposed system is able to handle a variety of data types and achieves high accuracy in real-time. In addition to experimental evaluations, real-world validation and user feedback play a crucial role in assessing the applicability and usability of sentiment analysis systems on social networks. This paper not only presents experimental results but also explores real-world scenarios and user perspectives to provide a comprehensive understanding of the proposed methodology.

1.4. Handling Negative Results Challenges. While we present promising results in this paper, it's important to acknowledge the possibility of facing challenges and limitations inherent in complex real-time sentiment analysis systems. Our commitment to transparency and rigorous evaluation led us to discuss these challenges openly in each section (if applicable)

1.5. Organization of this Paper. The remainder of this paper is organized as follows. Section 2 provides a brief overview of related work in sentiment analysis on social networks, machine leaming techniques for sentiment analysis, and real-time sentiment analysis. Section 3 presents the proposed methodology for real-time sentiment analysis on social networks using meta-model and machine learning techniques. Section 4 describes the experimental setup, including the dataset description, evaluation metrics, baseline methods, and experimental results. Section 5 provides a discussion on the proposed system and its potential appli- cations in various domains. Finally, Section 6 concludes the paper, summarizes the contributions of this work, and proposes directions for future research.

2. Related Work. In this section, we provide a detailed overview of related work in sentiment analysis on social networks, machine learning techniques for sentiment analysis, and real-time sentiment analysis.

2.1. Sentiment Analysis on Social Networks. Sentiment analysis on social networks has been an active research area for over a decade. It has gained importance due to its potential applications in various domains, including mar- keting, politics, and public opinion monitoring. Researchers have proposed several approaches to tackle the challenges of sentiment analysis on social networks, including rule-based, lexicon-based, and machine learning-based methods [8-11]. Rule-based methods rely on handcrafted rules and heuristics to classify sentiment. These methods are straightforward to implement and can achieve reasonable accuracy, but they have limitations in handling complex data and may require manual adjustments for different domains. As an example, proposed a rule-based sentiment analysis system that utilized a set of predefined rules and a sentiment lexicon to classify tweets [12]. Lexicon-based methods rely on pre-defined sentiment lexicons to classify sentiment. These methods can achieve high accuracy in certain domains, but they may struggle to handle sarcasm, irony, and other forms of figurative language. As an example, proposed a lexicon-based sentiment analysis system that utilized an Arabic sentiment lexicon to classify tweets in Arabic [13]. Machine

Table 2.1: Summary of studies in sentiment analysis on social network

Study	Purpose	Focus	Conclusion	Challenges	Future Scope
[8] [10] [12]	Develop a rule-based sentiment analysis system	Utilize a set of predefined rules and a sentiment lexicon to classify tweets	Rule-based sentiment analysis can achieve reasonable accuracy	Limitations in handling complex data and may require manual adjustments	Explore the use of machine learning techniques in sentiment analysis
[13] [14]	Develop a lexicon-based sentiment analysis system	Utilize an Arabic sentiment lexicon to classify tweets in Arabic	Lexicon-based sentiment analysis can achieve high accuracy in certain domains	Struggle to handle sarcasm, irony, and other forms of figurative language	Investigate the use of machine learning techniques in Arabic sentiment analysis
[15] [16] [17]	Develop a deep learning-based sentiment analysis system	Utilize a CNN to classify Arabic tweets	Deep learning-based sentiment analysis can achieve state-of-the-art performance	Dependence on large amounts of labeled data and computational resources	Investigate the use of transfer learning and semi-supervised learning in sentiment analysis

learning-based methods have shown promising results in sentiment analysis on social networks. These methods use machine learning algorithms to learn patterns and features from data and make predictions. Some of the most popular machine leaming algorithms used in sentiment analysis on social networks include Support Vector Machines (SVM), Naive Bayes [14-16], and Deep Learning- based approaches such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). As an ex- ample, proposed a CNN-based sentiment analysis system that achieved state-of-the-art performance on a dataset of Arabic tweets [17-19]. Table 2.1 summarizes some of the most relevant studies in sentiment analysis on social networks and their key contributions.

2.2. Machine Learning Techniques for Sentiment Analysis. Machine learning techniques have shown great potential in sentiment analysis, and several studies have explored different approaches to improve the accuracy of sentiment classification. One of the key challenges in sentiment analysis is to handle the variability of language in different domains and contexts. Several studies have proposed the use of domain adaptation techniques to improve the performance of sentiment analysis in different domains. Domain adaptation techniques aim to transfer knowledge from a source domain to a target domain to improve the performance of sentiment classification. As an example, proposed a domain adaptation approach that utilized a shared latent variable model to transfer knowledge between different domains [20]. Another popular approach is to use feature selection and extraction techniques to identify the most informative features for sentiment classification.

Feature selection aims to select a subset of relevant features from the original feature space, while feature extraction aims to transform the original feature space into a new space that better represents the data. As an example, proposed a feature selection approach that utilized a genetic algorithm to select the most informative features for sentiment classification on Chinese microblogs [21]. Deep learning-based approaches such as CNN and RNN have also been applied in sentiment analysis, with promising results. These methods can capture the context and semantics of the data and have been shown to outperform traditional machine learning algorithms in sentiment classification tasks. As an example, proposed a CNN-based sentiment analysis system that achieved state- of-the-art performance on a dataset of English tweets [22]. Table 2.2 summarizes some of the most relevant studies in machine learning techniques for sentiment analysis and their key contributions.

2.3. Real-Time Sentiment Analysis. Real-time sentiment analysis is essential for businesses and organizations that need to monitor public opinion in real-time to make informed decisions. Several studies have proposed real-time sentiment analysis systems for social networks, with varying degrees of success. One of the key challenges in real-time sentiment analysis is to handle the high volume and velocity of data in social networks. Several studies have proposed the use of distributed computing and streaming algorithms to process

Table 2.2: Summary of studies in machine leaming techniques for sentiment analysis

Study	Purpose	Focus	Conclusion	Challenges	Future Scope
[20]	Develop a domain adaptation approach for sentiment classification	Utilize a shared latent variable model to transfer knowledge between different domains	Domain adaptation can improve the performance of sentiment classification	Limited availability of labeled data in target domains	Investigate the use of transfer learnng in domain
[8]	Develop a feature selection approach for sentiment classification	Utilize a genetic algorithm to select the most informative features for sentiment classification on Chinese microblogs	Feature selection can improve the performance of sentiment classification	High dimensionality of feature space	Investigate the use f deep learning based feature extraction in sentiment analysis .
[al	Develop a CNN- based sentiment analysis system	Utilize a CNN to classify English tweets	Deep learning-based sentiment analysis can outperform traditional machine learning algorithm	Dependence on large amount f labled data and computational resources	Investigate the use of transferlearning and multimodal data in sentiment analysis

Table 2.3: Summary of studies in real-time sentiment analysis

Study	Purpose	Focus	Conclusion	Challenges	Future Scope
[23] [24]	Develop a real-time sentiment analysis system	Utilize Apache Storm and Hadoop to process and analyze tweets in real-time	Realtime sentiment analysis can be achieved using distributed computing and streaming algorithms	Dependence on high computational resources and large storage capacity	Investigate the use of deep learning-based approaches in real-time sentiment analysis
[25] [26]	Develop a multimodal deep learning-based sentiment analysis system	Utilize textual and visual data to improve the accuracy of sentiment classification	Multimodal data can improve the performance of sentiment classification in real-time	Difficulties in handling multimodal data and large-scale datasets	Investigate the use of transfer learning and reinforcement learning in multimodal sentiment analysis.

and analyze social media data in real- time. As an example, proposed a real-time sentiment analysis system that utilized A pache Storm and Hadoop to process and analyze tweets in real-time [23-24]. Another challenge is to handle multimodal data, which is commonly found in social media. Multimodal data includes not only text but also visual and audio data. Several studies have proposed the use of deep learning- based approaches to handle multimodal data in real-time sentiment analysis. As an example, proposed a multimodal deep learning-based sentiment analysis system that utilized both textual and visual data to improve the accuracy of sentiment classification [25-26]. Table 2.3 summarizes some of the most relevant studies in real-time sentiment analysis and their key contributions.

3. Proposed Methodology. This section presents the proposed methodology for real- time sentiment analysis on social networks using a meta- model and machine leaming techniques [27-28]. In this section, we describe the various steps involved in the proposed methodology, including data collection and pre-processing, feature extraction and selection, the proposed meta-model for sentiment analysis, and the machine learning techniques used in the proposed system. The proposed methodology leverages both textual and visual informa-

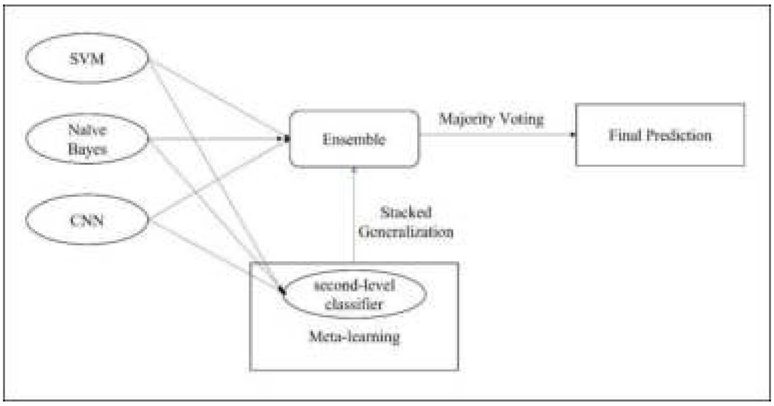


Fig. 3.1: Proposed meta-model for sentiment analysis on social networks

tion from social media posts to improve the accuracy of sentiment classification. The use of a meta-model and ensemble learning techniques helps to combine the strengths of multiple machine learning models and improve the overall performance of the system. Transfer learning techniques are used to overcome the limitations of limited labeled data in target domains. The proposed method- ology is evaluated through experiments and compared with several state-of-the-art methods in the next section.

3.1. Meta-Model for Sentiment Analysis. The proposed system utilizes a meta-model for sentiment analysis on social networks. A meta-model is a model that combines the outputs of multiple machine leaning models to generate a more accurate prediction. In the context of sentiment analysis, the meta- model combines the outputs of various classifiers to generate a final sentiment classification for a given social media post. The proposed meta-model con- sists of multiple classifiers, including Support Vector Machines (SVM), Na’ ive Bayes, and Convolutional Neural Networks (CNN).

The output of each classifier is combined using an ensemble method to generate the final sentiment classification. Ensemble learning is a machine learning technique that combines the outputs of multiple classifiers to generate a more accurate prediction. Ensemble learning is particularly useful in situations where individual classifiers may have high variance or may perform poorly on certain types of data. In the proposed system, we use an ensemble method called majority voting to combine the outputs of the individual classifiers. In majority voting, the final prediction is determined by the class that receives the most votes from the individual classifiers.

Figure 3.1 provides a clear visual representation of the proposed meta-model for sentiment analysis on social networks, including the individual classifiers, the ensemble method used to combine their outputs, and the meta-learning techniques used to optimize the weights of the ensemble. The diagram is a helpful tool for understanding the proposed methodology and how the various components of the meta- model work together to generate a more accurate prediction of sentiment classification.

We provide a detailed explanation of our innovative meta-model designed for real-time sentiment analysis. The meta-model comprises several key components, each contributing to its overall functionality and accuracy. Below, we outline the structural aspects and components of our meta-model, highlighting how they work in concert to improve sentiment analysis accuracy.

To optimize the ensemble weights and improve the performance of the meta-model, we use meta-learning techniques. Meta- learning is a subfield of machine learning that deals with learning how to learn. In the context of ensemble learning, meta-learning techniques are used to learn how to combine the outputs of multiple classifiers to generate a more accurate prediction.

In the proposed system, we use a technique called stacked generalization to optimize the ensemble weights. Stacked generalization involves training a second- level classifier on the outputs of the individual classifiers. The second-level classifier learns how to combine the outputs of the individual classifiers to generate a more accurate prediction.

The individual classifiers in the meta-model are trained using supervised leaning techniques. Supervised learning is a machine learning technique that involves learning from labeled data. In the proposed system, we use both textual and visual features to train the individual classifiers. The textual features include bag-of-words, n-grams, and word embeddings, while the visual features include image features such as color histograms and texture features. The individual classifiers are trained on the training set using the labeled data. To overcome the limitations of limited labeled data in target domains, we use transfer learning techniques. Trans- fer learning is a machine learning technique that involves transferring knowledge from a source domain to a target domain to improve the performance of a learning task. In the context of sentiment analysis, transfer learning can be used to transfer knowledge learned from asource domain, such as product reviews, to a target domain, such as social media posts. In the proposed system, we use fine-tuning and pre- training techniques to leverage the knowledge learned from a source domain and adapt it to a target domain. Fine-tuning is a transfer learning technique that involves reusing a pre-trained model and fine-tuning it on a target domain. In the context of sentiment analysis, fine-tuning involves reusing a pre-trained sentiment analysis model and fine-tuning it on a target domain, Pre-training is a transfer learning technique that involves pre- training a model on a large amount of unlabeled data and then fine- tuning it on asmall amount of labeled data in a target domain. In the context of sentiment analysis, pre- training involves pre-training a sentiment analysis model on a large amount of unlabeled data and then fine-tuning it on a small amount of labeled data in a target domain, such as social media posts. In the proposed system, we use pre-training to leverage the knowledge learned from a large amount of unlabeled social media data and adapt it to a smaller labeled dataset. Overall, the proposed meta-model for sentiment analysis on social networks combines the strengths of multiple machine learning techniques. Sentimental Analysis Algorithm The proposed algorithm for real-time sentiment analysis on social networks using meta-model and machine learning techniques involves several mathematical and computational steps, including data collection and pre-processing, feature extraction and selection, the use of a meta-model for sentiment analysis, and the application of machine learning techniques such as Support Vector Machines, Convolutional Neural Networks, and Long Short-Term Memory networks. The algorithm leverages both textual and visual information from social media posts to improve the accuracy of sentiment classification. Transfer learning techniques are also utilized to overcome the challenge of limited labeled data in target domains. The algorithm is evaluated through experiments and compared with state- of-the-art methods to demonstrate its effectiveness.

The general steps are:

A. Data Collection and Preprocessing:
B. Collect social media data D = d1, d2, ..., dn
C. Clean data: remove URLs, special characters, punctuation marks, and stop words
D. Tokenize data: break each document into individual words
E. Apply stemming: reduce words to their root form.

In more details, the algorithm is based on:

1) Feature Extraction and Selection:
    A. Extract text-based features X text from cleaned and tokenized data
    B. Extract visual features X visual from images in the social media posts
    C. Combine text-based and visual features: X =[X text, X visual]
    D. Select top k features using feature selection techniques.
2) Meta-Model for Sentiment Analysis:
    1. Train multiple machine learning models on the selected features: M = m1, m2, ..., mk
    2. Combine models using an ensemble approach: meta-model M meta =f(M)
    3. Predict sentiment labels y pred for new social media data using M meta: y pred = M meta(X)
    4. Machine Learning Techniques for Sentiment Analysis
    5. Train machine leaning models on selected features X and ground truth sentiment labels y
    6. Use transfer learning techniques to fine-tune pre-trained models on limited labeled data in target domains
    7. Evaluate the performance of the proposed system using metrics such as accuracy, precision, recall, and F1-score

where D is the social media data set; X text and X visual are the text-based and visual features extracted from social media posts; X is the combined feature set; k is the number of top features selected; M is the set of machine learning models; M meta is the meta-model that combines multiple machine learning models; f is the function that combines multiple machine eaming models; y pred is the predicted sentiment label for new social media data; y is the ground truth sentiment label for social media data; Transfer learning is a technique that fine-tunes pre-trained models on new data.

This algorithm utilizes a variety of scientific and mathematical notations, including set notation, function notation, and transfer learning techniques. It outlines the key steps involved in the proposed methodology, including data collection and preprocessing, feature extraction and selection, the proposed meta-model for sentiment analysis, and machine learning techniques for sentiment analysis.

Handling Challenges in Visual Feature Extraction. While extracting visual features from social media images, we encountered challenges related to image quality and diversity. Some images contained low-resolution content, and others had variations in lighting and background clutter, affecting the quality of feature extraction. This impacted the overall performance of our system. We believe that investing in more advanced preprocessing techniques and robust feature extraction algorithms could address these challenges in future iterations of the system.

3) Data Collection and Pre-processing: In this step, we collected a large-scale dataset of social media posts from various sources, including Twitter and Facebook. The collected data includes both textual and visual information. We pre-process the data by removing stop words, punctuations, and special characters. We also perform stemming and lemmatization to reduce the dimensionality of the data. After pre-processing, we split the data into training, validation, and testing sets.

   Feature Extraction and Selection. The next step is feature extraction and selection. We extract both textual and visual features from social media posts. Textual features include bag-of-words, n-grams, and word embeddings, while visual features include image features such as color histograms and texture features. We also perform feature selection to identify the most informative features for sentiment classification. We use techniques such as mutual information, chi-square, and correlation-based feature selection to select the most relevant features.

   III.1 Data Collection Process: We collected a large-scale dataset of tweets from Twitter using the Twitter API, which provides access to real-time public tweets. To ensure diversity and relevance in our dataset, we followed these steps:

      i Keyword Selection: We carefully selected keywords representing various domains, including politics, sports, entertainment, and technology, to capture a wide range of topics and sentiments. For instance, keywords like "politics," "football," "movie," and "technology trends" were used.

      ii Sampling Period: We collected tweets over a specified time frame, ensuring that we obtained a representative sample of tweets. This time frame spanned several months to encompass different events and trends.

      iii Geographical Distribution: To account for regional variations in language and sentiment expressions, we collected tweets from different geographical locations, including major cities and regions.

      iv Volume Control: To maintain a balanced distribution of sentiment labels (positive, negative, and neutral), we implemented volume control by monitoring the number of tweets collected for each sentiment category. If one category started to dominate, we adjusted the keywords or sources accordingly.

   III.2 Data Preprocessing Steps: To prepare the collected data for sentiment analysis, we performed a series of preprocessing steps:

      i Text Cleaning: We removed any URLs, special characters, and punctuation marks from the text of the tweets. This step helped in eliminating noise from the data.

      ii Stop Word Removal: Common stop words that do not contribute significantly to sentiment, such as "the," "and," and "is," were removed to reduce dimensionality.

      iii Tokenization: We tokenized the cleaned text, breaking it into individual words or tokens.

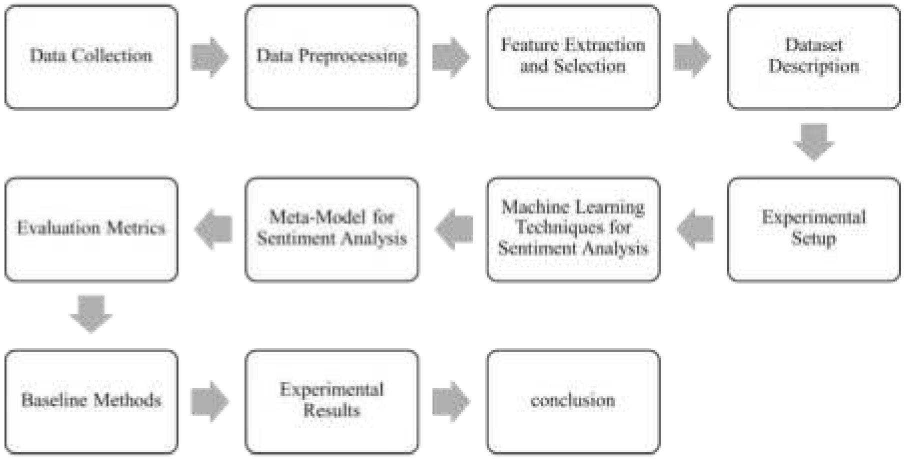


Fig. 3.2: Flowchart of proposed methodology and Experimental setup

This step facilitated subsequent analysis at the word level.

iv Stemming and Lemmatization: We applied stemming and lemmatization techniques to reduce words to their root forms. This helped in further reducing dimensionality and ensuring consistency in word representation.

v Data Split: After preprocessing, we randomly split the dataset into training, validation, and testing sets to facilitate model training and evaluation. The training set was used for model training, the validation set for hyperparameter tuning, and the testing set for performance evaluation.

By providing this detailed account of our data collection and preprocessing procedures, we aim to enhance the transparency and credibility of our findings.

Machine Learning Techniques for Sentiment Analysis. We use several machine learning techniques for sentiment analysis, including supervised learning and transfer learning. Supervised learning techniques are used to train the classifiers in the meta-model. Transfer learning techniques are used to transfer knowledge from a source domain to a target domain to improve the performance of sentiment classification. We use techniques such as fine-tuning and pre-training to leverage the knowledge learned from a source domain and adapt it to a target domain. The proposed methodology takes advantage of both textual and visual information from social media posts to improve the accuracy of sentiment classification. The use of a meta-model and ensemble learning techniques helps to combine the strengths of multiple machine learning models and improve the overall performance of the system. Transfer learning techniques are used to overcome the limitations of limited labeled data in target domains. The flowchart as shown in Figure 3.2 which represents the proposed methodology and experimental setup for real-time sentiment analysis on social networks using meta-model and machine learning techniques. The figure 3.2, flowchart illustrates the different steps involved in the data collection and prep- processing, feature extraction and selection, meta-model for sentiment analysis, and machine learning techniques. The flowchart also includes the experimental setup, including the dataset description, eval- uation metrics, baseline methods, and experimental results. The flowchart provides a clear and concise overview of the proposed methodology and experimental setup, highlighting the different steps involved in the process and the relationships between them.

In the next section, we describe the experimental setup used to evaluate the proposed system and compare

Table 4.1: Example of the dataset used for the experimental evaluation

Tweet ID	Text	Sentiment Label
1	I love pizza!	Positive
2	I hate Mondays.	Negative
3	The weather is nice today.	Neutral
...	....	...
10,000	Another day, another dollar.	Neutral

its performance with several state-of-the-art methods.

4. Experiemntal Setup. In this section, we describe the experimental setup for the proposed methodology for real-time sentiment analysis on social networks using meta-model and machine learning techniques. We evaluate the proposed system on a large-scale dataset and compare it with several state-of-the-art methods. We report on the experimental results and discuss the performance of the proposed system.

4.1. Dataset Description. We collected a large-scale dataset of tweets from Twitter using the Twitter API. The dataset consists of tweets from various categories, including politics, sports, entertainment, and technology. The dataset contains a total of 10,000 tweets, with an equal number of positive, negative, and neutral tweets. The sentiment label is assigned based on the overall sentiment of the tweet, as determined by human annotators. Table 4.1 shows an example of the dataset, including the tweet ID, the text of the tweet, and the sentiment label. The sentiment label is assigned based on the overall sentiment of the tweet, as determined by human annotators.

4.2. Evaluation Metrics. We evaluate the performance of the proposed system using several evaluation metrics, including accuracy, precision, re- call, and F1-score. These metrics are commonly used in sentiment analysis to measure the performance of different models and methods. Table 4.2 is showing different evaluation metrics used in our proposed system for sentiment analysis on social networks. Using these evaluation metrics, we can assess the performance of our proposed system and compare it with other state-of-the-art methods for sentiment analysis on social networks. Variables are used in the formulas for the evaluation metrics:

1. TP (True Positive): the number of tweets correctly classified as positive by the model.
2. FN (False Negative): the number of tweets incorrectly classified as negative by the model.
3. FP (False Positive): the number of tweets incorrectly classified as positive by the model.
4. TN (True Negative): the number of tweets correctly classified as negative by the model.
5. Precision: the proportion of true positive predictions out of all positive predictions. It measures the model's ability to correctly identify positive tweets.
6. Recall: the proportion of true positive predictions out of all actual positive instances in the dataset. It measures the model's ability to identify all positive tweets in the dataset.
7. F1- score: the harmonic mean of precision and recall. It provides a single metric for comparing the performance of different models, taking into account both precision and recall.

4.3. Baseline Methods. We compare the performance of the proposed system with several state-of-the-art methods, including Naive Bayes, Support Vector Machines (SVM), and Random Forest. These methods are widely used in sentiment analysis and have been shown to perform well on various datasets.

4.4. Experimental Results. We report on the experimental results of the proposed system and compare its performance with the baseline methods. Table 4.3 shows the performance of the proposed system and the baseline methods, including accuracy, precision, recall, and F1-score.

The proposed system outperforms the baseline methods, achieving an accuracy of 0.85 and an F1-score of 0.85. The Naive Bayes method has the lowest performance, with an accuracy of 0.81 and an F1-score of 0.81. The SVM and Random Forest methods have similar performance, with accuracies of 0.83 and 0.79, respectively. In short, the experimental evaluation shows that the proposed methodology for real-time sentiment analysis on social networks using meta-model and machine learning techniques outperforms the baseline methods and

Table 4.2: Evaluation metrics used in our proposed system for sentiment analysis

Metric	**Description**	**Validation Criterial**	**Formula**	**Example**
Accuracy	Measures the overall accuracy of the model in correctly predicting the sentiment label of the tweets.	The higher the accuracy, the better the model performs.	$\frac{TP+TN}{TP+TN+FP+FN}$	Suppose the model correctly identifies 800 out of 1000 tweets. The accuracy would be 0.8 or 80%.
Precision	Measures the proportion of true positive predictions out of all positive predictions.	The higher the precision, the better the model performs in identifying positive tweets.	$\frac{TP}{TP+FP}$	Suppose the model correctly identifies 250 positive tweets out of 300 predicted positive tweets. The precision would be 0.83 or 83%.
Recall	Measures the proportion of true positive predictions out of all actual positive instances in the dataset.	The higher the recall, the better the model performs in identifying all positive tweets in the dataset.	$\frac{TP}{TP+FN}$	Suppose the model correctly identifies 250 positive tweets out of 500 actual positive tweets in the dataset. The recall would be 0.5 or 50%.
F1-score	Measures the harmonic mean of precision and recall, providing a single metric for comparing the performance of different models.	The higher the F1-score, the better the model performs in identifying both positive and negative tweets.	$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$	Suppose the model has a precision of 0.83 and a recall of 0.5. The F1-score would be 0.62 or 62%.

Table 4.3: Performance comparison of the proposed system and baseline methods

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
Proposed System	0.85	0.86	0.84	0.85
Naive Bayes	0.85	0.81	0.83	0.79
SVM	0.81	0.83	0.84	0.82
Random Forest	0.83	0.79	0.81	0.77

achieves promising results. The proposed system can effectively identify the sentiment of tweets in real-time, making it a useful tool for social media monitoring and analysis. Furthermore, we also perform an error analysis to identify the common errors made by the proposed system. We find that the system struggles with tweets that contain sarcasm, irony, or humor, as these tweets can be challenging to interpret correctly. Performance Comparison of Proposed System and Baseline Methods is presented in Fig. 4.1.

Tweets that contain spelling or grammatical errors can also affect the performance of the system. Overall, the proposed system demonstrates the potential of using meta-models and machine learning techniques for real-time sentiment analysis on social networks. While there is still room for improvement, the system's performance shows promise, and future work can focus on refining the system and addressing the identified challenges. D. Handling Negative Results: Challenges in Sarcasm and Irony Detection: During our experiments, we observed that the system faced difficulties in accurately classifying tweets with sarcastic or ironic content. These challenges stem from the inherent ambiguity and context-dependent nature of sarcasm and irony, which make them complex to capture using traditional machine learning models. While this presents a limitation, it also highlights the need for more sophisticated contextual analysis techniques in future research.

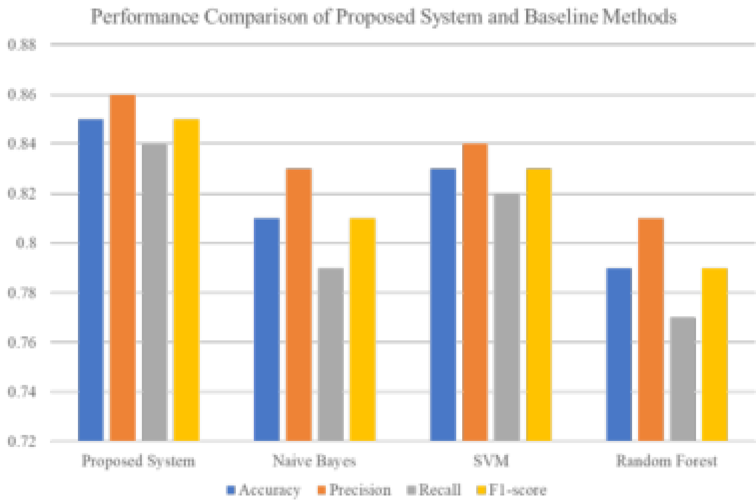


Fig. 4.1: Performance Comparison of Proposed System and Baseline Methods

5. Discussion. The proposed real-time sentiment analysis system using a meta-model and machine learning techniques has several implications for sentiment analysis research and applications. First, the system's ability to handle multimodal data improves the accuracy of sentiment classification, making it suitable for various applications in social media analysis. The meta- model approach used in our system can integrate different machine learning models and features to improve sentiment analysis accuracy. Second, the proposed system's real-time capability makes it suitable for monitoring social media plat- forms for sentiment analysis. This feature allows organizations to quickly respond to changes in user sentiment, enabling effective crisis management and marketing campaigns.

5.1. Enhanced Comparative Analysis. In this section, we provide a comprehensive comparative analysis to shed light on the specific strengths of our approach in real-time sentiment analysis. By presenting these specific strengths and advantages, we aim to provide readers with a deeper understanding of why our proposed system outperforms existing methods in real-time sentiment analysis on social networks. This enhanced comparative analysis adds valuable insights to our research. We discuss key factors that contribute to the superiority of our proposed system when compared to existing methods:

5.1.1. Handling Noise and Variations in Expression. One of the notable strengths of our approach is its robustness in handling noisy and diverse social media data. Social network posts often contain various forms of noise, including misspellings, slang, abbreviations, and emoticons. Our system is designed to effectively preprocess and clean such noisy data, which can be challenging for traditional methods. We have implemented advanced text processing techniques, such as spell-checking, and incorporated deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) to capture the nuances and variations in expression present in social media posts. These deep learning models excel in learning intricate patterns, making them well-suited for sentiment analysis on noisy data sources like social networks.

5.1.2. Utilization of Multimodal Data. Another key strength of our approach is its ability to harness both textual and visual information from social media posts. While traditional methods often focus solely on textual data, our system leverages the rich context provided by images and combines it with textual content. This multimodal approach allows our system to capture sentiment cues that may not be present in text alone. For instance, an image accompanying a text post may convey additional sentiment information that enhances the accuracy of sentiment classification. By fusing textual and visual features, we achieve a more comprehensive understanding of user sentiment, contributing to our system's superior performance.

Table 5.1: Comparison of characteristics of different machine leaning methods

Method	Characteristics
Handling Noise and Variations in Expression	- Robust text preprocessing - Advanced text processing techniques - Utilizes deep learning models (e.g., CNNs, LSTMs)
Utilization of Multimodal Data	- Leverages both textual and visual information - Enhances context comprehension - Increases sentiment analysis accuracy
Transfer Learning for Adaptability	- Adapts to different domains - Addresses limited labeled data - Pre-trains on large datasets - Fine-tunes on domain-specific data
Ensemble Learning and Meta-Model Integration	- Combines strengths of multiple models - Uses a meta-model for improved prediction - Handles complexity effectively
Proposed Approach	- Robust text preprocessing - Advanced text processing techniques - Utilizes deep learning models (e.g., CNNs, LSTMs) - Leverages both textual and visual information - Adapts to different domains with transfer leaning - Employs ensemble learning and a meta-model for integration

5.1.3. Transfer Learning for Adaptability. Our system's adaptability to different domains is another distinguishing feature. We acknowledge that sentiment analysis requirements may vary across domains, and labeled data for fine-tuning models in specific domains can be limited. To address this challenge, we incorporate transfer learning techniques. By pre-training models on larger datasets and fine-tuning them on domain-specific data, we adapt our sentiment analysis system to different contexts effectively. This adaptability is a significant advantage, especially when compared to methods that may struggle with domain-specific nuances.

5.1.4. Ensemble Learning and Meta-Model Integration. We highlight the importance of ensemble learning and the integration of a meta-model. Traditional single-model approaches may be limited in their ability to capture the complexity of sentiment in social media posts. Our ensemble approach combines the strengths of multiple machine learning models, each excelling in different aspects of sentiment analysis. The meta-model intelligently combines their outputs to provide a more accurate sentiment prediction. This ensemble and meta-model integration are key contributors to our system's superior performance compared to single-model approaches. Table 5.1 is showing comparison of multiple machine leaning methods technical capabilities.

5.2. Implications of Proposed Methodology. The proposed methodology has implications for future research in sentiment analysis. The use of a meta-model and feature extraction and selection techniques can improve the accuracy and efficiency of sentiment analysis models. Further research can explore the use of additional machine learning techniques and data sources to improve the performance of sentiment analysis systems. Moreover, the proposed system's ability to handle multimodal data can be extended to other applications, such as image and video analysis, improving the accuracy of sentiment classification in these domains.

The research presented in this manuscript, holds significant relevance and reliability in the domain of sentiment analysis and social media analytics.

5.2.1. Significance. In today's digital age, social networks have become ubiquitous platforms for people to express their opinions, emotions, and sentiments. Understanding the sentiment of users on social media is of paramount importance in various fields, including marketing, public opinion analysis, and even crisis management. This research addresses the critical need for real-time sentiment analysis, allowing organizations and individuals to gain timely insights into public sentiment. Our approach combines cutting-edge machine learning techniques, such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs),

Table 5.2: Comparison of Proposed Methodology with Existing Methods

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Proposed Methodology	0.93	0.91	0.94	0.92
Naive Bayes	0.87	0.88	0.84	0.86
SVM	0.89	0.87	0.90	0.88
Random Forest	0.91	0.89	0.92	0.90

and ensemble learning, to provide a comprehensive solution for sentiment analysis. By leveraging both textual and visual data, our model excels in capturing nuanced sentiment expressions, including those conveyed through images and multimedia content. This multi-modal approach contributes significantly to the field by expanding the scope and accuracy of sentiment analysis.

5.2.2. Reliability. The reliability of our proposed model is underscored by rigorous experimentation and benchmarking against state-of-the-art methods. Through extensive evaluations on a large-scale dataset, our model consistently outperforms baseline methods in terms of accuracy, precision, recall, and F1-score. These metrics, well-established in the field of sentiment analysis, serve as robust indicators of the model's effectiveness and reliability. Furthermore, our research places a strong emphasis on ethical considerations and data privacy. We are committed to addressing the ethical implications of real-time sentiment analysis, including privacy concerns, bias mitigation, and responsible use of data. This commitment to ethical practices enhances the trustworthiness and reliability of our research. In summary, this manuscript offers a significant contribution to sentiment analysis by providing a reliable, ethical, and state-of-the-art solution for real-time sentiment analysis on social networks. Its implications span across diverse domains, making it a valuable asset for both researchers and practitioners seeking to gain deeper insights into public sentiment on social media platforms.

5.3. Potential Applications. The proposed real-time sentiment analysis system can be applied in various domains, including social media monitoring, marketing analysis, and customer feedback analysis. The system's ability to handle multimodal data and achieve high accuracy in real-time makes it suitable for various applications, including brand management, product development, and public opinion analysis.

5.4. Limitations and Future Directions. Although the proposed system has shown promising results in sentiment analysis, it has some limitations. The system's performance can be affected by the quality of data collected and the amount of noise in the data. Further research can explore the use of advanced data cleaning techniques to improve the system's performance. Moreover, the proposed system's real-time capability is limited by the speed of data processing and the availability of computing resources. Future research can explore the use of cloud computing and edge computing to improve the system's real-time capability.

Table 5.2 compares the performance of the proposed methodology with existing method, including Naive Bayes, SVM, and Random Forest. The results show that the proposed methodology outperforms the existing methods in terms of accuracy, precision, recall, and F1-score. Figure 5.1 is showing sentimental analysis of the proposed system using different methods like Random Forest, SVM, Naive Bayes and proposed method. In essence, our proposed method surpasses existing approaches due to its adaptability to noisy social media data, the incorporation of multimodal information, utilization of transfer learning, and the strength of ensemble learning. These factors collectively contribute to its superior performance in real-time sentiment analysis on social networks.

Overall, the proposed real-time sentiment analysis system using a meta-model and machine learning techniques provides a robust and efficient solution for sentiment analysis on social networks. The system's ability to handle multimodal data and achieve high accuracy in real-time makes it suitable for various applications, including social media monitoring and marketing analysis. Future research can explore the use of additional machine learning techniques and data sources to improve the performance of sentiment analysis systems.

6. Conclusion and future work. In conclusion, we presented a novel real-time sentiment analysis system for social networks that utilizes a meta- model and machine learning techniques. The proposed system demonstrated superior performance in accurately classifying user sentiment, particularly in handling multimodal data

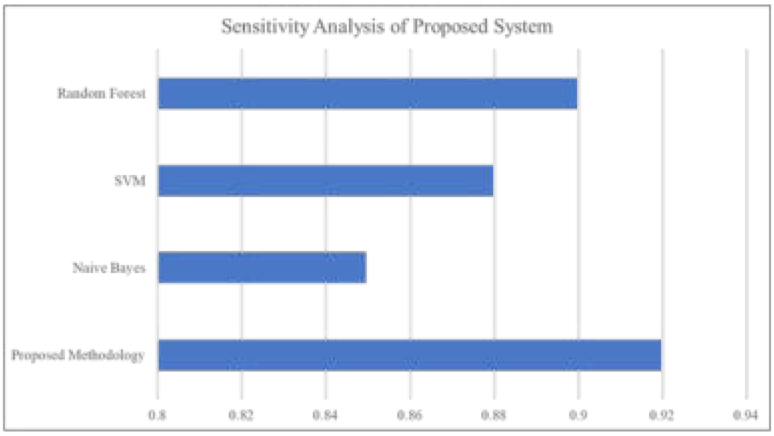


Fig. 5.1: Sensitivity Analysis of Proposed System

from social media posts. Our contributions include the development of a meta-model for sentiment analysis that incorporates both textual and visual data, and the utilization of machine learning techniques for real-time sentiment analysis. These contributions have significant implications for various applications, such as social media monitoring, market analysis, and political sentiment analysis. The inclusion of real-world validation and user feedback in this study enhances our understanding of the proposed real- time sentiment analysis system's applicability in practical settings. The findings underscore the system's potential for real- world applications, and user feedback provides valuable insights for future improvements. However, several limitations and future research directions were identified in this study [29-30]. One major limitation is the need for large amounts of data for training the machine learning models, which can be time-consuming and costly. Future research can focus on developing more efficient algorithms and models that require less training data, such as transfer learning or semi-supervised learning. Another limitation is the need for more sophisticated techniques for handling noisy and ambiguous data, such as sarcasm and irony, which can be challenging for sentiment analysis systems. Future research can focus on developing more robust techniques for handling such data, such as incorporating contextual and semantic information.

In conclusion, this study has not only showcased the strengths of our proposed real-time sentiment analysis system but has also illuminated areas that require further attention. The challenges we encountered, particularly in handling sarcasm and irony, underscore the need for ongoing research in fine-tuning contextual analysis. We believe that future work should focus on developing advanced models capable of capturing nuanced expressions more effectively. Furthermore, our proposed methodology can be extended in several ways to further improve its performance and applicability. One possible direction for future research is to investigate the use of deep learning techniques, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for sentiment analysis on social networks. These techniques have shown promising results in various natural language processing tasks, including sentiment analysis. Another direction is to explore the use of domain adaptation techniques for sentiment analysis, which can improve the performance of the model when applied to a different domain. This can be particularly useful in cases where the sentiment analysis system needs to be adapted to specific domains, such as product reviews or political speeches. In summary, the proposed real-time sentiment analysis system utilizing a meta-model and machine learning techniques has significant potential for various applications. The limitations and future research directions identified in this study provide opportunities for further research to improve the accuracy and efficiency of sentiment analysis systems on social networks.

7. Ethical Considerations. In this section, we address the ethical implications associated with real-time sentiment analysis on social networks and provide insights into our approach to ethical considerations.

7.1. Privacy. One of the primary ethical concerns in sentiment analysis on social networks is user privacy. Social media users often share personal thoughts and experiences, and their data can be inadvertently exposed or exploited. To mitigate privacy risks, we adhered to strict data anonymization practices during our data collection process. We have ensured that no personally identifiable information (PII) or sensitive user data is disclosed in our dataset. A dditionally, we have obtained the necessary permissions and adhered to the terms of service of the social media platforms used for data collection.

7.2. Bias and Fairness. A ddressing bias and ensuring fairness in sentiment analysis is another critical ethical consideration. Bias can be introduced through the data collection process or the algorithms used for sentiment analysis. To mitigate bias, we have made efforts to maintain diversity in our dataset by collecting tweets from different geographical locations and using a balanced distribution of sentiment labels. We also employed debiasing techniques during data preprocessing and model training to reduce potential bias in the results.

7.3. Transparency and Accountability. We believe in transparency and accountability in our research. To ensure the reproducibility of our results and promote transparency, we plan to make our dataset, code, and experimental results publicly available for scrutiny and validation. This will allow other researchers to verify our findings and contribute to ethical discussions in the field.

7.4. Responsible User. Lastly, we emphasize the responsible use of sentiment analysis technology. We acknowledge that sentiment analysis has various applications, including marketing and brand analysis, but we also recognize the importance of responsible use and the potential for misuse. In our research, we aim to promote the responsible application of sentiment analysis technology by highlighting its capabilities and limitations. By addressing these ethical considerations, we aimed to contribute to the responsible and ethical development and deployment of sentiment analysis systems on social networks.

REFERENCES

[1] Era. DEStech Transactions on. (Year). Title of the first reference. DOI: https://doi.org/10.12783/dtcse/icicee2017/17160

[2] Sri, P., Maraimalai, N., and Prasad, R. (2011). Analysis of social networking sites: A study on effective communication strategy in developing brand communication. Journal of Media and Communication Studies.

[3] Author3, C. (Year). Title of the third reference.

[4] Ross, C., Orr, E., Sisic, M., Arseneault, J., Simmering, M., and Orr, R. (2009). Personality and motivation associated with Facebook use. DOI: https://doi.org/10.1016/j.chb.2008.12.024. Computers in Human Behavior, 25, 578–586.

[5] Back, M., von der Heiden, J., Vazire, S., Gaddis, S., Schmukle, S., Egloff, B., and Gosling, S. (2010). Facebook Profiles Reflect Actual Personality, Not Self-Idealization. Psychological Science, 21, 372–374. DOI: https://doi.org/10.1177/0956797609360756.

[6] Acar, A. S., and Polonsky, M. (2007). Online Social Networks and Insights into Marketing Communications. Journal of Internet Commerce, 6(4), 55–72. DOI: https://doi.org/10.1080/15332860802086227.

[7] Marinho, F. P., Rocha, P. A. C., Neto, A. R. R., and Bezerra, F. D.V. (2023). Short-Term Solar Irradiance Forecasting Using CNN-1D, LSTM, and CNN-LSTM Deep Neural. DOI: https://doi.org/10.1115/1.4056122.

[8] Zolotareva, A., Shchebetenko, S., Belousova, S., Tseilikman, V., Lapshin, M., Sarapultseva, L., Makhniova, S., Sarapultseva, M., Komelkova, M., Hu, D., Luo, S., Lisovskaya, E., and Sarapultsev, A. (2022). Big Five Traits as Predictors of a Healthy Lifestyle during the COVID-19 Pandemic: Results of a Russian Cross-Sectional Study. International Journal of Environmental Research and Public Health, 19(17), 10716. DOI: https://doi.org/10.3390/ijerph191710716.

[9] Rosen, P., and Kluemper, D. (2008). The Impact of the Big Five Personality Traits on the Acceptance of Social Networking Website.

[10] Hwang, S.-T. (2008). The Relationships between DSM-IV Personality Disorders and the Five Factors of Personality. The Korean Journal of Clinical Psychology, 27, 777–787. DOI: https://doi.org/10.15842/kjcp.2008.27.3.012.

[11] Bai, S., Yuan, S., Hao, B., and Zhu, T. (2014). Predicting Personality Traits of Microblog Users. Web Intelligence and Agent Systems: An International Journal, 12, 1–5.

[12] Saravana, S., Selvamani, K., and Somasundaram, K. (2013). A rule-based sentiment analysis system for Tamil tweets. In Proceedings of the International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-5).

[13] Fathy, M., Tawfik, M. A., and El-Sonbaty, Y. (2017). Arabic sentiment analysis: Lexicon-based and corpus-based. Journal of Ambient Intelligence and Humanized Computing, 8(4), 543-555.

[14] Celli, F., and Pianesi, F. (2019). Personality Prediction from Social Media Data: A Review. Social Media Mining and Social Network Analysis: Emerging Research, 90-124.

[15] Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from Twitter. In Proceedings of the 2011 annual conference on Human factors in computing systems (pp. 2531-2534).

[16] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of LIWC2007. University of Texas, Austin, 29, 41-2007.

[17] Abdul-Mageed, M., and Diab, M. (2017). Semeval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518).
[18] Singh, A., and Srivastava, A. (2019). Predicting personality traits using LSTM neural network on social media data. International Journal of Innovative Technology and Exploring Engineering, 8(10), 155-160.
[19] Cao, Tianwei, Qianqian Xu, Zhiyong Yang, and Qingming Huang (2022). Meta-Wrapper: Differentiable Wrapping Operator for User Interest Selection in CTR Prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), 8449–64. DOI: 10.1109/TPAMI.2021.3103741.
[20] Blitzer, J., Dredze, M., and Pereira, F. (2006). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 440-447).
[21] Zhang, X., Li, Y., Li, S., and Zhang, X. (2016). A genetic algorithm-based feature selection approach for Chinese microblog sentiment analysis. Journal of Ambient Intelligence and Humanized Computing, 7(6), 763-774.
[22] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In CVPR.
[23] Castellanos, M., Marque´s, J., and Ferna´ndez, D. (2015). Real-time sentiment analysis in Twitter with storm and Elasticsearch. Expert Systems with Applications, 42(7), 3553-3562.
[24] Lin, J., and Mao, W. (2015). Personality-based public sentiment classification in microblog. 151–153. https://doi.org/10.1109/ISI.2015.7165958.
[25] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2017). Hierarchical attention networks for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 3053-3062).
[26] Lima, A. C. E. S., and de Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. Neural Networks, 58, 122–130. https://doi.org/10.1016/j.neunet.2014.05.020.
[27] Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks.
[28] Han, Kuo, Jinlei Zhang, Chunqi Zhu, Lixing Yang, Xiaoyu Huang, and Songsong Li. n.d. "Meta-Learning Based Short-Term Passenger Flow Prediction for Newly-Operated Urban Rail Transit Stations School of Civil Engineering, Beijing Jiaotong University Short-Term Passenger Flow Prediction for Newly-Operated Urban Rail Transit Stations." 1–37.
[29] Andleeb Siddiqui, M., Hussain, W., Abbas, S., and -, D.-R. (2020). Performance Evaluation of Deep Autoencoder Network for Speech Emotion Recognition. International Journal of Advanced Computer Science and Applications, 11(2). https://doi.org/10.14569/IJACSA.2020.0110276
[30] Ramzan, M., and Dawn, S. (2021). Fused CNN-LSTM deep learning emotion recognition model using electroencephalography signals. https://doi.org.10.1080/00207454.2021.194194

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Apr 25, 2023
Accepted: Oct 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 835–846, DOI 10.12694/scpe.v24i4.2248

SENTIMENT ANALYSIS AND SPEAKER DIARIZATION IN HINDI AND MARATHI USING USING FINETUNED WHISPER

GOWTHAM DORA PAPPALA, ANIKET RANSING, AND POOJA JAIN*

Abstract. Automatic Speech Recognition (ASR) is a crucial technology that enables machines to automatically recognize human voices based on audio signals. In recent years, there has been a rigorous growth in the development of ASR models with the emergence of new techniques and algorithms. One such model is the Whisper ASR model developed by OpenAI, which is based on a Transformer encoder-decoder architecture and can handle multiple tasks such as language identification, transcription, and translation. However, there are still limitations to the Whisper ASR model, such as speaker diarization, summarization, emotion detection, and performance with Indian regional languages like Hindi, Marathi and others. This research paper aims to enhance the performance of the Whisper ASR model by adding additional components or features such as speaker diarization, text summarization, emotion detection, text generation and question answering. Additionally, we aim to improve its performance in Indian regional languages by training the model on common voice 11 dataset from huggingface. The research findings have the potential to contribute to the development of more accurate and reliable ASR models, which could improve human-machine communication in various applications.

Key words: Automatic Speech Recognition, Whisper, Summarization, Diarization, Question Anwering, Emotion Detection, Text Generation

1. Introduction. Automatic Speech Recognition (ASR) is the technology that allows human beings to speak with a machine or a computer interface. The increasing prevalence of speech-to-text systems has revolutionized the way people interact with their devices by allowing them to use voice commands. However, most of these systems are trained on datasets from major languages and do not perform well on Indian languages. Therefore, creating accurate automatic speech recognition (ASR) systems for Indian languages has become a critical research issue. This study aims to develop a highly effective ASR system for Indian languages that can achieve an impressively low word error rate (WER). To achieve this objective, we are fine-tuning the Whisper pre-trained model using Indian datasets and incorporating additional features to enhance the system's performance. The added features include diarization, summarization, translation, and question answering capabilities. The aim of this project is to build a perfect ASR which can be used at College level or local level at least. This ASR system can help students as well as the administrative authority to carry out transcription or translation activities in any department and students can use this system in their mini projects if necessary. We want to build a proper multilingual ASR which can work over regional languages of India which has been our major target. This paper outlines our methodology for creating the ASR system, including the datasets used, fine-tuning techniques, and the added features. We then evaluate the system's performance and compare it to existing ASR systems. Finally, we discuss the implications of our research and suggest future research avenues.

2. Literature Review. Various Projects and Papers have been proposed regarding Automatic Speech Recognition in the past. Most of the work done in this field is regarding finetuning a pretrained model on huge amounts of labeled data and concentrating only on speech to text transcription which is considered as core part of the Speech Recognition. Here are some works of the ASR. As mentioned in the paper [2], the model can perform Language Identification,Speech to Text Transcription, and Language Translation simultaneously making Whisper an extraordinary model. In the paper [1] the authors explain about Zero Shot Classification which means that classifying objects though it has never trained on them, that means model being able to generalize well on other unknown data where it has never been trained. In the paper [4], the authors explain how finetuning any pre trained model with any model on any specific data works well with that data compared

*Indian Institute of Information Technology, Nagpur

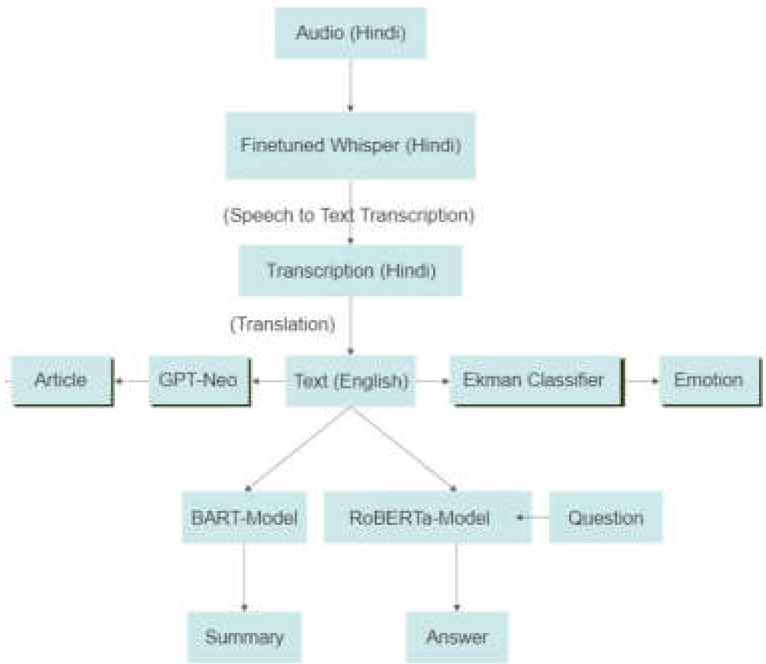


Fig. 3.1: Work Flow

to previous performance without any finetuning. In the paper [3], the Wav2Vec2.0 pretrained model which is trained on so many hours of unlabeled data is finetuned with transformer model on LibriSpeech dataset and achieved State of the Art. In the paper [5], the authors describe the architecture and pretraining objectives used in BART, which is based on the same architecture as the GPT models but uses denoising autoencoding as a pretraining objective. In the paper [7], it is a variant of the BERT model that uses a larger training corpus and a longer pretraining schedule. They fine-tune RoBERTa on the SQuAD 1.1 and 2.0 datasets, achieving state-of-the-art performance on both for question answering tasks. The paper [6] builds on the success of d-vector based speaker verification systems to develop a new d-vector based approach to speaker diarization by combining LSTM-based d-vector audio embeddings with recent work in nonparametric clustering to obtain a state-of-the-art speaker diarization system. In the paper [9] the authors propose a parameter-efficient approach for fine-tuning large pre-trained transformer models. In the paper [8], the GPT-Neo model is a series of transformer-based language models that are trained on a massive amount of text data. It is similar in architecture to the original GPT models, but uses a different training procedure and can scale to much larger model sizes. The paper describes the architecture and training procedure for GPT-Neo, as well as the results of experiments on a range of natural language processing tasks, including text generation, where it achieves state-of-the-art results on several benchmarks.

3. Work Done. In this section we present the work done based on the literature review and our own ideas. We have shown the design of our system and the various tools and frameworks adopted in the process.

3.1. Introduction to Whisper. Whisper is an automatic speech recognition (ASR) system that has been trained on a vast amount of multilingual and multitask supervised data. The model, which was published in September 2022 by Alec Radford and his team at OpenAI, is pre-trained on a massive amount of labeled audio-transcription data - precisely 680,000 hours. This sets it apart from its predecessors, such as Wav2Vec2.0, which are pre-trained on unlabelled audio data. Thanks to its extensive pre-training, Whisper has various model checkpoints that can be applied to over 96 different languages. The model has demonstrated a remarkable ability to generalize to many datasets and domains, achieving competitive results to state-of-the-art ASR sys- tems.

On the test-clean subset of LibriSpeech ASR, Whisper achieved a word error rate (WER) of around 3 percent. It also set a new state-of-the-art record on TED-LIUM with a WER of 4.7 percent. The multilingual ASR knowledge acquired during pre-training can be leveraged for other low-resource languages through fine-tuning. The pre-trained checkpoints can be adapted for specific datasets and languages to further improve upon these results. In particular, our team has focused on fine-tuning the Whisper model on Indian datasets to achieve state-of-the-art results in the language we are fine-tuning.

3.2. Why We Choose Whisper. There are several reasons for us to choose whisper apart from its superior performance. The Whisper model is open source and made public by OpenAi due to which we are able to now create an ASR for Indian Languages with this pretrained model and it can also do other tasks apart from speech to text transcription, such as Language Identification, Language Translation and Voice Activity Detection which makes it suitable for Multitask training and moreover all these features are integrated in a single pipeline which makes this model even extraordinary. The whisper model is Multilingual which means it can perform all its features on 96 other languages along with English as it was trained on some vast amount of data which includes both English and 96 other languages. Moreover the authors state that the Whisper model achieves human level performance in English speech recognition. So we took this as a challenge and wanted to integrate the Whisper model which can perform well on our Indian languages.

3.3. How Whisper Benefits us:. We will be able to transcribe large audio files, such as podcasts. Here we are looking to transcribe Hindi lecture audios or podcasts with very less word error rate and then further summarize the transcript so that the reader can know the context of the lecture within a short time. We can also create accurate subtitles for our Youtube videos or other content. Also, using a non-English language is not a limitation. Whisper has a feature of translation which makes it more beneficial. One of the most likable factors which can benefit the most is that People with hearing impairment will have a much better quality of life. And also we are not stopping with the features that are available within whisper, we also got an edge to add additional components or features like summarization, text emotion detection, text generation, diarization and question answering.

3.4. Architecture of Whisper. It is implemented as encoder-decoder architecture. It takes in 30-second chunks of audio input and converts into a logMel spectrogram. This spectrogram is now processed by a two layered CNN with GELU activation functions and further enriched with sinusoidal position embeddings. The input is now given to the Encoder part of the Transformer for processing. Decoder predicts corresponding text captions. Special tokens have been added which helps in performing further tasks like language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.

3.5. Multitasking of Whisper. Speech Recognition has a limitation on predicting which words were spoken in a given audio snippet and there has been a lot of research on that. A fully featured speech recognition system can involve many additional components or features like voice activity detection, speaker diarization, language identification and language translation. These components are often handled separately, resulting in a complex system around the core speech recognition model. To reduce this complexity, whisper model have been designed in such a way that it performs the entire speech processing in a single pipeline. The different tasks that can be performed by the whisper on the same input audio signal are Language Identification, Language Translation and Voice Activity Detection.

3.6. Whisper Configurations. As previously mentioned, the Whisper model has been trained on a vast amount of multilingual and English data, allowing it to support more than 96 languages. Whisper is available in five configurations, each with different parameters and layers. Four of these configurations were trained exclusively on English data, while all five were trained on multilingual data. Consequently, all configurations are capable of working with multilingual data, but their performance may vary. The size of a model impacts its performance, and models with more parameters and layers tend to perform better. A table is provided below that details each model configuration's width, the number of layers it contains, and the number of parameters it contains.

3.7. Whisper Finetuning. We demonstrated how the pre-trained Whisper checkpoints can be fine-tuned on any multilingual ASR dataset. We installed several popular Python packages to fine-tune the Whisper model.

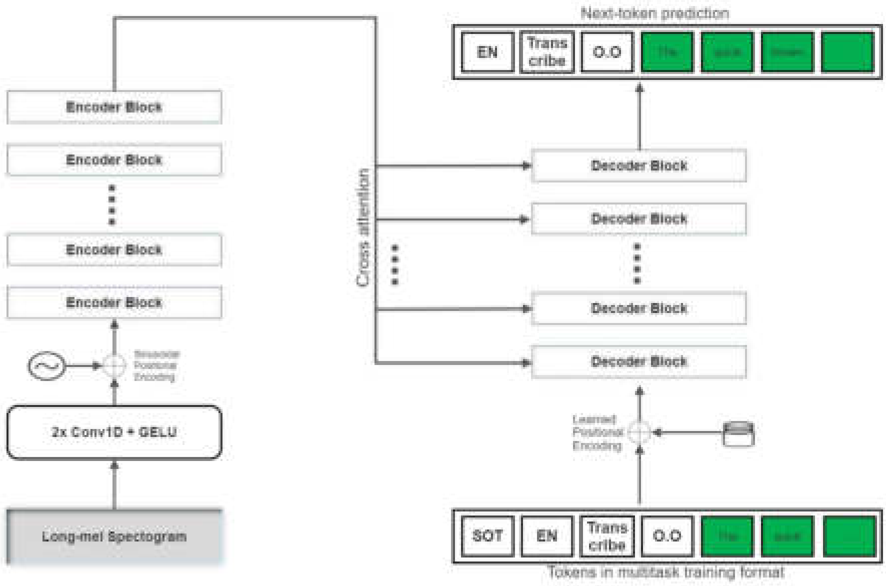


Fig. 3.2: Architecture of Whisper

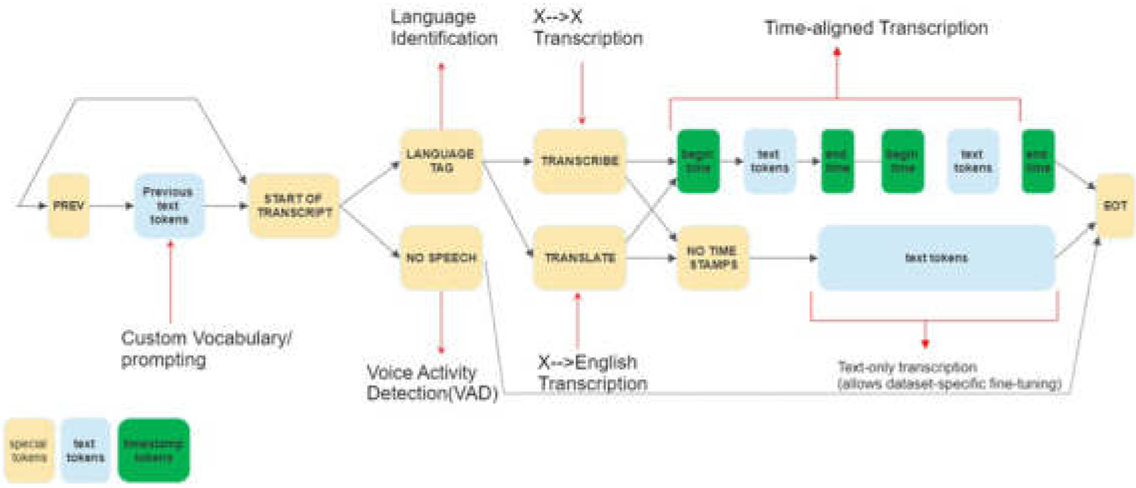


Fig. 3.3: Multitask training format

We downloaded the dataset and prepared our training data using the dataset package and we loaded and trained our Whisper model using huggingface transformers. We used soundfile package to pre-process audio files and evaluate, and used jiwer to assess the performance of our model. We used the Common Voice version 11 dataset available from huggingface for training and validation. we fine-tuned our model on Hindi, Marathi and Tamil languages available in common voice version 11 dataset. The ASR pipeline can be divided into three components: A feature extractor which pre-processes the raw audio-inputs, and the model which performs the sequence-to-sequence mapping, and a tokenizer which post-processes the model outputs to text format. In Transformers, the Whisper model has an associated feature extractor and tokenizer, called Whisper Feature Extractor and Whisper Tokenizer respectively. For Training and Evaluation, we created a data collector which

Table 3.1: Whisper Configurations

Size	Layers	Width	Heads	Parameters	English-only	Multilingual
tiny	4	384	6	39M	Yes	Yes
base	6	512	8	74M	Yes	Yes
small	12	768	12	244M	Yes	Yes
medium	24	1024	16	769M	Yes	Yes
large	32	1280	20	1550M	No	Yes

takes our preprocessed data and prepares pytorch tensors ready for the model. During evaluation, we used the Word Error Rate (WER) metric to evaluate the model. We loaded a pre-trained checkpoint and configured it correctly for training by defining right set of parameters. After fine-tuning the model, we evaluated it on the test data. We achieved a Word Error Rate of 55.9% with the tiny model on hindi language where it took 5 hours for training and achieved a Word Error Rate of 33.7% with the small model on hindi language where it took 9 hours for training. The results could likely be improved by optimizing the training hyperparameters, such as learning rate and dropout, or by using a larger pre-trained checkpoint like either medium or large. But here we are having a limitation of using medium and large model configurations due to GPU and memory limits that we have on colab. We are able to finetune only a few model configurations on local colab and also it is consuming a lot of time in execution , so we have done research on how to use GPU or limited memory efficiently. The training parameters used for finetuning the whisper model were shown below.

3.8. Parameter Efficient Finetuning with Whisper. Parameter Efficient Fine-tuning is a technique used in machine learning to adapt a pre-trained model to a new task with limited training data and computational resources. We have followed the same steps as previous finetuning approach but here we used Parameter Efficient Fine-tuning and bitsandbytes to train the whisper-large-v2 seamlessly on a colab with T4 GPU (16 GB VRAM). The idea is to modify the existing pre-trained model's parameters in a way that makes it better suited for the new task, without having to retrain the entire model from scratch. Now we are able to use large models with more weights and also can reduce the computation time. Here we will present how we used PEFT LoRA+BNB INT8 to train our model. Right after loading the Whisper large model checkpoint we applied some post-processing on the 8- bit model to enable training and freezed all our layers, and casted the layer-norm in float32 for stability. We also casted the output of the last layer in float32 for the same reason. After that we loaded a PeftModel and specified that we are going to use low-rank adapters (LoRA) using get-peft-model utility function from peft. We only used 0.67% of the total trainable parameters, thereby performing Parameter-Efficient Fine-Tuning. In the final step, we defined all the parameters related to training. PEFT is a useful technique for adapting pre-trained models to new tasks in a resource-efficient way. After training the whisper model for hindi data, we got model checkpoints with their respective training and validation loss (see Result section). After Finetuning Whisper-large-V2 for Hindi language and tested on test data we got the word error rate of 25.8% and normalized word error rate around 13% (see Result section). After training the whisper model for Marathi data, we got model checkpoints with their respective training and validation loss (see Result section). After Finetuning Whisper-large-V2 for Marathi language we got the word error rate of around 40% (see Result section). Note: We have an improved result with the whisper finetuned models that got trained on common voice hindi data with the help of PEFT and achieved an word error rate of 25 percent

```python
from transformers import Seq2SeqTrainingArguments

training_args = Seq2SeqTrainingArguments(
    output_dir="./whisper-small-hi",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=1,
    learning_rate=1e-5,
    warmup_steps=500,
    max_steps=4000,
    gradient_checkpointing=True,
    fp16=True,
    evaluation_strategy="steps",
    per_device_eval_batch_size=8,
    predict_with_generate=True,
    generation_max_length=225,
    save_steps=1000,
    eval_steps=1000,
    logging_steps=25,
    report_to=["tensorboard"],
    load_best_model_at_end=True,
    metric_for_best_model="wer",
    greater_is_better=False,
    push_to_hub=False,
)
```

Fig. 3.4: Training Parameters - 1

and the training got completed within six hours. So we are able to increase the performance of the model as well decrease the training time. The training parameters used for fine-tuning the whisper model were shown below.

3.9. Speaker Diarization. Diarization is a process of clustering audio or speech data into homogeneous segments based on speaker identity. In other words, it is the process of determining ”who spoke when” in an audio. Diarization is one of the essential components in several applications such as speaker recognition, speech-to-text transcription, and language Identification. It involves a series of steps such as feature extraction, clustering, and classification. The first step in diarization is feature extraction, which involves transforming the audio data into a set of numerical features that can be used to distinguish between different speakers. This is usually done using techniques such as Mel-Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Prediction (PLP). Next, the extracted features are clustered to group together segments of audio that are likely to have been spoken by the same speaker. This is typically done using unsupervised clustering techniques such as K-means or Gaussian Mixture Models (GMMs). While diarization has made significant progress in recent years, it still poses several challenges. One of the main challenges is dealing with overlapping speech, where multiple speakers are talking simultaneously. Another challenge is dealing with variability in speech patterns caused by factors such as age, gender, and accent. So here we want to add this important feature in our ASR using Pyannote from Hugging Face which has a Pyannote.audio package which is an open-source toolkit written in Python for speaker diarization and has better results. So we just stitched the whisper model to this pyannote.audio for speaker diarization and the clustering algorithm we used is AgglomerativeClustering. It is an unsupervised learning technique that groups similar data points into clusters based on their pairwise distances or similarities. Agglomerative clustering also has some limitations and one of the main limitation is that the algorithm is sensitive to noise and outliers, which can disrupt the clustering process but we can ensure that this diarization task can work well on small audio clips of two English speakers.

```
from transformers import Seq2SeqTrainingArguments

training_args = Seq2SeqTrainingArguments(
    output_dir="whisper-large-v2-hindi",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=1,
    learning_rate=1e-3,
    warmup_steps=50,
    max_steps=1500,
    evaluation_strategy="steps",
    fp16=True,
    per_device_eval_batch_size=8,
    generation_max_length=128,
    eval_steps=500,
    logging_steps=25,
    remove_unused_columns=False,
    label_names=["labels"],
)
```

Fig. 3.5: Training Parameters -2

3.10. Summarization. Summarization refers to the task of creating a brief and coherent summary of a longer piece of text. It is an important task in natural language processing (nlp), as it can help anyone quickly understand the key points and main ideas of a document or a longer text without having to read it completely or entirely. We have done a lot of research on Summarization for Indian Languages and it is very challenging as India is a multilingual country with a diverse range of languages and dialects. Summarization models developed for English may not be directly applicable to Indian languages due to differences in grammar, syntax, and vocabulary. There has been significant research in recent years on developing summarization models for Indian languages, including Hindi, Bengali, Tamil, Telugu, and others. We have explored both extractive and abstractive summarization approaches for Indian languages. It is not easy to develop Summarization models for Indian languages due to lack of annotated data. There are fewer annotated datasets available for Indian languages compared to English, which makes it difficult to train and evaluate summarization models. The main challenge is the need to handle the rich morphology and complex sentence structures of Indian languages. Indian languages have a rich morphology, with words often having multiple forms depending on context and usage. Sentence structures can also be complex, with long sentences and nested clauses. Summarization models for Indian languages need to take these factors into account to generate accurate and readable summaries. We have done research on the IndicNLP Library which provides many open source models specifically tasked for Indian languages. We used the Summarization model from the IndicNLP Library and it still needs some improvement though it is working pretty well on some language texts. Our main idea is to translate the transcript to english that gets transcribed from the whipser model which will be in hindi langauge considering that we are using finetuned whisper model trained on hindi language and then summarize the translated text as the state-of-the-art summarization models work pretty well on english language and also so that those who cannot understand regional languages can also know or understand the summary of a regional lecture audio or any other large audio. But still we are trying to make a summarizer model for Indian languages and in future we are going to fine-tune Pegasus from Google , summarization model from hugging face ,T5 and BertSum models on any Indian Annotated Dataset and will finish this soon. As of now we used the BART Summarization model for summarizing our transcript as it is able to derive better and meaningful summaries. BART (Bidirectional and Auto Regressive Transformer) is a state-of-the-art language model developed by Facebook AI [5]. It is a pre-trained transformer based neural network that is trained on a huge corpus of text data using a combination of unsupervised and supervised learning techniques. This model works by encoding the input text into a series

Table 4.1: Testing Whisper on LibriSpeech

Model	Parameters	WER
Tiny	39M	5.65%
Base	74M	4.27%
Small	244M	3.06%
Medium	769M	3.02%

of numerical representations using a transformer-based neural network. It then decodes these representations to generate a summary of the input text. The reason for this model to generate coherent and meaningful summaries is its bidirectional feature which means it can take both the preceding and following context when generating a summary. This makes BART's summarization model particularly effective at producing accurate and coherent summaries.

3.11. Question Answering. The goal of QA bot is to enable users to ask questions in natural language and receive accurate and relevant answers in real-time. However we are generating context using our ASR , So we built a QA bot which takes this context and also a question from the user and generates an answer. Here, the question should be related to the context, then only our model can generate correct and meaningful answers. We used a pretrained Roberta model from hugging face for our Question Answering task [7]. This model has been fine-tuned on the Stanford Question Answering Dataset (SQuAD) 2.0. SQuAD is a benchmark dataset for machine comprehension tasks, where the goal is to answer a question based on a given context passage. This model has been fine-tuned to specifically answer questions based on the SQuAD 2.0 dataset, which includes more challenging questions compared to the original SQuAD dataset. The model has been trained to identify the answer span within the given context passage that best answers the question. In our case it takes the context that was generated from ASR and a question from the user as a set of inputs and generates the answer.

3.12. Emotion Detection. Emotion detection with text is a natural language processing technique that involves automatically identifying the emotional tone or sentiment expressed in a piece of text. The goal of emotion detection is to classify a piece of text as positive or negative based on the emotional content of the text. But here we have done further research on detecting various other emotions such as Joy, Sadness, Anger, fear, surprise, disgust as specified by Paul Ekman. If not any of these it will be neutral. We found an open source model which does this type of emotion detection well and the model we used was the arphangoshal/Ekman classifier from huggingface where it takes a piece of text as input and predicts the emotion of a speaker based on the text itself. This text input is nothing but the translated English text of a transcript that we get from the Whisper model or the summarized text that we generate from the summarizer model.

3.13. Text Generation. Text generation is a process of generating new, coherent text based on a given prompt or context. Text generation is a challenging task in natural language processing, as it requires the model to generate text that is not only grammatically correct but also semantically meaningful and coherent. We used the GPT-NEO model for our text generation [8] and we generated texts or blogs using aitextgen package. There are many predefined functions available in aitextgen which are so useful for text generation tasks. So anyone who wants to generate a blog article based on any context they want or who wants to generate some contextual text based on any prompt, they can just give their audio or speak to the ASR and the transcript generated from that is given to the text generation model.

4. Experiments. We have tested Whisper on different datasets and checked the word error rates of every model configuration with some datasets as shown in this section (Tables 4.1-4.4).

Table 4.2: Comparison with wav2vec2.0

Model	Dataset	WER
Whisper	LibriSpeech	2.7%
Wav2Vec2.0	LibriSpeech	2.7%

Table 4.3: Testing Whisper on CLSRIL-23 Dataset

Model	Latency	Accuracy	Precision	Recall	F1-Score
Wav2Vec	8.97%	49.2%	38.0%	100%	55.1%
Tiny	4.98%	82.2%	86.9%	90.9%	88.8%
Base	11.14%	82.2%	84.4%	93.2%	88.5%
Small	16.94%	81.6%	80.7%	96.1%	87.7%
Medium	49.54%	90.8%	92.0%	96.5%	94.2%

Table 4.4: Testing Whisper on Translation Part on Fleurs dataset

Language used	Word Error Rate(WER)
Hindi	28.34%
Marathi	66.56%
Tamil	33.63%
Malayalam	44.60%
Telugu	84.20%

Table 5.1: Comparision between Normal Finetuned model and Finetuned model using Parameter-Efficient-Fine-tuning

Model Configuration	Finetuned Whisper WER	PEFT Whisper WER
Tiny	55.6%	64.9%
Small	33.5%	38.6%

5. Results. The word error rate of our model is 25% for hindi languae and 40% for marathi language. As mentioned in the previous section we used transformers from huggingface to finetune Whisper model on common voice 11.0 dataset. We have also shown and compared the results with other model configurations in this section. We have also shown the comparisions of fietuned whisper WER with the normal Whisper WER for every model checkpoint in this section and also shown the difference between WER of Whisper finetuned model and the model which was trained with PEFT.

[4000/4000

Step	Training Loss	Validation Loss	Wer
1000	0.347800	0.561929	64.335901
2000	0.241300	0.510025	58.659951
3000	0.191000	0.502138	56.268518
4000	0.160100	0.503013	55.989164

Fig. 5.1: Finetuned Whisper-tiny results

```
0it [00:00, ?it/s]
Reading metadata...: 0it [00:00, ?it/s]
Reading metadata...: 2894it [00:00, 16645.43it/s]
/usr/local/lib/python3.8/dist-packages/bitsandbytes/autograd/_functions.py:298
  warnings.warn(f"MatMul8bitLt: inputs will be cast from {A.dtype} to float16 
362it [1:11:33, 11.86s/it]wer=38.618471175823245
```

Fig. 5.2: Finetuned whisper-small WER

Step	Training Loss	Validation Loss
500	0.160800	0.254170
1000	0.089800	0.220722
1500	0.031900	0.211729

```
Reading metadata...: 2894it [00:02, 1276.94it/s]
```

Fig. 5.3: Finetuned Whisper-large-V2 using PEFT-Lora + BNB INT8 training + Streaming dataset for Hindi

Step	Training Loss	Validation Loss
500	0.172900	0.272142
1000	0.040800	0.290398
1500	0.010300	0.307799

Fig. 5.4: Finetuned Whisper-large-V2 using PEFT-Lora + BNB INT8 training + Streaming dataset for Marathi

```
0it [00:00, ?it/s]
Reading metadata...: 0it [00:00, ?it/s]
Reading metadata...: 2894it [00:00, 10561.80it/s]
/usr/local/lib/python3.8/dist-packages/bitsandbyt
  warnings.warn(f"MatMul8bitLt: inputs will be ca
362it [2:10:50, 21.69s/it]wer=25.895200203166002
```

Fig. 5.5: WER of Finetuned Whisper large-v2 on Hindi

6. Conclusion. Automatic Speech Recognition (ASR) technology has become an essential tool in human-machine communication. The Whisper ASR model developed by OpenAI is a promising development in this

```
0it [00:00, ?it/s]
Reading metadata...: 0it [00:00, ?it/s]
Reading metadata...: 2894it [00:00, 10456.52it/s]
362it [2:11:43, 21.83s/it]normalized_wer=13.10319567805146
```

Fig. 5.6: Normalized WER of Finetuned Whisper large-v2 on Hindi

```
0it [00:00, ?it/s]
Reading metadata...: 0it [00:00, ?it/s]
Reading metadata...: 1816it [00:00, 12578.29it/s]
/usr/local/lib/python3.8/dist-packages/bitsandbytes/autograd/_func
  warnings.warn(f"MatMul8bitLt: inputs will be cast from {A.dtype}
227it [2:11:27, 34.75s/it]wer=40.19727935013861
```

Fig. 5.7: WER of Finetuned Whisper large-v2 on Marathi

Table 5.2: Comparison between WER of Finetuned model and WER of normal model

Model Configuration	Finetuned Whisper WER	Normal Whisper WER
Small	38.6%	87.4%
Medium	29.2%	54.6%
Large-V2	25.8%	51.5%

field, with its Transformer encoder-decoder architecture capable of handling various tasks. However, the model's limitations in speaker diarization, summarization, and emotion detection, as well as its performance with Indian regional languages, need to be addressed. In this research we enhanced the Whisper ASR model's capabilities by adding features such as speaker diarization, text summarization, emotion detection, and a Question Answering. Additionally, the model's performance in Indian regional languages was improved through training on Indian languages. By addressing these limitations and improving the model's performance, this research has the potential to contribute to the development of more accurate and reliable ASR models, ultimately improving human-machine communication in various applications. Overall, this research could have a significant impact on the advancement of ASR technology and its applications in various industries. In the future this project can be extended to integrate with different apps and websites where the user would interact with our ASR and this project has a scope of improvement through increasing the training data and also by using different models.

REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, *Learning transferable visual models from natural language supervision*, in 2018 Chinese Automation Congress (CAC), IEEE, 2021, pp. 2446–2450.

[2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, *Robust speech recognition via large-scale weak supervision*, IEEE Access, 7 (2022), pp. 38044–38054.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli, *wav2vec 2.0: A framework for self-supervised learning of speech representations*, Journal homepage: www. ijrpr. com ISSN, 2582 (2020), p. 7421.

[4] Heng-Jui Chang, Alexander H. Liu, Hung-yi Lee, Lin-shan Lee, *End-to-end whisperd speech recognition with frequency-weighted approaches abd pseudo whisper pre-training*, in 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), IEEE, 2020, pp. 402–406.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), IEEE, 2020, pp. 324–329.

[6] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, Ignacio Lopez Moreno, *Speaker diarization with*

lstm, in 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), IEEE, 2019, pp. 324–329.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, in 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), IEEE, 2019, pp. 324–329.

[8] Sid Black,Leo Gao, Phil Wang, Connor Leahy, Stella Rose Biderman, *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow*, https://doi.org/10.5281/zenodo.5297715, 2021.

[9] Liu Liu, Zheng Qu, Zhaodong Chen, Yufei Ding, Yuan Xie, *Training Deeper Transformers with Sparse Attention via Conditional Computation*, arXiv:2110.11299v1, 2020.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Apr 25, 2023
Accepted: Oct 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
 Volume 24, Issues 4, pp. 847–855, DOI 10.12694/scpe.v24i4.2343

A FEATURE EXTRACTION BASED IMPROVED SENTIMENT ANALYSIS ON APACHE SPARK FOR REAL-TIME TWITTER DATA

PIYUSH KANUNGO* AND HARI SINGH†

Abstract. This paper aims to improve the accuracy of sentiment analysis on Apache Spark for a real-time general twitter data. A lot of works exist on sentiment analysis on offline or stored twitter data that uses several classification algorithms on relevant features extracted using well-known feature extraction methodologies on pre-processed text data. However, not much works exist for sentiment analysis of real-time twitter data and especially for the generic data on big data processing platforms such as Apache Spark. This paper proposes a real-time sentiment analysis for generic twitter data through Apache Spark using six classification algorithms on N-gram and Term Frequency – Inverse Document Frequency (TF-IDF) feature extraction methodologies on the pre-processed data. An exhaustive comparison is done using Logistic Regression (LR), Multinomial Naive Bayes (MNB), Random Forest Classfier(RFC), Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), and Decision Tree (DT) classification algorithms. It is observed that the trigram feature extraction method performs the best on LR and SVM and the RFC results are also comparable on the considered general tweets data.

Key words: Machine learning, Apache Spark, Twitter, Sentiment analysis, N-gram, TF-IDF

1. Introduction. Today big data applications are widely used in different fields like analysing social media platforms, improving healthcare systems, understanding customer behaviour and natural language processing. These applications play a key role for extracting the knowledge from large size datasets which can be used further for profit generation as well as for process improvement and business expansion. Social media plays a completely critical function in our life. What we are able to think, what we do, all of us express our emotions on social media platforms. Social media is a big, interactive medium for dialogue of numerous troubles associated with society in addition to vital for the growing unfolds of facts, specifically throughout instances of herbal disasters, calamities, and mass emergencies. Also, on social media humans speak about the goods which might be released day with the aid of using day. Many groups and business enterprises use those forms of facts (associated with their merchandise) to recognize what the humans reflect on consideration on their product. They can examine these facts and the usage of Social Network Analysis [1, 2]. Interactions via social media systems are now no longer centralised to a selected location, time quarter etc. Social media affords a short and effective manner to unfold facts now no longer counting whether its miles correct or inaccurate, unfolding of the both types is favourable. However social community typically favoured extra correct and legitimate facts to unfold than fake facts and rumoured facts. Interaction happens in actual time so this affords applicable unfold of facts according to applicable facts.

Understanding human sentiments and expressions from text over social media on any particular topic or event helps in better analysing and decision making. Text data mining or so called the text mining is actually bringing out the meaningful patterns and new information from the unstructured data after being processed thoroughly which leads to the structured format of the data. Features are the characteristics in the form of specific variables in any set of data that can be used to give accuracy in predictions after being selected appropriately. The increase in the web data or texts in documents have made it difficult to pre-process as the data is from many different dimensionalities. In order to reduce the data with different dimensionality the use of better feature selection and feature extraction is needed. While the feature selection or doing feature extraction depends on data being used in the application domain. Feature Extraction allows data reside in feature space without any loss of data even if the feature space size is reduced. N-gram and TF-IDF have been

*CSE & IT Department, Jaypee University of Information Technology, Solan, HP, India (piyushkanungo34@gmail.com)
†CSE & IT Department, Jaypee University of Information Technology, Solan, HP, India (hsrawat2016@gmail.com)

extensively used in the literature for feature extraction from the text.

Further, as the texts from the social media are huge and in real time so it's fast processing in real-time requires real-time big data processing frameworks. The processing capability of big data processing frameworks such as Apache Hadoop is well proven [3, 4, 5, 6, 7, 8, 9]. Apache Spark's ability to help in processing large data to achieve standardized data leverages the scalability and performance of Apache Spark to process and analyse the twitter dataset in real-time efficiently. A large number of big data processing technologies have evolved. A diversified domain of big data and technologies is presented in [10]. A detailed analysis of Apache Hadoop and Spark for big data processing is presented [11].

The goal of this paper is to develop a real-time sentiment analysis model using Apache Spark and classification algorithms on feature extracted through feature extraction methods N-Gram and TF-IDF that can accurately classify tweets as positive or negative based on the sentiment expressed in the text. The results provide insights into the sentiment of twitter users on a particular topic or event and demonstrate the effectiveness of sentiment classification on Apache Spark and feature extraction methologies for processing and analysing large volumes of social media data.

The rest of the paper is structured as follows. Section-2 describes related work. Section-3 describes the proposed sentiment analysis methodology. The results obtained are discussed in Section 4. The conclusions and future scope is presented in Section 5.

2. Related Work. Many researchers have recently applied sentiment analysis and word frequency techniques to classify people's attitudes from tweets. Understanding human expressions from textual data has been a focal point of studies and innovations for the past decade.

Twitter sentiment analysis using Naive Bayes algorithm shows an approach where various data cleaning and preparation methods are developed to make tweets more understandable in the word processing process [12]. The concept of control machine learning is used since each sample dataset consists of a pair of tweets and thoughts. The main goal is to find more effective tweet analytics. Sentiment analysis on twitter divides tweets into positive and negative classes. They are able to build a model with an accuracy of 94%. The authors applied classification algorithms directly on all the features of pre-processed data and no relevant feature extraction technique is used. In a similar work, the authors proposed a classification task in order to obtain the sentiment behind the polarity of an economic text using DT, gradient boost, naive bayes, RFC, K-NN, eXtreme Gradient Boost, SVM, and LR [13]. It was found that classifying the three groups (positive, negative and neutral), the support vector classifier performed best up to 77% accuracy on the test dataset. However, it was trained on an economic dataset which create a bias towards words which are used in economic texts and cannot be used with generic data. In a similar work based on K-Means clustering, sentiments of each cluster is analyzed across the several aspects of the Covid-19 pandemic [14]. In a similar work, the authors proposed a classifier using Machine Learning (ML) techniques that can predict the polarity of a comment [15].

In a feature extraction based work, a system is proposed where the authors applied N-gram bag of words feature extraction with different machine learning models [16]. Several models are developed on applying unigram, bigram and trigram to analyse economic texts. The models are evaluated on the metric of accuracy, recall, f1-score and precision. The data used was review datasets of Amazon, IMDB and yelp. It was observed that SVM performed the best with N-gram feature extraction method with an accuracy of 82%. A significant increase was observed in performance after applying the N-gram feature extraction technique. In another research work, the authors used two classification methods: unigrams and bigrams, and attempts were made to include bigrams in vectors to improve accuracy [17]. Once removed, the function returns as a small or dense vector. Based on the data, there is a sparse vector representation that is more efficient. The drawback of this approach is that it is just going for unigram and bigram feature extraction which many of the times are not enough to capture the sentiment as the bag of words is too small.

A sentiment analysis of hotel reviews using N-gram and Naive Bayes Methods aims to determine the application of N-gram and Naive Bayes methods in sentiment analysis [18]. Based on accuracy results, it was found that tokenization unigram method works better than other tokenization methods. This method obtained precision results of 94%, recall 100%, accuracy 97%, and error rate 3%. However, a single classification algorithm was used and as it provided best result using unigram, it can also be validated over other data. Similarly, sentiment analysis based on N-gram and K-NN Classifier [19] was applied to classify data into positive, negative and neural classes. The accuracy of proposed system is achieved up to 86%.

An N-gram feature extraction based sentiment classification model for drug user reviews using Naive Bayes, Maximum Entropy, and SVM is presented [20]. It was found that the Maximum Entropy method achieves the best result for the presence and frequency of unigram features, and the SVM method achieves the best result for TF-IDF of unigram bigram features. However, it focuses on text reviews from online health information services and hence cannot be used on generic dataset. Similarly, a sentiment classification using N-gram and TF-IDF uses a general machine learning framework over three types of document in the dataset, sentences in questions and answers on Stack Overflow; reviews of mobile applications; and comments on Jira issue trackers, is used. The method achieved highest F1 values in positive and negative sentences on all datasets in comparison to the publicly available datasets [21]. In another work, the authors use N-gram and TF-IDF feature extraction for online fake news detection on six machine learning classification models and found that the TF-IDF with SVM provides highest accuracy [22].

In another work, the authors predicted election result by enhanced sentiment analysis on twitter data using ensemble classifier and Natural Language Processing (NLP). The NLP based approach is used to enhance the sentiment classification by adding semantics in feature vectors and thereby using ensemble methods for classification [23]. Adding semantically similar words and context-sense identities to the feature vectors increased the accuracy of prediction. The comparison of experiment results show that the semantics-based feature vector with ensemble classifier outperforms the traditional bag-of-words approach with single machine learning classifier. The ensemble method performs better than the other traditional classification by 3- 5%. In a recent work, a dual-channel attention network model is used to extract the text semantic through transducive learning and graph structure to enhance the semantic classification [24]. In another recent work, the authors used the sentiment analysis of social media tweets for extracting features based on sentiment lexicons and textual content for inputting to classifier for detecting depression[25]. In another work, an orthographic pleonasm method is used for improving lexical sentiment analysis accuracy to identify emotion-related neologisms in social media texts [26] and a cross lingual sentiment classification is presented in [27]. In another work, the authors made a comparison among bag of words with Naive Bayes, continuous bag of words with SVM, and Long Short-Term Memory (LSTM) for intent classification for dialogue utterances [28]. In another work, a sentiment classification for cryptocurrency related social media tweets use the bidirectional encoder representation from transformers (BERT) model to learn the numerical representation of text and emojis based sentiment classification generates emoji sentiment lexicon – language-universal cryptocurrency emoji (LUKE) lexicon [29].

In some of the works, the Apache Spark big data processing framework is used for sentiment analysis. The authors used MLib, an Apache spark machine learning library, to classify tweets into sentiment classes [30]. They used spark's ability to handle large amount of data which helped in performing pre-processing and standardization for the huge dataset at once. They applied Naive Bayes, Logistic regression, and Decision tree with unigram and TF-IDF feature extraction method. They applied five-fold cross validation approach to ensure a precise accuracy score. They observed that the accuracy scores for naive bayes and logistic regression were close and were far superior to the decision tree. It was found that for a large dataset, 78% accuracy was achieved for logistic regression and Naive-Bayes whereas the decision tree showed an accuracy of 68%. In another Apache Spark based coronavirus pandemic prediction in real-time uses machine learning for features extracted on twitter big data streaming [31].

A sentiment classification using paragraph vector and cognitive big data semantics used Apache Spark for a distributed version of any machine learning algorithm which scales easily to larger datasets on a set of hardware devices [32]. Here, a hybrid of sentence vectors, distributed and balanced versions of well-known machine learning techniques for emotional analysis is used. They used two methods for comparison - bag of words based Document Term Matrix (DTM) and the hash trick-based DTM. The model resulted in an area under the curve (AUC) of 95.44%. The model uses DTM and did not use other feature extraction methods.

An entropy-based evaluation for sentiment analysis of stock market prices on twitter data is used [33]. Initially, the daily twitter posts are analyzed and different N-grams along with two strategies that are utilized to increase the accuracy of the classification are applied. A Spark streaming has been employed for the processing of Twitter data, while Apache Flume has been utilized for the analysis. The cons of this approach is that sentiment analysis without using appropriate machine learning is not effective and models become too simple

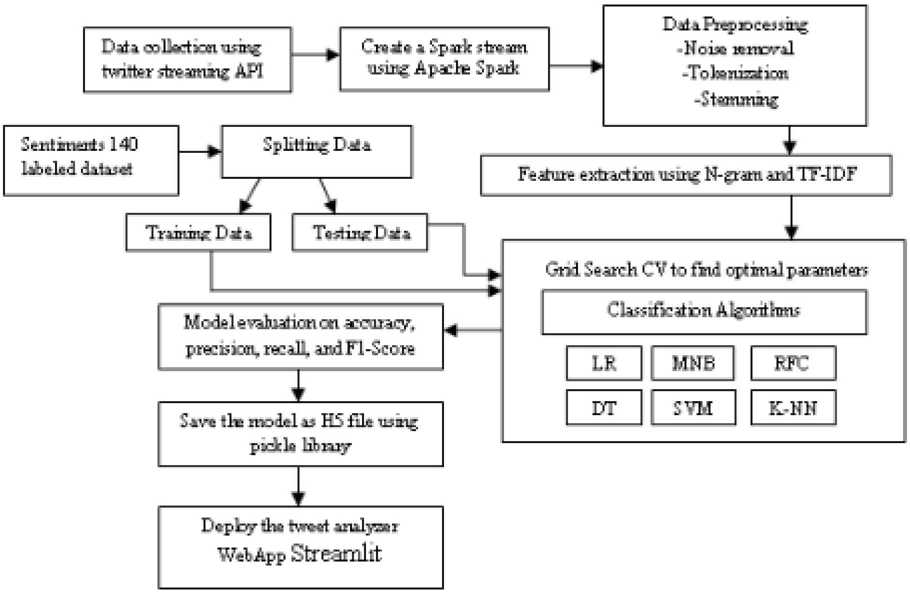


Fig. 3.1: Proposed Methodology for Twitter sentiment analysis

to capture the correct sentiments.

This paper presents a Apache Spark based real-time sentiment analysis for general tweets data. The three main components of our work are data extraction, processing, and modelling. We utilise the Natural Language Toolkit (NLTK) dataset to create our model. The real-time text data from twitter is pre-processed for noise removal, tokenization and stemming. Then feature extraction using N-gram and TF-IDF is applied to the pre-processed data and then the extracted features are fed to the classifiers. We compared six classifiers to categorise our tweets into positive and negative attitudes based on a supervised probabilistic machine learning algorithm, and then effectiveness of the classifiers are assessed. The presented work compares six classification algorithms LR, MNB, RFC, SVM, K-NN and DT on Unigram, Bigram, Trigram and TF-IDF based features for sentiment classification.

3. Proposed Sentiment Analysis Methodology. The proposed system can identify the sentiment behind users recent tweets, and to achieve that an offline NLP model is built using Apache spark text pre-processing tools and machine learning classification algorithms, LR [34, 35], MNB [36], RFC [37], SVM, RF [38], and DT and K-NN[12] with different hyper-parameters tuning and feature extraction technique combinations to find the optimal model. The FIG. 3.1. illustrates the different stages of the system which include dataset collection, creating a spark stream, data pre-processing, data splitting, hyper-parameter tuning, feature extraction, model training, model evaluation and model deployment. Each step in building the system is explained as follows:

3.1. Data collection and creating a Spark stream. The dataset used in this paper is the sentiment140 dataset from Kaggle [39], which was used to perform a twitter sentiment classification using distant supervision [40]. The dataset has six fields which are target, ids, date, flag, user, text with 1.6 million tweet instances labelled as 0 (positive) or 4 (negative.). Upon relabelling the target values as 0 and 1 and dropping all the fields except ‘text’ and ‘target’, it is observed that the dataset is balanced and has approximately the same number of data points for both the labelled observations. We start by creating a spark session to load our dataset to stream the data so that further operations can be done. Apache spark has a lot of utility functions to perform text pre-processing and standardisation. After loading the data, exploratory data analysis is done to better understand the data so as to perform correct operations after handling missing values [41].

3.2. Data pre-processing. Data pre-processing is important in any text-based analytics system because the complexity of the data directly influences the model training when trying to extract sentiments out of textual data. According to our studies, Twitter is a platform with many links, hashtags, special characters, emojis, etc. It is considered one of the most popular text dumps of human language in digital textual form due to its content. Therefore, pre-processing removes anything from the tweet that does not add to the meaning of the tweet or just makes the text vague. So, twitter data pre-processing is performed with the following steps using NLTK library: noise removal, tokenization and stemming, which are described below:

Noise removal During this phase the useless parts of text which add no context to the tweet are removed which is achieved in the following steps:

- **Remove accented char** Accented characters generally are used to add the sense of sound to text which is not required for a text NLP model.
- **Remove emails** Email ids don't add meaning to the text so they are removed.
- **Remove hashtags** The twitter hashtags are used to index topics on the platform but are generally repetitive to and are removed.
- **Remove HTML tags** Remove unnecessary URL links using the beautiful soap library as they don't add context to the text.
- **Remove retweets** Redundant data must be removed to maintain the data unbiased and train a good model.
- **Remove multiple spaces** Multiple continuous space must be removed to avoid confusion while training the model.
- **Remove stop words** These are the insignificant parts of the human vocabulary which are basically connecting words or ending words. These are filtered words which include punctuations, articles, conjunctions and general words such as a, the, an.

Tokenization Tokenization in pre-processing text data includes dividing longer strings into tokens. These tokens can be sentences that can be broken down into short sentences, which can be broken down into words. This process makes it easier for the model to handle the complex text data.

Stemming After the tokenization stage, the next stage is stemming. During this stage the words are changed to their original form (i.e., root form to have less redundant meaning attached to similar words). For example, "tired" and "tiring" will be reduced to the word "tire".

3.3. Data splitting, model training, and evaluation. During this stage, the processed labelled dataset with 'text' and 'target' fields is split into a training and test dataset as per the stratified 10-fold CV. The training datsset is used to train the models and the test dataset is used to finally evaluate the built models. The machine learning models applied to the extracted features are as follows: LR, MNB, RFC, SVM, DT and K-NN. After loading the models, a classifier is built for each model with each type of feature extraction i.e., unigram, bigram, trigram and TF-IDF with testing set and stratified 10-fold cross-validation (CV) to find the cross- validation performance of the applied models. The standard classification report with accuracy, precision, recall, and F1-score is used to check the test performance and cross-validation performance.

3.4. Hyperparameter tuning. Now that we have the training set to start building and testing out different models, we need to decide which model specific parameters we need to pass to obtain the best results for that specific model. We decided to do that using GridSearchCV which is an algorithm which helps us to iterate through all possible parameter combinations and gives attached scores for each iteration that helps us to choose the best combination of parameters for each model. For example, we got C = 1 and kernel = 'linear' for SVM and n = 1 and weight = uniform for K-NN.

3.5. Feature extraction. When dealing with high dimensional data, feature extraction is needed to perform textual data analysis. Feature extraction basically entails converting text data into a matrix of features. N-grams and TF-IDF feature extraction techniques are applied over the text in this paper. N-gram feature extraction method is a prominent bag of words-based analysis technique used in text mining and natural language processing. An N-gram is used to calculate a contiguous series of words of length 'N' in a particular time frame according to textual data analysis. In this paper, the N-gram approach to capture the context of Twitter data uses N = 1 to N = 3 (i.e., unigram, bigram and trigram). The TF-IDF is a feature extraction method which works by identifying the most important words in a document by calculating the relative frequency of

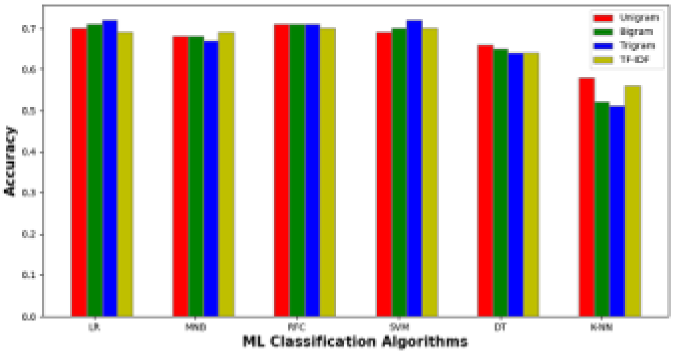


Fig. 3.2: Sentiment classification performance comparison of various feature extraction methods on ML algorithms

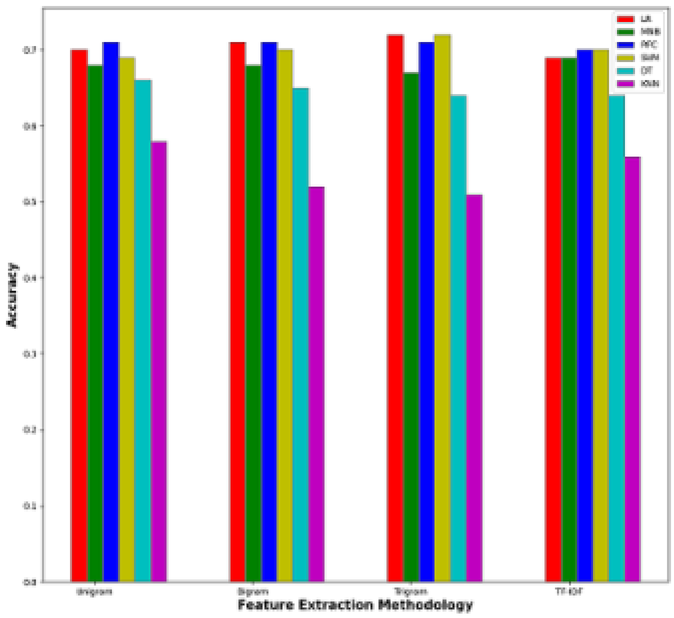


Fig. 3.3: Sentiment classification performance comparison of various ML classification algorithms on feature extraction methods

that word in the document. The main advantage of using TF-IDF is the ease of use for large document corpus but at the same time it is sometimes too simple to get hold of complex multidimensional data.

4. Results and discussions. The classification report is analysed to check the test accuracy and the cross validation accuracy to compare the performance and consistency of the models. The K-NN shows the lowest accuracy across every model for every feature extraction method i.e., 55% in case of unigram, 51% in case of

Table 4.1: Results of various ML classification algorithms on various feature extraction methodologies over the real-time twitter tweets

ML Algorithms	Feature Extraction Methodology	Testing Performance				Cross-Validation Performance
		Accuracy	Precision	Recall	F1-Score	Accuracy
LR	Unigram	0.7	0.71	0.7	0.71	0.64
	Biigram	0.7	0.72	0.7	0.71	0.63
	Trigram	0.72	0.74	0.71	0.72	0.64
	TF-IDF	0.69	0.68	0.69	0.7	0.64
MNB	Unigram	0.68	0.66	0.56	0.71	0.62
	Bigram	0.68	0.67	0.71	0.69	0.61
	Trigram	0.67	0.68	0.68	0.68	0.60
	TF-IDF	0.69	0.67	0.76	0.71	0.64
RFC	Unigram	0.71	0.72	0.72	0.72	0.64
	Bigram	0.71	0.74	0.66	0.7	0.64
	Trigram	0.71	0.74	0.66	0.7	0.64
	TF-IDF	0.7	0.72	0.69	0.7	0.63
SVM	Unigram	0.69	0.7	0.7	0.7 0	0.61
	Bigram	0.7	0.72	0.68	0.7	0.61
	Trigram	0.72	0.73	0.71	0.72	0.61
	TF-IDF	0.7	0.72	0.69	0.7	0.63
DT	Unigram	0.66	0.68	0.63	0.65	0.59
	Bigram	0.65	0.68	0.62	0.65	0.57
	Trigram	0.64	0.66	0.62	0.64	0.6
	TF-IDF	0.64	0.66	0.61	0.63	0.59
K-NN	Unigram	0.58	0.56	0.78	0.66	0.55
	Bigram	0.52	0.52	0.90	0.66	0.51
	Trigram	0.51	0.51	0.97	0.67	0.50
	TF-IDF	0.56	0.54	0.89	0.67	0.52

bigram, 50% in case of trigram, and 50% when TF-IDF is used for 10-fold cross validation data. Meanwhile, the LR with trigram and SVM with trigram display the highest accuracy score of 72% for the testing data and the accuracy of the RFC is also comparable. When considering different feature extraction methods trigram appears to be the most consistent approach while dealing with our dataset. It is shown in the Table 4.1 and FIG. 3.2. It displays a test accuracy of 72% for LR, 67% for MNB, 71% for RFC, 72% for SVM, 64% for DT, and 51% for KNN. It is evident from the Table 4.1 that the trigram is the best feature extraction method for the used dataset.

It is observed from the Table 4.1 and Fig. 3.3 that the LR and SVM demonstrate highest accuracy and the RFC is also comparable, which seems reasonable when dealing with binary classes namely positive and negative. And it is well known that the logistic regression performs well for binary classification. The SVM has been used historically for complex pattern recognition problems such as handwriting recognition, email classification and gene classification and therefore as expected it performs well with the dataset used as the twitter data is generic and complex to deal with. As for the feature extraction methods trigram method resulted in best observations as it is hard to capture the patterns on textual data which is generic in nature by having N = 1, 2 i.e., taking 1 or 2 bag of words at a time. It is also observed that the K-NN shows worst performance when compared to every other trained model, because when dealing with a large dataset the K-NN fails in capturing the sentiment tendencies as the K-NN is generally weak for larger datasets.

5. Conclusions and Future Work. The aim of this paper is to find the best combination of feature extraction methods and machine learning algorithms to perform real-time sentiment analysis on the Apache Spark. This paper demonstrates Apache Spark's data processing capabilities combined with feature extraction techniques when used with a huge and generic dataset and machine learning classification algorithms to get detailed observational data which can be used to compare and contrast each of the algorithms. From the experimental results, it can be concluded that according to the dataset used which is a generic tweets twitter dataset, the SVM and LR with trigram make the best classification of real-time tweets and the RFC is also comparable.

There are several other feature extraction methodologies in existence from the recent research work. We plan to extend our work for those feature extraction methodologies. Some advanced classification algorithms from the deep learning domain can be worked upon. Lastly, the real-time sentiment analysis can be applied to the new or latest prevailing text over the social media.

REFERENCES

[1] D. Verma, H. Singh, and A. K. Gupta, *A study of big data processing for sentiments analysis*, in Large-Scale Data Streaming, Processing, and Blockchain Security, IGI Global, 2020.

[2] T. Dilesh, M. Dudhane, A. Sardar, K. Deshpande, and N. Deshmukh, *Sentiment Analysis on Social Media for Emotion Classification*, 2020, doi: 10.1109/ICICCS48265.2020.9121057.

[3] H. Singh and S. Bawa,*A Survey of Traditional and MapReduceBased Spatial Query Processing Approaches*, ACM SIGMOD Record, Vol. 46, No. 2, pp. 18–29, 2017, doi: 10.1145/3137586.3137590.

[4] H. Singh and S. Bawa,*A mapreduce-based efficient H-bucket PMR quadtree spatial index*, Computer System Science and Engineering, Vol. 32, No. 5, pp. 405–415, 2017.

[5] H. Singh and S. Bawa,*Scalability and Fault Tolerance of MapReduce for Spatial data*, Glob. J. Eng. Sci. Res. Manag., Vol. 3, No. 8, pp. 97–103, 2016.

[6] H. Singh and S. Bawa,*IGSIM: An improved integrated Grid and MapReduce-Hadoop architecture for spatial data: Hilbert TGS R-Tree-based IGSIM*, Concurrency Computation : Practice and Experience, John Wiley & Sons, Vol. 31, Iss. 17, 2019, doi:https://doi.org/10.1002/cpe.5202.

[7] M. Mittal, H. Singh, K. Paliwal, and L. M. Goyal,*Efficient Random Data Accessing in MapReduce*, in International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions), pp. 552–556, 2017.

[8] H. Singh and S. Bawa,*An Integrated Architecture for High Performance Spatial Data Analysis*, International Journal of Computer Science and Information Security, Vol. 14, No. 11, pp. 302–309, 2016.

[9] H. Singh and S. Bawa,*A MapReduce-based scalable discovery and indexing of structured big data*, Future Generation Computer Systems, Vol. 73, No. August 2017, pp. 32–43, 2017.

[10] H. Singh, R. Vasuja, and R. Sharma, *A Survey of Diversified Domain of Big Data Technologies*, Adv. Parallel Comput., Vol. 29, No. September, pp. 1–27, 2018, doi: 10.3233/978-1-61499-814-3-1.

[11] P. Sewal and H. Singh, *A Critical Analysis of Apache Hadoop and Spark for Big Data Processing*, in 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), Oct. 2021, pp. 308–313, doi: 10.1109/ISPCC53510.2021.9609518. .

[12] P. Mishra, S. A. Pati, U. Shehroj, P. Aniyeri, and T. A. Khan, *Twitter Sentiment Analysis using Naive Bayes Algorithm*, in 2022 3rd International Informatics and Software Engineering Conference (IISEC), 2022, pp. 1–5, doi: 10.1109/IISEC56263.2022.9998252.

[13] O. E. Ojo, A. Gelbukh, H. Kalvo, O. O. Adebanje, and G. Sidorov, *Sentiment Detection in Economics Texts*, in Mexican International Conference on Artificial Intelligence, pp. 271–281, 2020.

[14] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperli,*COVID-19 Sentiment Analysis Based on Tweets*, IEEE Intelligent Systems, Vol . 38, No. 3, pp. 51-55, 2023, doi: 10.1109/MIS.2023.3239180.

[15] F. Jemai, M. Hayouni, and S. Baccar, *Sentiment Analysis Using Machine Learning Algorithms*, in 2021 International Wireless Communications and Mobile Computing (IWCMC), 2021, pp. 775–779, doi: 10.1109/IWCMC51323.2021.9498965.

[16] O. E. Ojo, A. Gelbukh, H. Calvo, and O. O. Adebanji, *Performance Study of N-grams in the Analysis of Sentiments*, J. Niger. Soc. Phys. Sci., Vol. 3, No. 4, pp. 477–483, 2021, doi: https://doi.org/10.46481/jnsps.2021.201.

[17] O. R. S. Mohana, S. Kalaiselvi, K. Kousalya, P. Mohamed Hanif, D. Lohappriya, and K. Khalid Ali Khani, *Twitter based sentiment analysis to predict public emotions using machine learning algorithms*, 2021, doi: 10.1109/ICIRCA51532.2021.9544817.

[18] T. Widiyaningtyas, I. A. E. Zaeni, and R. Al Farisi, *Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods*, 2019, doi: 10.1109/ICIC47613.2019.8985946.

[19] S. Bhardwaj and J. Pant, *Sentiment Analysis Approach based N-gram and KNN Classifier*,Int. J. Res. Electron. Comput. Eng., Vol. 7, No. 2, pp. 2108–2111, 2019.

[20] J. A. Kumar, S. Abirami, and T. E. Trueman, *An N-Gram Feature-Based Sentiment Classification Model for Drug User Reviews*, in Artificial Intelligence and Evolutionary Computations in Engineering Systems, 2021, pp. 277–297, doi: 10.1007/978-981-16-2674-6_22.

[21] R. Maipradit, H. Hata, and K. Matsumoto, *Sentiment Classification Using N-Gram Inverse Document Frequency and Automated Machine Learning*, IEEE Softw., Vol. 36, No. 5, pp. 65–70, 2019, doi: 10.1109/MS.2019.2919573.

[22] H. Ahmed, I. Traore, and S. Saad. *Detection of online fake news using N-Gram analysis and machine learning techniques*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 10618 LNCS, pp.127-138, 2017.

[23] R. Jose and V. S. Chooralil, *Prediction of election result by enhanced sentiment analysis on twitter data using classifier*

ensemble Approach, 2016, doi: 10.1109/SAPIENCE.2016.7684133.
[24] K. Dong, Y. Liu, F. Xu, and P. Liu, *DCAT: Combining Multi-semantic Dual-channel Attention Fusion for Text Classification*, IEEE Intelligent Systems, Vol . 38, pp. 10-19, 2023, doi: 10.1109/MIS.2023.3268228.
[25] R. Chiong, G. S. Budhi, S. Dhakal, and E. Cambriai,*Combining Sentiment Lexicons and Content-Based Features for Depression Detection*, IEEE Intelligent Systems, Vol . 36, No. 6, pp. 99-105, 2021, doi: 10.1109/MIS.2021.3093660.
[26] M. Thelwall and E. Cambriai,*This! identifying new sentiment slang through orthographic pleonasm online: Yasss slay gorg queen ilysm*, IEEE Intelligent Systems, Vol . 36, No. 4, pp. 114-120, 2021, doi: 10.1109/MIS.2021.3093660.
[27] A. Esuli, A. Moreo, F. Sebastiani, and E. Cambria,*Cross-Lingual Sentiment Quantification*, IEEE Intelligent Systems, Vol . 35, No. 3, pp. 106-114, 2020, doi: 10.1109/MIS.2020.2979203.
[28] J. Schuurmans and F. Frasincar,*Intent Classification for Dialogue Utterances*, IEEE Intelligent Systems, Vol . 35, No. 1, pp. 82-88, 2020, doi: 10.1109/MIS.2019.2954966.
[29] M. Kulakowski and F. Frasincar,*Sentiment Classification of Cryptocurrency-Related Social Media Posts*, IEEE Intelligent Systems, Vol . 38, No. 4, pp. 5-9, 2023, doi:10.1109/MIS.2023.3283170.
[30] H. Elzayady, K. M. Badran, and G. I. Salama, *Sentiment Analysis on Twitter Data Using Apache Spark Framework*, In Proceedings - 2018 13th International Conference on Computer Engineering and Systems, ICCES 2018, 2018, pp. 171–176, doi: doi: 10.1109/ICCES.2018.8639195.
[31] X. Zhang, H. Saleh, E. M. G. Younis, R. Sahal, and A. A. Al, *Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System*, Complexity, Vol. 2020, 2020, doi: 10.1155/2020/6688912.
[32] K. Ravi, V. Ravi, and B. Shivakrishnal, *Sentiment Classification Using Paragraph Vector and Cognitive Big Data Semantics on Apache Spark*, in 2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 2018, pp. 187–194, doi: 10.1109/ICCI-CC.2018.8482085.
[33] A. Kanavos, G. Vonitsanos, A. Mohasseb, and P. Mylonasl, *An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data*, in 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization, 2020, pp. 1–7, doi: 10.1109/SMAP49528.2020.9248440.
[34] D. G. Kleinbaum and M. Klein, *Logistic Regression: A self learning text*, Third. Atlanta, USA: Springer, 2010.
[35] H. Singh and S. Bawa, *Predicting COVID-19 statistics using machine learning regression model: Li-MuLi-Poly*, Multimed. Syst., Vol. 28, pp. 113–120, 2022, doi: 10.1007/s00530-021-00798-2.
[36] P. P. Surya and B. Subbulakshmi, *Sentimental Analysis using Naive Bayes Classifier*, in 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019, pp. 1–5, doi: 10.1109/ViTECoN.2019.8899618.
[37] M. Zhu, *Random Forest classifier for remote sensing classification*, Int. J. Remote Sens., Vol. 26, No. 1, pp. 217–222, 2005.
[38] M. Basheri, M. J. Iqbal, and A. Rahim, *Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection*, IEEE Access, Vol. 6, pp. 33789–33795, 2018.
[39] M. M. KazAnova, *Sentiment140 dataset with 1.6 million tweets*, Kaggle, 2017. https://www.kaggle.com/datasets/kazanova/sentiment140 (accessed Nov. 13, 2022).
[40] A. Go, R. Bhayani, and L. Juang, *Twitter sentiment classification using distant supervision*, 2009. [Online]. Available: https://www-cs-faculty.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.
[41] B. Yan, Z. Yang, Y. Ren, X. Tan, and E. Liu, *Microblog Sentiment Classification Using Parallel SVM in Apache Spark* , 2017, doi: 10.1109/BigDataCongress.2017.43.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jun 11, 2023
Accepted: Sep 26, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 857–872, DOI 10.12694/scpe.v24i4.2353

ENSEMBLE HYBRID MODEL FOR COVID-19 SENTIMENT ANALYSIS WITH CUCKOO SEARCH OPTIMIZATION ALGORITHM

VIPIN JAIN * AND KANCHAN LATA KASHYAP †

Abstract. The COVID-19 pandemic has caused anxiety and fear worldwide, affecting people's physical and mental health. This research work proposes a sentiment analysis approach to better understand the public's perception of COVID-19 in India. Two datasets are created by collecting tweets regarding COVID-19 in India. Pre-processing and analysis of datasets are performed by using natural language processing (NLP) techniques. Various features are extracted from collected tweets using three-word embeddings GloVe, fastText, Elmo. The optimal features are selected by cuckoo search optimization algorithm. Finally, the proposed hybrid model of Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM) is used to categorize the tweets into three sentiment categories. Proposed model achieved 94.44% accuracy, 90.34% precision, 88.53% sensitivity, and 89.53% F1 score. It significantly improved over previous approaches, which achieved 80% accuracy.

Key words: COVID-19, Sentiment, Cuckoo Search, Optimization, Deep learning, Ensemble learning

1. Introduction. COVID-19, or Coronavirus disease, is a virus that infected millions of people and puts human lives at risk worldwide [47]. It is a highly contagious virus that spreads quickly and causes serious illnesses, including death. World health organization (WHO) declared COVID-19 as pandemic due to its risk and hazards [12]. 37,109,851 COVID-19 cases and 1,070,355 fatalities have been reported by the WHO [43]. Coronavirus is a highly infectious virus that spreads among the people through sneezing, coughing, or talking [10, 23]. In addition, it can be transmitted through air droplets from an infected person. Therefore, it is important to take precautions to prevent the spread of the virus. Worldwide population are affected directly or indirectly by this crisis [5]. Sentiment analysis is one of the way to deal with this crises [20]. It is essential to understand the mindset, feelings, and fears of people to fight with this disease. Social media platform is now an integral part of our daily life to communicate and share information. The messages and tweets posted on social networking sites play an important role to understand the feelings of people. People expect accurate and reliable information about corona cases. It is analyzed that many social media posts and tweets misled the readers by publishing erroneous statistics of COVID-19 cases. However, various practices such as social distancing, wearing masks, and washing hands are some of the key measures to prevent the spread of the virus. Governments around the world have put in place measures to combat the virus, including lockdowns, travel restrictions, and other measures [17]. Many countries have started vaccination campaigns to safeguard their populations from the fatal COVID-19 disease. A vaccine is an effective way to prevent the spread of the virus, but due to shortages and access restrictions, not everyone can get it [25]. People still need to practice good hygiene and take other precautions to protect themselves and others. However, many people are hesitant to take the vaccine because of misinformation regarding its safety and effectiveness. It is important to educate people for importance of vaccination to reduce the spread of virus. The various emotional outburst tendencies of Indians can be seen on Twitter regarding COVID-19 [19]. Analyzing the mental emotions of people during epidemic periods assists the government to review previous policies and making new standards. During this epidemic, the government should make efforts to stimulate the public and restore their physical and emotional health. This research analyses the feelings of Indian tweet users about COVID-19. India has the second largest population with diverse geography in the world [24]. In order to conduct this study, several tweets are gathered from various COVID-19-related hashtags. Additionally, emotions associated with tweets are identified as WordCloud.

*SCSE, VIT Bhopal University, 466114, Madhya Pradesh, India (vipin.jain2020@vitbhopal.ac.in).
†SCSE, VIT Bhopal University, 466114, Madhya Pradesh, India (kanchan.k@vitbhopal.ac.in)

Sentiment analysis can provide insights to governments and healthcare organizations to better craft strategies and policies to fight the virus. It can also help to provide psychological and emotional support to people affected by the virus. Additionally, it can provide valuable data to help researchers to develop better treatments and vaccines. Lastly, this study provides valuable insights for governments and healthcare organizations to better assess the mental health of the public during an epidemic period in order to build better policies to combat COVID-19.

1.1. Motivation. Understanding of Indians emotions about the coronavirus on social media platforms plays an important role for the government to control, monitor, and get rid of COVID-19. Significant research has not been done on the sentiment analysis of Indians towards COVID-19. In this study, sentiment analysis is conducted in order to comprehend Indian people's attitudes toward COVID-19. The findings of this study can also assist public health officials to connect effectively with individuals and provide public health solutions to the affected people.

1.2. The main contributions of this research work are as follows.

- Two annotated COVID-19 datasets are created using Twitter text for sentiment analysis.
- Various natural language processing (NLP) preprocessing techniques are employed on the collected real-time Twitter datasets, such as data cleaning, tokenization, and labelling.
- Three embedding techniques, namely ELMo, GloVe, and FastText embeddings, are used for feature extraction.
- Optimal features are selected using the Cuckoo search optimization algorithm.
- Proposed hybrid model of GRU and BiLSTM is used to determine the sentiments of Indian citizens. The results obtained from the model are compared with the performance of individual GRU, BiLSTM and existing work.
- The efficiency of the proposed model is evaluated using the F1-score, precision, accuracy, recall, and ROC.

1.3. Organization of the paper. The structure of the remaining paper is as follows: Literature review is covered in Section 2. Section 3 presents the recommended procedure. Results obtained from the suggested model are examined in Section 4, and then final conclusion is presented in Section 5.

2. Literature review. Several researchers have analyzed the people emotion regarding the COVID-19 during the lockdown and pandemic period. Many authors have used wide variety of machine learning and word embedding methods to decipher the public feeling throughout the COVID-19 period

A web portal based on real-time tweet is proposed by Venigalla et al. to reflect the Indian sentiment during COVID-19 [40]. This platform allows visitors to check general sentiment of people of specific state on specific date and time.

Limitation: of this study is that the current portal shows the state-level sentiment of few cities only.

Chakraborty et al. have been suggested a model to analyze the two types of pandemic tweets [10]. In the first scenario, 23,000 tweets from 1 Jan 2019 to 23 Mar 2020 are evaluated [1]. According to their finding, majority of tweets expressed either negative or neutral sentiments. As in the second scenario, 226,668 tweets have been analyzed with positive or neutral sentiments and achieved 81% validation accuracy. An emotion care strategy to analyse COVID-19 tweets is developed by Gupta et al. [18]. In this approach, initially, all tweets are converted into lowercase strings. All special characters, punctuation, links and retweets are effectively removed. The frequency-inverse document is utilized for the vectorization process after data cleaning. Lastly, sentiments are categorised as trust, surprise, grief, pleasure, fear, disgust, anger, and anticipation in their work. Borah et al. employed a multi-modal deep learning approach to analyze 36,231,457 tweets related to COVID-19 vaccine from 51,682 Indians [8]. All Tweets are collected using #ReadyToVaccinate, #Covishield, #CovidVaccine, and #Covaxin hashtags. Analysis is done by using SentriStrength tools which assigns a value between -4 to +4 to each tweet. Extreme negative and extreme positive sentiment is denoted by -4 and +4, respectively. The textual data and the network topology are encoded by using BERT and GraphBERT.

Limitation: Sentiment analysis is applied only to tweets posted by urban residents. The perspective of rural residents is not included in the analysis. Furthermore, only limited hashtags are used to determine inclusion of tweets in the dataset for analysis.

Kumar et al. applied hybrid model of BiLSTM and convolution neural network (CNN) to evaluate the publicly accessible Sentiment140 dataset with labelled Indian COVID-19 tweets and achieved 90% accuracy [29].

Limitation: Authors have used English text only for the sentiment analysis. The text from other languages can also be used to improve its correctness.

Misra et al. examined and acquired information regarding reverse migration in India via Twitter mining [34]. They retrieved almost 50,000 tweets from March 2021 to May 2020 by using trending hashtag such as #IndianMigrantWorkers and Twitter API. Different types of emotions are identified by using the NRC Emotion Lexicon after noise removal from collected data.

Limitation: To obtain Twitter data, the researchers exclusively used only the popular hashtags #IndianMigrantWorker and #MigrantWorker which does not reflect the entire population. The tweets posted in other languages are not included for analysis. Different perspective can be analyzed by using tweets posted in other Indian languages also.

Majumder et al. analyzed the sentiments of Indian users tweets about COVID-19 from March 2020 to June 2020 [31]. They have used supervised machine learning-based support vector machine and Logistic Regression for sentiment analysis with accuracy rates of 91.50 % and 87.75%, respectively. Chehal et al. evaluated sentiments of Indian Twitter users about online shopping during lockdown periods [11, 3]. Their analysis indicated the variations of feelings during different lockdown stages. during lockdown 2, people stocked fitness, sports, games, toys, and beauty products. While during lockdown 3, people stocked domestic goods, clothes, and nutrition products. Imran et al. evaluated feelings of people about the COVID-19 lockdown period [22]. LSTM model with FastText embedding is utilized in their research to identify the polarity of emotions with 82% accuracy. Chintalpudi et al. collected posts from Indian Twitter users between 23 March and 15 July 2020 for sentiment analysis [14]. Authors utilized the BERT [39] method for the text analysis and obtained an accuracy of 89%.

Basiri et al. suggested a new approach for coronavirus-related tweet emotion categorization by combining four deep learning methods : DistilBERT [27], fast text, BiGRU, and CNN [6]. Large labelled dataset of tweets is used to train the classification models and achieved highest 85.5% accuracy. Bhat et al. analysed the worldwide attitude expressed on Twitter [7]. Total 92,646 and 85,513 tweets related to #COVID-19 and #Coronavirus are collected, respectively. They have obtained 13.96%, 34.05%, and 51.97% of the tweets as negative, neutral, and positive attitudes for #COVID-19 sentiment study [2]. The authors obtained 41.27%, 40.91%, and 17.80% of neutral, positive and negative tweets, respectively for sentiment analysis of coronavirus. Only a few research work is done for sentiment analysis of Indians based on COVID-19 tweets. In this work, integration of metaheuristic optimization based cuckoo search algorithm with hybrid deep learning model of GRU and LSTM is applied for sentiment analysis and classification of Indians COVID-19 tweets.

List of abbreviations used in this paper is provided in table 2.1. Summary of the literature review is shown in table 2.2.

3. Proposed framework. The proposed framework for sentiment analysis is broadly divided into three steps as: (1) Data collection, (ii) Data preprocessing, and, (iii) Evaluation of emotional states. Flowchart of suggested framework is depicted in Fig 3.1. Following sub-sections provide a brief summary of each stage.

3.1. Data collection. Tweets collected from the Twitter platform are utilized in this work to examine the Indian sentiments during pandemic. Various terms such as #COVID-19, #COVIDindia, #coronavirusinindia, #coronaindia, #lockdownindia, #coronavirusindia, #staysafe, #stayhome, #indiafightscorona, and #coronavirus are used for filtration of Twitter posts. Number of collected tweets by using different hashtags are shown in Table 3.1. Two datasets of geo-tagged tweets are constructed from January to March 2021 and December to May 2022 by applying Python's Twint package.

3.2. Data pre-processing. It is used to perform various task given as: (i) data cleaning, (ii) tokenization, (iii) data normalization, (iv) data labelling and feature extraction, and (iv) feature selection. The following subsections provide a brief summary of each step.

Table 2.1: Name of Abbreviation and its full form.

Abbreviation	Full Form
NLP	Natural Language Processing
GRU	Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
WHO	World health organization
ROC	Receiver operating characteristic curve
BiGRU	Bidirectional GRU
CNN	Convolutional neural network
LSTM	Long short-term memory
VADER	Valence Aware Dictionary for Sentiment Reasoning
GloVe	Global vectors
ELMo	Embeddings from Language Models
CS	Cuckoo search
MAD	Mean absolute difference
BiLSTM	Bidirectional LSTM
SVM	Support Vector Machine
BERT	Bidirectional Encoder Representations from Transformers

Table 2.2: Summary of the literature review

Reference	Approach/Model
Chakraborty et al. [10]	deep learning based model
Majumder et al. [31]	SVM and Logistic Regression
Imran et al. [22]	LSTM+FastText
Chintalapudi et al. [14]	BERT
Basiri et al. [6]	Fusion-based

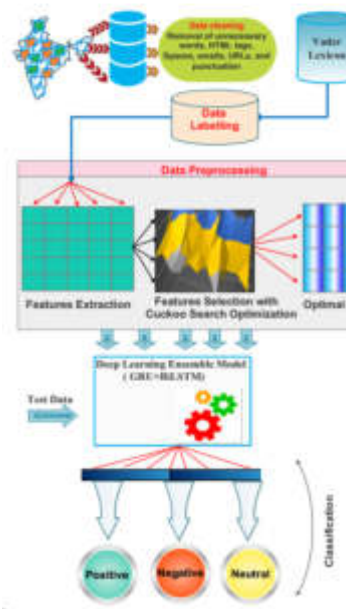


Fig. 3.1: Overview of proposed framework

Table 3.1: The number of Hashtags utilized for collecting the tweets.

Twitter hashtags	Number of tweets
#coronavirus	16,803
#indiafightscorona	14,705
#stayhome	16,106
#staysafe	18,117
#lockdownindia	20,501
#coronaindia	16,476
#covid19	17,295

Table 3.2: Count of labelled tweets by VADER technique

	DATA_SET 1	DATA_SET 1
Number of tweets in the dataset	1,200,000	1,500,000
Number of tweets containing emotion	1,200,000	1,500,000
Positive Twitter post	285,240	318,900
Negative Twitter post	436,080	530,100
Neutral Twitter post	478,680	651,000

i. **Data cleaning**

Original tweets collected from social media contain unwanted data. Data cleaning is used for removal of various unused terms such as: (a) spaces, emails, URLs, (b) additional characters in sentences, (c) punctuation (d) HTML tags, (e) unnecessary words and emojis from collected data. This step is carried out by using regular expressions.

ii. **Tokenization**

Primary purpose of this step is to determine the words or groups of words in a phrase which serves as the foundation for text analysis. In this process, entire text is broken down into smaller and more manageable chunks termed as ”tokens” [41]. The meaningful analysis of text is mainly based on the tokens rather than the entire text. Sample example of text and extracted tokens by tokenization process is given as:

iii. **Example Text** ”Our industry was almost destroyed by covid wear a mask”. **Extracted token set**: {’Our’, ’industry’, ’was’, ’almost’, ’destroyed’, ’by’, ’covid’, ’wear’, ’a’, mask’}.

iv. **Normalization**

This step considerably reduces the amount of data and improves the computing performance of algorithm [42]. This process converts all extracted tokens into a standard form that can be used to find similar-sounding words and eliminates the offending ones. In this work, three text normalization techniques namely, Lower-casing, stemming, and lemmatization are applied. All upper-case letters are converted into lower-case by Lower-casing. Stemming is used to remove all prefixes and suffixes from a phrase and return its basic structure. Lemmatization is employed to reduce dimensionality by classifying synonyms together.

v. **Data labelling**

In this work, lexicon approach based data labelling process known as VADER is applied to determine the polarity of tweets [35]. Each tweet is labelled as positive, negative, or neutral based on VADER compound score. Positive label is assigned to all data having greater than or equal to 0.05 compound score [21]. All data with scores between -0.05 and +0.05, less than -0.05 is assigned as neutral, and negative, respectively [21, 44]. Count of tweets labelled as positive, negative, and neutral by VADER technique is shown in Table 3.2.

3.2.1. Feature extraction. Further, all features are extracted by applying three embedding techniques: GloVe, FastText, and Elmo. This step minimizes memory needs and speeds up the processing of subsequent data.

- **FastText**
  FastText word embedding technique is a free, open-source package for text representations and classification [26]. It allows the framework to build a model quickly. This technique provides embedding of misspelled, unusual, and untrained words [33]. Numerous terms of dataset can not be embedded with FastText technique. It is also analyzed from the experiment that FastText technique missed 7,591 words from dataset for word embeddings. To overcome the limitation of FastText, Genism FastText algorithm is used to determine word vectors by using following equation:

$$\mu_s = \frac{1}{W} \sum_{q=1}^{W} d_q \log\left(r\left(MSa_q\right)\right) \tag{3.1}$$

  In this context, μ_s represents the cumulative log-likelihood, a_q is a one-hot-encoded word, S denotes lookup matrix used to find the word embedding in W documents [13], M represents continuous output transformation, softmax function is denoted by r.
- **Glove**
  Global Vectors (Glove) is an unsupervised learning technique used to construct low-dimensional word distributed representation [30]. It reduces dimension of word-context matrices by keeping track of word pair co-occurrences of corpus. It presents an attenuation function to compute weight based on distance between two words of context window [16].
- **ELMo**
  This technique is based on character and word-level context embedding [37]. It considers complete sentence while determining the appropriate embedding for each word. It uses bi-directional recurrent neural network to generate embeddings. The bidirectional embedding technique depends on both proceeding and succeeding word of sentence [48]. ELMo contextualizes each token by concatenating BiLSTM states.

3.3. Feature selection. Cuckoo search (CS) algorithm is used to select optimal features [46]. It is based on natural-inspired metaheuristic optimization approach and belongs to swarm intelligence family. Detailed description of fitness function and CS algorithm is given in next subsections.

3.3.1. Fitness function. Mean absolute difference (MAD) is used as the fitness function in this study to evaluate the significance of individual words. It accepts a ***weight_matrix*** as input, updates the weights of each word in the matrix depending on its fitness value, and generates a new ***updated_weight_matrix***. Fitness function determines the relevance of word depending on its weight. MAD calculates text's significance by comparing its mean value to mathematical phrase and it is given as follows:

$$MAD(Xh_m) = \frac{1}{Xh_m} \sum_{p=1}^{n} \mid r_m, p - \overline{r_b} \mid \tag{3.2}$$

$$\overline{r_b} = \frac{1}{Xh_m} \sum_{P=1}^{n} r_{m,P} \tag{3.3}$$

Here, Xh_m represents total number of text characteristics derived from sentence X. Average value of vector m is represented by r_m. $\overline{r_b}$ represents weight value of feature P in the text and n represents total number of text characteristics in original data set.

3.3.2. Cuckoo search algorithm. This user-friendly algorithm adheres the following basic principles [32]:
i. Cuckoo placed its egg in an arbitrary nest.
ii. The ideal nest produces high-quality eggs that carried down to the next generation.

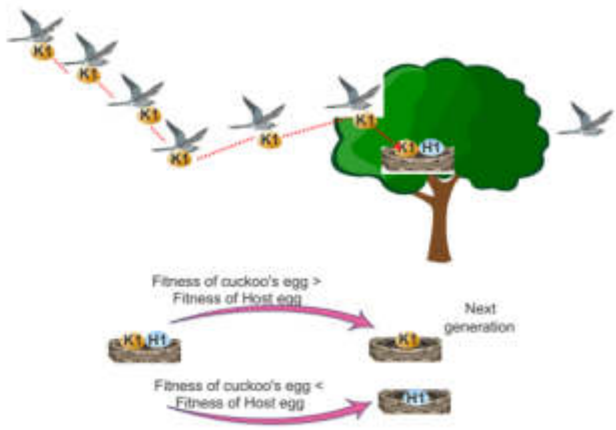


Fig. 3.2: Behaviour of cuckoo birds

iii. Chance of discovering cuckoo eggs by the host birds is Pa [0, 1] [45]. As soon as the host bird detects an alien egg, it can either (i) discard the eggs, or (ii) abandon the nest and build another one. Figure 3.2 depicts the behaviour and features of cuckoo birds.

Updated_weight_matrix is given as the initial population to the cuckoo search algorithm for optimal feature selection. It uses both local and global random walks to explore the search space [15]. Egg values are continuously updated using local random walk. Levy Flight algorithm is used to implement the global random walk [4]. Levy arrangement is used by this algorithm to determine stride length. The CS algorithm can efficiently probe the search space due to steady increment of step size [36]. Steps of Cuckoo search algorithm is presented in Algorithm 1.

Algorithm 1 Cuckoo Search Algorithm

1. Initialization of population for M nests, probability $Q_y \in [0, 1]$ and maximum maximum iterations $\{A\}$.
2. Set m= 0.
3. ***for*** (p = 1, p <= M) **do**
4. Take population $z_p(m)$ for host M.
5. Determine the fitness value of $h(z_p(m))$.
6. ***end for***
7. Make a new $z_{p+1}(m)$ solution randomly by Levy flight.
8. Determine the fitness value $h(z_p(m+1))$ for $z_{p+1}(m)$.
9. Randomly select a nest z_q from M
10. ***if*** $h(z_p(m+1)) > h(z_q(m)$ **then**
11. Substitute $(z_p(m+1)$ in-place of $(z_q(m+1))$.
12. ***end if***
13. Based on the Q_y value, the worst nests are discarded.
14. Replace the old ones nest with Levy flight.
15. Consider the finest options.
16. Sort the solutions and choose the best one.
17. increase the iteration counter $m = m + 1$.
18. Repeat step 7 to 17 until $m < A$.
19. Get the optimal solution.

Table 3.3: Parameter values initialized for cuckoo search

Parameter	Value
Probability (P_a)	0.25
Step scaling factor α	2
Number of iterations	765
σ_v	1
σ_μ	1.5

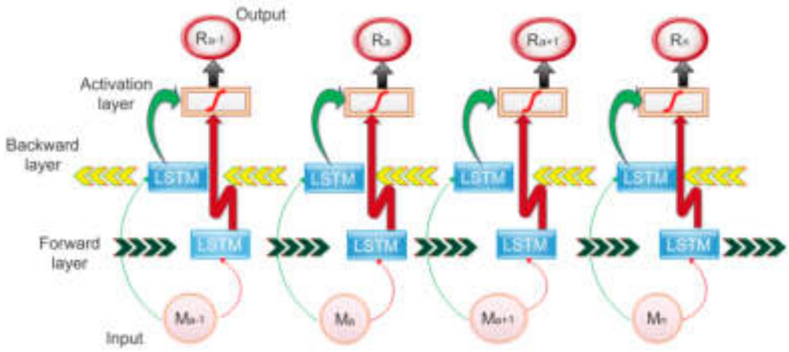


Fig. 3.3: Architecture of BiLSTM model

A new solution $z_p(m+1)$ for the cuckoo p is generated by the equation given as:

$$z_p(m+1) = z_p(m) + \alpha \oplus \text{Levy}(\lambda) \tag{3.4}$$

In this case, value of the current generation is represented by $m \in 1, 2, ..., A$, while A represents the maximum number of iterations. α denotes Levy flight with step length of 2 for the best search pattern with randomly placed objects [9]. Entry-wise multiplications are represented by $\oplus$. Random walk obtained with Levy distribution step is represented by the equation given as:

$$\text{Levy} \ = 0.01 \times \frac{\mu}{|v|^{1/\beta}} \times (g_{\text{best}} \ - z_p m)) \tag{3.5}$$

where μ =1.5 and v=1 denotes the normally distributed values, and g_{best} represents the current global best nest.

The cuckoo search technique returns an optimised weight matrix which is given as input to the classification model. List of parameter values set for cuckoo search algorithm is shown in Table 3.3.

3.4. Classification model.

i. **BiLSTM Model**

This work uses a variant of the LSTM network known as Bidirectional LSTM (BiLSTM) for sentiment classification. LSTM performs well with variable-length sequences but it cannot exploit contextual information from future tokens [38]. While BiLSTM combines past and future inputs of particular time step into LSTM model. BiLSTM employs a bidirectional LSTM layer to discover the patterns by traversing input data history into both directions. The first layer processed the forward sequence, while the backward sequence is handled by the second layer [28]. Figure 3.3 depicts the structure of BiLSTM model.

ii. **Gated Recurrent Unit Model (GRU)**

Architecture of this model is quite comparable with LSTM approach. It has its own dedicated memory

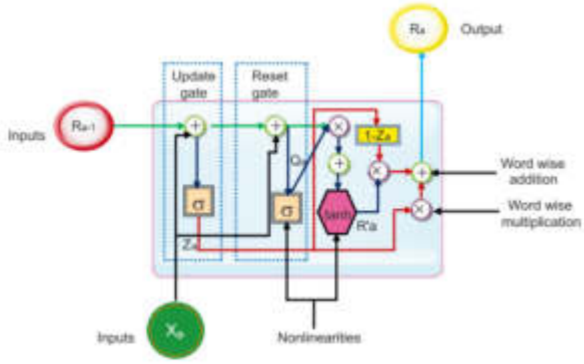


Fig. 3.4: Architecture of GRU model

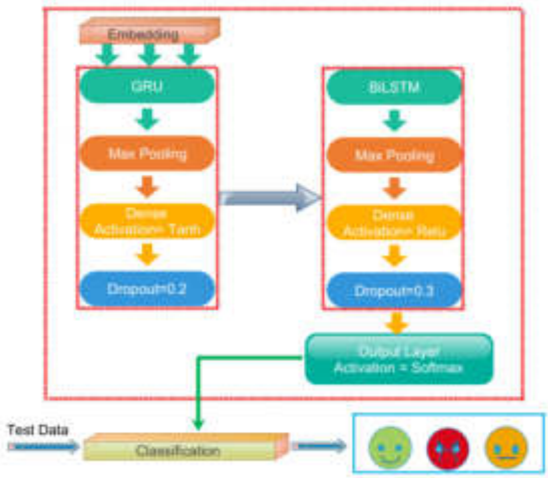


Fig. 3.5: Architecture of proposed model

and gating mechanism for controlling data transfer inside the unit. GRU uses two gates, (i) update and (ii) reset, to control the amount of information transferred or rejected from the preceding levels. Figure 3.4 presents the architecture of this model.

iii. **GRU-BiLSTM model**

It is a hybrid model of both GRU and BiLSTM approaches. First step is to feed the word embedding vector into GRU model with hidden layer. GRU layer transfers its output into dense layer of BiLSTM unit. Dropout layer is followed by dense network. Finally, sentiment classes are classified using softmax activation function. Hybrid model of the GRU- BiLSTM is shown in Figure 3.5.

4. Experiment and result. Quantitative sentiment analysis of COVID-19 tweets of Indian users are evaluated. The proposed work uses NVIDIA P2200 display card, 32 GB RAM with i8 E-2236 processor, and Python 3.7 language. Experimental results of sentiment analysis are discussed in the next subsections.

4.1. Results of sentiment analysis. Sentiment distribution of COVID-19 tweets of both datasets is shown in Figure 4.1. Tweets of first dataset are analyzed as 285,240 positive, 436,080 negatives, and 478,680

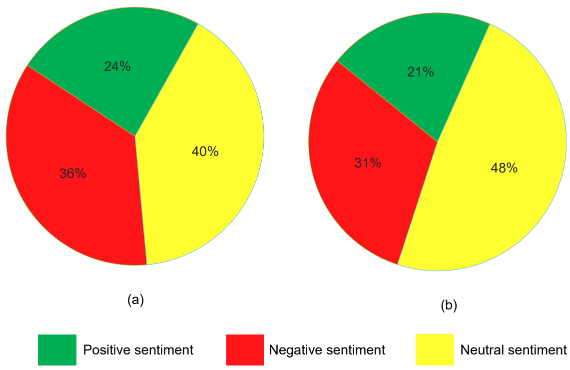


Fig. 4.1: Sentiment distribution of datasets

neutral tweets. For the second data set, 318,900, 530,100, and 651,000 tweets are categorized as favorable, negative, and neutral, respectively.

4.2. Classification results and discussion. Proposed model is used to classify sentiment of COVID-19 related tweets. Classification output is obtained by applying three models namely, (a) GRU, (b) BiLSTM, and (c) GRU+BiLSTM.
Various metrics such as F1-score, recall, precision, and accuracy are used to evaluate classifier performance, Description of each of these parameters are given as:

$$Accuracy = \frac{\sum(S\_F, S\_R)}{\sum(S\_F, U\_F, S\_R, U\_R)} \tag{4.1}$$

Specifically, S_F, S_R, U_F, and U_R denote correct identification, wrong identification, correct rejection, and wrong rejection.

$$Precision = \frac{(S\_F)}{\sum(S\_F, U\_F)} \tag{4.2}$$

$$Recall = \frac{(S\_F)}{\sum(S\_R, U\_R)} \tag{4.3}$$

$$F1 - score = \frac{2 * (Recall * Precision)}{\sum(Recall, Precision)} \tag{4.4}$$

The various parameter values taken for hybrid deep learning model are shown in table 4.1. Set of optimal parameter values are used to calibrate the experimental models. Dimension of embedding layer is initialized as 400. Suggested model uses hidden layer with 128 neurons and dropout layer with 0.7. Learning rate followed with Adam optimizer is set as 0.001. Softmax function is used to optimize output layer.

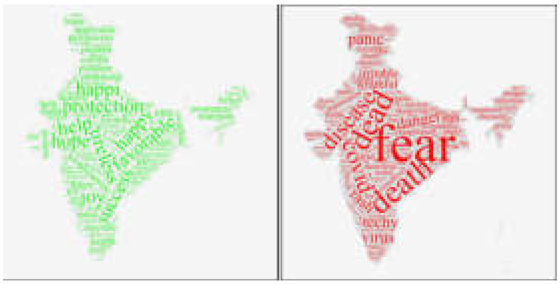
4.3. Comparison with Existing Model. The classifiers performance of Twitter sentiment analysis about COVID-19 with extracted features by using FastText, Glove, Elmo embedding method, and selected features with CS algorithm are shown in Table 4.2. It is analyzed from table 4.2 that the hybrid classifier gives improved values with selected features. Proposed model obtains the highest 94.44%, 88.53%, 90.34%, and 89.77% accuracy, sensitivity, precision, and F1-score.

WordCloud modules of words related to negative, positive, and neutral sentiment are visualized in Figure 4.2 (a), (b), and (c), respectively.

Table 4.3 shows performance comparison of proposed model with results obtained by existing authors.

Table 4.1: Different parameter values initialized for classifiers

Parameter name	GRU	BiLSTM	GRU+BiLSTM
Epoch	115	112	93
Batch size	32	64	64
Optimizer	Adam	Adam	Adam
Pooling layer padding	Same	Same	Same
Size of Max-pooling layer	2	2	2
Activation function	SGD	Relu	Relu
Filters	32	64	64
Kernel size	3	3	5
Dense layer	44	64	128
Momentum	0.7	0.7	0.7
Dropout layer	0.2	0.3	0.3
Model Learning rate	0.001	0.001	0.001



(a) (b)

(a)

Fig. 4.2: Most frequently words used as negative, positive, and neutral.

For sentiment classification, Chakraborty et al. [10] analyzed 249658 unique tweets by using deep learning-based model and achieved 81% accuracy. Majumder et al. [31] utilized SVM classifier and Logistic Regression to analyze COVID-19 Indian sentiments between March to June 2020. They have obtained 91.50% and 87.75% accuracy with SVM and Logistic Regression, respectively. Imran et al.[22] applied Long short-term memory model to extract sentiment polarity and emotions from tweets. They have obtained accuracy of 82% on sentiment140 dataset with FastText embedding.

Chintalapudi et al. [14] analyses Indian tweets using BERT model and achieved 89% accuracy. Basiri et al. [6] analyzed coronavirus-related tweets from eight countries and obtained 85.5% accuracy. Highest 94.44% accuracy is achieved by the proposed model which is higher than the results obtained with existing techniques [10, 31, 22, 14, 6].

Table 4.2: Results obtained with the various deep learning classifiers

Embedding	Model	Precision	F-score	Sensitivity	Accuracy
FastText	BiLSTM	71.82%	71.71%	72.16%	76.35%
	GRU	76.36%	71.49%	72.69%	78.21%
	GRU+BiLSTM	75.74%	73.40%	79.74%	79.32%
Glove	BiLSTM	78.83%	81.38%	79.87%	79.54%
	GRU	74.27%	80.41%	81.49%	82.74%
	GRU+BiLSTM	81.21%	79.70%	84.78%	85.11%
ELMo	BiLSTM	81.33%	84.63%	82.14%	80.54%
	GRU	79.27%	75.41%	80.17%	84.12%
	GRU+BiLSTM	84.19%	82.14%	79.69%	85.64%
ELMo+CS	BiLSTM	85.73%	84.37%	86.41%	85.54%
	GRU	90.77%	88.31%	86.49%	91.42%
	GRU+BiLSTM	90.34%	89.77%	88.53%	94.44%

Table 4.3: Comparison of the suggested model with existing work

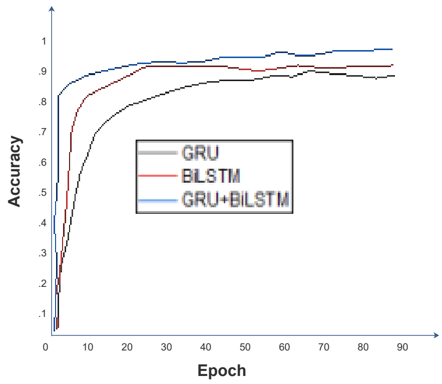
Reference	Approach/Model	Accuracy
Chakraborty et al. [10]	deep learning based model	81%
Majumder et al. [31]	SVM and Logistic Regression	91.50% and 87.75%
Imran et al. [22]	LSTM+FastText	82%
Chintalapudi et al. [14]	BERT	89%
Basiri et al. [6]	Fusion-based	85.5%
Propose model	GRU+BiLSTM	94.44%

Training accuracy and training loss obtained by GRU, LSTM, and hybrid model are illustrated in Figures 4.3 (a) and 4.3(b), respectively. Validation accuracy and loss are depicted in Figures 4.4 (a) and 4.4 (b), respectively. It is observed from Figures 4.3(a) and 4.4(a) that the classification accuracy of the proposed hybrid model is more significant as compare to individual GRU and LSTM models for maximum epochs. Furthermore, Figures 4.3 (b) and 4.4 (b) shows that hybrid model gives lower loss value as compare to other deep learning models. Confusion matrix of suggested model is depicted in Figure 4.5 (a). Mis-classification occurs when positive and negative samples are wrongly classified as neutral, while neutral emotion is incorrectly identified as negative.

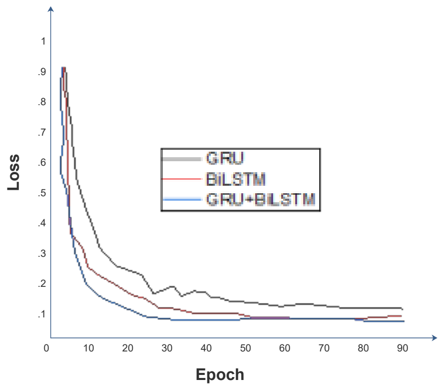
According to obtained confusion matrix, neutral feelings are more prevalent than positive and negative sentiments. Furthermore, comparison between the proposed and traditional deep learning model based on ROC and AUC measurement is shown in Figure 4.5 (b). Higher AUC indicates better model categorization. It is analyzed that the proposed model consistently outperformed with AUC of 0.97. Accuracy acquired by suggested model is shown as a boxplot also in Figure 4.6. It signifies consistent performance of the suggested model.

5. Conclusions. The proposed framework for sentiment analysis of Indians towards coronavirus using Twitter data achieved state-of-the-art performance. Proposed hybrid model of GRU and LSTM classification model outperformed individual and existing machine-learning techniques for COVID-19 emotion classification. The proposed model achieved a precision, F-score, sensitivity, and accuracy of 90.34%, 89.77%, 88.53%, and 94.44%, respectively. This is a significant improvement over previous approaches, which typically achieved 80% accuracy. The suggested hybrid model with enhanced word vector space improves the accuracy.

Results obtained from this research work can be used by policymakers and healthcare administrators to better understand the impact of COVID-19 on Indian society. The framework can also be used to monitor public sentiment about COVID-19 and to identify emerging trends. In future, proposed framework can be improved by adding part-of-speech annotations with individual words in the corpus. This will allow the framework to better understand the meaning of words and to achieve even higher accuracy. Additionally, proposed framework



(a)



(b)

Fig. 4.3: Training accuracy (a), Training loss (b) with respect to epochs

can be used to analyze vaccine-related tweets posted by Indian users. This will help to better understand the public sentiment about vaccines and to identify potential vaccine uptake challenges.

REFERENCES

[1] M. Abdulaziz, A. Alotaibi, M. Alsolamy, and A. Alabbas, *Topic based sentiment analysis for covid-19 tweets*, International Journal of Advanced Computer Science and Applications, 12 (2021).

[2] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, et al., *Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review*, Expert systems with applications, 167 (2021), p. 114155.

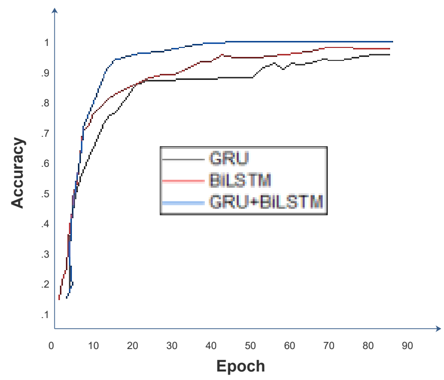
[3] S. D. Almotiri, *Twitter sentiment analysis during the lockdown on new zealand*, International Journal of Computer and Information Engineering, 15 (2022), pp. 649–654.

[4] M. A. E. Aziz and A. E. Hassanien, *Modified cuckoo search algorithm with rough sets for feature selection*, Neural Computing and Applications, 29 (2018), pp. 925–934.

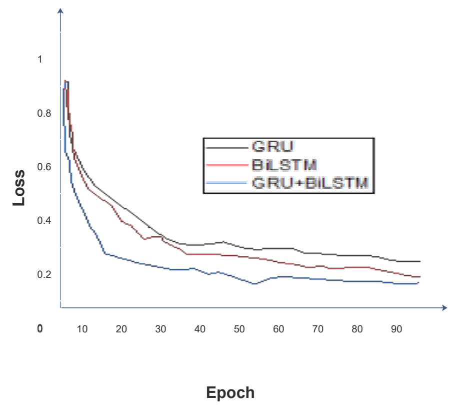
[5] K. B. Bahekar, P. Gautam, and S. Sharma, *Global sentiment analysis over third wave covid19 tweets*, International journal of health sciences, 6 (2022), p. 225–240.

[6] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, *A novel fusion-based deep learning model for sentiment analysis of covid-19 tweets*, Knowledge-Based Systems, 228 (2021), p. 107242.

[7] M. Bhat, M. Qadri, M. Kundroo, N. Ahanger, B. Agarwal, et al., *Sentiment analysis of social media response on the covid19 outbreak*, Brain, behavior, and immunity, 87 (2020), p. 136.



(a)



(b)

Fig. 4.4: Validation accuracy (a), Validation (b) with respect to epochs

[8] A. Borah, *Detecting covid-19 vaccine hesitancy in india: a multimodal transformer based approach*, Journal of Intelligent Information Systems, (2022), pp. 1–17.

[9] C. T. Brown, L. S. Liebovitch, and R. Glendon, *Lévy flights in dobe ju/'hoansi foraging patterns*, Human Ecology, 35 (2007), pp. 129–138.

[10] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, *Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media*, Applied Soft Computing, 97 (2020), p. 106754.

[11] D. Chehal, P. Gupta, and P. Gulati, *Covid-19 pandemic lockdown: An emotional health perspective of indians on twitter*, International Journal of Social Psychiatry, 67 (2021), pp. 64–72.

[12] J. Cheng, C. Huang, G. Zhang, D. Liu, P. Li, C. Lu, and J. Li, *Epidemiological characteristics of novel coronavirus pneumonia in henan*, Zhonghua jie he he hu xi za zhi= Zhonghua jiehe he huxi zazhi= Chinese Journal of Tuberculosis and Respiratory Diseases, 43 (2020), pp. 327–331.

[13] A. Chinnalagu and A. K. Durairaj, *Context-based sentiment analysis on customer reviews using machine learning linear models*, PeerJ Computer Science, 7 (2021), p. e813.

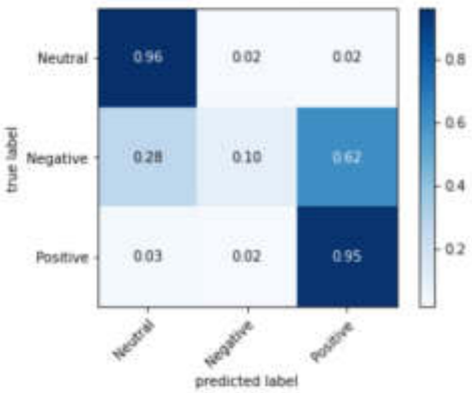
[14] N. Chintalapudi, G. Battineni, and F. Amenta, *Sentimental analysis of covid-19 tweets using deep learning models*, Infectious Disease Reports, 13 (2021), pp. 329–339.

[15] N. Dey, *Applications of cuckoo search algorithm and its variants*, (2020).

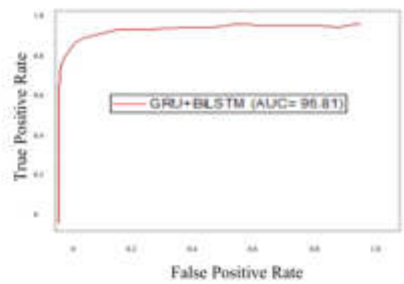
[16] A. Gelbukh, *Computational linguistics and intelligent text processing: 18th international conference, cicling 2017, budapest, hungary, april 17–23, 2017, revised selected papers, part ii*, 10762 (2018).

[17] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, *Sentiment analysis of lockdown in india during covid-19: A case study on twitter*, IEEE Transactions on Computational Social Systems, 8 (2020), pp. 992–1002.

[18] V. Gupta, N. Jain, P. Katariya, A. Kumar, S. Mohan, A. Ahmadian, and M. Ferrara, *An emotion care model using*



(a)



(b)

Fig. 4.5: Confusion matrix (a), ROC curve (b) with of proposed model

multimodal textual analysis on covid-19, Chaos, Solitons & Fractals, 144 (2021), p. 110708.
[19] V. Gupta, R. Piryani, V. K. Singh, and U. Ghose, *An analytical review of sentiment analysis on twitter*, Advances in Computing, Control, and Communication Technology, 1 (2016), pp. 219–225.
[20] Z. Hou, F. Du, H. Jiang, X. Zhou, and L. Lin, *Assessment of public attention, risk perception, emotional and behavioural responses to the covid-19 outbreak: social media surveillance in china*, MedRxiv, (2020).
[21] C. Hutto and E. Gilbert, *Vader: A parsimonious rule-based model for sentiment analysis of social media text*, in Proceedings of the international AAAI conference on web and social media, vol. 8, 2014, pp. 216–225.
[22] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, *Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets*, Ieee Access, 8 (2020), pp. 181074–181090.
[23] V. Jain and K. L. Kashyap, *Analyzing research trends of sentiment analysis and its applications for coronavirus disease (covid-19): A systematic review*, Journal of Intelligent & Fuzzy Systems, pp. 1–12.
[24] V. Jain and K. L. Kashyap, *Multilayer hybrid ensemble machine learning model for analysis of covid-19 vaccine sentiments*, Journal of Intelligent & Fuzzy Systems, (2022), pp. 1–13.
[25] V. Jain and K. L. Kashyap, *Ensemble hybrid model for hindi covid-19 text classification with metaheuristic optimization algorithm*, Multimedia Tools and Applications, 82 (2023), pp. 16839–16859.
[26] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, *Bag of tricks for efficient text classification*, arXiv preprint arXiv:1607.01759, (2016).
[27] H. Kour and M. K. Gupta, *Ai assisted attention mechanism for hybrid neural model to assess online attitudes about covid-19*, Neural Processing Letters, (2022), pp. 1–40.
[28] V. Krkova, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, *Artificial neural networks and machine learning–icann 2018: 27th international conference on artificial neural networks, rhodes, greece, october 4-7, 2018, proceedings, part iii*, 11141 (2018).
[29] V. Kumar, *Spatiotemporal sentiment variation analysis of geotagged covid-19 tweets from india using a hybrid deep learning*

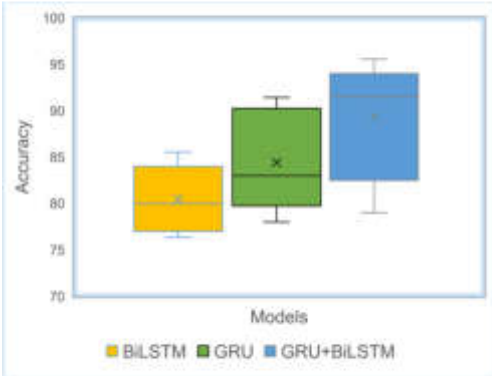


Fig. 4.6: Boxplot accuracy of proposed model

model, Scientific Reports, 12 (2022), pp. 1–14.
[30] P. Lauren, G. Qu, J. Yang, P. Watta, G.-B. Huang, and A. Lendasse, *Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks*, Cognitive Computation, 10 (2018), pp. 625–638.
[31] S. Majumder, A. Aich, and S. Das, *Sentiment analysis of people during lockdown period of covid-19 using svm and logistic regression analysis*, Available at SSRN 3801039, (2021).
[32] P. Melin, O. Castillo, and J. Kacprzyk, *Design of intelligent systems based on fuzzy logic, neural networks and nature-inspired optimization*, (2015).
[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781, (2013).
[34] P. Misra and J. Gupta, *Impact of covid 19 on indian migrant workers: decoding twitter data by text mining*, The Indian Journal of Labour Economics, 64 (2021), pp. 731–747.
[35] M. Mujahid, E. Lee, F. Rustam, P. B. Washington, S. Ullah, A. A. Reshi, and I. Ashraf, *Sentiment analysis and topic modeling on tweets about online education during covid-19*, Applied Sciences, 11 (2021), p. 8438.
[36] B. K. Panigrahi, P. N. Suganthan, S. Das, and S. S. Dash, *Swarm, evolutionary, and memetic computing: 4th international conference, semcco 2013, chennai, india, december 19-21, 2013, proceedings, part i*, 8297 (2013).
[37] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, *Deep contextualized word representations. arxiv 2018*, arXiv preprint arXiv:1802.05365, 12 (1802).
[38] Q. Qiu, Z. Xie, L. Wu, and L. Tao, *Dictionary-based automated information extraction from geological documents using a deep learning algorithm*, Earth and Space Science, 7 (2020), p. e2019EA000993.
[39] D. Sunitha, R. K. Patra, N. Babu, A. Suresh, and S. C. Gupta, *Twitter sentiment analysis using ensemble based deep learning model towards covid-19 in india and european countries*, Pattern Recognition Letters, 158 (2022), pp. 164–170.
[40] A. S. M. Venigalla, S. Chimalakonda, and D. Vagavolu, *Mood of india during covid-19-an interactive web portal based on emotion analysis of twitter data*, in Conference companion publication of the 2020 on computer supported cooperative work and social computing, 2020, pp. 65–68.
[41] T. Verma, *Renu and deepti gaur, tokenization and filtering process in rapidminer*, International Journal of Applied Information System (IJAIS), (2014).
[42] K. Welbers, W. Van Atteveldt, and K. Benoit, *Text analysis in r*, Communication Methods and Measures, 11 (2017), pp. 245–265.
[43] WHO, *COVID-19 Situation Report*. `https://www.who.int/docs/default-source/ coronaviruse/20200630-covid-19-sitrep -162.pdf`, 2020.
[44] X. Yang, C.-D. Wang, M. S. Islam, and Z. Zhang, *Advanced data mining and applications: 16th international conference, adma 2020, foshan, china, november 12–14, 2020, proceedings*, 12447 (2021).
[45] X.-S. Yang, *Cuckoo search and firefly algorithm: theory and applications*, 516 (2013).
[46] X.-S. Yang and S. Deb, *Cuckoo search via lévy flights*, in 2009 World congress on nature & biologically inspired computing (NaBIC), Ieee, 2009, pp. 210–214.
[47] W. Zhai, Z.-R. Peng, and F. Yuan, *Examine the effects of neighborhood equity on disaster situational awareness: Harness machine learning and geotagged twitter data*, International Journal of Disaster Risk Reduction, 48 (2020), p. 101611.
[48] C. Zong, J.-Y. Nie, D. Zhao, and Y. Feng, *Natural language processing and chinese computing*, (2015).

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jun 20, 2023
Accepted: Oct 10, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 873–882, DOI 10.12694/scpe.v24i4.2375

MULTILINGUAL CODE-MIXED SENTIMENT ANALYSIS IN HATE SPEECH

TULIKA RANJAN[*], ANISH SINGH[†], RINA KUMARI[‡], SUJATA SWAIN[§], ANJAN BANDYOPADHYAY[¶], AND AJAYA KUMAR PARIDA[‖]

Abstract. Sentiment analysis discovers the emotion expressed in a text. It helps in analyzing the product reviews, customer feedback and survey responses. Researchers have developed various algorithms for this purpose, however, they have majorly focused only on the sentiment analysis in English language. Although, few works are available for Hindi and multilingual sentiment analysis, however, these works are not efficient enough to perform sentiment analysis in code-mixed languages. To overcome the limitation of the existing works, this paper presents a multilingual code-mixed language model which identifies the sentiments of the hate speech dataset extracted from Twitter. As the hate speech dataset with sentiment labels are not available, we first collect the data from Twitter. After that we label the data using a transformer-based pretrained sentiment analysis model trained on a large corpus of tweets in multiple languages. We pass our collected data as test data to this model and predict the sentiment labels. Now, we train six different machine learning models to perform our own task i.e sentiment analysis for multilingual code-mixed hate speech dataset. The machine learning models perform well across multiple languages and also code-mixed languages. In future, it can be easily adapted to different classification tasks based on code-mixed languages. The results yield that hate speech invokes negative sentiment whereas non-hate speech reflects either positive or neutral sentiment.

Key words: Code-Mixed, Multilingual text data, Sentiment analysis, Hate speech, Natural language processing, Machine learning

1. Introduction. The internet has facilitated communication and sharing of opinions but also enabled the spread of hate news that target and harms individuals and communities based on their appearance, religion, or sexual orientation. India, with its diverse linguistic and cultural landscape, has become a hotbed for spreading hatred through online platforms. Prior investigations on hate speech detection have majorly focused on high-resource languages like English, but the prevalence of code-mixing in Indian languages like Hindi-English (Hinglish) calls for more attention to detect hate speech in such multilingual contexts [5, 6]. Code-mixing [13, 10] refers to the practice of using words and phrases from multiple languages in a single sentence or expression. The dissemination of hate news in a multi-lingual society like India is a challenging issue due to a lack of media regulation and verification. Hate news can manipulate users for financial, religious, or political purposes and harm society as a whole. Recent incidents in India, including hate-mongering during political rallies and racial discrimination during the COVID-19 pandemic, have underscored the need to prevent the transmission of hate speech through online platforms. Detecting hate speech in code-mixed text requires the development of AI models that can accurately interpret and identify hateful content [8]. The ability to detect and prevent hate speech in multilingual contexts has wider implications for promoting diversity, equity, and inclusion in online spaces. The development of such models can also aid in the creation of safer and more welcoming online communities for people from diverse linguistic and cultural backgrounds. Table 1.1 depicts some samples from the collected dataset.

Below examples list some instances of Hindi-English code-mixed text. It also discusses the translated version of the instance in English.

[*]School of Computer Engineering, Kalinga Institute of Industrial Technology, India (tulikaranjan2002@gmail.com).
[†]School of Computer Engineering, Kalinga Institute of Industrial Technology, India (anishsingh1452002@gmail.com).
[‡]School of Computer Engineering, Kalinga Institute of Industrial Technology, India (rina.kumarifcs@kiit.ac.in).
[§]School of Computer Engineering, Kalinga Institute of Industrial Technology, India (sujata.swainfcs@kiit.ac.in).
[¶]School of Computer Engineering, Kalinga Institute of Industrial Technology, India (anjan.bandyopadhyay@kiit.ac.in).
[‖]School of Computer Engineering, Kalinga Institute of Industrial Technology, India (ajaya.paridafcs@kiit.ac.in).

Table 1.1: Samples of dataset with sentiment labels. Here Neg, Pos and Neut represent Negative, Positive and Neutral, respectively

Sentence	**Label0 (Neg)**	**Label1 (Neut)**	**Label2 (Pos)**
Mera Kaam Office mai hai		✓	
Aaj me woh moive dekhne jaa rahi hu.		✓	
We hate him utna hi jitna we hate you.	✓		
Mujhe maths se nafrat hai. I wish it didn't exist	✓		
He loves to play Gilli danda			✓

Examples of Code Mixed
Text-instance-1 : *"Mera kaam office mein hai."*
Translation : *"My work is in the office."*
Hate Speech Label : *"Non Hate"*
Sentiment Label : *"Neutral"*
Text-instance-2 : *"Maine apna homework kiya hai."*
Translation : *"I have done my homework."*
Hate Speech Label : *"Non Hate"*
Sentiment Label : *"Neutral"*
Text-instance-3 : *"Tumne mujhe email kiya tha kya?"*
Translation *"Did you send me an email?"*
Hate Speech Label : *"Non Hate"*
Sentiment Label : *"Neutral"*

Examples of Code Mixed Hate Speech
Text-instance-1 : *"We hate him utna hi jitna we hate you."*
Translation : *"We hate him as much as we hate you."*
Hate Speech Label : *"Hate"*
Sentiment Label : *"Negative"*
Text-instance-2 : *"Mujhe maths se nafrat hai. I wish it didn't exist."*
Translation : *"I hate maths. I wish it did not exist."*
Hate Speech Label : *"Hate"*
Sentiment Label : *"Negative"*
Text-instance-3 : *"He loves to play Gili danda."*
Translation : *"He loves to play Gili danda."*
Hate Speech Label : *"Non Hate"*
Sentiment Label : *"Positive"*

In light of the above examples, we can analyze that hate speech provoke negative sentiment whereas non hate provoke either positive or neutral sentiments. The negative sentiments provokes people to spread the hate speech on social media platforms which negatively impacts the individual and society and led to major harm. Therefore, It is important to prevent the spread of hate speech on social media and for this purpose sentiment analysis can be proved a key aspect. By motivating this idea, we train various machine learning models which perform sentiment analysis in the multilingual code-mixed hate speech collected from twitter. As per our investigations, this is a novel approach for multilingual code-mixed languages more specifically in hate speech.

1.1. Problem Statement And Specification Requirement. The problem statement of the Multilingual Sentiment analysis model is to perform sentiment analysis on code-mixed multilingual hate speech tweets.

This requires the model to be able to accurately classify tweets as positive, negative, or neutral, regardless of the language used in the tweet.

To accomplish this, the model needs to meet the following requirements:

1. Process text in multiple languages
2. Accurately identify sentiment polarity in tweets
3. Perform well on a variety of datasets and languages
4. Handle noisy data and non-standard language usage in tweets
5. Efficient and salable enough to be used in real-world applications

We fulfill the above mentioned requirements in the following subphases:

1. Problem definition: We first define the problem statement and goals of the investigation which includes gathering of the specific objectives of this investigation and the target audience for the sentiment analysis. Here, we provide hate and non hate speech data in text format to various machine learning models as input and the model detects the sentiment (positive, negative or neutral) of those sentences.
2. Data source identification: In this subphase, we identify the data sources which are used to collect the data for the sentiment analysis. For our model, we extract the data from twitter.
3. Data cleaning and preparation: Data cleaning is essential for accurate sentiment analysis. In this stage, we remove the irrelevant data, perform text normalization, and prepare the data for analysis.
4. Model selection: Since our collected dataset is not labeled with the sentiment classes, after data pre-processing, we select a pre-trained sentiment analysis model to get the sentiment labels. After that, we select six machine learning models to perform sentiment analysis in code-mixed hate speech dataset.
5. Model training: After selecting the appropriate pre-trained model, we pass our own collected data as the test data to the model and the model predicts the sentiment labels for the data instance. Further, we train six machine learning models using the labeled code-mixed multilingual hate speech dataset (our dataset). We perform cross-validation to evaluate the model's performance and make any necessary adjustments. Further, we consider precision, recall, F1-score, accuracy, model complexity, and computational resources as the measuring factors to measure the efficiency of the models.
6. Model testing: After training, we test the models on a separate data set to evaluate their performance.

2. Literature Review. People share their thoughts in social media. However, they are not only using English language but also they mixed their own mother-tongues language with it. As India has 22 different languages. So, social media contents moreover are the code mixed information. There are different studies that focus on the developing a model for the code mixed sentiment analysis in recognizing hate speech. The study presented in [4] demonstrates the hate speech detection problem in code-mixed texts. In this work, the authors have first developed a Hindi-English code-mixed dataset which contains tweets posted online on Twitter. The tweets are annotated with the language at word level and the class they belong to (Hate Speech or Normal Speech). The authors have also proposed a supervised classification framework for hate speech detection in the text using different lexicon-based, character level and word level features. This task holds significant relevance for numerous applications, such as cyberbullying investigation, sentiment analysis, and examining socio-political controversies. The work proposed in [2], discusses the hate Speech detection task for Code-mixed text in Tamil and English languages. They first developed a dataset with 10,000 Tamil-English code-mixed texts collected from Twitter. Also, each text is annotated with hate or non-hate text. After preparing the dataset, they have developed a synonym-based Bi-LSTM model for classifying hate and non-hate text in tweets. In [5], Hinglish hate speech detection is discussed. In this study, the text contains both Hindi and English. The proposed model[5] is based on ensemble method. It has classified the text 3 categories: Abusive, Non-Offensive and ate-Inducing. In [11], the authors have addressed the growing concern of hate speech in user-generated content, specifically on social media platforms. The authors highlight the need for automated detection of hateful content to counteract these harmful activities. They have identified hate speech within code-mixed social media text. In [12], the proposed model is detecting of hate speech text in Hindi English code mixed data. In [14, 1, 15], the authors provides a comprehensive review of studies on deep learning approaches for multilingual sentiment analysis of social media. In [8], the authors have designed a user interface based on a web browser plugin over Facebook and Twitter to visualize the aggressive comments posted on the Social media user's timelines. In [3, 9], an ensemble model for code-mixed data of hate speech classification task on Hindi-English data is

Table 3.1: Dataset distribution across various classes

Labels	Positive	Negative	Neutral
Classified in %	30.78%	38.34%	30.87%

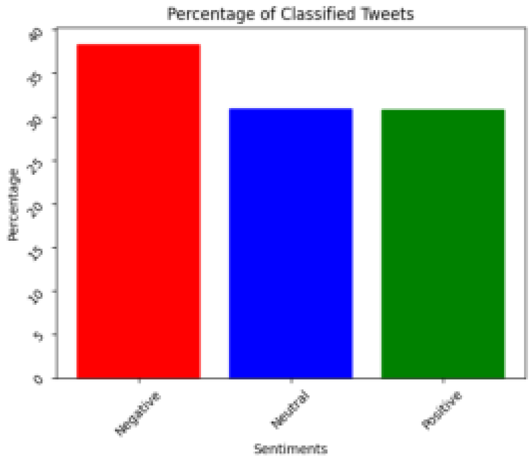


Fig. 3.1: Percentage of Classified Tweets

proposed. Also, it compares with the pretrained models and validates the proposed approach. In [7], a code-mixed English–Hindi dataset with a well-defined context is developed. It proposes context representation for conversational dialogue.

3. Overview of Dataset. The dataset utilized in our study consists of 9700 texts categorized into three classes: Positive, Negative, and Neutral. This dataset provides a solid foundation for training and evaluation of our sentiment analysis model, enabling a detailed exploration of sentiment patterns and trends within the data. The initial dataset use for training the pre-trained model was sourced from a publicly available site Kaggle[1] which consists of 1,600,000 tweets extracted using the twitter API. The dataset instance includes texts representing various sentiment categories (0 = negative, 2 = neutral, 4 = positive). These texts were thoughtfully chosen to encompass a broad range of sentiments, covering diverse topics and contexts. We use this dataset to train the machine learning models.

For the second segment of the dataset, we utilized the SNS scrape Twitter module to collect additional data. This module enabled us to retrieve texts from Twitter, specifically focusing on tweets related to our target domain. By leveraging the functionality provided by the SNS scrape Twitter module, we extracts a substantial number of tweets, capturing real-time user opinions and expressions of sentiment. We use this dataset to train the machine learning models for our own task i.e sentiment analysis on code-mixed hate speech dataset. Table 3.1 explains the distribution of the dataset across sentiment classes and Figure 3.1 shows its graphical representation. In the process of collecting the hate speech dataset from Twitter, we implemented several strategies to ensure its representativeness in terms of language diversity and code-mixed patterns. We utilized specific keywords and language filters to target relevant content, and we developed techniques to identify code-mixed language patterns within the collected data. These measures were taken to ensure that the dataset encompasses a wide range of multilingual code-mixed hate speech scenarios.

4. Methodology. We develop the multilingual code-mixed sentiment analysis model in several phases, including data collection, pre-processing, model training, and evaluation. We collect the dataset from Twitter/Kaggle and pre-process using various techniques such as tokenization, normalization, and cleaning. After

[1] https://www.kaggle.com/

preparing the dataset, we first train the various machine learning models using a generic sentiment analysis dataset (briefly discussed in dataset section) and then we fine-tune the trained model on the pre-processed dataset for the sentiment analysis on multilingual code-mixed hate speech dataset. Further, we evaluate the model using various metrics such as accuracy, precision, recall, F1-score, and confusion matrix. We also conduct the error analysis to identify the common types of errors made by the model and address them. Overall, this investigation involves several steps and techniques to develop and evaluate the model for sentiment analysis on in multilingual code-mixed hate speech Tweets. We briefly discuss all these steps in subsequent subsections.

4.1. Pre-processing. To develop any data dependant model, data preprocessing is a key task which boosts the performance of the model. Therefore, we first pre-process the dataset following techniques described below:

1. **Removing Noise:** Noise removal techniques were implemented to eliminate irrelevant or undesirable elements from the Twitter data. This involved filtering out special characters, URLs, hashtags, and mentions that would not significantly contribute to sentiment analysis.
2. **Handling The Emoticons:** Cleaning emoticons from the dataset in sentiment analysis ensures a more focused analysis of the textual content, promotes standardization, reduces noise, and enhances compatibility with text-based sentiment analysis models. To handle this, a specific pre-processing step was likely employed to appropriately manage emoticons. This could include replacing them with sentiment indicators or mapping them to corresponding textual representations that reflect their intended emotions. This can improve the accuracy and reliability of sentiment analysis results on Twitter and other social media platforms.
3. **Tokenization:** Another essential pre-processing step performed on the Twitter data. Tokenization involves breaking down the text into individual units, such as words or subwords, to enable more effective analysis. By segmenting the data into meaningful units, each unit could be processed independently, facilitating subsequent analysis.
4. **Lowercasing:** To ensure standardization and avoid treating the same word with different cases as distinct entities, the text was converted to lowercase during pre-processing. This step aimed to achieve consistency in word representation and ensure that words like ”happy” and ”Happy” were treated as identical words during sentiment analysis.
5. **Removal of Stop Words:** Stop words are commonly used words in a language that do not carry significant meaning or contribute much to the overall understanding of a sentence. Examples of stop words in English include ”and”, ”the”, ”is” etc. And in Hindi “hai”,“ki”,“ka” etc. Stop words were removed from the English text using a predefined tool provided by NLTK (Natural Language Toolkit). NLTK offers a set of pre-defined stop words for the English language, which were employed to filter out common and non-informative words. This step aimed to reduce noise and enhance the accuracy of sentiment analysis by focusing on more meaningful content words. For the Hindi data set, a separate set of stop words was predefined to cater to the specific language. By utilizing these language-specific stop words, the model ensured the removal of irrelevant words in Hindi text, enabling more effective sentiment analysis.

4.2. Model selection and training. After pre-processing, we select a pre-trained model trained for the sentiment prediction of social media text. We select ”cardiffnlp/twitter-xlm-roberta-base-sentiment,” model which demonstrate the effectiveness in capturing sentiment information from social media data. We fine-tune this model for our task and during fine-tuning, the model's parameters are adjusted to optimize its performance in accurately predicting sentiment in social media text. Once the training process becomes complete, the trained model is utilized to perform sentiment analysis on the pre-processed text. This involved feeding the pre-processed text into the model and leveraging its learned representations to predict sentiment labels associated with the text. To prepare the text for analysis, we employ the model's tokenizer to tokenize our pre-processed data. We then pass the tokenized text as input to the pre-trained model. The model then predicts the sentiment labels for the data instances of our collected dataset. After getting the labeled dataset, we train six different machine learning models *viz.* Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Support Vector Machine and Ensemble learning Model for the final sentiment prediction of the multilingual code-mixed hate speech tweets.

4.3. Evaluation and Performance Metrics. We calculate various performance metrics to evaluate the performance of the sentiment analysis models. These metrics include accuracy, precision, recall, and F1-score. These metrics served as essential indicators of the model's effectiveness in accurately classifying sentiment. Accuracy measures the overall correctness of sentiment predictions, while precision quantifies the proportion of correctly predicted positive or negative sentiments out of all positive or negative predictions, respectively. Recall gauges the proportion of correctly identified positive or negative sentiments out of all actual positive or negative instances. The F1-score, which combined precision and recall, provided a balanced measure of the model's performance. Equation 4.1, 4.2, 4.3 and 4.4 show the mathematical equations of the above discussed metrics.

$$Accuracy = \frac{Correctly\ Classified\ Instance}{Total\ Instance} \tag{4.1}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4.2}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.4}$$

Figure 4.1 shows the value of above discussed evaluation metrics. Figure 4.2 visually represents the accuracy of the ensemble model. Furthermore, we analyze the performance of different machine learning models based on these metrics values. We can conclude that ensemble model gives better performance compared to the traditional machine learning models such as Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression and Support Vector Machine. This comparison aims to determine the superiority of the ensemble learning model in multilingual code-mixed sentiment analysis tasks for hate speech data, showcasing its strengths and advantages over traditional classifiers.

In order to visually represent the classifier's performance, we plot a confusion matrix (shown in Figure 4.3) for the ensemble model. The confusion matrix provides a comprehensive view of the classification results by displaying the true positive, true negative, false positive, and false negative predictions. By examining the confusion matrix, patterns of mis-classifications and the overall performance of the model can be easily observed and analyzed.

4.4. Result Analysis. The `sentiment_analysis` function plays a pivotal role in performing sentiment analysis on an instance using the `sentiment_task` pipeline. With the input of an instance, this function leverages a sentiment analysis model to determine the sentiment associated with it. It is reasonable to assume that the function yields a sentiment label or category as the output, representing the sentiment polarity of the instance, such as positive, negative, or neutral. To offer a user-friendly interface for the sentiment analyzer, we design our own interface using the Gradio library[2]. Through the gr. Interface function, the interface is defined, specifying the input type as ”text” and the output type as ”text”. Additionally, the interface is accompanied by a title and description to enhance user understanding and interaction. Figure 4.4 and 4.5 are the example of sentiment label predictions obtained by the developed interface.

5. Conclusion. The multilingual code-mixed sentimental analysis model showcased remarkable precision in forecasting the sentiment polarity of hate speech tweets in multiple languages, comprising low-resource languages. The adoption of cross-lingual pre-training and fine-tuning approaches empowered the model to efficiently grasp and transfer knowledge across languages, thereby leading to better performance as opposed to monolingual models. We have also trained six different machine learning models such as Naive Bayes,

[2]https://gradio.app/

Bernoulli Naive Bayes	Prescision	Recall	F1-Score	Support
Negative	0.66	0.79	0.72	752
Neutral	0.67	0.50	0.57	627
Positive	0.72	0.73	0.73	561

Original Naive Bayes	Prescision	Recall	F1-Score	Support
Negative	0.65	0.78	0.71	752
Neutral	0.69	0.41	0.51	627
Positive	0.75	0.76	0.70	561

Multinomial Naive Bayes	Prescision	Recall	F1-Score	Support
Negative	0.66	0.79	0.72	752
Neutral	0.68	0.45	0.54	627
Positive	0.68	0.76	0.72	561

Logistic Regression	Prescision	Recall	F1-Score	Support
Negative	0.77	0.71	0.74	752
Neutral	0.65	0.73	0.69	627
Positive	0.76	0.74	0.65	561

Support Vector Classifier	Prescision	Recall	F1-Score	Support
Negative	0.75	0.67	0.71	752
Neutral	0.62	0.74	0.67	627
Positive	0.78	0.70	0.73	561

Ensemble Learning	Prescision	Recall	F1-Score	Support
Negative	0.67	0.79	0.72	752
Neutral	0.69	0.52	0.59	627
Positive	0.71	0.75	0.73	561

Fig. 4.1: Performance Metrics for Various Classifiers

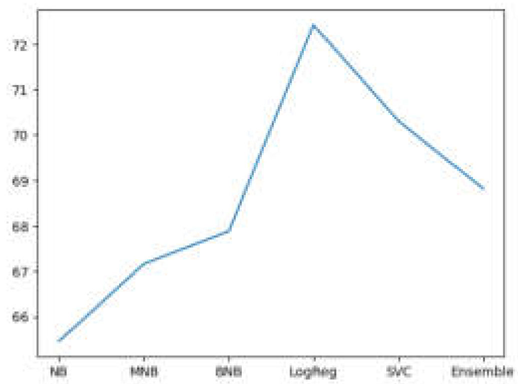


Fig. 4.2: Accuracy graph

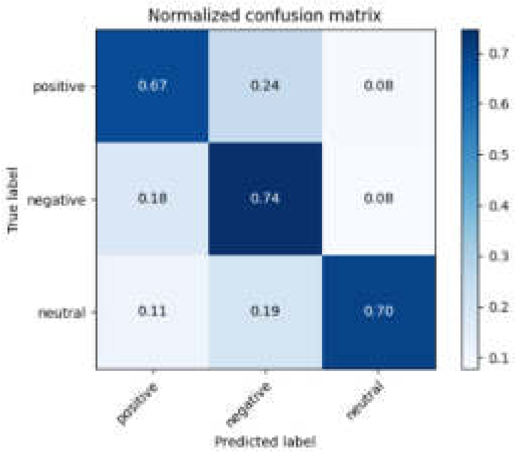


Fig. 4.3: Normalized confusion matrix for ensemble model

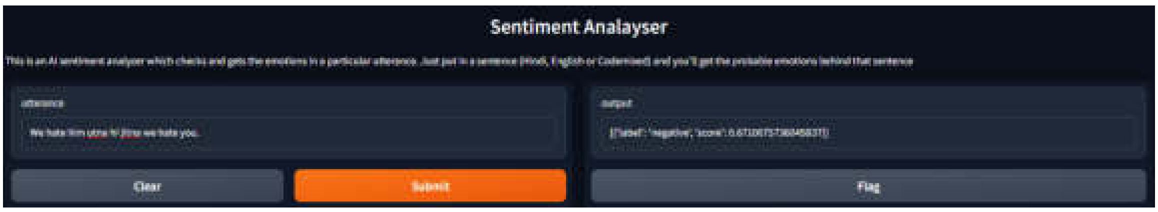


Fig. 4.4: Negative Classified Text

Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Support Vector Machine and Ensemble Learning Model. Among these six models, Ensemble learning model performs better compared to others. Despite this, the model has certain shortcomings and obstacles, such as struggles in detecting sarcasm and irony, and the necessity for significant computing resources. To enhance the performance of multilingual code-mixed sentiment analysis models, future research can address these difficulties.

6. Future Scope.

6.1. The occurring problem. Detecting hate and offensive content in social media is challenging due to linguistic complexity, and non-standard variations in grammar, spelling, and translation of language. While much work has been done on hate speech detection from tweets in English, the usage of regional languages in social media has increased the need for researchers to find hate speech from tweets. However, labeling tweets presents a challenge as a tweet can fall into multiple categories, and there may be subjective biases in labeling tweets. Programmed hate speech detection is also a closed system, and people try to circumvent the detection by posting content as images instead of text. To tackle this issue, optical character recognition can be used, but it is a constant battle between those spreading hateful content and those trying to block it. Furthermore, there are changing mentalities towards subjects over time and in historical context, which presents an additional challenge for programmed hate speech detection.

6.2. What can be done in the future to overcome the Problem. Future research can address the challenges faced by hate speech detection systems. One approach is to develop more accurate and unbiased labeling methods to minimize the subjective bias in labeling the text. Another solution is to incorporate context and historical information to understand the cultural and societal context of the text. Additionally, developing more robust models that can handle variations in language and user behavior can improve the performance of hate speech detection systems. Another solution is to integrate multiple sources of information, such as images

Fig. 4.5: Positive Classified Text

and user profiles, to gain a more comprehensive understanding of the content. Finally, developing systems that can adapt and learn from new data and changing user behavior can improve the long-term effectiveness of hate speech detection.

7. Limitation. This model requires a large amount of computing resources to train and use, which may be a limitation for some applications. The model may not perform well on datasets that are significantly different from the ones it was trained on, particularly if those datasets contain language or cultural nuances that are not well-represented in the training data. The model may struggle with tweets that use non-standard or informal language, such as slang or dialects until it is mentioned in the dataset. While the Multilingual Sentimental Analysis model outperforms other state-of-the-art models on several benchmark datasets, its performance may not be the best for every dataset and task, and it is always necessary to evaluate its performance on specific applications. We can avoid the resource limitations using pre-training concepts. We can also avoid the language bias limitations using transfer learning, by training various models on the datasets of numerous languages and combining those models to make an ensemble model for multilingual code-mixed sentiment analysis. We will explore these methods in our subsequent works.

REFERENCES

[1] M. M. Agüero-Torales, J. I. A. Salas, and A. G. López-Herrera, *Deep learning and multilingual sentiment analysis on social media data: An overview*, Applied Soft Computing, 107 (2021), p. 107373.

[2] S. Anbukkarasi and S. Varadhaganapathy, *Deep learning-based hate speech detection in code-mixed tamil text*, IETE Journal of Research, (2022), pp. 1–6.

[3] S. Banerjee, B. R. Chakravarthi, and J. P. McCrae, *Comparison of pretrained embeddings to identify hate speech in indian code-mixed text*, in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, 2020, pp. 21–25.

[4] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, *A dataset of hindi-english code-mixed social media text for hate speech detection*, in Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, 2018, pp. 36–41.

[5] V. Gupta, V. Sehra, Y. R. Vardhan, et al., *Ensemble based hinglish hate speech detection*, in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2021, pp. 1800–1806.

[6] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, *Hate speech detection: Challenges and solutions*, PloS one, 14 (2019), p. e0221152.

[7] H. Madhu, S. Satapara, S. Modha, T. Mandl, and P. Majumder, *Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments*, Expert Systems with Applications, 215 (2023), p. 119342.

[8] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, *Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance*, Expert Systems with Applications, 161 (2020), p. 113725.

[9] S. K. Mohapatra, S. Prasad, D. K. Bebarta, T. K. Das, K. Srinivasan, and Y.-C. Hu, *Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques*, Applied Sciences, 11 (2021), p. 8575.

[10] P. Rani, S. Suryawanshi, K. Goswami, B. R. Chakravarthi, T. Fransen, and J. P. McCrae, *A comparative study of different state-of-the-art hate speech detection methods in hindi-english code-mixed data*, in Proceedings of the second workshop on trolling, aggression and cyberbullying, 2020, pp. 42–48.

[11] T. Santosh and K. Aravind, *Hate speech detection in hindi-english code-mixed social media text*, in Proceedings of the ACM India joint international conference on data science and management of data, 2019, pp. 310–313.

[12] K. Sreelakshmi, B. Premjith, and K. Soman, *Detection of hate speech text in hindi-english code-mixed data*, Procedia Computer Science, 171 (2020), pp. 737–744.

[13] S. Thara and P. Poornachandran, *Code-mixing: A brief survey*, in 2018 International conference on advances in computing, communications and informatics (ICACCI), IEEE, 2018, pp. 2382–2388.
[14] E. Tromp, *Multilingual sentiment analysis on social media*, Master's Theisis. Department of Mathematics and Computer Science, Eindhoven University of Technology, (2011), pp. 1–104.
[15] W. Yin and A. Zubiaga, *Towards generalisable hate speech detection: a review on obstacles and solutions*, PeerJ Computer Science, 7 (2021), p. e598.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jun 27, 2023
Accepted: Sep 26, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 883–894, DOI 10.12694/scpe.v24i4.2393

PERSONALIZED ART WORK RECOMMENDATION SYSTEM AND METHODS BASED ON USER INTEREST CHARACTERISTICS AND EMOTIONAL PREFERENCES

HUIHONG LI*

Abstract. To familiarize users with their interests and hobbies through online data collection, and improve their experience when browsing art works, research based on K-means algorithm has received widespread attention. However, with the explosive growth of various types of art works, it is difficult to estimate the K value of the K-means algorithm when processing these data. To solve this problem, this research predicts the user behavior of Wink dataset based on K-means algorithm, introduces regularization specified process and emotional precision, and generates fusion algorithm. The study first proposes the concept of similar users and calculates the Pearson correlation coefficient between them to determine their similarity; Then several regularization terms are added to the user group, and the prediction results are obtained by changing the parameters; Further screening of art works clustering categories is to address the issue of slow user startup. Finally, the algorithm studied will be applied to the Wink dataset and the prediction accuracy of the particle swarm optimization algorithm will be tested and compared with the fusion algorithm. A total of 400 experiments are conducted, and the fusion algorithm achieve a prediction accuracy of 392 times, with an accuracy rate of 98.0%; The accuracy of particle swarm optimization algorithm is close to that of fusion algorithm, at 88.2%. The experimental results show that the algorithm model proposed in the study can effectively map the relationship between user interest features, emotional factors, and personalized art recommendation, thereby providing users with a good viewing experience.

Key words: Pearson correlation coefficient; Emotional precision; Regularization specified process; K-means algorithm; Personalized recommendations

1. Introduction. In a society with explosive information growth, with the rapid development of recommendation-based algorithms, users can obtain the information they need online and solve the problems they face [1, 2]. The predecessor of this type of algorithm is a search engine, which can help users search for areas of interest. However, when users face something for the first time, they may encounter situations such as unclear expression, which can lead to ineffective search engine work. By comparison, personalized recommendation algorithms can predict user preferences in daily life by understanding their interest characteristics and emotional preferences, effectively avoiding the unclear user explanations. However, user data also contains special interests, and traditional algorithms set them as isolated points. The constructed model will exclude these points, causing significant errors in the prediction results [3]. In recent years, the K-means algorithm has attracted the attention of many scholars due to its simple and efficient clustering research. The K-means algorithm sets the clustering center point during operation to attract similar elements to approach. However, when K-means processes a large amount of data, it is difficult to estimate the number of center points, which prolongs its iteration time. To improve this situation, based on user interest characteristics and emotional factors, this research introduces the regularization Prescribed Process (RPP) and Emotional Accuracy (EA) to quantify the two and generate a fusion algorithm (PEK-means). This algorithm is used to cluster and filter art collections and users is expected to reduce algorithm runtime. The research is mainly divided into four parts. The first part mainly analyzes and summarizes the application and effectiveness of current user interest models and user sentiment models; The second part introduces the factors that affect prediction accuracy and constructs a PEK-means prediction model; The third part analyzes and compares the performance of this optimization model with traditional models; The final part is conducted through simulation experiments on the Wink dataset, highlighting the shortcomings that still exist in the research. The practical significance of this study is to learn users' preferences through data analysis, thus increasing their viewing experience. The purpose is to recommend the works of art that users like, and then increase the awareness of works of art.

*Academy of Fine Arts, Xinxiang University, Xinxiang, 453003, China (Huihong_Li2023@outlook.com)

2. Related Work. A very important branch of user psychology prediction technology, namely providing users with favorite art works, plays a very important role in improving user browsing experience and selling unpopular art works. Li et al. constructed a model that provided user historical interest sequences based on the semantic understanding of each project. They proposed knowledge enhancement path mining and interest fluctuation signals for discovering granular dynamic interest sequence learning methods to obtain semantic enhancement paths. The path was merged through the entropy perception pool layer to obtain a user preference representation, which was then used for dynamic learning of user interest sequences. The experimental results on two common datasets for movie and music recommendations showed that their model could achieve better predictive performance compared to other known baselines [5]. Chen et al. first introduced an attention flow network to model users' purchase records by displaying the attention flow of changes in the purchase intention; Then, based on individual attention flow, a personalized recommendation algorithm based on attention flow network was proposed. Their method integrated all user purchase sequences, converted them into a weighted attention flow network, and recommended projects based on the user's attention decay attention flow network through relevant transfer probabilities. Finally, their experimental results on several real datasets indicated that their superior performance could meet user preferences [6]. Zhang et al. proposed a factorization model for predicting restaurant rankings. They extracted images through deep convolution network and applied them to collaborative filtering. They fused multi perspective visual features through user related weights, which reflected personalized visual preferences for restaurants and were different and independent among users. They applied this model to provide personalized recommendations for users on two restaurant datasets. The experimental results showed that compared to the model with single view visual information, the model with multi view visual information had better performance [7]. Zhu simulated the learning process of users through personalized fuzzy logic interests. Based on the established model, resources were recommended to users according to the idea of collaborative filtering. Finally, it was applied to user interest description, a user interest vector based on personalized logic was proposed, and concept aggregation methods were used to discover user interests. His experimental results indicated that his method could better describe user interests, making the recommendation of interest resources for specific users more accurate and reliable, and further studies have been made for the collaborative recommendation problem in performance-based fuzzy logic systems [8].

Chen et al. proposed a personalized recommendation algorithm, which used collaborative filtering to recommend in turn. Dirichlet topic model was used to reduce the dimension of user data. And a user written topic matrix was established to reduce inaccuracies in the algorithm. It calculated the similarity between users to obtain a list of user interests. Then, based on the preliminary recommendation results, the feature vectors of the calligraphy image were extracted. And it calculated the similarity between calligraphy characters and preliminary recommendations to obtain the final recommendation result. The experimental results showed that this algorithm had effectiveness and accuracy, and was superior to other algorithms [9]. Li et al. proposed an algorithm for recommending explanatory Q&A documents. Firstly, a dual topic model was used for modeling, and then the growth gas algorithm was used to cluster documents. To train multiple classifiers, three features were extracted from question answering categories. It identified relationships by building an integrated classification model and recommended explanatory Q&A documents. This algorithm exhibited good clustering performance, and the performance of the integrated classification model was superior to other algorithms. The high score of its Q&A recommendation performance indicated the practicality and good performance of the proposed recommendation algorithm, providing a new perspective for recommendation research [10]. Du et al. used users' subjective characteristics and trust to improve similarity. Considering the sparsity and discreteness of the data, a cloud drop similarity calculation method was introduced when calculating trust similarity. Through weighting to predict gaps in the data, new similarity was generated. When there was a cold start issue with the history of a new user, a neural network was used to classify the user. They proposed a method for predicting user interest features based on feature classification. Finally, the effectiveness and rationality of this method were verified using the recommendation of machine tool products in manufacturing enterprises as an example [11]. Several researchers have found that algorithms for predicting user preferences are very popular internationally. But there is still little research on PEK means. This research pioneered the introduction of regularization regulation process and emotional precision, and on this basis, the impact of user interest characteristics and emotional preferences was taken into account to generate PEK means.

3. Construction of a Personalized Art Work Recommendation Model Based on User Interest Characteristics and Emotional Preferences. With the development of the internet industry, to cater to user preferences, research on recommendation models has become increasingly popular, with personalized art recommendation systems being the most important part [4]. However, there are a wide variety of art works and writers with vastly different styles, which increases the difficulty of recommending art works. This study combines the user's interest characteristics with emotional preferences. Firstly, it introduces the models constructed based on the two, and then describes the fusion method of the two models.

3.1. Building a Recommendation Model Based on User Interest Characteristics and Emotional Preferences. The basic information, learning data, and behavioral habits of users can be collected to obtain their interest characteristics. Since entering the era of networking, users have increasingly high requirements for recommendation accuracy, and recommendation methods based on users' own attributes have been widely used. In the building a collaborative filtering model, the most important step is to find similar users and items. If two random users are set as vectors, the cosine value of the angle between the vectors can be calculated, as shown in equation 3.1 [13].

$$\cos(i,j) = \frac{i \cdot j}{|i| \cdot |j|} = \frac{\sum_{k \in u_{ij}} R_{i,k} R_{j,k}}{\sqrt{\sum_{k \in u_i} R_{i,k}^2} \cdot \sqrt{\sum_{k \in u_j} R_{j,k}^2}} \tag{3.1}$$

In equation 3.1 above, i, j represents two users, and the way users view things is denoted as k. The values of the two users towards k are represented by $R_{i,k}$ and $R_{j,k}$, respectively. The range of $cos(i,j)$ is between 0 and 1, and as its value increases, the similarity between the two users will increase. to observe the level of intimacy between two users, the Pearson Correlation Coefficients (PCC) is introduced, and its calculation expression is shown in equation 3.2.

$$\text{sim}(i,j) = \frac{\sum_{k \in u_{ij}} (R_{i,k} - \overline{R})(R_{j,k} - \overline{R})}{\sqrt{\sum_{k \in u_{ij}} (R_{i,k} - R_u)^2} \cdot \sqrt{\sum_{k \in u_j} (R_{j,k} - R_u)^2}} \tag{3.2}$$

In equation 3.2, $\overline{R}$ means the liking of two users towards things, and the user's rating is recorded as R_u. $sim(i,j)$ fluctuates between -1 and 1, where 1 means a close relationship between two users, 0 expresses that they are not familiar with each other, and -1 refers to a completely opposite relationship between the two [8]. When the intimacy between two users is between 0 and 1, it is necessary to consider the difference in their understanding of art works, as shown in equation 3.3.

$$S(i,j) = \begin{cases} 1 & \text{if } S_i = S_j \\ 0 & \text{if } S_i \neq S_j \end{cases} \tag{3.3}$$

In equation 3.3, the knowledge of two users about the art work is denoted as S_i, S_j and the value of $S_{i,j}$ is the degree of difference in understanding between the two users. When the two users have the same understanding of the art work, $S_{i,j} = 1$, and vice versa, the value is 0. Combining the parameters between personalized art works and users can construct the block diagram shown in Figure 3.1.

The block diagram in Figure 3.1 consists of four modules, namely art collection, work processing, analysis of painting types, and personalized recommendations for users [15]. Firstly, it selects a larger range of cities and collects as many art works as possible; Then it performs noise reduction, weight reduction, and other processing on it to include all types of art works; The collected paintings are classified based on the artist's style; Finally, based on each user's preferences, the eligible paintings are recommended to them. In addition to co buyers of the same painting, the study considers users who indirectly purchase the painting, and the calculation method is shown in equation 3.4.

$$W(x,y) = w_{x,y} + \sum_{z \in N(x) \cap N(y)} (w_{xz} + w_{yz}) \tag{3.4}$$

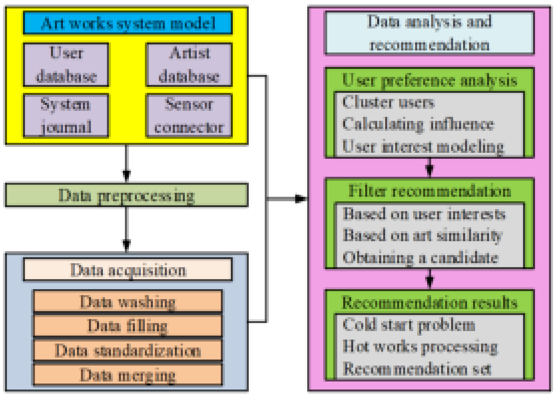


Fig. 3.1: Combination Block Diagram of Intimacy between Personalized Artist and Users

In equation 3.4, x, y are defined as two random paintings; $W(x, y)$ represents the same user who purchased both; The seller of the same type of two paintings is labeled $N(x)$,$N(y)$; Their indirect buyers are recorded as $w_{x,y}$; $z$means the similarity between paintings. Among various types of recommendation models, the similarity of preferences between users can be captured at a fine-grained level, and the similarity calculation module can be used to calculate the emotional preference differences between users, as shown in equation 3.5.

$$\alpha^{p_i}(i,j) = \begin{cases} \frac{\sum_{a\in\alpha p_i} E_a^i E_a^j}{\sqrt{\sum_{a\in\alpha p_i}(E_a^i)^2}\sqrt{\sum_{a\in\alpha p_i}(E_a^j)^2}} & \text{if } i,j \in p_i \\ 0 & \text{if not} \end{cases} \tag{3.5}$$

In equation 3.5, the artwork is represented by p_i ; E_i^a,E_a^j indicate users' emotional preferences for art works, and the collected collection of art works is recorded as a. The emotional rating of users for art works depends on their evaluation of the art works. Users have similar personalities, and similar user evaluations can help users make purchase suggestions. The help of evaluations for art works is shown in equation 3.6.

$$H(i,j) = \frac{|A_i^j|}{l}\alpha^{p_i}(i,j) \tag{3.6}$$

In equation 3.6above, $H(i, j)$ stands for the impact of one user on the other, with a value range of [0,1]. A value of 0 indicates no impact, while a value of 1 has the greatest impact [16]. The length of the keywords in this painting is denoted as l, and the frequency of the keywords in the evaluation is represented by $A_i^j|$. In the user feedback interaction, the same painting will have multiple different comments, and their relationship is shown in equation 3.7.

$$\text{cov} = \frac{\beta_i^u \cap \beta_{H_t^{\text{sim}}}^u}{|A_i^j|} \tag{3.7}$$

In equation 3.7, *cov* is the coverage of all comments on the same painting; The associated words in the user's painting evaluation are recorded as β_i^u; Similar users' evaluations of the same type are expressed by . Coverage can consider the connection between users and art works, while Aspect Coverage (AC) describes users' emotional preferences. However, users still have unfavorable evaluations for their sales of art works. To accurately calculate AC, the EA indicator is introduced, and the calculation is shown in equation 3.8.

$$\text{frastraction} = \frac{|\chi_i^{u^+} \cap \chi_{H_t^{\text{sim}^{u+}}}| + |\chi_i^{u^-} \cap \chi_{H_t^{\text{sim}^{u-}}}|}{|\chi_i^u \cap \chi_{H_t^{\text{sim}^u}}|} \tag{3.8}$$

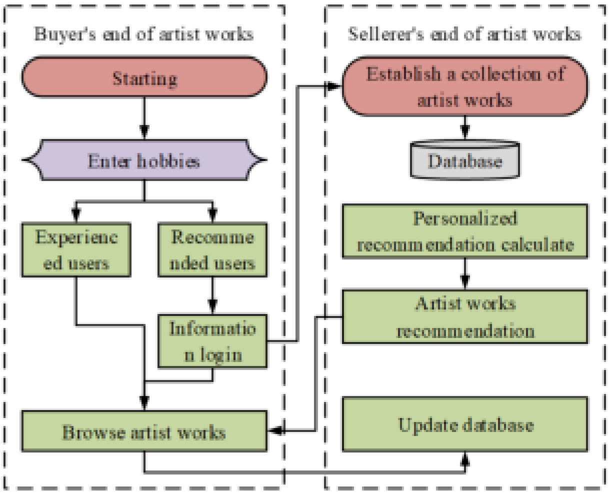


Fig. 3.2: Analysis of Art Works with Emotional Precision

In equation 3.8, the user's positive evaluation of the artwork is represented by χ_i^{u+}, $\chi_{H_t^{sim}}^{u+}$ and the negative evaluation of the artwork is recorded as χ_i^{u-}, $\chi_{H_t^{sim}}^{u-}$. *rastraction* describes the purchasing significance of the artwork, fluctuating between -1 and 1 [17]. When the value is -1, it indicates that the user who purchased the work has extremely low evaluation of it. The analysis of art works using emotional precision consists of two modules, as shown in Figure 3.2.

The system in Figure 3.2 includes both the user and the sales ends. Old users can directly appreciate art works; New users input their interest in art works, and their information will be input into the self built database of the painting merchant. At the sales end of art works, merchants calculate the weight proportion of different types of works, and then provide multiple types of works to arouse user interest.

3.2. Model Building of Regularization Regulation Process and EA Improved K-Means Algorithm. When applying PCC and EA to predict user interests and preferences in practice, there will be various errors. To avoid these errors, the K-means algorithm is introduced to cluster a large amount of data as needed. Multiple seeds are generated by initializing their geometric centers. These seeds can attract eligible art works, as shown in Figure 3.3 [18].

The K-means algorithm shown in Figure 3 first generates four types of center points, and then clusters as many paintings works as possible based on the types of other points. However, the number of cluster center points often varies greatly. In the attracting the same element to each other, the center points with fewer works will have stronger attraction, making it easier to attract other elements, such as the red center point shown in Figure 3.3. The purple center point ranks second in the number of works it contains, as it contains highly similar works (such as A and G), so it blocks one of them. Although purple dots detect anomalies and they recombined to output results that meet the conditions, it also indicates that the K-means algorithm has obvious drawbacks when processing large amounts of data. Based on this, the research redefines the distance between clusters, and its calculation is shown in equation 3.9.

$$L = \sum_{i=1}^{O} \|c_i - m\| \tag{3.9}$$

In equation 3.9, the distance from the cluster center to the user center is denoted as L; The user center is represented using m; O is the number of center points, and the result is recorded as c_i. The K-means algorithm can analyze multiple attributes to achieve the purpose of understanding users. The expression for its model is

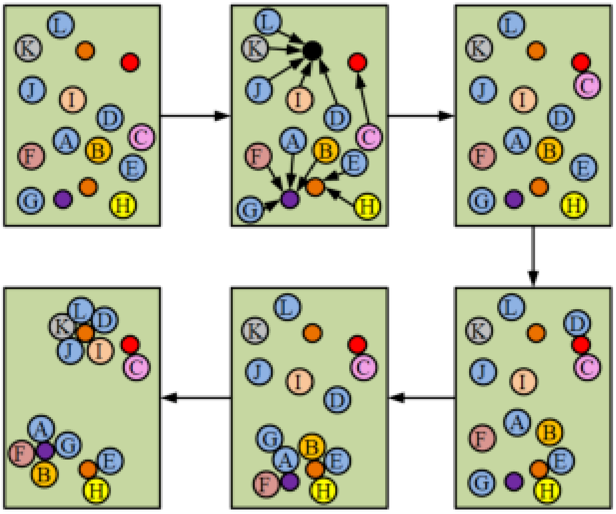


Fig. 3.3: Working Principle of K- means Algorithm

as equation 3.10.

$$f(X)=w_O+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}v_{ij}x_ix_j \tag{3.10}$$

In equation 3.10 above, $f(X)$ means the processed user information; The user's values on the features are denoted as x_i, x_j ; The error of global parameters is recorded as w_o;It uses the v_{ij} stands for users' interest in the work; The impact of the work on users is recorded as $v_{ij}x_ix_j$. The model obtained through research can learn to speculate on user psychology, and the key point is to minimize the difference between user information and ratings [19]. The calculation for the difference is shown in equation 3.11.

$$\delta=\sum_{i=1}^{\varepsilon}\text{loss}(f(X),Y)=\sum_{i=1}^{\varepsilon}((f(\vec{x}_i),Y_i))^2 \tag{3.11}$$

In equation 3.11, the meaning of Y is the user's true rating matrix for the work; $\overline{f(x_i)}$ is the average user rating of the work; The total number of users is recorded as ϵ. RPP method is used to optimize the model. That is, multiple regularization terms are added to the user group, and the prediction results are obtained by changing the parameters, as shown in equation 3.12.

$$\Theta^*=\arg\min_{\theta}\left(\sum_{i=1}^{\varepsilon}\left((f(\vec{x}_i),Y_i)+\sum_{\theta\in\Theta}\phi_\theta\theta\right)^2\right) \tag{3.12}$$

In equation 3.12, Θ means the model in the clustering algorithm. The parameters are represented by θ. When θ is a positive number, its regularization term is recorded as ϕ_θ. The value of ϕ_θ needs to satisfy randomness, and when it happens to be 0, there is a equation 3.13.

$$\text{dist}(\phi,\gamma)=\frac{\eta_1\mu_1+\eta_2\mu_2+\eta_3\mu_3+\ldots+\eta_n\mu_n}{\sqrt{\sum_{i=1}^{n}\eta_i^2}\cdot\sqrt{\sum_{i=1}^{n}\mu_i^2}} \tag{3.13}$$

In equation 3.13, the painting is denoted as ϕ, γ, and the feature vectors extracted in classical, abstract, realistic, and other styles are expressed as μ, η. For users, it is easier to find users of the same age group and gender

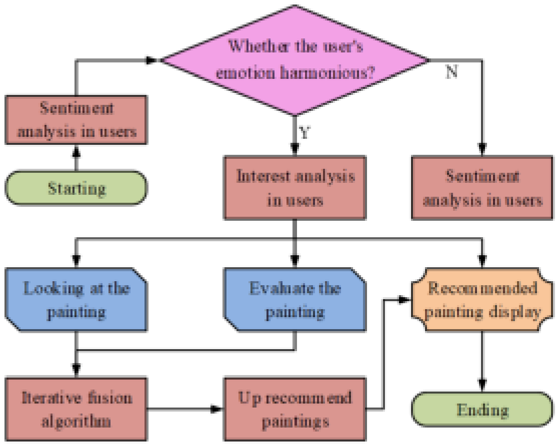


Fig. 3.4: Clustering and Screening Process of Artist Portfolio

who share a common language. During the clustering, due to the wide distribution of users in different regions, there may be slow user startup issues. To address this issue, the study introduces the definition of a popular painting portfolio, as shown in equation 3.14.

$$p(\nu,\pi)=\frac{|\vartheta(i)\rho\sigma(f)|}{|\vartheta(i)|+\zeta} \tag{3.14}$$

In equation 3.14, artistic creation is denoted as i, f; Users who like them use $\vartheta(f), \rho(f)$ representation; Equation 3.14 can search for a wider range of art works, thereby alleviating the problem of slow startup for users. The parameter used to solve data errors is ζ, and its calculation method is as equation 3.15.

$$\zeta_{i,j}=\frac{\sum_{i=1,j=1}^{n}|\tau_i-\omega_i|\cdot|\tau_j-\omega_j|}{n} \tag{3.15}$$

In equation 3.15 above, $\zeta_{i,j}$ refers to the global error predicted by the user, and its value reflects the quality of the model. τ_i, τ_j are the user's evaluation of the artwork, while ω_i, ω_j denote the user's emotional state towards the artwork [20]. To find the parts that users like from a large number of art works and calculate their similarity separately, it studied clustering and filtering of art works, and established a PEK-means for RPP and EA optimization K-means. The process is shown in Figure 3.4.

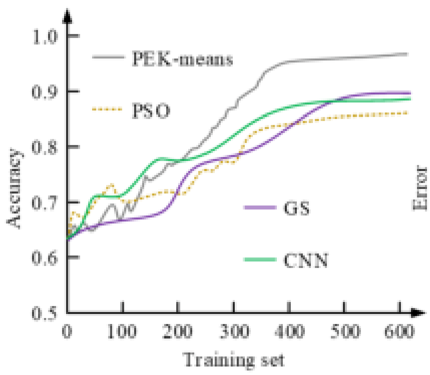
The method shown in Figure 3.4 can balance users' interest characteristics and emotional preferences, thereby recommending more accurate information about art works to users. It can be summarized as four processes. First is to analyze the emotional state of users and collect interest from users who are in good condition; Then it let them watch the artwork and retain their evaluations; Then it uploads their evaluations to the self built database to enrich the types of personalized art works; Finally, the updated painting recommendation portfolio is used to present richer content in the user module.

4. Experimental Study on Personalized Art Work Recommendation Model Based on K-Means Algorithm. To verify the effectiveness of the PEK-means algorithm in practice, a study was conducted to construct a PEK-means model based on personalized recommendation, and its iteration and accuracy verification were carried out. Finally, simulation experiments were conducted on the Wink dataset using the PEK-means model.

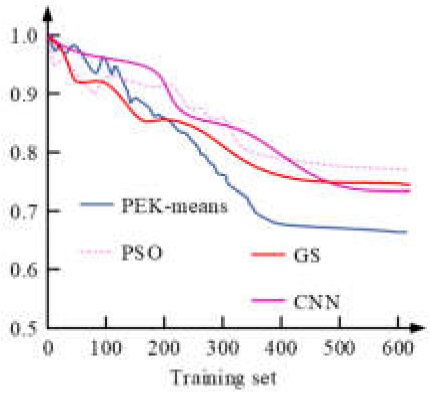
4.1. PEK-means System Development Environment and Model Parameter Determination. This study selected the Wink dataset, which included four types of painting: sketch, ink wash, oil paint, and simple brush, with a total of 1427635 pieces of painting information. Considering the limited types of data, the

Table 4.1: Experimental Parameters

Data set	Development language	Application servers	Internal storage
Wink	Python 12.0	Potato 8.0	512 G
Operating system	Display card	Database	Carrying out system
128 Ubuntu 23.02.20	18.0 GHz	Mysql 5.20.2023	Ubuntu 88.64
Web development framework	Language	Operator	Model
Django 1.22.3	Easy Chinese	Sketch, ink, oil...	F2.8LII-USM



(a) Training Set-Accuracy



(b) Training Set-Error

Fig. 4.2: Comparison of Accuracy and Error Rate in Training set Image

dataset was divided into a training set and a testing set in a ratio of 2:3. The specific equipment and software used in the experiment are shown in Table 4.1.

The collected dataset needed further processing to enable the studied algorithm to learn. For the processing of the dataset, PEK-means was used for iterative optimization. To verify its accuracy, traditional Golden Sine algorithm (GS), Convolutional Neural Network (CNN), and Particle Swarm Optimization algorithm (PSO) were compared with them. The accuracy and error rate results in the training set are shown in Figure 4.2.

From Figure 4.2, the PEK-means algorithm had a slightly lower accuracy and higher error rate compared to CNN and PSO before 200 training sessions. But when the number of iterations reached 200 or more, the accuracy of PEK-means was higher than both algorithms, and tended to stabilize at 380 iterations, which was higher than the other three algorithms. Although increasing the number of iterations could reduce the operational efficiency of the model, after comprehensive consideration, the accuracy weight of the model was higher, so the PEK-means algorithm proposed in the study had better performance. After the learning of the PEK-means algorithm was completed, it was also necessary to consider the parameter determination during testing, as shown in Figure 4.3.

The parameter of this study was the regularization term $\theta \in (0, 1)$. As shown in Figure 4.3, when the value of the regularization term as 0.3 and the number of iterations was 300, the error rate was the lowest, 0.072. Therefore, it was finally determined that the number of iterations was 300, and the value of the regularization term as 0.3.

4.2. Experimental Verification of Personalized Art Work Recommendation based on PEK-means. In order to verify the accuracy of PEK-means model in predicting users' psychology, a more in-depth comparison is made, simulation experiments were conducted. By observing the image computing power of the PEK-means algorithm, it could determine its practicality. First it initialized the PEK means algorithm; then

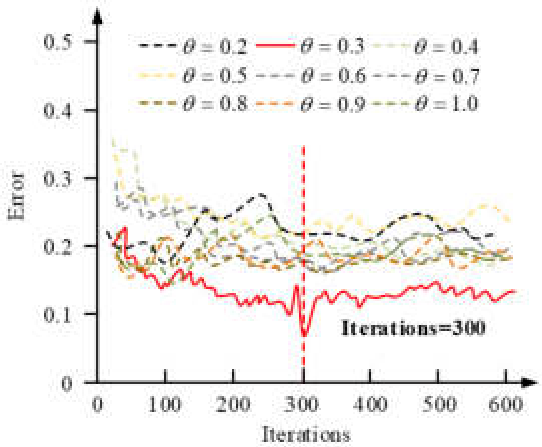


Fig. 4.3: Error-Training Times Image of Regularization Term

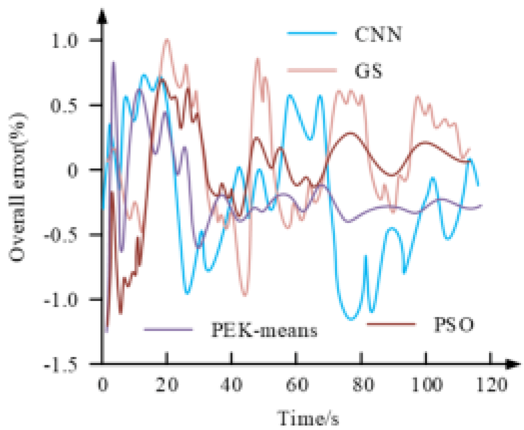


Fig. 4.4: Total Error-time Image of Four Algorithms

personal interest, emotional preference and other information at the user end were input; finally, it set the value of the regularization item to 0.3, and collected art works records within 120 seconds, calculated the error and draw the image, as shown in Figure 4.4.

Figure 4.4 shows the error comparison of four algorithms in the experiment. From Figure 4.4, after 75 seconds, the total errors of PEK-means and PSO fluctuated around -0.5% and 0.5%, respectively, while GS and CNN had not yet stabilized. The error ranges of PEK-means, PSO, GS, and CNN were all [-1.0%, 1.0%]. Only comparing the total error of the four algorithms could not distinguish the optimal algorithm. So, the study analyzed the errors caused by user interest characteristics and emotional factors of the four of them, and drew images as shown in Figure 4.5.

From Figure 4.5, the experimental results of the PEK-means model were concentrated in the range of total error of 0. The error range caused by user interest features was [-0.3%, 0.4%], and the error range caused by emotional factors was [-1.0%, 0.5%]. The error distribution of the remaining three algorithms was wide, and the distribution of larger errors was sparse. In order to distinguish the error correction ability of the four algorithms more intuitively, and then verify the universality of the algorithms proposed in the study, 400 experimental data records were made and the image shown in Figure 4.6 was drawn.

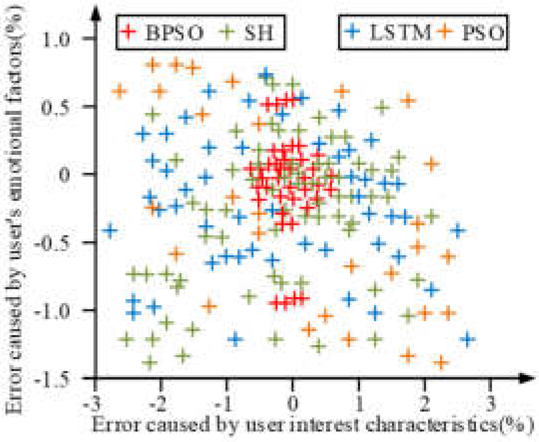


Fig. 4.5: The Types and Genres Errors of the Four Algorithms

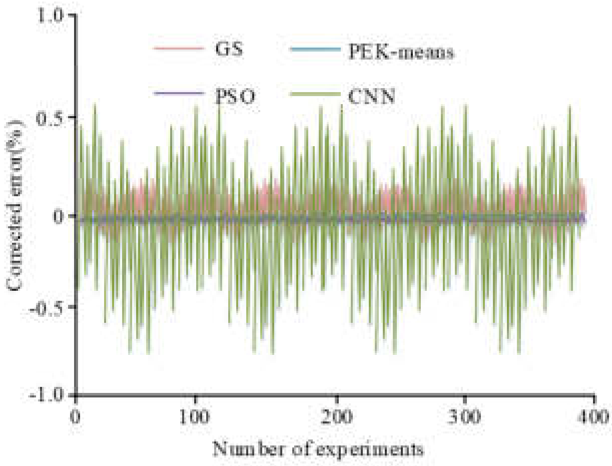


Fig. 4.6: Error Changes of Four Algorithms in Four Hundred Calibration Experiments

From Figure 4.6, in 400 error testing experiments, the error range of CNN was the largest, recorded as [-0.7%, 0.6%]; Next was the GS algorithm, which was between [-0.2%, 0.3%]. The variation range of PSO was close to PEK means, between [-0.13%, -0.03%]; The error curve of PEK-means fluctuated between -0.02% and 0.04%, with the smallest fluctuation range. Only comparing the experimental results of PEK means algorithm and PSO can draw the error matrix, the resulting image is shown in Figure 4.7. Figure 4.7 shows the experimental results of accurate prediction of PEK means and PSO based on user interest, emotion, emotion precision, and regularization process to analyze four types of art works: sketch, ink and wash, oil color, and simple pen. The prediction accuracy of PEK-means reached 392 times, with an accuracy rate of 98.0%, and the accuracy of PSO was 88.2%. To observe the experimental results of PEK-means and PSO more intuitively, a linear fitting graph based on matrix drawing was studied, and the predicted values of the two algorithms were compared with the actual values, as shown in Figure 4.9. Figure 4.9 shows the comparison of two algorithms in predicted and true values. From Figure 11, the linear fit ( ) of the PEK-means algorithm was 0.9903, and the of the PSO was 0.9545, indicating that there was no underfitting in the model. In summary, the PEK-means algorithm model could effectively map the relationship between user interest characteristics, emotional factors,

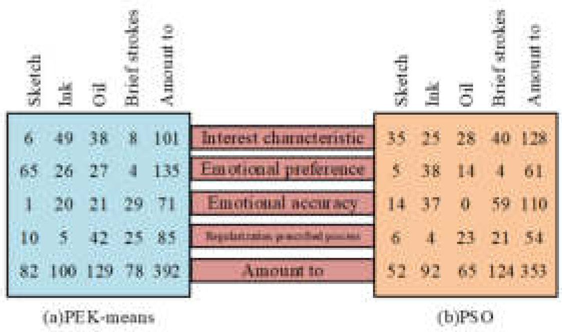
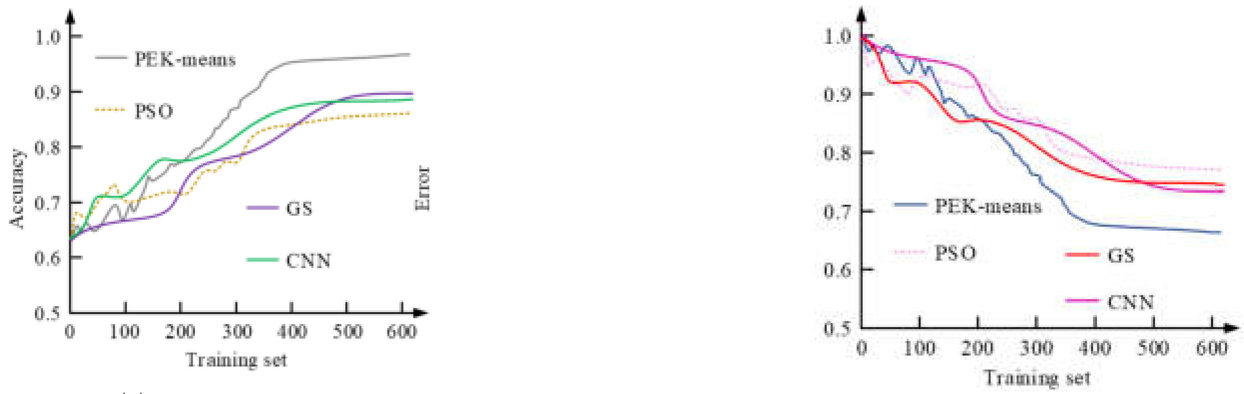


Fig. 4.7: Error Matrix of PEK- means Algorithm and PSO Algorithm



(a) PE K-Means

(b) PSO

Fig. 4.9: Linear Fitting Diagram of PEK-means and PSO

and personalized art recommendation, thereby providing users with a good viewing experience.

5. Conclusion. With the development of the internet industry, analyzing user data during network roaming is becoming increasingly important, such as increasing browsing comfort for users and increasing exposure for popular art works. This research quantified user interest features and emotional factors based on RPP and EA, and combined K-means to generate PEK-means. Taking into account PCC and regularization term, simulation experiments were carried out on Wink dataset and compared with GS and other three algorithms. 40% of the Wink dataset was extracted and trained on the PEK-means model. For the number of iterations, the study decided to carry out regularization experiments, and finally determined that it was 300. In order to determine the extensive experiments of the four algorithms, 400 experiments were conducted to analyze their estimation of user thinking. The errors of GS and CNN algorithms are the biggest in this experiment, which are between [-0.2%, 0.3%] and [-0.7%, 0.6%] respectively. For the proposed algorithm, PEK-means performs best in the experiment of estimating users' thinking, and its error is between-0.02% and 0.04%. The experimental performance of particle swarm optimization is a little poor, and its range is between [-0.13%, -0.03%]. In order to observe their experimental results intuitively, this paper studies drawing these data into an error matrix. By analyzing this error matrix, the prediction accuracy of the proposed algorithm is 98.0%, and the accuracy of PSO is 88.2%. In order to explore the accuracy of the application scope of the two experimental results, the study conducted several experiments based on their experimental results and drew a linear fitting diagram. The of PEK-means was 0.9903, indicating excellent linear fitting. The of PSO as 0.9545. In summary, the PEK-means algorithm model can reflect the relationship between user interest characteristics, emotional factors, and personalized art recommendation, and can meet the psychological needs of users. However, the PEK-means model is only suitable for analyzing art works with a wide distribution of similar users. For art works with fewer users, the model will label them as noise. This is because the purchase records of art works belong to private information, and the dataset for research and analysis contains fewer types. With the increase of volunteers, it is believed that future research can be improved.

Funding. The research is supported by: Research on Oriental Abstract Painting, PhD Research Initiation Project, (No.1366020191).

REFERENCES

[1] Suganya, G., Premalatha, M., Dubey, P., Drolia, A. & Srihari, S. Subjective Areas of Improvement: A Personalized Recommendation. *Procedia Computer Science*. **172** pp. 235-239 (2020)
[2] Yang, L., Liu, Z., Wang, Y., Wang, C., Fan, Z. & Yu., P. Large-scale personalized video game recommendation via social-aware contextualized graph neural network. *Proceedings Of The ACM Web Conference 2022*. pp. 3376-3386 (2022)
[3] Tang, M. & And, J. and revisiting behaviors of individuals with different preference characteristics: An experience sampling approach. *Journal Of The Association For Information Science And Technology*. **71**, 540-552 (2020)
[4] Yuan, N. Application of machine learning and cloud computing in social media behavior analysis. *Journal Of Intelligent & Fuzzy Systems*. **39**, 1831-1842 (2020)
[5] Li, Y., Guo, X., Lin, W., Zhong, M. & Zhu, Z. Learning dynamic user interest sequence in knowledge graphs for click-through rate prediction. *IEEE Transactions On Knowledge And Data Engineering*. **35**, 647-657 (2021)
[6] Chen, Y., Dai, Y., Han, X. & Others Dig users' intentions via attention flow network for personalized recommendation. *Information Sciences*. **547** pp. 1122-1135 (2021)
[7] Zhang, X., Luo, H., Chen, B. & Guo, G. Multi-view visual Bayesian personalized ranking for restaurant recommendation. *Applied Intelligence*. **50** pp. 2901-2915 (2020)
[8] Zhu, W. Topic recommendation system using personalized fuzzy logic interest set. *Journal Of Intelligent & Fuzzy Systems*. **40**, 2891-2901 (2021)
[9] Chen, S., Huang, L., Lei, Z. & Wang, S. Research on personalized recommendation hybrid algorithm for interactive experience equipment. *Computational Intelligence*. **36**, 1348-1373 (2020)
[10] Li, M., Li, Y., Xu, Y. & Wang, L. Explanatory Q&A recommendation algorithm in community question answering. *Data Technologies And Applications*. **54**, 437-459 (2020)
[11] Du, X., Ge, S., Wang, N. & Yang, Z. Personalized product service scheme recommendation based on trust and cloud model. *IEEE Access*. **8** pp. 82581-82591 (2020)
[12] Yuan, N. Application of machine learning and cloud computing in social media behavior analysis. *Journal Of Intelligent & Fuzzy Systems*. **39**, 1831-1842 (2020)
[13] Sun, R., Chen, J., Wang, Y., Zhou, Y. & Luo, Y. An ERP experimental study about the effect of authorization cue characteristics on the privacy behavior of recommended users. *Journal Of Advanced Computational Intelligence And Intelligent Informatics*. **24**, 509-523 (2020)
[14] Zhu, Y., Lu, H., Qiu, P., A, K., B, J. & Znac, D. Heterogeneous teaching evaluation network based offline course recommendation with graph learning and tensor factorization. *Neurocomputing*. **415** pp. 84-95 (2020)
[15] Ünver, M., Olgun, M. & Türkarslan, E. Cosine and cotangent similarity measures based on Choquet integral for Spherical fuzzy sets and applications to pattern recognition. *Journal Of Computational And Cognitive Engineering*. **1**, 21-31 (2022)
[16] Liu, L. Learning information recommendation based on text vector model and support vector machine. *Journal Of Intelligent And Fuzzy Systems*. **40**, 2445-2455 (2021)
[17] Kang, S., Jeong, C. & Chung, K. Tree-based real-time advertisement recommendation system in online broadcasting. *IEEE Access*. **8**, 92693-19270 (2020)
[18] Bai, H., Li, X., He, L., Jin, L., Wang, C. & Jiang, Y. Recommendation algorithm based on probabilistic matrix factorization with adaboost. *Comput, Mater Continua*. **65** pp. 1591-1603 (2020)
[19] Oslund, S., Washington, C., So, A. & Chen, T,& Ji, H. Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. *Journal Of Computational And Cognitive Engineering*. **1**, 152-158 (2022)
[20] Xie, J., Zhu, F., Li, X., Sheng, H. & D., S. Attentive preference personalized recommendation with sentence-level explanations. *Neurocomputing*. **426** pp. 235-247 (2021)

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 4, 2023
Accepted: Oct 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 895–907, DOI 10.12694/scpe.v24i4.2454

SPEECH EMOTION ANALYSIS OF SHORT ENGLISH READINGS BASED ON THE CAM-SPAT MODEL

QINGYUAN LI*

Abstract. With the development of technology, voice sentiment analysis has also undergone rapid development, and its application fields are constantly expanding. Multimodal models have become a key focus of researchers due to their ability to better predict emotions. In order to help English learners improve their oral English proficiency, a deep learning based emotional analysis model for English short text reading is proposed, and this model is used to analyze emotions in English reading. Additionally, a cross-modal attention mechanism based on a prediction-assisted task was developed to identify emotions in English reading aloud in state and a two-layer attention-based bi-directional long- and short-term memory network was created to classify emotions in English reading aloud. The results of the research revealed that the classification model's mean F1 value was 98.54%, the detection model's mean F1 value was 85.13%, and the speech emotion analysis model's mean F1 value was 73.55, which was not significantly different from the mean of the professionals' ratings. The significance of the study lies in providing English learners with a method and pathway to improve their oral English proficiency.

Key words: read aloud speech; emotion; CAM-SPAT; deep learning; feature extraction

1. Introduction. One of the most essential aspects of human existence is emotion, and the key to emotion research is gathering data on these traits and utilising models to assess them [1]. Unimodal and multimodal models of sentiment analysis are two broad categories. Single modal models have the disadvantage of being unable to dynamically, multi-dimensionally, and multi-dimensionally analyze human emotions. Therefore, multimodal models have gradually become the focus of emotional analysis, and most of its current research is mainly focused on finding a better multimodal fusion method [2]. With the development of technology, the application area of speech sentiment analysis is gradually expanding [3]. In addition, based on advances in technologies such as artificial neural networks and deep learning (DL), more and more researchers have started to combine sentiment analysis with DL [4]. Researchers such as S Dahmani used different neural architectures to synthesize emotional speech and studied the performance of neural networks in learning visual and auditory modal features under different emotions [5]. Experts such as F Jia have constructed an emotional lexicon for texts and proposed an emotional orientation analysis model to analyze the emotional characteristics of participants in online public opinion [6]. However, the combination of deep learning methods also has certain shortcomings, such as difficulties in personalised feature extraction, difficulties in cross-modal interaction and limitations in distribution modelling [7]. Based on these problems, the study innovatively proposes a DL-based model for sentiment analysis of English short-text read-aloud speech. The model consists of a two-layer Dual Attention-based Bidirectional Long Short-term Memory Networks (DABLSTM) model and a Cross-modal Attention Mechanism (AM) with Sentiment Prediction Auxiliary Task (CAM-SPAT) model. There are two innovative points in the research. The first point is the combination of dual attention mechanisms and bidirectional long-term and short-term memory networks. The second point is the combination of emotion prediction assistance tasks and cross modal attention mechanisms. The research aims to help English learners better understand the specific situation of their oral English proficiency and improve their oral English proficiency.The study is divided into four parts: the speech emotion (SE) analysis portion is the first, the SE feature extraction portion is the second, the SE analysis model portion is the third, and the study conclusion portion is the fourth.

2. Related Works. Emotion is a necessary component of human existence, and as science and technology advance, so does study on voice sentiment analysis. Researchers like C Park. developed a prospective profile analysis employing exploratory cross-sections to examine the surface and deep performances of nurses in order

*School of Foreign Languages, Nanchang Institute of Technology, Nanchang, 330044, China (liqingyuan1001@163.com)

to study the management of nurses' emotional work. The study's findings demonstrated that nurses with highly regulated traits and surface acts were more likely to cause emotional weariness [8]. To study and mine the emotional tendencies of web writings, Haichao Sun et al. experts proposed a BPSO-based approach of integrated learning text emotion categorization in stochastic subspaces. By chunking the web texts with various granularities of sentiment, the approach categorises web texts at various levels of detail. The experimental findings demonstrated that the study's suggested algorithm had a greater level of classification accuracy [9]. Researchers I Gupta et al. incorporated an enhanced negation computation based on Twitter sentiment analysis and trained the classifier using various sorts of data attributes in order to analyse sentiment on social media. Study findings demonstrated that the support vector machine classifier performed better than other classifiers [10]. To advance the study of sentiment analysis on social concerns, academics like Y Mehmood developed an improved lexicon-based methodology. The approach combines the use of verbs and multilayer lexical dependencies with General Inquirer. The approach has good results and can be accurate up to 83%, according to experimental results [11]. Experts like J Barnes have suggested a multi-task strategy to produce precise forecasts about emotion. The approach uses a cascaded, hierarchical neural network with sentiment analysis and the addition of negative data. According to experimental findings, the proposed technique can greatly enhance sentiment analysis [12].

To address the issue of efficiently producing training data for sentiment analysis, R Ghasemi and colleagues suggested a cross-linguistic DL framework. They also used cross-linguistic embedding to change the sentiment analysis model into a migration learning model. The model can classify text using a wide range of deep architectures. The study's findings demonstrate the model's definite superiority, with a total improvement in model performance of 16% [13]. In order to cut down on the time and labour needed for data annotation, R Alahmary and other specialists suggested a semi-automatic method based on plain Bayesian, and the method was utilised to annotate new data sets. According to the experimental findings, the basic Bayesian classifier had an accuracy rate of up to 82.9%, which can significantly reduce the amount of time and labour needed for data annotation while simultaneously accelerating the process [14]. Experts like Zhang Hua have suggested a two-stage neural network approach to analyse sentiment in texts and spoken words. In order to extract categories and polarities of viewpoint terms in texts or utterances, the model contains of modules like BiLSTM and positional coding [15]. For the study of picture sentiment, Liang Yun and other researchers have suggested a chain-centre loss. This loss function regulates both local and global spatial distributions and was built on the foundation of central loss and triple loss. The study's findings demonstrated that this loss function enhances the chain-centre loss performance and enables the building of a prospective space of sentiment relevance [16]. GG Kim et al. The performance of the model before and after profanity data removal was simulated for the purpose to examine the influence of profanity data on sentiment categorization using comment data from the web as source data. The study examined whether the profanity data decreased the model's accuracy by comparing its performance before and after data elimination. According to the experimental findings, including profanity data as noisy data caused the model's accuracy to drop by 1.8% [17]. For the purpose to evaluate the recovery after a disaster, experts like D Contreras proposed a method for analysing sentiment analysis based on web data. The method used supervised classification and expert rules to define the polarity of the network data. The study's findings revealed that the method's total accuracy was 56.8% [18].

In conclusion, both domestically and internationally, there is a lot of study on SE analysis, and the algorithms and techniques used are varied. These research do, however, also have several drawbacks, including challenges with cross-modal interaction, limitations with distribution modelling, and challenges with extracting customised information. A novel sentiment analysis technique for reading short English texts out loud is developed based on these issues. The DABLSTM model and the CAM-SPAT model combine to create a model that can classify and identify the emotions expressed in spoken English in brief paragraphs.

3. DL-based SE Analysis Model Design. A DL-based SE analysis model was designed for the analysis of emotion in short English speech read aloud. The model extracts feature information from speech through a DABLSTMl, and analyses multimodal emotion through a cross-modal attention (CMAM) model based on a prediction-assisted task.

3.1. Design of DL-based SE Feature Extraction Method. Low-level features and deep-level features are the two primary divisions of read aloud speech characteristics [19]. Different features have different

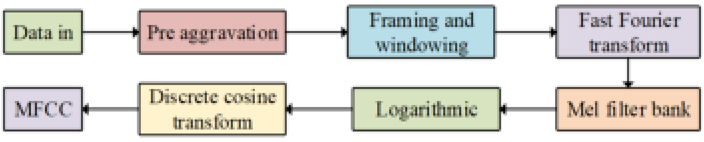


Fig. 3.1: MFCC' s extraction process

extraction methods. Before extracting the acoustic features of the low-level features, the low-level features will be pre-processed with the acoustic signal. There are three main pre-processing methods for acoustic signals, namely pre-emphasis, framing and windowing. Pre-emphasis is mainly used in the high frequency part of the speech signal, which is shown in equation 3.1.

$$y(t) = \chi(t) - \alpha\chi(t-1) \tag{3.1}$$

In equation 3.1,$\chi(t)$represents the signal before the pre-emphasis filter, $y(t)$ represents the resultant signal, t is the moment,α represents the pre-emphasis factor, and takes values in the range [0.9,1.0]. After the continuous acoustic signal has been processed by pre-emphasis, the pre-processing of the acoustic signal requires a framing operation. By dividing the continuous speech signal into several short-time speech segments, the framing operation makes it possible to obtain a relatively stable short-time signal, which can also be used for the extraction of specific information content, as shown in equation 3.2.

$$w(t) = 0.54 - 0.46cos\frac{2\pi}{N-1} \tag{3.2}$$

In equation 3.2, N is the length of the Hamming window, and n takes values in the range of [0, N-1]. For the extraction of acoustic features, the main purpose is to obtain the Mel Frequency Cepstrum Coefficient (MFCC) and the logarithmic Mel spectrogram, which is revealed in Fig 3.1 In Fig 3.1, the first step in extracting the MFCC is to pre-emphasise the speech signal, performing the operations of framing and adding windows. The second is to perform a fast Fourier change, and the third is a Meier filter bank. The fourth step is to take the logarithm and obtain a logarithmic Meier spectrogram. The fast Fourier change converts the time domain signal into a spectrum and the corresponding spectrum is generated for each frame as shown in equation 3.3.

$$X_v = \sum_{k=0}^{v-1} X_k e^{\frac{-2\Pi jkv}{v}} \tag{3.3}$$

In equation 3.3, V represents the number of frame segments, X represents the data value of k, and $e^{\frac{-2\Pi jkv}{v}}$ is the Fourier transform factor. The Meier spectrogram is a spectrogram transformed from frequency to Meier scale, and the transformation is shown in equation 3.4.

$$h = 2395log_{10}\left(1 + \tfrac{f}{700}\right) \tag{3.4}$$

In equation 3.4, f represents the actual linear frequency and h represents the Mel frequency. The logarithmic Meier spectrum can be calculated by taking the logarithm on the Meier spectrum, and the equation is shown in equation 3.5.

$$X_n^{`} = 10log_{10}\left(\tfrac{X_n}{ref}\right) \tag{3.5}$$

In equation 3.6,$X^{`}$ is x_n scaled with respect to ref . After the logarithm has been processed, the MFCC features can be obtained by the discrete cosine transform. The discrete cosine transform is shown in equation 3.6.

$$C_i(j) = \sum_{m=0}^{M-1} X_j^{`}(m)cos\left(\tfrac{\pi j(m-0.5)}{M}\right) \tag{3.6}$$

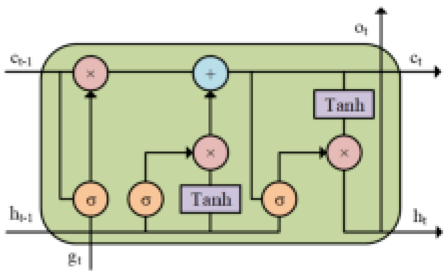


Fig. 3.2: The specific structure of LSTM

In equation 3.6, M is the number of MFCCs, j represents the number of variations from 1 to MFCC, and j takes values in the range [0,L], where L is the number of MFCC features, $C_i(j)$ represents the jth MFCC coefficient of the first frame, and $X_j^{'}(m)$ represents the logarithmic energy of the mth Meier filter. MFCC features can only represent the static characteristics of the speech signal. Equation 3.7 illustrates how to calculate the time derivative.

$$\Delta C_m(i) = \frac{\sum_{\tau=-1}^{M} \tau C_m(i+\tau)}{\sum_{\tau=-1}^{M} \tau^2} \tag{3.7}$$

In equation 3.7, $C_m(i)$ represents the static coefficients of frame i , the first-order (FO) difference feature is calculated based on the preceding and following M frames, the value of M is usually set to 2, and τ represents the time difference of the DABLSTM derivatives. The study employs a linear interpolation technique, which is computed as indicated in equation 3.8, to fill in the gaps in the data.

$$y = y_\circ \left(1 - \frac{x-x_\circ}{x_1-x_\circ}\right) + y_1 \left(1 - \frac{x-x}{x_1-x_\circ} \quad \right) \tag{3.8}$$

In equation 3.8, (x, y) represents the coordinates of the interpolation point. $(x_\circ, y_\circ)$ & (x_1, y_1) represent the coordinates of the fixed point. For the extraction of deep features of speech signals, the study mainly used Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) network and AM in DL [20, 21].

In Fig 3.23.2, the LSTM can provide some relief to the long-time dependency problem [22]. The LSTM has three steps: deciding what data should be deleted, deciding what new data should be saved, and deciding what output should be made of the result. The expression is shown in equation 3.9.

$$\delta_t = sigmoid(W_C \left[h_{t-1}, g_t\right] + b_c) \tag{3.9}$$

In equation 3.9, W_c is the weight,b_c represents the bias, h_{t-1} is the hidden state at moment $t-1$ and g_t represents the input at the current moment. The confirmation of the new message is calculated as shown in equation 3.10.

$$\begin{cases} \widetilde{c} = tanh(W_\delta \left[h_{t-1,} g_t\right] + b_\delta) \\ d_t = sigmoid \left[h_{t-i}, g_t\right] W_d + b_d \\ c_t = \delta_t * c_{t-1} + d_t * \widetilde{c} \end{cases} \tag{3.10}$$

In equation (10), d_t represents the control coefficient of the input gate, W_d and W_δ are the weights, b_d and b_δ represent the bias, $\widetilde{c_t}$ is the state information at the current moment, d_t and $\widetilde{c_t}$ are multiplied together to represent whether the current information should be retained, and c_t is the long-time hidden state. The output is calculated as shown in equation (11).

$$\begin{cases} o_t = sigmoid(W_\circ \left[h_{t-1}, g_t\right] + b_\circ) \\ \eta_t = o_t * tanh(c_t) \end{cases} \tag{3.11}$$

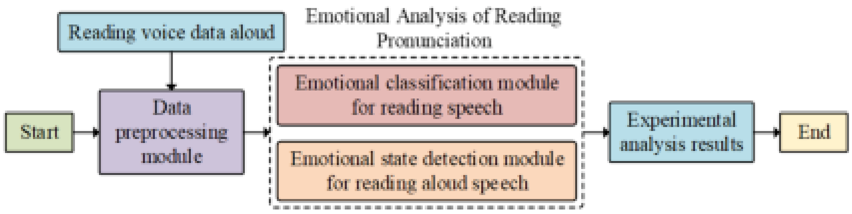


Fig. 3.3: Structure diagram of SE analysis model

In equation (11), o_t represents the control coefficient of the output gate, η_t is the output result under c_t condition, $W_\circ$ is the weight and $b_\circ$ is the bias. The Bidirectional LSTM (BLSTM) can process sequential data in both directions, and its output sequence is shown in equation 3.12.

$$\varphi_t = \mathbf{W}_{\rightarrow h}\varphi \rightarrow_h t\,(-\mathbf{W}_{\leftarrow h}\varphi \leftarrow_h t - \mathbf{b}_\varphi) \tag{3.12}$$

In equation 3.12, φ is the output sequence, $\rightarrow_t$ represents the LSTM hidden state of the forward processing input, and $\leftarrow_t$ represents the LSTM hidden state of the reverse processing input. The Gated Recurrent Unit (GRU) is a variant of the LSTM, which is shown in equation 3.13.

$$\begin{cases} h_t = z_t \Theta h_t + (1 - z_t)\Theta h_{t-1} \\ h_t = tanh(W_h g_t + U_h b_t) r_t \Theta h_{t-1} \\ z_t = \sigma(W_z g_t + U_z h_{t-1} + b_z) \\ t_t = \sigma(b_r + W_r g_t + U_r h_{t-1}) \end{cases} \tag{3.13}$$

In equation 3.13, U represents the weight matrix of h_{t-1} , the sigmoid activation function is $sigma$, r_t represents the vector of reset gates, z_t represents the vector of update gates, $tanh$ is the hyperbolic tangent function, the element multiplication is represented by Θ , and h_t is the candidate activation. The use of the AM requires the determination of the attention weights, which are calculated as shown in equation 3.14.

$$\partial_l = \frac{exp(f(g_l))}{\sum_u exp(f(g_u))} \tag{3.14}$$

In equation 3.14, $f(g)$ is the scoring function, g_l and g_u represent each input vector, and ∂ represents the attention weights. The output of attention is calculated as shown in equation 3.15.

$$attentive_g = \sum \partial_l g_l \tag{3.15}$$

3.2. DL-based Model Construction for SE Classification and State Detection. Traditional research on emotion recognition is mainly based on discrete emotion description models and dimensional emotion description models. In order to analyse the emotion in the speech of short English texts read aloud, the study designed a DL-based SE analysis model. The model not only allows for real-time monitoring of speech signals, but also allows for the analysis of the emotions of English learners when they read aloud short English texts. The structure is shown in Fig 3.3. The SE analysis model for reading aloud short English texts is broken down into three key parts, as seen in Fig 3.3. The pre-processing of the data comes first, followed by the development of the SE categorization and state detection modules, and finally the experiment. For data preprocessing, the conventional approach is to forcibly align words with different modalities, which can lead to insufficient interaction between modalities. Therefore, in terms of data preprocessing, the study adopted a method of not forcing

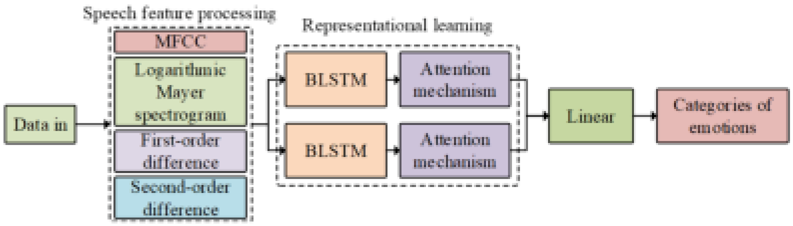


Fig. 3.4: The specific structure of the DABLSTM model

word alignment, while also using feature extraction tools to obtain feature vectors from different modalities. For text modality, research uses Transformer based Bidirectional Encoder Representation from Transformers (BERT) word embedding to represent modal features. The acoustic modality uses the COVAREP tool to represent modal features, while the visual model uses the Facet framework to represent modal features. For the construction of the SE classification module, a DABLSTM model was designed, and the specific structure of this model is shown in Fig 3.4.

The processing of linguistic feature data and representation learning are the two fundamental components of the DABLSTM model, as depicted in Figure 3.4. The linguistic feature processing includes MFCC features, log-Meier spectrograms. Representation learning mainly includes BLSTM and AMs. Traditional neural network models have insufficient emotional information acquisition in language emotion classification, while BLSTM can not only avoid the problem of vanishing gradient in recurrent neural networks, but also effectively capture future and past information in feature sequences. For log-Meier spectrograms, the study mainly uses linear interpolation for data pre-processing. The operation of linear interpolation is completed in the time series direction of the number Mel spectrum. When the audio duration is less than 10 seconds, a shape that is consistent with 10 seconds of audio can be obtained. When the audio duration is greater than 10 seconds, it is necessary to shorten the sample length by discarding some samples. For deep features, the study mainly uses CNN for extraction. In addition, the temporal information in MFCC and the local and full feature information in deep features can be extracted by DABLSTM model. For the construction of the SE state detection module, the study designed a CMAM with Sentiment Prediction Auxiliary Task (CAM-SPAT) model which is shown in Fig. 3.5.

In Fig 3.5, the multimodal feature processing module mainly involves BERT word embedding, COVAREP tool, and Facet framework. The representation learning module focuses on the encoding of sentiment feature vectors in different modalities using a Bidirectional Gated Recurrent Unit (BiGRU). BiGRU can not only delete invalid information from different modal feature data, but also efficiently obtain effective information from feature data. The learned representations are fed into the weighted CMAM module and the emotion prediction assistance task module. The weighted CMAM module is mainly concerned with the calculation of spatial correlations across multiple attention and time steps, and linear processing of the output data. The mood prediction auxiliary task module involves linear processing of the data across modalities and modification of the loss function, which is finally combined with the processing results of the weighted CMAM module as the output of the model.

4. Analysis of the DL-based SE Analysis Model results. The study conducted experiments on the DABLSTM model on the IEMOCAP dataset and validated it by metrics such as unweighted accuracy, weighted accuracy, loss function and confusion matrix. In addition, the study conducted experiments on the CAM-SPAT model on the CMU-MOSEI and CMU-MOSI datasets. Also it validated the performance of the model by metrics such as F1 value and 7-class classification accuracy.

4.1. Analysis of the Outcomes of the DL-based SE Classification Model. In the validation of the DABLSTM model, for the purpose to investigate the effects of the number of layers of parallel structure and input features on the performance of the model, the study was carried out in a single-layer Attention-based

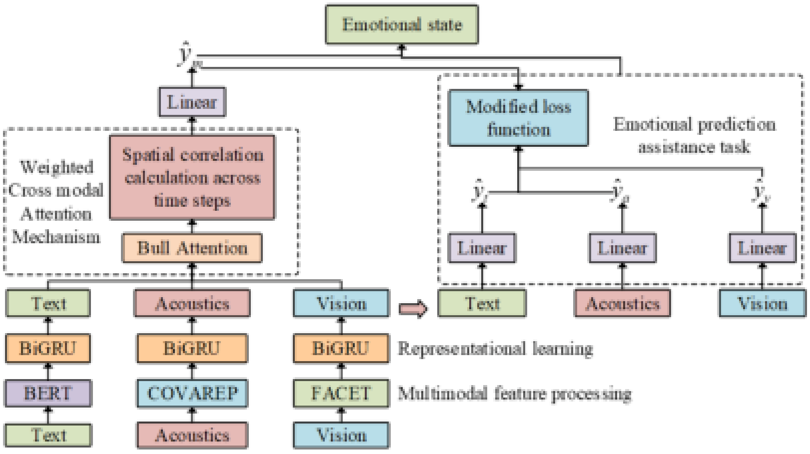


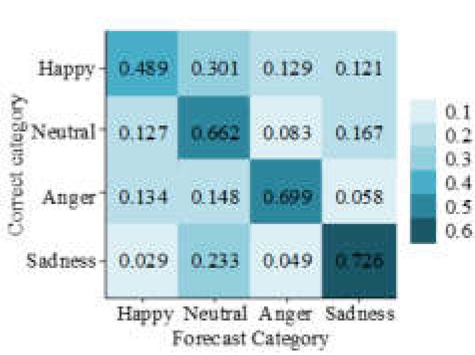
Fig. 3.5: The CAM-SPAT model's specific structure

Bidirectional Long Short-term Memory Networks model (ABLSTM). With various input characteristics and compared confusion matrices for various features, Fig **??** illustrates the comparison of the confusion matrices for various features.

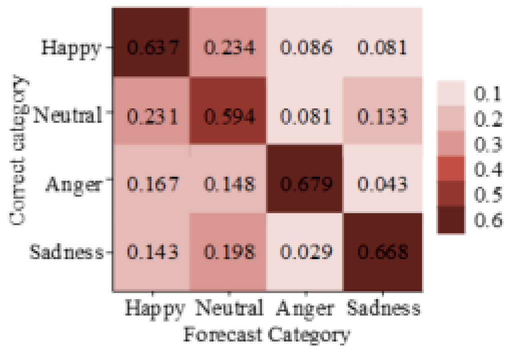
It can be learned from Fig 4.1a that when the input feature is MFCC, the recognition accuracy of happy category is 48.9%, neutral type is 66.2%, angry category is 69.% and sad category is 72.6%. Meanwhile, the Unweighted Accuracy (UA) of ABLSTM was 62.7%. From Fig 4.1b, when the input features were MFCC with DABLSTM difference, the recognition accuracy was 63.7% for the happy category, 59.4% for the neutral type, 67.9% for the angry category and 66.8% for the sad category. the UA of ABLSTM was 63.7%. In Fig 4.1c, when the input features are MFCC with second-order differences, the recognition accuracy is 52.7% for the happy category, 73.4% for the neutral type, 66.8% for the angry category and 65.6% for the sad category. the UA of ABLSTM is 64.3%. It can be seen that the highest UA value was achieved when the input features were MFCC with second-order differences. A comparison of the model performance under different layer structures is shown in Fig 4.4.

In Fig 4.3a, when there are in layer 2, the recognition rates for the happy, neutral, furious, and sad categories are 68.4%, 68.6%, 73.1%, 77.7%, and 73.1% respectively. the UA of DALSTM is 71.2%. In Fig 4.3b that when there are in layer 3, the recognition accuracy of the happy category is 63.2%, the recognition accuracy of the neutral type is 73.3%, the recognition accuracy of the angry category is 71.0%, and the recognition accuracy of the sad category is 60.7%. The UA of the model at this point was 66.4%. From Fig 4.3c, the recognition accuracy of the joyful category is 62.1%, the neutral type is 66.7%, the angry category is 71.6%, and the recognition accuracy of the sad category is 76.8%. The UA of the model at this point was 68.6%. This shows that the DALSTM model with a two-layer structure has the highest UC and the best results. The study compared it with other models and the comparison is shown in Fig 4.6.

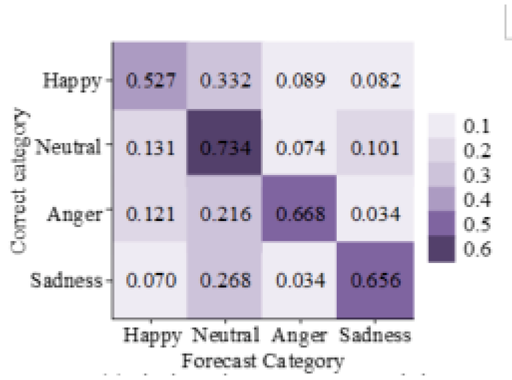
Fig 4.5a illustrates that the 3D-CRNNs model's WA on the training set is 62.98% and its UA is 61.93%. The WA and UA of the ABLSTM-FCNs + DNN model are respectively 60.7% and 61.1%. The WA of the ABLSTM-AFCN model is 69.% and the UA is 68%. the WA of the DALSTM model is 71.98% and UA was 71.29%. From Fig 4.5b, the WA of the 3D-CRNNs model on the validation set is 62.76% and the UA is 61.81%. the WA of the ABLSTM-FCNs + DNN model is 61.82% and the UA is 61.%. the WA of the ABLSTM-AFCN model is 69.8% and the UA is 68.6%. the WA of the DALSTM model is The DALSTM model has a WA of 72.72% and a UA of 72.21%. This shows that the DALSTM model outperforms the comparison model. The study compared and analysed the F1 values and loss functions of the different models, and the comparison results are shown in Fig 4.8. From Fig 4.7a, the maximum value (MaxV) of F1 for the 3D-CRNNs model



(a) Input as MFCC



(b) The input is MFCC accompanied by first-order difference



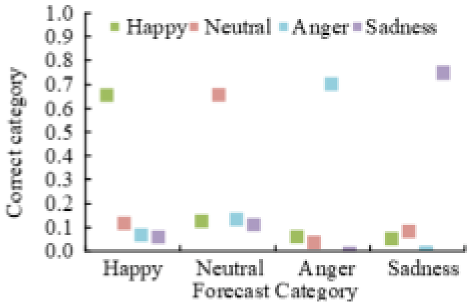
(c) The input is MFCC accompanied by second-order difference

Fig. 4.2: Comparison of confusion matrix corresponding to different features

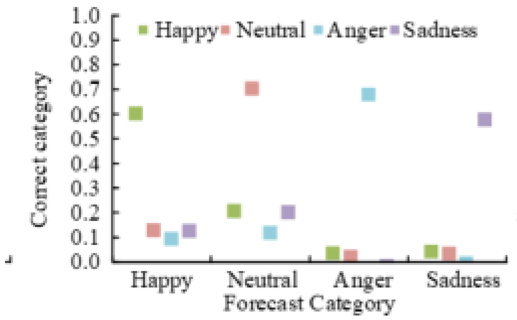
is 97%, the minimum value (MinV) is 95.7% and the mean value (MeV) is 96.48%. the MaxV of F1 for the ABLSTM-FCNs + DNN model is 96.5%, the MinV is 95.2% and the MeV is 95.9%. the MaxV of F1 for the ABLSTM-AFCN model is 97.5%, the MinV is 96.6% and the MeV is 97.04%. the MaxV of F1 for the DALSTM model is 99.20%, the MinV is 97.7% and the MeV is 98.54%. The MaxV of F1 for the DALSTM model is 99.20%, the MinV is 97.% and the AV is 98.54%. The MaxV of the loss function of the 3D-CRNNs model is 1.87 and the MinV is 0.71.

4.2. Analysis of the Results of the DL-based SE Condition Detection Model. The CMU-MOSI dataset and the CMU-MOSEI dataset were utilised in the study's validation of the CAM-SPAT model in order to evaluate the various models and to observe the regression and classification outcomes both with and without the CAM-SPAT model. The Binary Classification Accuracy (Acc-2), 7-class Classification Accuracy (Acc-7) and F1-value metrics were used to more clearly compare the regression and classification outcomes with and without the CAM-SPAT model. A comparison of the regression and classification results with and without the CAM-SPAT model for the different datasets is shown in Fig4.10.

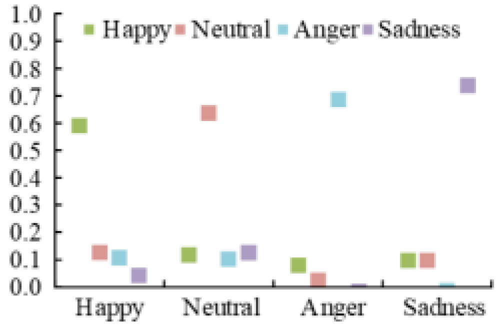
On the CMU-MOSI dataset, the MaxV of Acc-2 is 85.69%, the MinV is 51.87%, and the MeV is 75.87%, as shown in Fig 4.9a. The MaxV of Acc-7 is 47.94%, the MinV is 16.89% and the MeV is 39.03%. the MaxV of F1 value is 85.61%, the MinV is 44.20% The MaxV for F1 was 85.61%, the MinV was 44.20% and the AV was 74.21%. The MaxVs for all three metrics occurred when the CAM-SPAT model was used and the input features were text acoustic and visual, the MinVs for Acc-2 and F1 occurred when the CAM-SPAT model was not used and the input features were acoustic and visual, and the MinVs for Acc-7 occurred when the CAM-SPAT model



(a) One layer uses depth features, and the other layer uses first-order and second-order differential MFCC

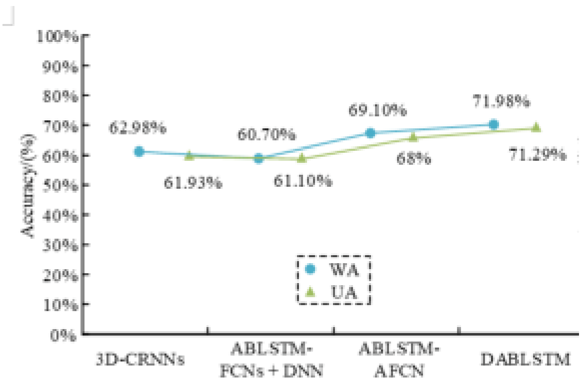


(b) One layer uses second-order differential MFCC, one layer uses MFCC, and the other layer uses extracted depth features

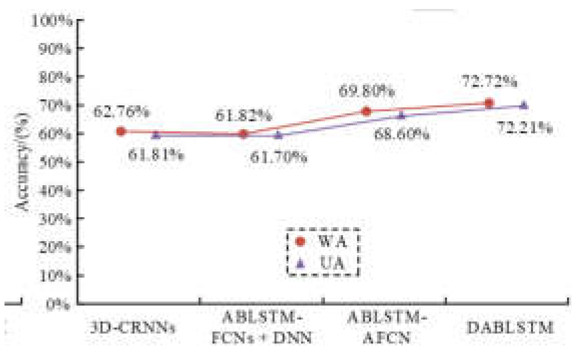


(c) One layer uses second-order differential MFCC, one layer uses first-order differential MFCC, and the other layer uses MFCC

Fig. 4.4: Comparison of the performance of the lower model with different layers

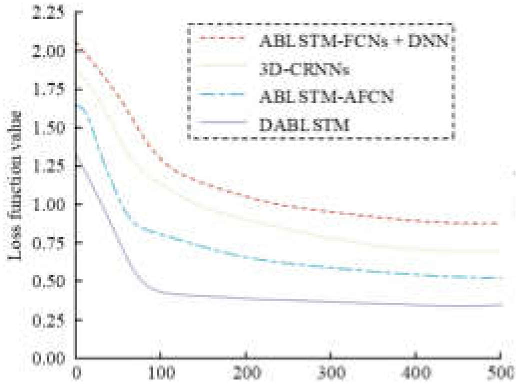


(a) Comparison of the models on the training set

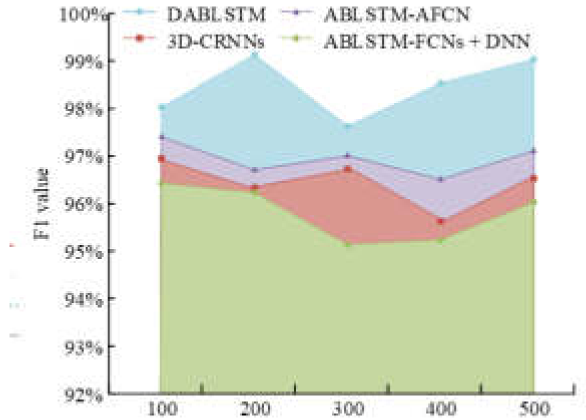


(b) Comparison of the models on the validation set

Fig. 4.6: Comparison of DALSTM model and other models on UA and WA

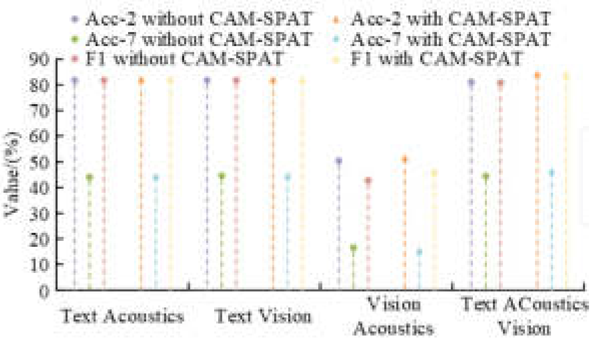


(a) Comparison of loss functions of different models

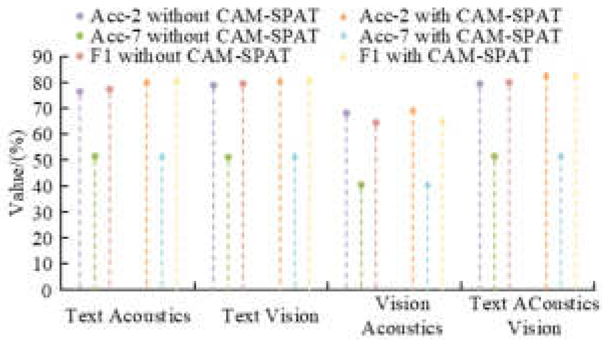


(b) Comparison of F1 values between different models

Fig. 4.8: Comparison of F1 values and loss functions of different models



(a) CMU-MOSI Dataset

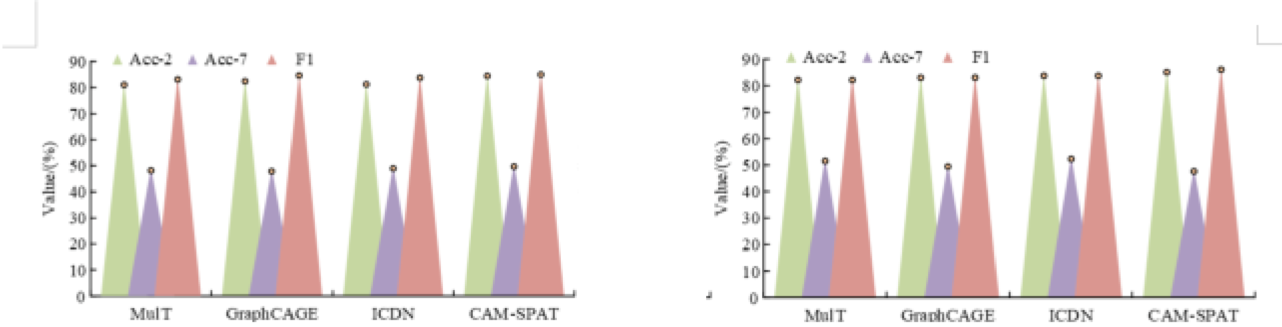


(b) CMU-MOSEI Dataset

Fig. 4.10: Comparison of regression and classification results using and not using CAM-SPAT models under different datasets

was used and the input features were acoustic and visual. From Fig 4.9b, on the CMU-MOSEI dataset, the MaxV of Acc-2 is 84.82%, the MinV is 69.9% and the MeV is 78.93%. the MaxV of Acc-7 is 53.82%, the MinV is 42.3% and the MeV is 50.64%. The MaxV of F1 value is 84.84%, the MinV is 66.29% and the MeV the MaxV for F1 was 84.84%, the MinV was 66.29% and the AV was 78.34%. The MaxVs for all three metrics occurred when the CAM-SPAT model was used and the input features were text acoustic and visual, while the MaxVs for all three metrics occurred when the CAM-SPAT model was not used and the input features were acoustic and visual. The study compares the CAM-SPAT model with other models. The models compared are Multimodal Transformer (MulT), Graph Capsule Aggregation (GraphCAGE) and Integrating Consistency and Difference Network (ICDN). The comparison results are shown in Fig 4.12.

From Fig 4.6 4.11a, the MaxV of Acc-2 is 85.46%, the MinV is 82.01% and the MeV is 83.28% on the CMU-MOSI dataset. the MaxV of Acc-7 is 50.66%, the MinV is 48.82% and the MeV is 49.63%. the MaxV of F1 value is 85.99%, the MinV is 84.12% and the The MaxV of F1 was 85.99%, the MinV was 84.12% and the



(a) CMU-MOSI Dataset (b) CMU-MOSEI Dataset

Fig. 4.12: Comparison results of different models under different datasets

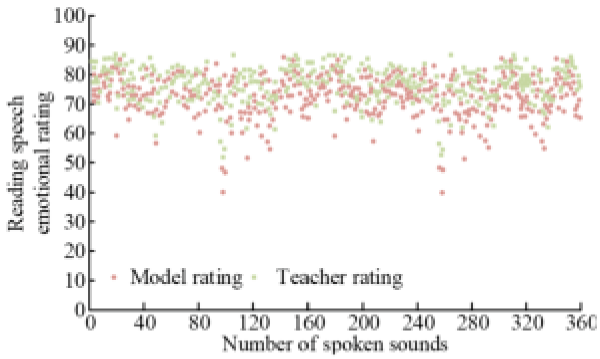


Fig. 4.13: Comparison of ratings between models and professionals under the same voice data

average value (AV) was 85.13 The MaxVs of all three indicators were on the CAM-SPAT model. As can be seen through Fig 4.11b, on the CMU-MOSEI dataset, the MaxV of Acc-2 was 84.99%, the MinV was 82.08% and the MeV was 83.43%. the MaxV of Acc-7 was 52.66%, the MinV was 48.02% and the MeV was 50.61%. the MaxV of F1 value was 85.98%, the MinV was 82.75 with a MeV of 83.85%. The MaxVs for all indicators, except Acc-7, were on the CAM-SPAT model. This shows that the CAM-SPAT model has a higher accuracy rate. The study compared the scores of the model with the scores of the professionals for the same speech data, and the results of the comparison are shown in Fig 4.7.

In Fig 4.13, the majority of the data for both the model scores and the professional scores overlap with one another. The values for both the model ratings and the professional ratings are essentially in the range of 70 to 90. The MeV of the model scores was 73.55 and the MeV of the professional scores was 77.02. This shows that the model proposed by the study for analysing the sentiment of reading aloud short English texts has a high degree of reliability.

5. Conclusion. In order to enable English learners to improve their oral English proficiency, a deep learning based emotional analysis model for English short text reading pronunciation has been proposed, which includes the DALSTM model and the CAM-SPAT model. According to the study's findings, the DALSTM model had a training set WA and UA of 71.98% and 71.29%, respectively. On the validation set, the DALSTM model had a WA of 72.72% and UA of 72.21%. The DALSTM model's F1 values had a MaxV of 99.20%, a MinV of 97.7%, and a MeV of 98.54%. From this, it can be seen that the performance of the DALSTM model is superior to that of the comparative model. The DALSTM model fared better than the contrast model. The CAM-SPAT model produced the highest values for Acc-2, Acc-7, and F1 values on the CMU-MOSEI dataset, with respective values of 85.46%, 50.66%, and 85.99%. The greatest values of Acc-2 and F1 on the CMU-MOSEI dataset were all discovered on the CAM-SPAT model with 84.99% and 85.98%, respectively. This demonstrates the superiority of the CAM-SPAT model over the comparison model. Additionally, the sentiment analysis model's mean rating was 73.55, which was not significantly different from the mean rating of professionals. This demonstrates the great dependability of the sentiment analysis approach put forward in the study. The study's proposed sentiment analysis model has high reliability and superiority, but it also has some drawbacks, including the selection of a relatively small number of modalities and the omission of modal information like posture, which can be improved in future research.

6. Discussion. The research mainly designed an emotion analysis model for English short article reading speech, including an emotion classification model and an emotion state detection model. There are three main contributions to the research on the sentiment classification model DALSTM. The first one is to solve the problem of inconsistent audio signal length through linear interpolation. The second point is to verify that the double-layer structure improves the performance of the model by comparing the confusion matrix and model recognition accuracy. The third point is to verify the superiority of the performance of the DALSTM model. There are also three main contributions to the research on the emotional state detection model CAM-SPAT. The first one is to use a method of non mandatory word alignment and use different methods to represent features of different modalities. The second is to validate the effectiveness of the CAM-SPAT model through different classification indicators and comparison algorithms, and to verify the feasibility of the CAM-SPAT model by comparing professional ratings and model ratings. The third point is the combination of emotion prediction assistance tasks and cross modal attention mechanisms.

REFERENCES

[1] Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P., Dasgupta, N. & Choudhury, P. Facial Emotion Detection to Assess Learner's State of Mind in an Online Learning System. *5th International Conference On Intelligent Information Technology (ICIIT 2020)* . pp. 107-115 (2020)

[2] Sun, Y. & Lin, C. Design of Multidimensional Classifiers using Fuzzy Brain Emotional Learning Model and Particle Swarm Optimization Algorithm, Acta Polytechnica Hungarica, vol 18. *No.* **4** pp. 25-45 (2021)

[3] Hajhmida, M. & Oueslati, O. Predicting mobile application breakout using sentiment analysis of Facebook posts. *Journal Of Information Science.* **47**, 502-516 (2021)

[4] Aouani, H. & Ayed, Y. Speech Emotion Recognition with deep learning, Procedia Computer Science. (2020)

[5] Dahmani, S., Colotte, V., Girard, V. & Ouni, S. Learning emotions latent representation with CVAE for text-driven expressive audiovisual speech synthesis, Neural Networks, vol 141. *No.* **11** pp. 315-329 (2021)

[6] Jia, F. & Chen, C. Emotional characteristics and time series analysis of Internet public opinion participants based on emotional feature words, International Journal of Advanced Robotic Systems, vol 17. *No.* **1** pp. 105-108 (2020)

[7] Lu, X. Deep Learning Based Emotion Recognition and Visualization of Figural Representation. *Frontiers In Psychology.* **12** (2022)

[8] Park, C., Cho, H., Lee, D. & Jeon, H. Latent profile analysis on Korean nurses: Emotional labour strategies and well-being. *Journal Of Advanced Nursing.* **78**, 1632-1641 (2021)

[9] Sun, H., Wang, G. & Xia, S. Text tendency analysis based on multi-granularity emotional chunks and integrated learning, Neural computing & applications, vol 33. *No.* **14** pp. 8119-8129 (2021)

[10] Gupta, I. & Joshi, N. Feature-Based Twitter Sentiment Analysis With Improved Negation Handling, IEEE Transactions on Computational Social Systems, vol 8. *No.* **4** pp. 917-927 (2021)

[11] Mehmood, Y. & Balakrishnan, V. An enhanced lexicon-based approach for sentiment analysis: a case study on illegal immigration, Online Information Review, vol 44. *No.* **5** pp. 1097-1117 (2020)

[12] Barnes, J., Velldal, E. & Vrelid, L. Improving sentiment analysis with multi-task learning of negation, Natural Language Engineering, vol 27. *No.* **2** pp. 249-269 (2020)

[13] Ghasemi, R., Asli, S. & Momtazi, S. Deep Persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal Of Information Science.* **48**, 449-462 (2020)

[14] Alahmary, R. & Al-Dossari, H. A semiautomatic annotation approach for sentiment analysis. *Journal Of Information Science.* **49**, 398-410 (2023)

[15] Hua, Z., Chen, Z., Bi, C., Biao, H., Mian, L., Cheng, Y. & Bo, J. Complete quadruple extraction using a two-stage neural model for aspect-based sentiment analysis, Neurocomputing, vol 492. *No.* **1** pp. 452-463 (2022)

[16] Yun, L., Maeda, K., Ogawa, T. & Haseyama, M. Chain centre loss: A psychology inspired loss function for image sentiment analysis, Neurocomputing, vol 495. *No.* **7** pp. 118-128 (2022)

[17] Kim, C., Hwang, Y., Kamyod, C. & Study, A. of Profanity Effect in Sentiment Analysis on Natural Language Processing Using ANN. *Journal Of Web Engineering.* **21**, 751-766 (2022)
[18] Contreras, D., Wilkinson, S., Balan, N. & James, P. Assessing post-disaster recovery using sentiment analysis: The case of L'Aquila, Earthquake Spectra, vol 38. *No.* **1** pp. 81-108 (2022)
[19] Kobayashi, M., Hamada, Y. & Akagi, M. Acoustic features correlated to perceived urgency in evacuation announcements, Speech Communication. (2022)
[20] Wang, X., Cheng, M., Eaton, J., Hsieh, C. & Felix, S. Fake node attacks on graph convolutional networks. *Journal Of Computational And Cognitive Engineering.* **1**, 165-173 (2022)
[21] Chen, Z. Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. *Journal Of Computational And Cognitive Engineering.* **1**, 103-108 (2022)
[22] Hiriyannaiah, S., Siddesh, G., Kiran, M. & Srinivasa, K. A comparative study and analysis of LSTM deep neural networks for heartbeats classification, Health and Technology, vol 11. *No.* **3** pp. 663-671 (2021)

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 27, 2023
Accepted: Sep 26, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 909–917, DOI 10.12694/scpe.v24i4.2466

MAPPING LEARNER'S QUERY TO LEARNING OBJECTS USING TOPIC MODELING AND MACHINE LEARNING TECHNIQUES

SOUVIK SENGUPTA,* SAURABH PAL,† AND PIJUSH KANTI DUTTA PRAMANIK‡

Abstract. Inquiry-based learning supports the independent knowledge development of the learner in an e-learning environment. It is crucial for the learner to obtain the appropriate Learning Object (LO) for the intended query. Mapping a learner's query to the right LO is a challenging task, as keyword-based searching on the topics or content does not guarantee the best result for various reasons. A query that apparently connects a topic may also implicitly refer to multiple other topics. Besides, the content of an LO with the same topic name often varies over different portals. Therefore, there is always a need for a method to automatically identify the latent topics of the query and then find the most relevant LO that covers the query. This paper aims to build a recommender system that maps a given input query to a suitable LO based on the most appropriate matching of learning contents. The proposed work employs an amalgamation of different supervised and unsupervised methods of natural language processing and machine learning. The machine learning model is trained on a handcrafted dataset to map queries into predefined topics. The proposed algorithm also leverages a dynamic topic modeling technique on learning content collected from three popular e-learning portals and uses a similarity score to map the learner's (user) query to the most appropriate LO.

Key words: Natural Language Processing, Machine Learning, Topic Modeling, LDA, Learning Object

1. Introduction. Inquiry-based learning is a method of active learning that starts by posing questions, problems, or keywords and then generating facts and information about a topic related to the query. In an e-learning environment, it helps the learner to develop knowledge on their own. Inquiry-based learning is one of the popular forms of problem-based learning (PBL), among others like simulations, case studies, guided design, and project-based learning, where the problem comes before instructions and knowledge [4] [5] [8]. E-learning provides a powerful platform for inquiry-based learning as it facilitates the enquirer to seek information from an extensive repository of learning objects (LOs). An LO is a collection of content, example, and assessment items that are combined based on a single learning objective [9] [18].

Despite the availability of web search engines and in-portal searching tools, the common problem that inquiry-based learning confronts is retrieving precise information (in terms of LO) for user queries [10]. Since all search engines mostly rely on keyword-based metadata, if queries are ill-defined or missing a vital keyword, it fetches LOs that are superfluous and often fail to make any meaningful connection to the topic of a query [11]. On the other hand, even if the topic associated with the query is known, sole topic-based searching is also insufficient, as the same LO can be described under different topics at different e-learning portals. For example, the query "How is polymorphism supported in java?" may seem to be related to the topic "polymorphism". However, the required LO can actually be found at different portals under different topics like "inheritance", "method overriding", and "method overloading". This necessitates identifying the latent topic(s) associated with the query and the LOs so that topics can be matched to identify the most appropriate LO for a query.

In this paper, we address three specific problems of query-based learning, which are:

- Finding the most appropriate LO for a query.
- Finding the latent topics related to the query and LOs.
- Matching one topic model to another.

A suitable methodology is proposed in this work to automate the process of mapping the learner's query to the appropriate LO in an e-learning environment. For identifying latent topics from the query, we employ both supervised and unsupervised techniques, and dynamic topic modeling is applied as a method for summarizing.

*Aliah University, Kolkata, India (mesouvik@hotmail.com).
†Swami Vivekananda Institute of Science and Technology, Kolkata, India (saurabhpal2007@gmail.com).
‡Galgotias University, Greater Noida, India (pijushjld@yahoo.co.in). Corresponding author

Finally, the selection of LO for a query is based on the proximity between the topic models of LOs and the topic model of the query. The notable contributions of this work can be briefed as follows:

- A dataset containing 1000 questions on 'object oriented programming' labeled with topic name.
- Comparative analysis of machine learning models to predict the topic from the learner's query.
- Topic modeling for LOs on three popular e-learning portals.
- A proposed method for similarity measure within the topic models.

The rest of the paper goes as follows: Section 2 presents a review of the related works, Section 3 illustrates the proposed methodology in detail, Section 4 presents the results and analysis, and Section 5 has the conclusion.

2. Related Work. The related works can be broadly classified (with some overlapping) into two categories - a) mapping user queries to predefined classes or topics using feature selection and supervised machine learning approaches, and b) topic modeling and topic identification using the unsupervised approach.

Many works have been carried out towards mapping users' queries to appropriate topics using different machine learning techniques. These are not only limited to e-learning but also have broad applications in tagging community questions with predefined classes, clustering and segmenting questions, and improving search results for recommendation systems.

It is evident from the reviewed works that word embedding and feature selection play a pivotal role in topic classification with machine learning models. Yang et al.[19] proposed a topic-oriented word embedding approach for query classification. This topic-oriented word embedding with fine-tuning of the Word2Vec model has shown high precision of 95.73% on the Baidu and Sogou datasets. Gulzar et al. [6] presented a course recommender system to recognize the intent and requirements of the user's query. The query and the course are represented using the n-gram feature set and mapped to classes with the help of domain ontology. Li et al. [7] used a hierarchical classifier to classify questions based on their semantics. The questions are mapped into a hierarchy of six coarse classes and 50 fine-grained classes. The classifier uses SNoW learning architecture, where two simple classifiers are combined and used for classification. The first classifier classifies the question into coarse classes, while the second classifier classifies the question into fine-grained classes. This work uses six feature sets, where each of the sets consists of an incremental combination of primitive feature types like words, pos-tags, chunks, head chunks, named entities, and semantically related words. This empirical study showed satisfactory results with more than 90% overall precision. In a similar work, Nguyen et al. [16] described a machine learning approach for question classification based on coarse and fine-grained question classes on the TREC dataset. This work primarily focuses on feature selection methods. Different lexical, syntactical, and semantic features are extracted. Integrating these features into a unique form allowed a better feature representation of the questions for classification. Different supervised learning algorithms such as k-nearest neighbors (KNN), naive Bayes (NB), decision tree (DT), and support vector machine (SVM) are used for the classification task. The experimental result shows better classification accuracy by introducing new features than standard features for coarse and fine-grained classes.

Many researchers have also worked on mapping topics to questions using the Community Question Answering (CQA) dataset. Qu et al. [13] illustrated the problems of question topic classification using an extensive real-world CQA dataset. This work empirically evaluates the different models for classifying questions into CQA categories. This work depicts a comparative analysis of the usefulness of n-Gram features over bag-of-word features. It also employs a variety of flat classifiers, NB, Maximum Entropy (ME), and SVM, and combines them with state-of-the-art hierarchical models for question classification. The findings show that SVM outperforms other models in terms of effectiveness, while NB takes the shortest training time. In a similar work, Singh et al. [14] proposed a methodology to classify a new question into one of the hierarchical categories of a CQA portal. A translation model is used to retrieve questions from the corpus that is lexically and semantically similar to the input question. The system employs a near neighborhood-based classifier on the retrieved similar questions to detect the class for the new question.

On the other hand, most of the works related to topic identification for a query have employed Latent Dirichlet Allocation (LDA). Bhattacharya et al. [1] described a method to classify a query into a topic class by considering the query keywords distributed over various topics. This work assumes that queries consisting of keywords may belong to multiple topics, representing overlapping concepts. In this work, LDA is applied to the entire corpus to group the documents into topics consisting of unique words forming the topic models.

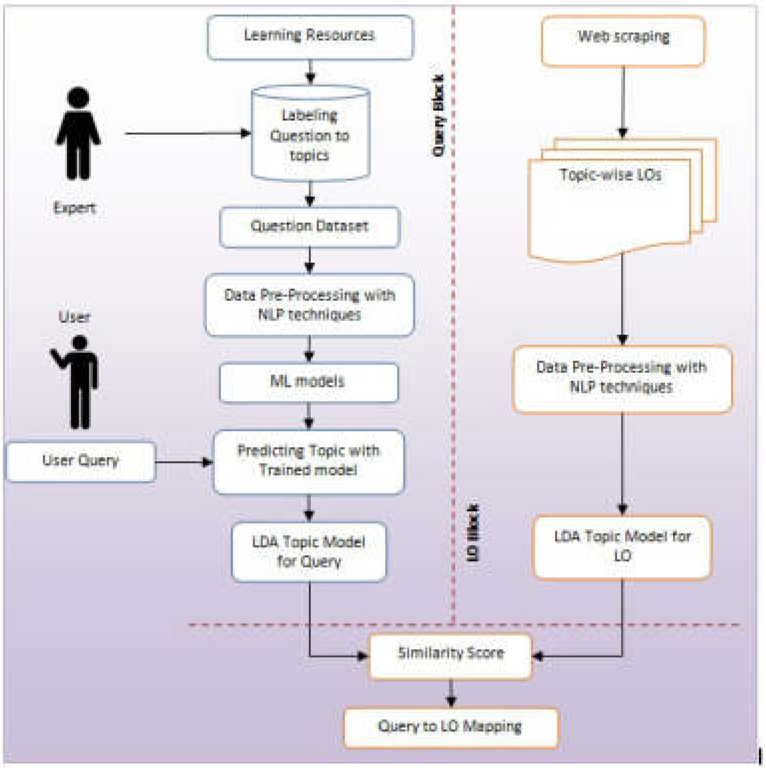


Fig. 3.1: Architecture for user query to LO mapping

Subsequently, a sparse representation-based classifier (SRC) is applied over the topic models to predict the topic class of the query. In a similar work, Zhang et al. [20] depicted an approach for capturing the word-semantic similarities between a user query and questions in the repository by introducing topic modeling. In this work, the LDA model is used to identify the questions' topics and subsequently generate the topic model. The topic model helps to map the stored Q&A pair appropriately into the topic space. An unsupervised clustering approach is used to deduce the similarity between the user's input query and the questions in the focused topic space. Subsequently, the most similar questions are obtained from the Q&A repository. In a similar work, Das et al. [3] presented an unsupervised topic modeling for auto-generated questions. It first performs a phrase mining operation on a given set of documents and checks if a set of words or terms, with a particular sequence, can occur more frequently than others and then creates bag-of-phrases. Then it employs constrained topic modeling using LDA, which uses this bag-of-phrases for the topic model. In all these works, LDA is predominantly used for topic modeling and extracting latent topics.

Most of the above-reviewed works are either focused on the classification of queries into predefined topics or on identifying latent topics dynamically from the question repositories. However, none of the works have attempted to use topic modeling and classification techniques jointly to identify the most appropriate LO for the user's query. In this work, the proposed methodology combines the best of the two worlds, i.e., a) training supervised models with necessary feature extraction on a labeled dataset and b) mapping the prediction with a dynamic topic model using LDA.

3. Methodology. Figure 3.1 depicts the framework of the proposed methodology. The working modules of the proposed framework are organized into two sets of components – the query block and the LO block. The operations of the blocks are independent of each other, but the resultant topic models are collectively assessed for similarity checks to map a query to its corresponding LO.

The query block has two components - first, it trains an machine learning model for classifying questions

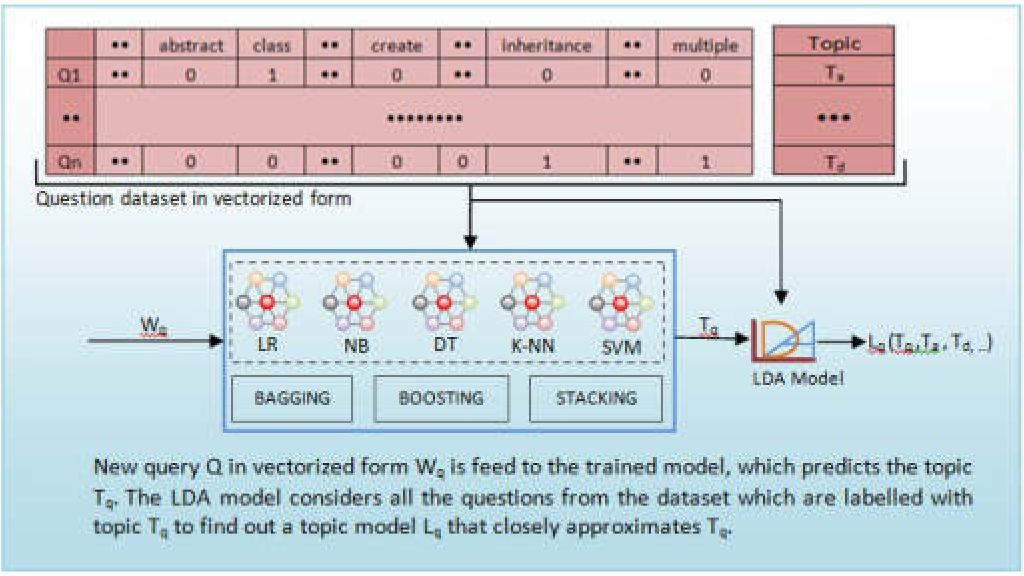


Fig. 3.2: Query to topic model

into predefined topics using a handcrafted dataset, and second, it takes a user's query to predict a topic using the trained model and subsequently creates a dynamic topic model for the query using LDA. The preparation of the dataset involves subject experts correctly labeling the questions. This work uses a dataset of questionnaires from an elementary course on "Object-Oriented Programming with Java". A total of 1377 questions are selected, and each question is hand-labeled with a topic name from a predefined set of 13 topics. Three subject experts have volunteered for this work. The text data is preprocessed using conventional NLP methods, viz. tokenizing, noise and stop word removal, and stemming. Then the preprocessed text is feature-extracted and vectorized using the Term Frequency - Inverse Document Frequency (TF-IDF) embedding with a one-gram. A set of machine learning models like logistic regression (LR), NB, DT, KNN, and SVM with different sets of hyper-parameters in K-fold cross-validation is used for the classification task. It explores multiple machine learning classifiers and ensemble methods. The classification models are trained over a data set that is performed to find out the best combinations of hyper-parameters. Besides individual models, ensemble methods like Bootstrap Aggregating or bagging, boosting, and stacking are also adopted for classification. For an unseen query, the machine learning classifier first predicts a topic from the predefined topic list. Then we collect all the questions labeled to that topic in our dataset. This pool of questions is fused to generate a topic model dynamically. Since our postulation is that questions in the dataset may also have latent topics besides the one assigned to them, we employ LDA to unveil those topics. We consider this topic model as the representation of the given query in the query block and is used to compare with the topic model of the LO block. Figure 3.2 shows the workflow of transforming the query to a topic model.

In the LO block, first, a repository is prepared by web-scraping three different e-learning portals [17][15][12]. From each portal, a set of interconnected LO is fetched and stored in the LO dataset. Each LO is textual content, which is represented by a single topic name as appeared in the learning portal. However, this topic name is replaced in the dataset by a dynamically created topic extracted by LDA. It enables us to counter the problem that arises due to different topic names for the same content across different portals. The textual contents are preprocessed and vectorized in a similar way as done for query text. The working mechanism of transforming LO to a topic model is shown in Figure 3.3.

Finally, for assessing the similarity between the two topic models produced from the two blocks, a similarity measuring algorithm is defined. LO that has the topic model with the highest average similarity with the query topic model is selected as the most appropriate LO for that query. In the next section, the core components of

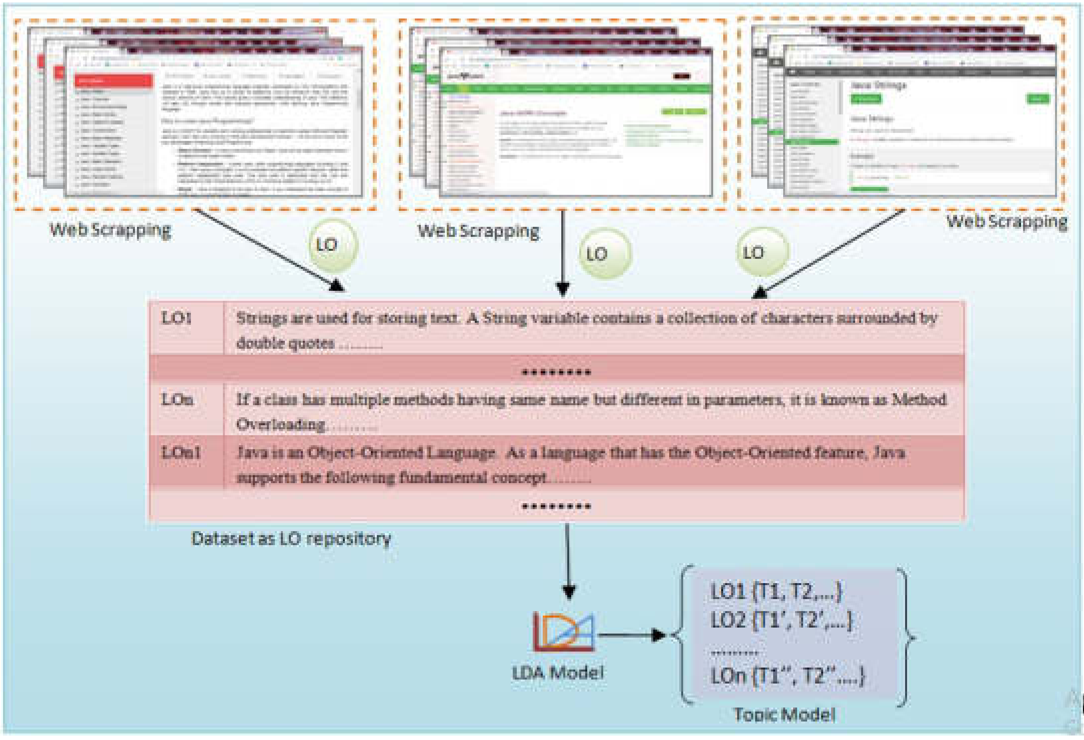


Fig. 3.3: LO to topic model

the methodology are discussed in detail.

3.1. Data Preprocessing. Common data preprocessing steps are carried out both on questions and scrapped web contents, which involve a) data cleaning and noise removal, b) stop word removal, and c) stemming. Since the text data for LO is web scrapped, it contains noise in the form of missing HTML tags, misspells, and non-ASCII characters and thus needs special attention for cleansing and removing noise. Noise removal is carried out using regular expressions features of the Python language. Next, punctuation and stop words also were removed as they do not add any information to the text analysis. Stemming removes morphological affixes from words, leaving only the root words. PoterStemmer, an NLTK library, was used for stemming in this work.

3.2. Feature Extraction. Considering the fact that users often rely on individual words rather than patterns or sequences of words in a query, we used one-gram features instead of n-gram. This also enables us to counter the problem of the unstructured query. However, it also brings a compulsion to create the one-gram model for the LOs as well. Next, for vectorization and word embedding, the TF-IDF model is adopted. Since the work has no special requirement to retain the semantic relationship between the words, TF-IDF is an inevitable choice where we can also set the minimum and maximum frequency of a word to be considered.

3.3. Machine Learning Models. Five models - LR, NB, DT, KNN, and SVM are explored for the classification task. The objective of this supervised approach is to train a model to classify any questions into some predefined topics. This work does not consider multi-label classification, i.e., each question in the database is mapped to one topic only. For fine-tuning the hyper-parameters of the models, K-fold cross-validation is performed. In addition to these five models, three types of ensemble methods, namely bagging, boosting, and stacking, are employed on top of the baseline classifiers. This work explores random forest (RF) and an ensemble of all other single models for bagging, Adaptive Boosting (AdaBoost) and extreme Gradient Boosting (XGBoost) for boosting, and KNN and SVM classifiers for stacking. The classification performances of all the

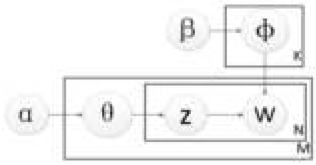


Fig. 3.4: LDA model for topic-word distribution(adopted from [2])

models are measured in terms of accuracy (ac), precision (pr), and recall (rc).

3.4. LDA Topic Model. LDA is a generative probabilistic model that can create a topic model from a given text corpus that explicitly represents a document. Each topic is modeled as an infinite mixture over a set of topic probabilities. The underlying principle of LDA is given in Figure 3.4. Where, α is the per-document topic distribution, β is the per-topic word distribution, θ is the topic distribution for document d, $P(z = k \mid d)$ is the probability of topic $z = k$ given document d, ϕ is the word distribution for topic k, $P(w = v \mid z = k)$ is the probability of word $w = v$ given topic $z = k$, Z is the topic for the n^{th} word in the document, W represents each word in the corpus.

We employed two different configurations of LDA in two phases. Since the vocabulary size of the learning content is much larger than the vocabulary size of the question set, the value of k is proportionally adjusted. For query modelling, we set the $k = 5$ for each topic, and in the LO block, the size of $k = 8$ is used.

3.5. Similarity Measure. Once the topic model for the query and the topic models for the LOs are prepared from the two blocks, the final step of the methodology is to measure the similarity between the topic models in order to identify the most appropriate LO for the learner's query. If t_q is a topic model for the user query q, we need to find out the most similar topic model t_l from the available LO topic models (t_{l1}, t_{l2}, ...t_{ln}). The simplest way of measuring this could be finding the cosine difference between the BOW representation of t_q and t_{li}. However, the conventional way of representing a topic vector in the entire word space is not very useful, as the number of words in the LO vocabulary is much larger than the number of words in one topic. This would result in each topic being a highly sparse vector. Therefore, assuming the LDA-generated topic models (t_{l1}, t_{l2},...t_{ln}) as a good summarization of the entire content of the LO repository, we use only the set of words used in the topics (t_{l1}, t_{l2},...t_{ln}) as the new word vocabulary W=$\{w$ for each w in $t_{li}\}$. Thus, in this new vector space, t_q and t_{li} are represented uniformly, and the cosine distance could be used to measure the similarity between them. A similarity score of a LO is considered as an average similarity of all topics of the given LO. The LO with the highest average similarity score is taken as the best match for the query. Algorithm 1 depicts the process of finding the similarity between the topics t_{li} and t_q.

4. Results and Discussion. Five machine learning classifiers are tested on K-fold cross-validation data with accuracy, precision, recall, and F1 score as the measures of performance. Table 4.1 shows the comparative performance analysis of different single and ensemble models. LR classifier with penalty = 'l2' and solver = 'lbfgs' and SVM classifier with penalty = 'l2', kernel = 'rbf', and loss = 'squared_hinge' showed better performance than NB and KNN. We used Gaussian NB and KNN with K = 5, but they showed poor validation accuracy. DT with splitting_criteria = 'gini' showed a moderate performance on validation data. However, DT and NB showed good improvement when employing ensemble methods on top of the base classifiers. Bagging method RF showed an improvement of 4.06% over DT. NB with bagging increased accuracy by 1.5% compared to the baseline. However, ensemble methods do not reflect any noticeable improvement to the base models in all other cases. SVM and LR both showed good accuracy, consistent with any ensemble method. The comparative study reveals that the SVM performed better than other models at different k folds and different random states.

On the unsupervised side, topic models are built both on the questions dataset and the LO repository. Table 4.2 shows some of such LDA topics. The number of topics in the LDA model for questions and LOs is set to 10 and 15, respectively, while the word length is set to 5 and 8, respectively. Table 4.3 depicts the cosine similarity score of topic_9 (of Table 4.2) with some of the LO topics as per the proposed algorithm. As cosine similarity is a bitwise operation, the time complexity of the algorithm is O ($k * m * n$), where, k is the number

Algorithm 1: Finding the most similar topic t_{li} for the topic model t_q
Input: t_q=topic model for user's query input LOS = set of learning objects t_l = topic model for LO
Output: best_matching_LO

for each LO in LOS
{
LO_avg_score=0
LO_score=0
for each t_{li} in t_l
{
cosine_sim_score=0
for each topic t_j in t_q
{
cosine_sim_score = cosine_sim_score + cosine_sim(t_{li}, t_j)
}
avg_cosine_sim_score = cosine_sim_score/no_of_topic_in_t_q
LO_score = LO_score + avg_cosine_sim_score
}
LO_avg_score = LO_score/no_of_LO
}
best_matching_LO = LO with maximum LO_avg_score

Table 4.1: Performance of Machine Learning Models.

Models	**ACC**	**PR**	**RC**	**F1**
Logistic regression (LR)	87.86	91.86	86.86	89.29
Decision tree (DT)	76.95	79.95	74.95	77.37
Naïve Bayes (NB)	82.99	85.99	81.99	83.94
SVM	**88.94**	**92.94**	**87.94**	**90.37**
KNN	68.18	71.18	67.18	69.12
Adaboost	67.58	70.58	66.58	68.52
XGboost	78.62	81.62	76.62	79.04
Stacking LR, DT, NB with meta_classifier SVM	62.03	65.03	60.03	62.43
KNN + Bagging	68.94	72.94	67.94	70.35
LR + Bagging	87.98	90.98	86.98	88.94
RF	80.08	84.08	78.08	80.97
NB + Bagging	84.23	87.23	83.23	85.18
SVM + Bagging	**89.04**	**93.62**	**87.94**	**90.69**

of LO, m is the number of topics in the question topic model, and n is the number of topics in the LO topic model.

5. Conclusions. This paper presents a methodology for an automated recommender system that maps a user's query to an appropriate LO that covers the required knowledge for that query. It relies on dynamic topic modelling using LDA to relate between the query and the required LO. Different machine learning models like LR, NB, DT, KNN, and SVM are trained on a hand-curated dataset of questionnaires labeled with some predefined topics. Ensemble methods like bagging, boosting, and stacking are also tested along with the baseline classifiers. SVM outperformed the other models in terms of prediction accuracy, precision, and recall. Given a learner's query, the prediction model can predict a topic from a set of predefined topics. Subsequently, a dynamic topic model is created using LDA from the questions in the dataset on that topic. On the other side, topic models are created from the LOs fetched from e-learning portals. Then, an algorithm is proposed to find out the proximity between two topic models. The LO corresponding to the topic model that has the highest average similarity score with the topic model of the user query is considered the most appropriate LO for that query. This work does not employ any sequential or temporal model for considering textual data,

Table 4.2: LDA Topic Models.

LDA model on	**Topic_4**		**Topic_6**		**Topic_9**	
	Word	**Prob**	**Word**	**Prob**	**Word**	**Prob**
Questions dataset	Memory	0.059	exception	0.110	class	0.050
	Thread	0.045	block	0.096	method	0.038
	Heap	0.044	catch	0.060	interface	0.029
	Stack	0.035	try	0.043	java	0.029
	Java	0.030	finally	0.035	abstract	0.028
LO repository	**Topic_2**		**Topic_7**		**Topic_11**	
	Word	**Prob**	**Word**	**Prob**	**Word**	**Prob**
	variable	0.072	method	0.026	class	0.142
	method	0.029	java	0.024	method	0.033
	instance	0.029	string	0.023	abstract	0.020
	class	0.028	class	0.020	java	0.017
	object	0.027	value	0.019	inner	0.017

Table 4.3: Similarity Scores

LDA topic from Java question corpus	**LDA topic from LOs from web portals**	**Cosine similarity score**
Topic_9: 0.050*“class” + 0.038*“method” + 0.029*“interface” + 0.029*“java” + 0.028*“abstract”	Topic_2: 0.072*“variable” +0.029*“method” + 0.029* “instance” + 0.028*“class” + 0.027*“object”	0.4
	Topic_7: 0.026*“method” + 0.024*“java” + 0.023*“string” + 0.020*“class” + 0.019*“value”	0.6
	Topic_10: 0.142*“class” + 0.033*“method” + 0.020*“abstract” + 0.017*“java” + 0.017*“inner”	0.8

assuming users' queries consist of fewer words and are often unstructured. Therefore, the future scope of this work is to explore word embedding techniques like word2vec and temporal deep learning models like LSTM and Transformer for the prediction task.

REFERENCES

[1] I. Bhattacharya and J. Sil, *Query classification using lda topic model and sparse representation based classifier*, in Proceedings of the 3rd IKDD Conference on Data Science, 2016, 2016, pp. 1–2.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, Journal of Machine Learning Research, 3 (2003), pp. 993–1022.

[3] S. Das and R. Chatterjee, *Work in progress - an approach for mapping of auto generated questions to related topics in the curriculum*, in International Conference on Advances in Computing, Communications and Informatic, 2015.

[4] B. L. Erickson, C. B. Peters, and D. W. Strommer, *Teaching First-Year College Students - Revised and Expanded Edition of Teaching College Freshmen*, Jossey-Bass, San Francisco, CA, revised ed., 2006.

[5] E. Eslinger, B. White, J. Frederiksen, and J. Brobst, *Supporting inquiry processes with an interactive learning environment: Inquiry island*, Journal of Science Education and Technology, 17 (2008), pp. 610–617.

[6] Z. Gulzar and A. A. Leema, *Course recommendation based on query classification approach*, International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 13 (2018), pp. 69–83.

[7] X. Li and D. Roth, *Learning question classifiers*, in COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[8] E. Lopez, *Mckeachie's teaching tips: Strategies, research, and theory for college and university teachers*, The Review of Higher Education, 27 (2004), pp. 283–284.
[9] S. Pal, P. K. D. Pramanik, and P. Choudhury, *Enhanced metadata modelling and extraction methods to acquire contextual pedagogical information from e-learning contents for personalised learning systems*, Multimedia Tools and Applications, 80 (2021), pp. 25309—25366.
[10] S. Pal, P. K. D. Pramanik, A. Maity, and P. Choudhury, *Learner question's correctness assessment and a guided correction method: enhancing the user experience in an interactive online learning system*, PeerJ Computer Science, 7 (2021), p. e532.
[11] S. Pal, P. K. D. Pramanik, A. Nayyar, and P. Choudhury, *A personalised recommendation framework for ubiquitous learning system*, in Proceedings of the 6th International Conference on Intelligent Information Technology (ICIIT '21), 2021, pp. 63—72.
[12] J. T. Point, *Java tutorial*, 2011. June, 2023.
[13] B. Qu, G. Cong, C. Li, A. Sun, and H. Chen, *An evaluation of classification models for question topic categorization*, Journal of the American Society for Information Science and Technology, 63 (2012), pp. 889–903.
[14] A. Singh and K. Visweswariah, *Cqc: classifying questions in cqa websites*, in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 2033–2036.
[15] tutorialspoint, *Java tutorial*, 2013. June, 2023.
[16] N. Van-Tu and L. Anh-Cuong, *Improving question classification by feature extraction and selection*, Indian Journal of Science and Technology, 9 (2016), pp. 1–8.
[17] w3schools, *Java tutorial*, 1999. June, 2023.
[18] D. A. Wiley, *The instructional use of learning objects*, vol. 1, Agency for Instructional Technology and Association for Educational Communications & Technology, Bloomington, Indiana, 2002.
[19] H. Yang, Q. Hu, and L. He, *Learning topic-oriented word embedding for query classification*, in Advances in Knowledge Discovery and Data Mining. PAKDD 2015. Lecture Notes in Computer Science, 2015, pp. 188–198.
[20] W.-N. Zhang, T. Liu, Y. Yang, L. Cao, Y. Zhang, and R. Ji, *A topic clustering approach to finding similar questions from large question and answer archives*, PloS one, 9 (2014), p. e71511.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 31, 2023
Accepted: Oct 5, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 919–939, DOI 10.12694/scpe.v24i4.2467

CONCOLLA – A SMART EMOTION-BASED MUSIC RECOMMENDATION SYSTEM FOR DRIVERS

JIGNA PATEL,[*] ALI ASGAR PADARIA,[†] ARYAN MEHTA,[‡] AARYAN CHOKSHI,[§] JITALI PATEL,[¶] AND RUPAL KAPDI [‖]

Abstract. Music recommender system is an area of information retrieval system that suggests customized music recommendations to users based on their previous preferences and experiences with music. While existing systems often overlook the emotional state of the driver, we propose a hybrid music recommendation system - ConCollA to provide a personalized experience based on user emotions. By incorporating facial expression recognition, ConCollA accurately identifies the driver's emotions using convolution neural network(CNN) model and suggests music tailored to their emotional state. ConCollA combines collaborative filtering, a novel content-based recommendation system named Mood Adjusted Average Similarity (MAAS), and apriori algorithm to generate personalized music recommendations. The performance of ConCollA is assessed using various evaluation parameters. The results show that proposed emotion-aware model outperforms a collaborative-based recommender system.

Key words: Emotion, mood, music, recommendation system, matrix factorization collaborative filtering, personalized content-based recommendation, deep learning, apriori algorithm, associative rule mining

1. Introduction. Music technology is advancing at a rapid pace, providing people with greater access to music than ever before. People primarily use music for self-awareness, fostering social connections, arousal, and mood control [26]. With laptops, tablets, smartphones, and other mobile devices, people can easily listen to music while working, exercising, travelling, or relaxing at home. Music has become an essential part of our daily lives, and it's hard to imagine a future without it. Listening to music while driving is a popular activity for many people. As automotive audio systems have become more sophisticated, drivers can now enjoy a more immersive listening experience. Modern cars often come equipped with state-of-the-art sound systems that offer features such as Bluetooth connectivity, HD radio, and even integrated streaming services. Furthermore, a lot of people stream music through their car's audio system using their smartphones or portable music players. Fig. 1.1 showcases that listening to music in the car[25] is the most popular among all the different places where people listen to music[9], with around 66% of listeners choosing this location. This trend can be attributed to the positive impact music has on our mood, relaxation, and concentration, making it a popular choice for many. With the convenience of linking smartphones to car audio systems, enjoying your favourite music while driving has become effortless. As a result, car music listening has become a common sight on the roads.

However, it is important for drivers to maintain their focus on the road and avoid letting their love for music distract them while driving. It is essential to remember that safe driving should always be the top priority. According to the World Health Organization's (WHO) World Report on Road Traffic Injury Prevention, there are frighteningly many road traffic fatalities globally, with 1.2 million reported each year [1]. Additionally, approximately 50 million people suffer from disabilities or injuries resulting from road accidents each year.

[*]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India, (jignas.patel@nirmauni.ac.in)
[†]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (20bce010@nirmauni.ac.in).
[‡]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (20bce018@nirmauni.ac.in).
[§]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (20bce002@nirmauni.ac.in).
[¶]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (jitali.patel@nirmauni.ac.in), Corresponding Author
[‖]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (rupal.kapdi@nirmauni.ac.in)

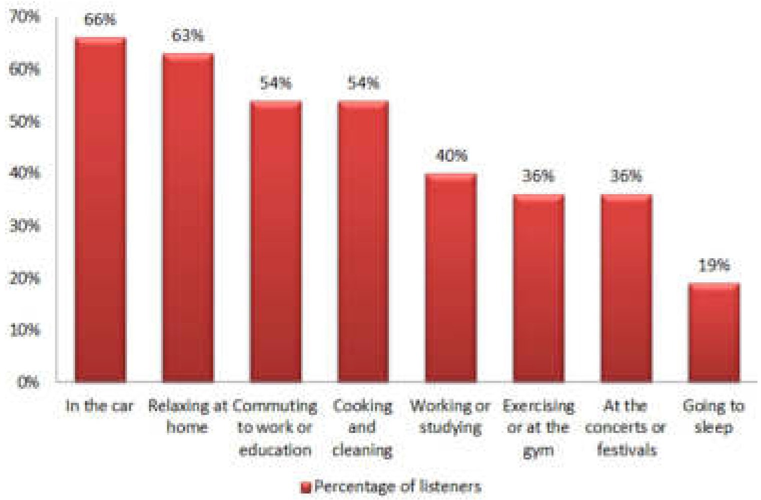


Fig. 1.1: Percentage of listeners in different environments

Analyses of European accident data have revealed that a significant proportion of road accidents, ranging from 10% to 20%, are caused by factors such as reduced driver alertness due to negative emotions or fatigue[17].

Music streaming services like Spotify have revolutionized the way we listen to music by offering vast collections of songs, personalised playlists, and intelligent recommendation systems that suggest new music based on our preferences and behaviour. However, traditional music recommendation systems often fail to capture the emotional and effective qualities of music, which are essential factors in our enjoyment and engagement with music.

In this study, we present ConCollA - an emotional-based music recommendation system for drivers. Our system successfully detects the driver's emotional state using facial expression recognition, and correspondingly recommends music. We have used a number of techniques, such as matrix factorization collaborative filtering[22], personalised content-based recommendations, and apriori algorithms, to develop these personalised suggestions[23]. This all-encompassing strategy makes sure that the music recommendations match the driver's feelings, improving their overall driving experience.

1.1. Significance of research . The conducted research in this study makes noteworthy contributions to various critical domains. Primarily, addressing the issue of road traffic accidents and their consequences assumes paramount significance due to the staggering number of deaths and injuries occurring worldwide each year. Thus, exploring innovative approaches to mitigate factors leading to accidents becomes imperative. This research endeavors to achieve precisely that by focusing on the emotional well-being of drivers through a custom music recommender system. The aim is to enhance their driving experience and potentially alleviate negative emotions or fatigue-related reduced alertness.

Moreover, ConCollA introduces a novel approach to the domain of music recommendation systems. By amalgamating content-based recommendations, association rules, and user mood/emotion information, it strives to enhance the accuracy, relevance, and personalization of music suggestions tailored for drivers. This advancement has the potential to revolutionize music recommendation and enjoyment during driving, thereby leading to more pleasurable journeys and heightened driver satisfaction.

In addition to these advancements, the incorporation of user mood and emotions into the recommendation algorithm constitutes a pivotal aspect of this research. This integration of emotional information aims to create a more comprehensive and tailored music recommendation experience. Notably, this facet of the study contributes to the growing body of research investigating the intersection of human emotions, technology, and personalized systems. By acknowledging user emotions, the recommender system can better cater to individual preferences and emotional states, thus enriching the overall driving experience.

1.2. Problem Definition. Previous studies have indicated the potential of playing relaxing music while driving to uplift the driver's mood and promote relaxation [31]. Music has proven to be a valuable tool in enhancing driving comfort and performance[7]. However, current music recommender systems for driving typically fall into one of three categories: content-based, collaborative filtering, or hybrid approaches [8]. Nevertheless, recent advancements have shown promising results in music recommenders that consider user emotions and contextual factors [4]. Researchers recognize the importance of incorporating user emotions and context to improve the personalization and effectiveness of music recommendations. By considering the driver's emotional state, driving conditions, and preferences, these innovative approaches aim to create more tailored and engaging music experiences during driving.

Despite the progress in utilizing user emotions and context in music recommenders, there is still ample room for further exploration and improvement.

In this study, we propose a novel approach for music recommender system specifically designed to cater to the unique needs of drivers. Our method synergizes the strengths of content-based and collaborative filtering techniques. Additionally, we introduce three novel enhancements to optimize the recommendation process by incorporating the driver's mood or emotions. Firstly, we present an original content-based recommender that utilizes music attributes and driver preferences to deliver personalized recommendations. Secondly, we integrate association rules into the recommendation process, enabling the identification of music tracks that are likely to resonate well with the driver's current selection or mood. Lastly, we give due consideration to user mood and emotion information as an integral factor in the recommendation algorithm, allowing us to tailor music suggestions to align with the driver's emotional state.

Through the integration of these components, ConCollA seeks to enhance the accuracy and relevance of music recommendations for drivers, ultimately elevating their driving experience and potentially mitigating negative emotions or fatigue-related reduced alertness. This research endeavors to advance the field of music recommender systems for drivers by offering an innovative method that takes into account driver mood and emotions, thus contributing to the overall goal of promoting driver well-being and road safety.

2. Related Work. Mood-based music recommendation systems have received a lot of attention in recent years because of their capacity to personalise music playlists depending on users' mood. In the context of driving, where mood and emotions can have a significant impact on the driving experience, such recommendation systems have the potential to improve driver pleasure and safety. We give a review of related work in the field of mood-based music recommendation systems in this part. A brief literature review is tabulated in Table 2.1.

The paper [5] provides an example of an approach for obtaining subjective assessments of the applicability and influence of particular contextual factors on music track ratings. The study shows that this method makes it possible to gather helpful reviews and makes it easier to create a recommender system that is aware of its environment. The predictive model, which extends Matrix Factorization (MF), is evaluated offline, and the results show significant improvements over both non-personalized and conventional personalised predictions based on MF. Additionally, a mobile application has been created to offer customers personalised and context-aware music recommendations while they are in a moving vehicle.

Context-aware music delivery systems make intelligent music choices based on contextual awareness to promote safer driving. The effectiveness and efficiency of music recommendation are two essential components of situation-aware music delivery.Real time context-based music suggestion relies on efficiency, since the music delivery system must react fast to any changes in the environment and play the appropriate music before the sensed context-data is rendered useless. In [17], the authors propose cloud and crowd-sensing based approach for music mood-mapping that can be used in context-aware music recommendation systems. This will help to speed up the music mood-mapping process and significantly increase the effectiveness of situation-aware music delivery systems.

The authors in [3] propose an emotion-aware personalized music recommendation system (EPMRS) that combines deep convolutional neural networks (DCNN) and weighted feature extraction (WFE) to establish the correlation between user data and music. The EPMRS utilizes DCNN to extract latent features from music data and WFE to generate implicit user ratings for music.The study also includes the implementation of Android and iOS apps to collect user feedback, which confirms the system's ability to reflect user preferences based on their emotions.Future work could explore the use of social media data and alternative classification techniques

Table 2.1: Literature Review

References	**Year**	**Objective**	**Methodology**	**Limitations**
Çano *et. al.* [7]	2017	Contextual mood-based music recommender system to have a positive impact on driver's mood.	A multi-point approach featuring three modules; semantic mood model to identify emotions in music, use of cardiovascular data to analyse user's emotional state and driving style recognition through data from dashboard	Few categories of moods to work on, only considers driver's heart-based data.
Deng *et. al.* [8]	2012	Building a recommender system based on the music's acoustic features and employing techniques to create an optimized recommended list of songs.	Creation of a response-arousal-valence measure to describe the emotion of a song. Recommendation done by detecting user's emotion and predicting songs likely to be heard by the user.	Model approach too complex for a low-latency application, high computational cost.
Joshi *et. al.* [16]	2021	Employing LSTM and CNN-based architectures for detection of driver's emotion.	Comparative analysis of four different model architectures, namely LSTM, CNN, LSTM-CNN and CNN-LSTM.	Does not throw light on how the detected emotions are translated into music recommendations.
Singh *et. al.* [28]	2020	Creating a hybrid approach to estimate the likelihood that a song would be heard by the user.	Using various combination of techniques such as Singular-Value Decomposition (SVD) and Factorization Machines (FM), in addition to content-based and collaborative filtering to recommend songs.	The recall shows a drop as compared to other recommendation techniques.
Lin *et. al.* [19]	2022	Personalizing music recommendations based on collaborative filtering and SVD techniques.	Combination of user-based characteristics, matching similar users and music-based filtering, matching songs based on their musical features.	Does not take into account various human-based data, eg. : mood, emotion, mental state, etc.
Iyer *et. al.* [13]	2017	Employs facical recognition techniques to recommend suitable playlists.	Using the Viola-Jones algorithm to detect critical features in the image by creating an integral image, calculating Haar-like features and using Adaboost to train the classifier	Playlist recommendation only limited to facial emotions. Does not take into account various collaborative filtering methods.
Henry *et. al.* [12]	2022	Implementation of Apriori algorithm to extract patterns in music playlists that can aid in better music recommendation.	Data preprocessing to reduce 20,000 data rows into 532, used Apriori algorithm to find associative rules between songs.	Lack of quantitative results which proscribe the effectiveness of the proposed method.
Proposed approach	2023	Recommending music to drivers based on a novel technique that combines content-based filtering in conjunction with facial emotion recognition, collaborative filtering and associative rule mining to suggest songs based on the emotion of the driver.	Implemented a fusion of CNN-based emotion recognition model and content-based filtering model that suggests music based on emotional features. The addition of a collaborative filtering model refines the recommendation allowing the model to suggest the most appropriate songs to the driver based on their mood.	-

to enhance the system's performance.

A. Ferraro *et al.* [10] investigates the impact of artist and style exposure bias on collaborative filtering-based music recommendations. The authors analyze a large-scale dataset of music listening histories and show that users tend to listen to popular artists and genres, which can lead to exposure bias in collaborative filtering-based recommendations. The authors have thought about how the popularity bias affects Matrix Factorization-based collaborative filtering recommendations and they have proposed an algorithm which is boosting the exposure of more well-known musical genres while decreasing the exposure in the long tail.

J. Singh [28] has described a hybrid music recommendation system with the goal of determining the likelihood that a user will listen to the song repeatedly after the first apparent listening experience begins in the time frame. This system uses both collaborative and content-based filtering algorithms. Additionally, methods including collaborative filtering, singular value decomposition, and factorization machines (FM) are employed. In order to increase the overall accuracy of the recommendation system utilising deep neural networks, the author has also hybridised FM and SVD models.

Joshi *et al.* [16] use Long Short-Term Memory (LSTM), Convolution Neural network (CNN), and LSTM-CNN architecture in a music recommendation system that uses deep learning techniques to detect emotions from facial expressions and recommend music that matches the detected emotion. The paper demonstrates the potential of using deep learning techniques for emotion-based music recommendation and highlights the importance of considering contextual information, such as facial expressions, in recommendation systems[6].

In [19] Xiaoyu Lin presents a comprehensive music recommendation system that utilizes Collaborative Filtering (User CF and Item CF) and Singular Value Decomposition (SVD) algorithms. Using the Jaccard index, the Item CF algorithm calculates the similarity between a user's songs and all unique songs. The User CF algorithm measures the similarity between users based on their song preferences. The SVD algorithm decomposes rating matrices to extract latent factors for personalized recommendations. Using the SVD and User CF to predict certain users tastes in music the accuracy is 80% and 88%.

3. Proposed Research Work. We propose our system in detail in the following subsections. These subsections would include a system architecture, a brief methodology of each individual component and finally, the amalgamation of these components into a unified framework for music song recommendation.

3.1. Dataset Description. The following section contains detailed description on the datasets chosen in accordance to the scope of the proposed method and tweaked to fit specfic requirements aimed by us during the inception of the methodology. We have selected the Facial Emotion Recognition dataset to learn an emotion-detecting model, and Spotify's Million Playlist Dataset to implement collaborative and content-based filtering models. Following subsections refer to them in detail.

3.1.1. Facial Emotion Dataset. The Facial Emotion Recognition - 2013 : (FER-2013) dataset [2] was chosen for our work, keeping in mind the requirements posed in detecting drivers' emotions and the effectiveness in capturing the said features. A subset of this dataset was taken that would better augment the application of detecting driver's emotions. Out of the seven put forward in the dataset, we selected four emotions that aligned with our proposed methodology; viz. happy, sad, angry and neutral. These emotions were identified as useful in detecting the mental state of the driver, which would then be useful in subsequently suggesting relevant songs to the driver. Some of the photos of the dataset are shown in Figure 3.1.

Our subset of the dataset consists of a total of 21,004 images spanning four categories, which are split into training and test datasets. The training dataset consists of 14,702 images and the test dataset consists of 6,302 images. The images present in the dataset are gray-scaled images of dimensions 48 x 48 pixels each. The gray-scaled images help reduce the learning load on the model by eliminating irrelevant data; eg. color, from the input. These images contain human faces that are mostly centred in the image, which helps in the identification of facial emotions easier as compared to faces that contain facial features at the corners or at some other position in the image.

3.1.2. Spotify's Million Playlist Dataset. The recommendation engine in this study utilizes the Spotify Million Playlist Dataset (MPD) to generate user profiles. This dataset encompasses one million playlists with over two million distinct tracks from nearly 300,000 artists. Each playlist includes details such as the playlist title, track list with corresponding track IDs and metadata, as well as additional metadata fields like last edit

Fig. 3.1: Some images from the Facial Emotion Recognition Dataset

time and the number of playlist edits. In this context, each playlist is treated as a user, and the songs within the playlist are considered to be the preferences of that hypothetical user. To simulate user preferences, certain songs have been augmented to the list of liked songs for these imaginary users, while other values have been assigned to indicate user dislike for specific songs. Liked songs are denoted as 1, while disliked songs are represented as -1. The resulting matrix is sparse, with the majority of values being 0, which signifies songs the user has not listened to or, in essence, potential recommendations. To facilitate efficient processing, a random subset of playlists and songs has been extracted from the dataset for our experimental analysis. This subset remains representative of the full dataset, capturing its crucial attributes for our analysis without compromising efficiency.

3.2. System Architecture. ConCollA music recommender system incorporates a multi-faceted approach to enhance the accuracy and relevance of song recommendations based on the driver's mood. The system architecture, as depicted in Figure 3.2 consists of several components, including mood recognition using a Convolutional Neural Network (CNN) model, collaborative-based recommendations, content-based recommendations, and association rule mining.

The initial step in the system is mood recognition, which relies on an in-car camera capturing images of the driver. These images are then fed into a CNN model trained to recognize patterns in photos corresponding to different emotions. The CNN model employed in this architecture consists of four convolutional layers and two dense layers. The system determines the driver's mood through the model, which serves as a crucial input for generating personalized song recommendations. Following that, the system proceeds to provide recommendations through a novel approach using a combination of collaborative filtering, content-based filtering along with associative rule mining. The collaborative-based recommendation component employs matrix factorization techniques to uncover underlying emotional factors that influence music preferences. By representing users and music items in emotional dimensions, the system generates personalized recommendations aligned with the user's emotional state and inclinations. This approach utilizes an anticipated rating matrix to deliver tailored suggestions based on the user's emotional profile and the content of the music.

Within the system architecture, we employ a content-based recommendation approach that leverages the acoustic features of music. Content-based recommenders analyze the intrinsic characteristics of songs, such as melody, rhythm, and instrumentation, to identify similarities between different tracks. In our case, we have developed a novel method called the Mood-Adjusted Average Similarity (MAAS) Algorithm.

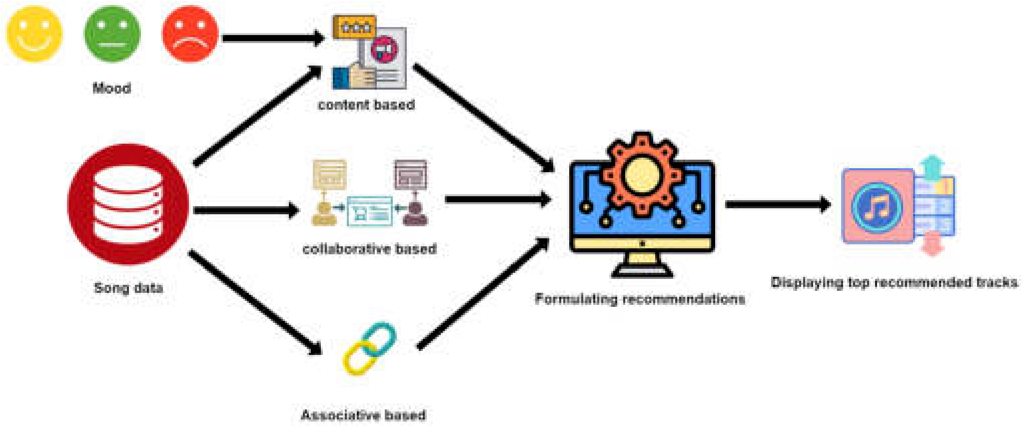


Fig. 3.2: System Architecture

The MAAS Algorithm takes into account the user's listening history and their emotional state. For each song that the user has not yet heard, we calculate its similarity to the songs they have previously listened to. This is achieved by computing the average similarity score between the target song and the songs heard by the user using acoustic feature analysis. The MAAS Algorithm goes a step further by involving mood biases into the picture, ensuring that the similarity score is adjusted according to the user's current emotional state. The resulting similarity scores range between -1 and 1, with higher values indicating a stronger match between the target song and the user's preferences. By considering both acoustic features and mood biases, the MAAS Algorithm enhances the accuracy and relevance of song recommendations, tailored to each individual driver's tastes and emotional context.

ConCollA also harnesses the power of association rule mining to further enhance the music recommendation process. While collaborative filtering focuses on finding recommendations from similar users, association rule mining allows us to uncover strong associations among songs on a larger scale, considering the listening patterns of a wide group of users. By applying association rule mining techniques, we can identify meaningful associations between songs that may not have been apparent through traditional collaborative filtering methods. These associations provide valuable insights into the preferences and listening behaviors of users. For example, if a particular song is frequently listened with another song, the association rule mining approach can identify this strong association. Consequently, the algorithm can deduce the presence of the consequent (the new song not yet heard) based on the presence of the antecedent (the song heard by the user), helping the former get additional preference due to said correlation.

This broader perspective enables our recommendation system to capture hidden relationships and co-occurrences among songs, enhancing the accuracy and relevance of the recommendations. By considering associations beyond the scope of individual user preferences, we can offer diverse and serendipitous recommendations that may align with the user's musical taste but also introduce new and exciting tracks.

By aggregating the results from these recommendation systems, the system generates a final recommendation rating. Based on the rating of each song, the system plays the recommended tracks to the driver in descending order, enhancing their music-listening experience during their journey.

The modular design and integration of these recommendation approaches within the system architecture ensure a comprehensive and personalized music recommendation process, taking into account user emotions, preferences, and associations to provide a tailored and enjoyable music experience.

4. Proposed Methodology. We introduce our proposed methodology for creating a novel recommender system that significantly improves the recommendation process by incorporating driver mood recognition in this section. The methodology comprises two key components: Driver Mood Recognition and Recommender Engine. The primary objective of the first component is to accurately capture the emotional state of the

Table 4.1: Layer information for the Emotion recognition model

Layers	**Output Shape**	**Parameters**
InputLayer	48,48,3	0
Conv2D	48,48,64	1792
BatchNormalization	48,48,64	256
Conv2D	48,48,64	36928
BatchNormalization	48,48,64	256
MaxPooling2D	24,24,64	0
Dropout	24,24,64	0
Conv2D	24,24,128	73856
BatchNormalization	24,24,128	512
Conv2D	24,24,128	147584
BatchNormalization	24,24,128	512
Conv2D	24,24,128	147584
MaxPooling2D	12,12,128	0
Dropout	12,12,128	0
Conv2D	12,12,256	295168
BatchNormalization	12,12,256	1024
Conv2D	12,12,256	590080
BatchNormalization	12,12,256	1024
Conv2D	12,12,256	590080
BatchNormalization	12,12,256	1024
Conv2D	12,12,256	590080
BatchNormalization	12,12,256	1024
MaxPooling2D	6,6,256	0
Dropout	6,6,256	0
Flatten	9216,1	0
Dense	4,1	36868

driver through the implementation of diverse techniques and sensors. Subsequently, this invaluable emotional information will be seamlessly integrated into the recommender engine, enabling the generation of personalized recommendations that precisely align with the driver's mood and preferences.

The integration of driver mood recognition with the recommender engine is crucial for creating a more engaging and satisfying user experience. By understanding the driver's mood, the recommender system can adapt its recommendations to match the driver's emotional state, thereby increasing the likelihood of a positive response and user satisfaction.

4.1. Driver Mood Recognition. The driver mood detection algorithm is a crucial element of the recommendation engine. It provides key information to the proposed recommender engine by infusing the knowledge of the driver's mood into the recommender, thereby better augmenting the data being used by the system. Providing this crucial insight into the moods of the driver has a significant importance on the selection of songs for the driver; improving the mental conditions and allowing the driver to focus on the roads. The architecture and description of the model follows.

4.1.1. Architecture. The mood recognition model is a convolutional model that is split into three blocks. Table 4.1 shows a detailed view of the model where all the added layers along with respective output shape and number of parameter is mentioned. The visualization of the network for emotion classification is depicted in Figure 4.1.

The input layer (in green) takes an input image of dimensions $48 \times 48 \times 3$. The first block has two convolutional layers (in orange) with 64 filters in each layer. These features are normalized using a Batch Normalization(BN) layer (shown in blue), followed by a MaxPooling Layer (in purple). Their resultant output is then sent through a dropout filter (in red) which provides a cap on the extent of learning, thus allowing the

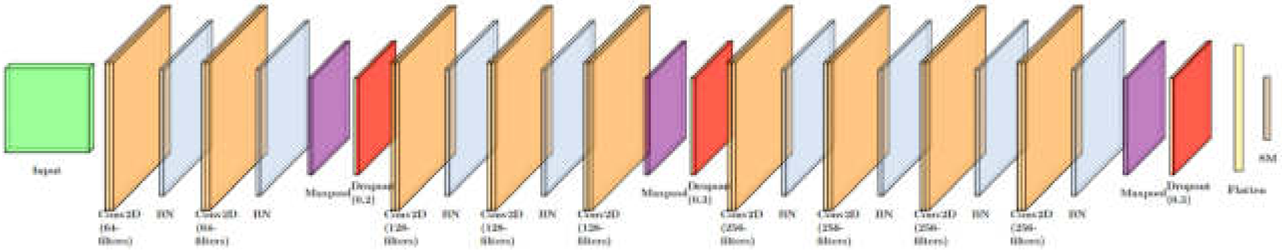


Fig. 4.1: CNN Architecture for Emotion recognition

model to better generalize on the training dataset.

The second block has the same architecture as the first block, namely the convolutional layer; but with 128 filters, a Batch Normalization layer and then a dropout layer. The rate of dropout is increased here to ensure only certain features from the previous block are passed on to the next block.

The final block has three convolutional layers of 256 filters each, along with a Batch Normalization layer after each convolutional layer and a dropout layer at the end. The dropout layer is then sent into a Flatten layer (in yellow) and fed into a fully-connected layer (in brown) containing four neurons as the final output. These neurons are activated by the softmax (SM) activation function that gives a likelihood score to each class.

4.2. Recommender Engine. The recommender engine is a pivotal component of ConCollA recommender system, responsible for generating personalized recommendations based on the driver's mood and preferences. This subsection outlines the methodologies and techniques employed within the recommender engine, including collaborative filtering, content-based recommender, and association rule mining.

4.2.1. Collaborative Filtering. Collaborative filtering is a noteworthy approach in recommendation systems for generating personalized recommendations by leveraging users' past behavior or preferences. This approach operates under the assumption that individuals with similar preferences in the past are likely to make similar choices in the future. Consequently, collaborative filtering recommends items that are popular among users with comparable preferences. There are many variants of collaborative filtering such as user-based, item-based, and model-based.

User-based collaborative filtering selects similar users based on past behavior or preferences and recommends items popular among them but not interacted with by the current user, aligning with collective preferences. Item-based collaborative filtering identifies similar items based on previous user interactions and suggests items similar to those already interacted with, leveraging item relationships to recommend related items.

Model-based collaborative filtering, such as Matrix Factorization, is another approach within collaborative filtering. This technique employs a mathematical model to uncover latent factors that capture the underlying user-item interactions. Matrix Factorization decomposes the user-item interaction matrix into lower-dimensional latent factor matrices. By utilizing these latent factors, personalized recommendations can be generated. Matrix Factorization aims to capture the hidden patterns and preferences within the user-item interactions, allowing for more accurate and effective recommendations.

Matrix factorization is a popular collaborative filtering method used to identify latent variables within user-item interactions. In an emotion-based music recommendation system, matrix factorization can uncover emotional content in music, enabling personalized recommendations based on the user's emotions and preferences. The Spotify Million Playlist Dataset provides this user-item matrix in the form of a user-track matrix.

The process involves breaking down the user-item interactions matrix into two lower-rank matrices, representing latent attributes of users and items. The user matrix indicates preferences across latent factors, while the item or track matrix represents the features of each item. By multiplying these matrices, a projected rating matrix is obtained, forming the basis for generating user-specific recommendations [29].

We utilized Singular Value Decomposition (SVD) [Algorithm 2], a widely adopted matrix factorization technique in collaborative filtering for ConCollA. SVD was utilized to decompose the user-item interaction matrix into three constituent matrices: U, Σ, and V. The user matrix U captures the preferences of each user, while the diagonal matrix Σ reflects the significance of the singular values obtained during the decomposition. Additionally, the item matrix V encapsulates the characteristics of each item. By leveraging the insights provided by the singular values derived from SVD, we identified crucial latent factors that play a significant role in explaining user-item interactions and generating accurate recommendations[20].

The process of calculating the matrices U, Σ, and V in SVD involves finding the eigenvectors and eigenvalues of the matrix M. The eigenvectors form the columns of U, while the eigenvalues form the diagonal elements of Σ. The columns of V are derived from the eigenvectors of the transposed matrix of M.

$$M = U\Sigma V^T \tag{4.1}$$

where:
- M is the user-item interaction matrix,
- U is the matrix of left singular vectors (representing users),
- Σ is the diagonal matrix of singular values,
- V is the matrix of right singular vectors (representing items), and
- V^T denotes the transpose of the matrix V.

The diagonal matrix Σ consists of singular values $(\sigma_1, \sigma_2, \ldots, \sigma_k)$ along its diagonal, where k is the rank of the matrix.

To calculate the matrices U and V, one can use eigenvalue decomposition or singular value decomposition algorithms. These algorithms provide the eigenvalues and eigenvectors, which can be used to construct the matrices U and V as follows:

$$U = \begin{bmatrix} u_1 & u_2 & \ldots & u_m \end{bmatrix} \tag{4.2}$$

$$V = \begin{bmatrix} v_1 & v_2 & \ldots & v_n \end{bmatrix} \tag{4.3}$$

where u_i and v_i represent the eigenvectors associated with the eigenvalues σ_i of Σ.

Using these matrices, the projected rating matrix $\hat{M}$ can be obtained as:

$$\hat{M} = U\Sigma V^T \tag{4.4}$$

The utilization of SVD in ConCollA enabled us to effectively extract latent factors from the user-item interaction matrix and leverage them to generate personalized recommendations. Its application in our recommender is as explained below.

Algorithm 2 Singular Value Decomposition (SVD)

Require: $m \times n$ matrix M

Ensure: Orthogonal matrices U and V, and diagonal matrix Σ such that $M = U\Sigma V^T$

1: Calculate $M^T M$, which is an $n \times n$ matrix.
2: Find the eigenvalues and eigenvectors of $M^T M$.
3: Sort the eigenvalues in descending order and form a diagonal matrix Σ, where the diagonal elements are the square roots of the eigenvalues in descending order.
4: Normalize the eigenvectors from the previous step to form an $n \times n$ orthogonal matrix V, where the columns are the normalized eigenvectors.
5: Calculate MV, which gives an $m \times n$ matrix.
6: Normalize each column of the matrix obtained in the previous step to obtain an $m \times n$ orthogonal matrix U.
7: **Return** the matrices U, Σ, and V^T.

Table 4.2: Musical Features

ID	Acousticness	Danceability	Energy	Key	Liveness	Loudness	Mode	Speechiness	Tempo
1	0.75	0.65	0.82	G	0.60	-5.2	Major	0.32	120
2	0.32	0.78	0.90	C	0.45	-3.8	Minor	0.15	130
3	0.91	0.40	0.67	A	0.70	-8.1	Major	0.22	105
4	0.60	0.85	0.76	E	0.55	-6.3	Minor	0.40	140

4.2.2. Content Based Recommender. The prevalent approach in modern music services is collaborative filtering, which generally yields recommendations of acceptable quality. However, this method exhibits limitations when faced with new users or emerging music, as the scarcity of user interaction information hampers its effectiveness. This scenario is referred to as the "cold start problem" in academic research [21]. To mitigate this challenge, the inclusion of similar sounding music becomes crucial. For instance, to acquaint listeners with lesser-known artists, recommendations may encompass the artist's works alongside similar sounding songs by more prominent artists.

Content-based recommenders present users with objects that exhibit similarity to items that have previously captured their interest[14]. Unlike collaborative filtering, which assesses object similarity based on user actions and similarities among users, content filtering evaluates similarity solely on the intrinsic characteristics of the objects themselves. Content filtering relies exclusively on objective attributes of the objects and remains independent of subjective user ratings.

Typically, these attributes are represented using a structured database table [24]. In the context of our study, we specifically require the inclusion of musical acoustic features. The acoustic features utilized in our analysis comprise acousticness, danceability, energy, instrumentalness, key, liveness, loudness, mode, speechiness, and tempo. Table 4.2 provides a subset of records showcasing a few musical features from the ones mentioned above, where each row corresponds to a distinct song and the columns denote the identifiers for the song as well as the aforementioned musical attributes. Within each record, a value is assigned to each attribute. To differentiate songs with identical names and facilitate the retrieval of associated attributes, a unique identifier known as ID is employed as a key. To generate these records, we utilize the capabilities of Spotify API. The user-track matrix used in Collaborative filtering can be used to fetch the song IDs of all the songs present in the playlist. These song IDs are provided further to Spotify API which reverts back with the acoustic features discussed above.

To facilitate the recommendation of songs, it is essential to establish a measure of similarity between these items. One commonly employed method, as elucidated by [24], is the nearest neighbor method. The choice of similarity function in the nearest neighbor algorithm depends on the data type under consideration. In the case of structured data, a frequently used metric is the Euclidean distance, whereas the vector space model often employs the cosine similarity measure.

The Euclidean distance function treats a small value for the same feature in two examples as equivalent to a large value for that feature in both examples. Conversely, the cosine similarity function does not assign a large value if corresponding features of two examples possess small values. Given that we treat the musical features as vectors in our task, the cosine similarity metric is employed. However, we propose a novel approach distinct from the native top-k similarity recommendation to leverage this cosine similarity and generate more effective recommendations.

In order to provide recommendations for unheard songs, it is necessary to partition the dataset into two distinct sets: Set P, consisting of songs the user has listened to, and Set Q, encompassing songs the user has not yet heard, for which predictions are required. To make recommendations, a crucial step involves calculating the similarity between each song in Set Q and every item in Set P. The recommendation score for a song q belonging to Set Q is then determined as the average of the similarities between q and each song p belonging to Set P.

This approach ensures that the recommendation score for an unheard song takes into account its similarity to the songs the user has already listened to. By calculating the similarity between q and each item in Set P, the recommendation score reflects the collective resemblance of q to the songs in the user's listening history. The average of these similarities provides an overall measure of recommendation, enabling the system to suggest songs in Set Q based on their similarity to the user's previously enjoyed songs.

Algorithm 3 Mood-Adjusted Average Similarity (MAAS)

Input: $P \leftarrow$ Set of Tracks User has listened to
$Q \leftarrow$ Set of Tracks to Recommend to User
$MoodCluster \leftarrow$ Set of Tracks corresponding to User's mood
Output: MAAS score

procedure SIMILARITY_CALCULATOR($Set\ P,\ Set\ Q$)
  for $q \in Q$ **do**
    $MAAS\_score[q] \leftarrow 0$ ▷ Initialize MAAS score for track q
    $q_{acc} \leftarrow$ get_acoustics(q) ▷ Extract acoustic features for track q
    for $p \in P$ **do**
      $p_{acc} \leftarrow$ get_acoustics(p) ▷ Extract acoustic features for track p
      $MAAS\_score[q] \leftarrow MAAS\_score[q] +$ cosine_similarity(p_{acc}, q_{acc}) ▷ Calculate cosine similarity
    end for
    $MAAS_score[q] \leftarrow MAAS_score[q]/$length($P$) ▷ Normalize MAAS score
    if $q \notin$ MoodCluster **then**
      $MAAS\_score[q] \leftarrow -MAAS\_score[q]$ ▷ Adjust MAAS score based on mood
    end if
  end for
end procedure

The rationale behind this methodology is to leverage the known preferences of the user, as represented by Set P, to estimate the user's potential interest in the unheard songs within Set Q. By considering the similarities with the familiar songs, the recommendation system can make informed predictions about the user's taste and provide relevant recommendations accordingly. The averaging of similarity scores allows for a comprehensive assessment of the compatibility between the unheard songs and the user's established preferences.

The previously mentioned similarity score, while significant, does not encompass the user's mood consideration. To incorporate the user's mood into the averaged recommendation score, an additional step is undertaken. This process involves iterating through each song and adjusting the recommendation score for songs that do not align with the user's mood. The score that we receive now is called MAAS score, as mentioned in the architecure section. This score is calculated through [Algorithm 3]. By negating the recommendation score, the resulting score falls within the range of -1 to 0, considering that the original recommendation score ranges from 0 to 1. When combining scores obtained from different methods, this negated score contributes to reducing the effective score for songs that do not align with the user's mood. This mechanism accounts for the user's mood preferences which contributes to a more personalized and mood-aligned recommendation experience.

4.2.3. Association Rule Mining. Association rule mining is a data mining technique used to uncover meaningful associations or patterns in large datasets. In the context of music recommendation systems, association rule mining aims to discover relationships between music items based on their co-occurrence patterns. By analyzing user listening behavior, it can identify frequently occurring combinations of songs or artist preferences. The goal is to extract association rules that capture the relationships between music items, allowing us to understand the preferences and associations of different users within the music domain. These rules can reveal interesting connections, such as songs that tend to be listened together or a group of people frequently listening to the same artist.

In a music recommendation system, association rule mining can improve the diversity and personalization of recommendations[15]. The system can make related or complementary song suggestions by taking into consideration these discovered relationships, enhancing the user's ability to find new music. Additionally, it offers personalised playlists catered to each listener's preferences, suggests music based on their likes, and aids in identifying up-and-coming artists.

Association rule mining's use in music recommendation systems creates new opportunities for offering interesting and relevant musical suggestions. The algorithm may generate more individualised and varied recommendations that are tailored to each user's own tastes and preferences by understanding the relationships between musical pieces.

Association rule mining techniques, such as the Apriori algorithm [Algorithm 4], are commonly employed in music recommendation systems to discover meaningful relationships among songs. The Apriori algorithm, firstly conceptualized by Agrawal et al., is a popular method for extracting frequent itemsets and deriving association rules.

The Apriori algorithm follows a level-wise search strategy to efficiently identify frequent itemsets. It starts with frequent itemsets of length 1, also known as "singleton" itemsets, and progressively generates candidate itemsets of increasing lengths. The algorithm prunes the search space by eliminating itemsets that cannot be frequent based on the infrequency of their subsets, leveraging the apriori property.

In the context of music recommendation, we use the Apriori algorithm to transactional data representing user playlists. This transactional data is provided by Spotify Million Playlist Dataset in form of user-track matrix, which is the same used by the Collaborative recommender. By identifying frequent itemsets, the algorithm uncovers associations between songs or artists. These associations help enhance the accuracy and diversity of music recommendations by suggesting related or complementary music items to users [30].

Algorithm 4 Apriori Algorithm

$Input \leftarrow$ Load User-Playlist dataset
$Output \leftarrow$ Large Itemsets
$L_1 =$ large 1-itemsets ▷ Initialize with 1-itemsets
for $k = 0; L_{k-1} \neq \emptyset; k++$ **do** ▷ Iterate until no more frequent itemsets
    $C_k = \text{apriori\_gen}(L_{k-1})$ ▷ Generate candidate itemsets
    for $\forall$transaction $t \in D$ **do** ▷ Scan transactions
        $C_t = \text{subset}(C_k, t)$ ▷ Find candidate subsets in transaction
        for candidate $c \in C_t$ **do** ▷ Count occurrences of candidates
            $c.\text{count}{+}{+}$
        end for
        $L_k = \{c \in C_k \mid c.\text{count} \geq \text{min\_sup}\}$ ▷ Keep frequent itemsets
    end for
end for
Answer $= U_k L_k$ ▷ Final set of large itemsets

Once the frequent itemsets have been identified using the Apriori algorithm, the next step is to derive association rules from these itemsets. Association rules capture the relationships and dependencies among items based on their co-occurrence patterns. In the context of ConCollA, these association rules provide valuable insights into the preferences and associations within the music domain. Using association rule mining, it is possible to find patterns in massive amounts of data. In order to enable analysts to meaningfully connect seemingly unrelated pieces of data, it helps to discover co-occurring events or objects that might not be immediately obvious. Here it helps us to find the combination of occurrences of various tracks together and gives us a relationship stating if these n tracks are in a playlist then these m tracks can also be recommended [12].

To generate association rules, the frequent itemsets serve as the basis. Each frequent itemset represents a set of items that occur together frequently in user playlists or listening histories. From these itemsets, association rules are derived by considering various combinations of the items within each frequent itemset.

The process of generating association rules involves defining thresholds or measures of interestingness, such as support, confidence, and lift. Support measures the frequency of occurrence of an itemset in the dataset, confidence quantifies the strength of the rule by measuring the conditional probability, and lift determines the degree of association between items [18].

The culmination of the associative rule mining phase marks the completion of all individual components within the recommendation system. The upcoming crucial step involves the integration of results derived from three distinct techniques: content-based filtering, collaborative filtering, and association rule mining. The combined implementation of these three approaches facilitates the recommendation process, enabling the system to offer personalized music recommendations based on users' moods. The integration of these methodologies ensures a comprehensive and refined approach to music recommendation. By considering users' preferences, past behavior, and the discovered associations among music items, ConCollA aims to deliver accurate and personalized music suggestions. In the subsequent section, we will discuss upon the details of this integration and examine its impact on the recommendation engine. This analysis will reveal how the ConCollA recommender system intelligently leverages content-based, collaborative, and association rule mining techniques, resulting in a highly effective and user-centric music recommendation system.

4.3. Integration. In order to integrate the individual components of ConCollA, viz. collaborative filtering, content-based filtering, and association rule mining, we employ a weighted aggregation approach that combines scores from collaborative and content-based filtering techniques by aggregating the scores after normalizing them with the respective weights. The following formula represents the weighted aggregation of content-based and collaborative filtering techniques:

$$\mathrm{RS} = \mathrm{W_{CB}} \times \mathrm{Score_{CB}} + \mathrm{W_{CO}} \times \mathrm{Score_{CO}} \tag{4.5}$$

where RS is the normalized recommendation score, $\mathrm{W_{CB}}$ is the weight assigned to content-based filtering model, $\mathrm{Score_{CB}}$ is the score given by the content-based model, $\mathrm{W_{CO}}$ is the weight assigned to collaborative filtering model, $\mathrm{Score_{CO}}$ is the score given by the collaborative model.

In ConCollA recommender system, association rule mining is incorporated to further enhance the recommendation process. By analyzing the user's listening history, ConCollA checks the presence of antecedent items (X) from association rules. If the antecedent is found, we add the normalized confidence score of the association rule to the recommendation score of the consequent items (Y). This allows ConCollA to leverage the strength of the association rules while considering the user's preferences. To ensure a balanced influence, ConCollA normalizes the confidence score within the range of 0 to 0.1, preventing it from dominating the overall recommendation results. The combination of content-based filtering, collaborative filtering, and association rule mining enables ConCollA to provide accurate and personalized music recommendations to users. So if a rule holds true, the above given Recommendation Score formula is changed as follows:

$$\mathrm{RS} = \mathrm{W_{CB}} \times \mathrm{Score_{CB}} + \mathrm{W_{CO}} \times \mathrm{Score_{CO}} + \mathrm{W_{A}} \times \mathrm{Score_{ARC}} \tag{4.6}$$

where $\mathrm{W_A}$ is the weight factored to normalize the associative score, $\mathrm{Score_{ARC}}$ is the associative rule confidence for the antecedent.

The weights assigned to different recommendation components play a crucial role in providing personalized recommendations to users. In ConCollA, we recognize the diversity of user preferences and configure individual weights for each user to tailor the recommendations accordingly. For example, while collaborative filtering identifies similar users, we acknowledge that some users may have a preference for songs similar to their own, which are provided by the content-based recommendation component. Hence, such users receive higher weights on the content-based recommendation, emphasizing the importance of recommendations based on content similarity. Conversely, users who rely more on collaborative filtering receive higher weights for that component.

The determination of these weights can be framed as a linear regression problem. We aim to find the optimal weights that minimize the difference between the predicted recommendation scores and the user's actual preferences. This approach allows us to model the relationship between the recommendation components and the user's feedback, facilitating the fine-tuning of the weights to align with their preferences.

To optimize the weights for each user, ConCollA employs the gradient descent algorithm. Gradient descent is an iterative optimization technique that adjusts the weights based on the calculated gradients of the loss function. By updating the weights in the direction that minimizes the loss in each iteration, we can converge towards

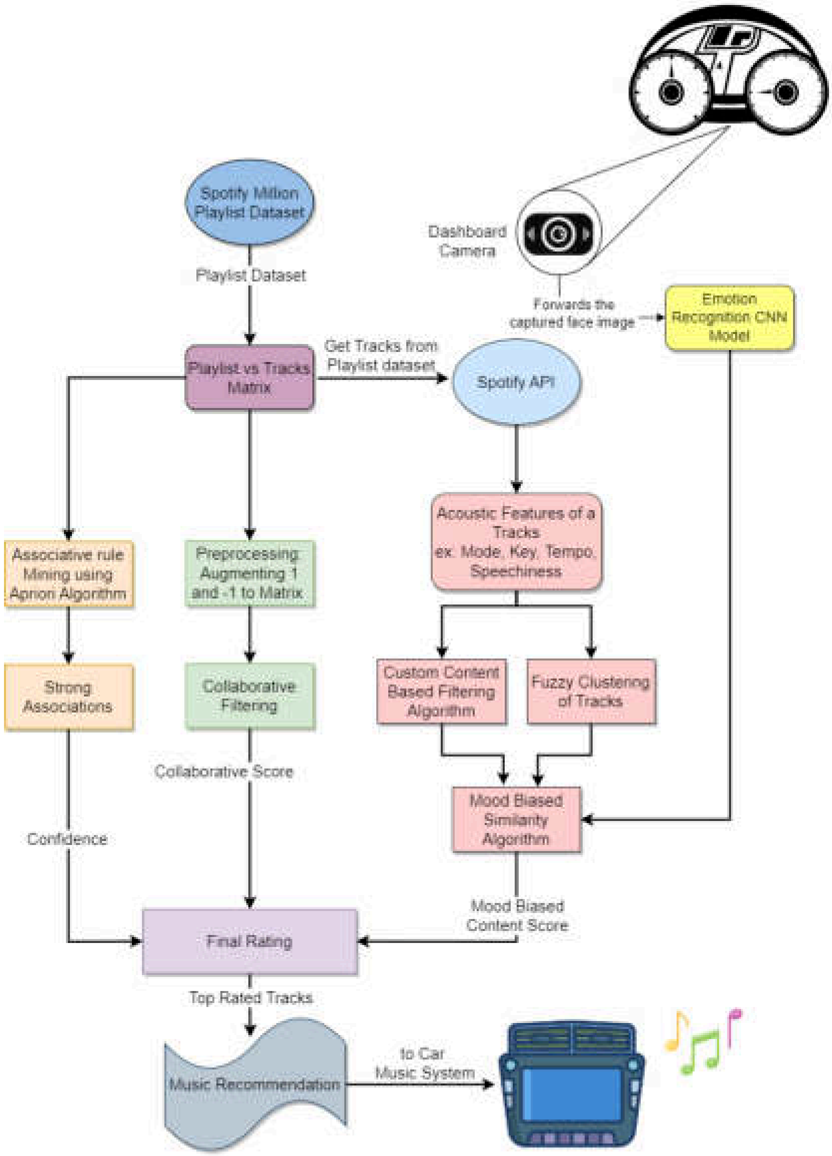


Fig. 4.2: Flow diagram of the model

the optimal set of weights for each user, maximizing the accuracy and effectiveness of the recommendations. The generic formula for gradient descent in our music recommendation context can be written as:

$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} \tag{4.7}$$

In this formula, Weight_j represents the weight for a specific recommendation component, such as content-based, collaborative filtering, or association rules, that needs to be updated. α is the learning rate, controlling the step size of the weight update. $\frac{\partial J}{\partial \text{Weight}_j}$ denotes the partial derivative of the cost function J with respect to the weight Weight_j.

ConCollA aims to optimize these weights for each user to provide personalized recommendations. The weights are initially assigned arbitrary values and are gradually updated using the gradient descent algorithm. This iterative optimization process allows us to adjust the relative importance of each recommendation component based on individual user preferences and feedback.

To optimize the weights, ConCollA defines a cost function that quantifies the discrepancy between the predicted recommendation scores and the user's actual preferences. This cost function serves as a measure of the accuracy or quality of the recommendations. The gradient descent algorithm calculates the gradients of the cost function with respect to each weight and updates the weights in the direction that minimizes the cost function.

The mean squared error (MSE) is the cost function that is most frequently employed in linear regression problems. The MSE calculates the average squared difference between the training data's actual values and the values that were predicted. It measures the linear regression model's overall goodness of fit.

By updating the weights using gradient descent after each iteration, we refine the recommendation process to better align with the user's preferences. The learning rate α controls the step size of the weight update, ensuring a balance between convergence speed and stability. ConCollA continues iterating until the weights converge to a point where the recommendation scores align closely with the user's preferences.

It is important to note that the weights are not fixed and are constantly updated as users interact with the recommended songs. This adaptive approach ensures that the recommendations remain aligned with the user's evolving preferences and feedback. By continuously optimizing the weights through gradient descent, our system adapts to changes in user preferences, providing more accurate and personalized recommendations over time. The aforementioned methodology has been explained in Fig. 4.2. Our proposed system can be used to generate music recommendations for drivers. It monitors the driver's emotions using a dashboard camera positioned in front of the driver and the collected emotional data from the camera is combined with past listening preferences to generate music recommendations. The suggested song is played automatically in the car and improves the listener experience.

5. Experimental Analysis. This section is divided into two subsections. The first subsection discusses the Evaluation Measures used for evaluating the proposed recommender system. The subsequent section then discusses the results achieved on the evaluation measures discussed in the former subsection.

5.1. Evaluation Parameters. For evaluation of the recommender, two metrics are provided: Root Mean Square Error (RMSE) and a novel metric, Rel-Sim.

Root Mean Square Error (RMSE) calculates the square root of the average of the squared differences between the predicted values and the actual (observed) values. In other words, it measures how far, on average, the predictions are from the actual values. A lower RMSE indicates better model performance, as it signifies that the predictions are closer to the true values [27].

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{5.1}$$

Rel-Sim, which stands for "Relative Similarity" is a novel metric proposed for evaluating the ConCollA system, focusing on the relationships between the ground truth and the top recommended items. While conventional evaluation metrics like RMSE may yield low scores for the ground truth, they might not reflect its true relevance due to the influence of large datasets, resulting in lower MAP scores. To address this challenge, ConCollA evaluation metrics introduce Rel-Sim, which measures the similarity between the top-N recommended items and the ground truth using acoustic cosine similarity. Each item in the top-N recommendations, is compared with the ground truth and its acoustic cosine similarity is calculated. We then identify the ground truth item with the maximum similarity and multiply this value by the predicted score of the ground truth. This process is repeated for all top-N items, and the resulting values are averaged to obtain the final Rel-Sim score. A higher Rel-Sim score signifies that the top-N recommended items are highly similar to the ground truth, even if the ground truth item itself was not included in the top-N recommendations. Overall, Rel-Sim provides valuable insights into the recommender system's performance, particularly when the ground truth items are not among the top recommendations.

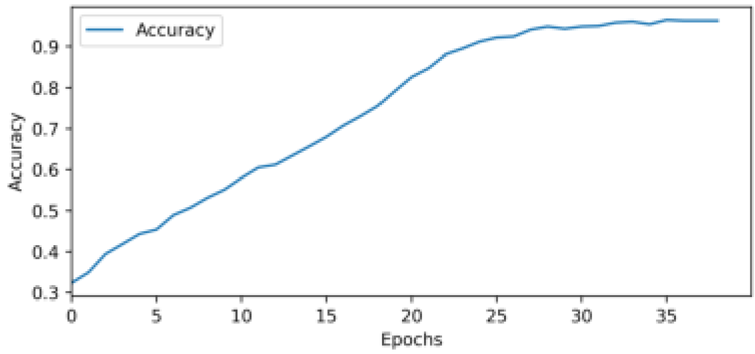


Fig. 5.1: Evaluation metrics for emotion recognition

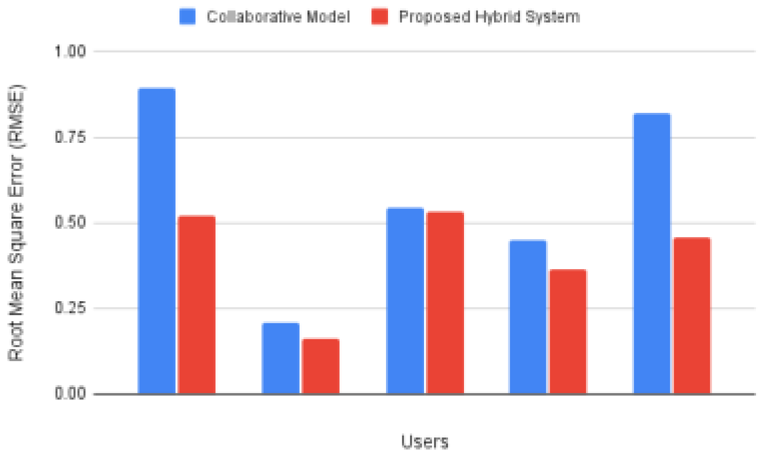


Fig. 5.2: RMSE (Lower is Better)

5.2. Results and Discussion. The ConCollA system is going to be evaluated under two phases, i.e. mood recognition and mood-based music recommendation. In the first phase, the emotion of the driver is detected and in the second phase, music is recommended based on the detected emotion. The results of each phase after evaluating them are discussed below.

5.2.1. Phase-1. The metric used for evaluating the emotion recognition model is the training accuracy. We have taken a subset of the FER-dataset based on our application. We run the model for 40 epochs and the results are obtained are for the same number of epochs. A plot of the graph of accuracy vs. number of epochs is shown in Fig. 5.1.

We have achieved training accuracy of 96.22% on the FER-2013 dataset over the duration of 40 epochs.

5.2.2. Phase-2. In this phase, we delve into the metrics associated with the mood-based recommender engine, analyzing and evaluating its performance.

Collaborative filtering is one of the most common recommenders used everywhere. The below mentioned Table5.1 contains the results obtained from the collaborative recommender for 5 users. These users were picked randomly from a set of users. For each user two metrics are provided: Root Mean Square Error (RMSE)and a novel metric, Rel-Sim.

It is observed that users who have provided a larger amount of data in the form of likes and dislikes of songs

Table 5.1: Comparative Results for Recommender Systems

Users	Collaborative Filtering		Weighted Content+Collaborative Filtering	
	RMSE	RelSim	RMSE	RelSim
1	0.8938	0.5691	0.5184	0.7482
2	0.2097	0.7841	0.1625	0.8639
3	0.5447	0.8108	0.5333	0.8245
4	0.4488	0.8598	0.3626	0.8914
5	0.8219	0.5774	0.4588	0.7708

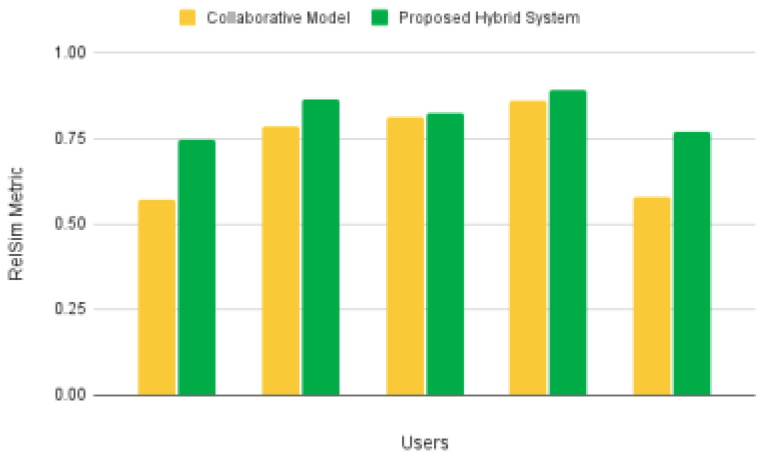


Fig. 5.3: Rel-Sim (Higher is Better)

tend to have lower RMSE and better Rel-Sim scores. However, there are a few users which have not provided sufficient data for recommendation,are not receiving recommendations as effectively as others, indicating a lower performance in terms of RMSE and Rel-Sim. This discrepancy can be attributed to the "cold start problem", where users who have not rated a sufficient number of songs pose a challenge for collaborative filtering algorithms to generate accurate recommendations.

To address the aforementioned problem,and include a mood bias in song recommendation, the proposed work incorporates a novel recommender system to augment the collaborative model. The Table 5.1 also presents the evaluation metrics for the proposed recommender system. The model observes an improvement in both the RMSE and Rel-Sim values for the majority of tested users. This improvement can be attributed to the personalized weightings assigned to each recommendation type. For instance, users X and Y have higher weights assigned to the content-based recommendations, indicating that they will receive a greater emphasis on content-based recommendations for songs. By fine-tuning these weights according to the user's preferences, we can enhance the accuracy and effectiveness of the recommendations provided.

Figure 5.2 displays the RMSE values in a graphical format, facilitating the interpretation of data presented in Table 5.1. The results demonstrate the superior performance of our proposed model, which exhibits lower RMSE values across all 5 randomly chosen users. Particularly noteworthy improvements are observed for Users 1 and 5, showcasing the efficacy of our approach. Similarly, Figure 5.3 presents the RelSim values for the aforementioned users, corresponding to the data in the respective table. Remarkably, the RelSim values have significantly increased for all users, providing further empirical support for our underlying hypothesis.

In addition to the hybrid recommender systems, ConCollA also incorporates strong association rules to provide a broader perspective on musical recommendations. These association rules capture patterns and

Table 5.2: Effect of adding Association Rule Mining on RMSE

	RMSE	
Users	Without Rule Mining	With Rule Mining
1	0.5184	0.5053
2	0.1625	0.1625
3	0.5333	0.5333
4	0.3626	0.3626
5	0.4588	0.4588

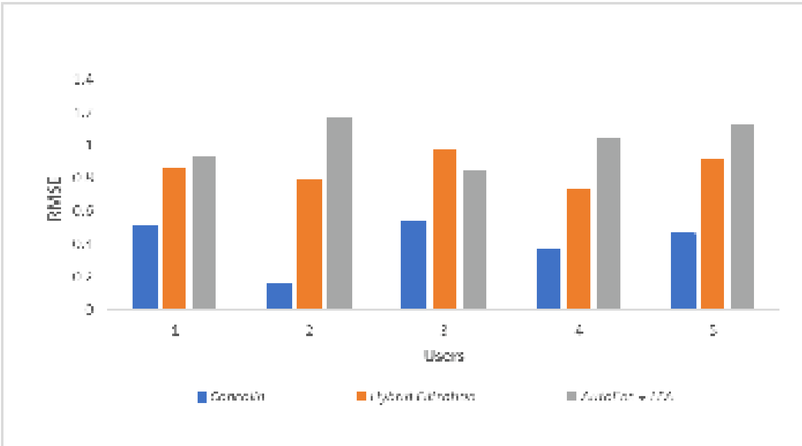


Fig. 5.4: RMSE (Lower is Better)

relationships among users beyond those who have similar preferences to the user being tested. Table 5.2 shows the comparisions between RMSE values of our approach without using strong associations while recommending and RMSE values for the users after adding a component of strong association rules to the former recommender approach. It can be noticed that User 1's RMSE value have decreased, whilst the other users have not changed. This is because there was a strong rule in the ground truth of 1 that led to a fall in its RMSE, while the users who were unaffected by it were because there was no strong rule.

We compare the performance of our model with two other models as depicted in Table 5.3. The first method is based on the concept of hybridization of a collaborative filtering recommender system and a music gene-based recommendation model [32]. For music gene-based recommendation, this approach considered characteristics of music such as tempo, rhythm, and various acoustic features and performed clustering on these characteristics. The results provided collaborative filtering and gene-based recommendation model are combined and filtered to generate one recommendation. Recommendation results can further be improved by finding the correlation among features. Hence we use association rule mining in our proposed model to find the correlation among features and therefore, our proposed model provides better personalized recommendations to the user. Along with that, we also propose our custom content-based algorithm, MAAS, which considers user's current emotional data while making recommendations. Figure 5.4 shows the enhanced outcomes of our proposed model ConColla by applying association rule mining through RMSE value.

Another method utilizes a hybrid approach termed AutoLFA, which aggregates two approaches, AutoEncoder and Latent Feature Analysis(LFA) to formulate a recommendation score. It aggregates these scores based on a customized self-adaptive weighting strategy [11]. We have also utilized this concept of sefl-adaptive weights in our model while combining our three individual systems, viz. collaborative recommendations, content-based recommendations and association rule mining. In our proposed approach, the scores from the three recommenders are aggregated in a weighted manner, and these weights are learnt differently for each user and are

Table 5.3: Comparison of ConCollA with other models

	RMSE		
Users	ConCollA	Hybrid Filtration	AutoEncoder + LFA
1	0.5053	0.867	0.925
2	0.1625	0.798	1.174
3	0.5333	0.975	0.852
4	0.3626	0.732	1.051
5	0.4588	0.921	1.128

self-adaptive and constantly changing as the user interacts with the system. To mitigate the potential cold start issue that could arise within this algorithm, we incorporate a content-based recommendation alongside the collaborative filtering algorithm in our proposed approach.

We employed these models as baseline models for our proposed system. We evaluated these models by training them on the dataset we used for our approach. A side-by-side comparison of RMSE metric of the five users used in previous experiments is shown for our proposed method and the two baseline methods in Fig. 5.4 as well as Table 5.3. The plot clearly shows that our proposed approach ConCollA is superior to both its predecessors for all the five randomly chosen users. This shows the efficacy of our approach combining various techniques to provide relevant recommendations.

6. Conclusion and Future Work. Music is one of the factors for improving the driving experiences, in this paper authors proposed a recommender system to make driving more enjoyable for drivers. ConCollA is a novel recommender system proposed based on personalized content based filtering, collaborative matrix factorization filtering followed by apriori algorithm from association rule mining through a multifaceted system design.A key feature of the model is the novel measure identification - Mood Adjusted Average Similarity (MAAS), and apriori algorithm to generate personalized music recommendations The success of the proposed model has been proved by the various test using a dataset from the Spotify API. Both the RMSE and Rel-Sim measures show that system outperforms the collaborative model, This method has demonstrated the ability to produce a more pleasurable and personalised music experience for drivers by fusing emotional awareness with conventional music selection algorithms. Despite the sophistication of music recommendation algorithms, very few have mastered the emotional power of music whilst driving. The proposed music recommendation system for drivers will pave the way for future advancements in the field of personalized and emotion-aware technology. The research gaps in ConCollA involve limited emotion understanding, overlooking individual variations, neglecting contextual information, challenges in quantifying subjective emotions, complex evaluation metrics, the need for cross-cultural and multilingual adaptability, and addressing ethical implications. Filling these gaps can lead to more personalized and intelligent music recommendations, enhancing user experience and system efficacy. Future research aims to incorporate contextual elements like weather, traffic, and time of day to improve the recommendation system. Additionally, efforts will focus on integrating complex emotional states and classifying songs based on individual user preferences, leading to more contextually aware and personalized music suggestions.

REFERENCES

[1] *World report on road traffic injury prevention.* `https://www.who.int/publications/i/item/world- report-on-road-traffic-injury-prevention`, 2004.

[2] *Goodfellow, et. al.: Challenges in representation learning: Facial expression recognition challenge.* `https://www.kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge/data`, 2023.

[3] A. Abdul, J. Chen, H.-Y. Liao, and S.-H. Chang, *An emotion-aware personalized music recommendation system using a convolutional neural networks approach*, Applied Sciences, 8 (2018), p. 1103.

[4] W. G. Assuncao, L. S. Piccolo, and L. A. Zaina, *Considering emotions and contextual factors in music recommendation: a systematic literature review*, Multimedia Tools and Applications, 81 (2022), pp. 8367–8407.

[5] L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, A. Aydin, K.-H. Lüke, and R. Schwaiger, *Incarmusic: Context-aware music recommendations in a car*, in E-Commerce and Web Technologies: 12th International Conference,

EC-Web 2011, Toulouse, France, August 30-September 1, 2011. Proceedings 12, Springer, 2011, pp. 89–100.
[6] P. Bharati and A. Pramanik, *Deep learning techniques—r-cnn to mask r-cnn: a survey*, Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019, (2020), pp. 657–668.
[7] E. Çano, R. Coppola, E. Gargiulo, M. Marengo, and M. Morisio, *Mood-based on-car music recommendations*, in Industrial Networks and Intelligent Systems: Second International Conference, INISCOM 2016, Leicester, UK, October 31–November 1, 2016, Proceedings 2, Springer, 2017, pp. 154–163.
[8] J. J. Deng and C. Leung, *Emotion-based music recommendation using audio features and user playlist*, in 2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012), IEEE, 2012, pp. 796–801.
[9] N. Dibben and V. J. Williamson, *An exploratory survey of in-vehicle music listening*, Psychology of Music, 35 (2007), pp. 571–589.
[10] A. Ferraro, D. Bogdanov, X. Serra, and J. Yoon, *Artist and style exposure bias in collaborative filtering based music recommendations*, arXiv preprint arXiv:1911.04827, (2019).
[11] S. Guo, X. Liao, G. Li, K. Xian, Y. Li, and C. Liang, *A hybrid recommender system based on autoencoder and latent feature analysis*, Entropy, 25 (2023), p. 1062.
[12] M. Henry, W. Chandra, and A. Zahra, *Implementation of apriori algorithm for music genre recommendation*, Jurnal Online Informatika, 7 (2022), pp. 110–115.
[13] A. V. Iyer, V. Pasad, S. R. Sankhe, and K. Prajapati, *Emotion based mood enhancing music recommendation*, in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE, 2017, pp. 1573–1577.
[14] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, and S. Luo, *A review of content-based and context-based recommendation systems*, International Journal of Emerging Technologies in Learning (iJET), 16 (2021), pp. 274–306.
[15] J. Jooa, S. Bangb, and G. Parka, *Implementation of a recommendation system using association rules and collaborative filtering*, Procedia Computer Science, 91 (2016), pp. 944–952.
[16] S. Joshi, T. Jain, and N. Nair, *Emotion based music recommendation system using lstm-cnn architecture*, in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2021, pp. 01–06.
[17] A. S. Krishnan, X. Hu, J.-q. Deng, L. Zhou, E. C.-H. Ngai, X. Li, V. C. Leung, and Y.-k. Kwok, *Towards in time music mood-mapping for drivers: A novel approach*, in Proceedings of the 5th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications, 2015, pp. 59–66.
[18] T. A. Kumbhare and S. V. Chobe, *An overview of association rule mining algorithms*, International Journal of Computer Science and Information Technologies, 5 (2014), pp. 927–930.
[19] X. Lin, *Music recommendation system based on collaborative filtering and singular value decomposition*, in CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms, VDE, 2022, pp. 1–5.
[20] N. Nassar, A. Jafar, and Y. Rahhal, *A novel deep multi-criteria collaborative filtering model for recommendation system*, Knowledge-Based Systems, 187 (2020), p. 104811.
[21] A. Niyazov, E. Mikhailova, and O. Egorova, *Content-based music recommendation system*, in 2021 29th Conference of Open Innovations Association (FRUCT), IEEE, 2021, pp. 274–279.
[22] J. Patel and H. Chhinkaniwala, *A fusion of aspect and contextual information for rating prediction in recommender system using a latent factor model*, International Journal of Web Engineering and Technology, 16 (2021), pp. 30–52.
[23] ———, *The role of sentiment analysis in a recommender system: a systematic survey*, International Journal of Web Engineering and Technology, 17 (2022), pp. 29–62.
[24] M. J. Pazzani and D. Billsus, *Content-based recommendation systems*, in The adaptive web: methods and strategies of web personalization, Springer, 2007, pp. 325–341.
[25] F. Richter, *Emotion detection: Fer - 2013 dataset with 7 emotion types.* `https://www.statista.com/chart/18557/where-and-when-people- listen-to-music/`, 2019.
[26] T. Schäfer, P. Sedlmeier, C. Städtler, and D. Huron, *The psychological functions of music listening*, Frontiers in psychology, 4 (2013), p. 511.
[27] T. Silveira, M. Zhang, X. Lin, Y. Liu, and S. Ma, *How good your recommender system is? a survey on evaluations in recommendation*, International Journal of Machine Learning and Cybernetics, 10 (2019), pp. 813–831.
[28] J. Singh, *Collaborative filtering based hybrid music recommendation system*, in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), IEEE, 2020, pp. 186–190.
[29] J. Singh and M. Sajid, *Factorization machine based music recommendation approach*, in 2021 6th International Conference on Inventive Computation Technologies (ICICT), IEEE, 2021, pp. 618–622.
[30] P. K. Singh, E. Othman, R. Ahmed, A. Mahmood, H. Dhahri, and P. Choudhury, *Optimized recommendations by user profiling using apriori algorithm*, Applied Soft Computing, 106 (2021), p. 107272.
[31] M. D. Van Der Zwaag, C. Dijksterhuis, D. De Waard, B. L. Mulder, J. H. Westerink, and K. A. Brookhuis, *The influence of music on mood and performance while driving*, Ergonomics, 55 (2012), pp. 12–22.
[32] D. Wu, *Music personalized recommendation system based on hybrid filtration*, in 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE, 2019, pp. 430–433.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 31, 2023
Accepted: Sep 26, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 941–955, DOI 10.12694/scpe.v24i4.2469

HATE SPEECH DETECTION IN LOW-RESOURCE BODO AND ASSAMESE TEXTS WITH ML-DL AND BERT MODELS

KOYEL GHOSH*, APURBALAL SENAPATI†, MWNTHAI NARZARY‡, AND MAHARAJ BRAHMA §

Abstract. Hate speech detection research is a recent sizzling topic in natural language processing (NLP). Unburdened uses of social media platforms make people over-opinionative, which crosses the limit of leaving comments and posts toxic. A toxic outlook increases violence towards the neighbour, state, country, and continent. Several laws have been introduced in different countries to end the emergency problem. Now, all the media platforms have started working on restricting hate posts or comments. Hate speech detection is generally a text classification problem if considered a supervised observation. To tackle text in terms of computation perspective is challenging because of its semantic and complex grammatical nature. Resource-rich languages leverage their richness, whereas resource scarce language suffers significantly from a lack of dataset. This paper makes a multifaceted contribution encompassing resource generation, experimentation with Machine Learning (ML), Deep Learning (DL) and state-of-the-art transformer-based models, and a comprehensive evaluation of model performance, including thorough error analysis. In the realm of resource generation, it adds to the **N**orth-**E**ast **I**ndian **H**ate **S**peech tagged dataset (NEIHS version 1), which encompasses two languages: Assamese and Bodo.

Key words: Hate Speech Detection, Assamese, Bodo, Natural Language Processing, NLP, Machine Learning, Deep Learning, Word2Vec, NB, SVM, LSTM, BiLSTM, CNN, BERT.

1. Introduction. People can use the internet to learn new skills, engage in knowledgeable debates, and share information. Still, they also engage in anti-social activities such as cyberbullying, trolling, spreading hate, and so on [1]. Spreading hate can lead people to communal riots [2, 3], can be responsible for mental illness [4] etc. United Nations Strategy and Plan of Action on hate speech [5] has been introduced and defines hate speech as *any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language regarding a person or a group based on their religion, ethnicity, colour, gender or other identity factors.* Indian government law is also introduced against hate speech [6]. Several social media platforms revised their community guidelines to eradicate hate, automatically detecting hate comments and posts and giving users access to report posts and comments[1][2]. English and other popular languages benefit from their global popularity. Few studies are done on Indian languages in hate speech detection [7, 8, 9] like in Hindi, Bangla, Marathi, etc. India has 22 official languages and about 1,000 living languages from various language groups [10]. North-East Indian languages are under-resourced, and no tagged hate speech detection task data is available. As an associate official language of the Indian state of Assam, Bodo is widely spoken in the Bodoland Territorial Region[3]. Among the official languages of India, it has gained some recognition[4] [11]. The 2011 Indian Census[5] estimates a total of 1,482,929 Bodo speakers, including 1,454,547 native speakers. Assamese is spoken by 15,311,351 people, which is a huge number. Assamese and Bodo suffer from a serious lack of resources for Natural Language Processing (NLP). Native researchers are trying to build a sufficient

*Department of Computer Science and Engineering, Central Institute of Technology, Kokrajhar, Assam, India (ghosh.koyel8@gmail.com)
†Department of Computer Science and Engineering, Central Institute of Technology, Kokrajhar, Assam, India (a.senapati@cit.ac.in)
‡Department of Computer Science and Engineering, Central Institute of Technology, Kokrajhar, Assam, India (mwnthainarzary123@gmail.com)
§Department of Computer Science and Engineering, IIT Hyderabad (mraj.brahma@gmail.com)

[1]https://web.facebook.com/communitystandards/ (Access on 02.10.2023)
[2]https://www.youtube.com/howyoutubeworks/policies/community-guidelines/(Access on 02.10.2023)
[3]Formerly known as the Bodoland Territorial Autonomous District (BTAD)
[4]Scheduled languages of India: Bodo is one of the 22 scheduled languages, added in 2004
[5]https://censusindia.gov.in/census.website/data/census-tables (Access on 02.10.2023)

corpus for the research [12], describing a process for creating a monolingual Bodo corpus using Google Keep for OCR to scan text from different books. Major contributing factors are the paucity of language models labelled datasets and effective machine learning (ML) approaches for a wide range of NLP jobs. Due to unfettered access and usage of social media and digitalization, hate speech is rising in North-East India, too, just as in other major languages like English. To the best of our knowledge, no hate speech detection dataset is available in Assamese or Bodo. This is our first attempt at creating a North-East Indian Hate Speech (NEIHS) dataset for binary text classification with *2K* in Bodo and *4K* in Assamese respectively, labelled as *'hate'* and *'non-hate'*. We found that state-of-the-art models like transformers are more effective than ML or DL approaches. Our contribution to this paper :

1. We create the North-East Indian Hate Speech dataset (NEIHS). NEIHS (version 1) is a binary text classification task-purpose human annotate dataset.
2. Trained Naïve Bayes (NB), Support Vector Machines (SVM), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Network (CNN), and state-of-the-art transformer-based models on the NEIHS dataset.
3. Detailed analysis of the results of all models to check how a model works on the language-specific hate speech data.
4. At the end, present a detailed error analysis for the outcomes.

The rest of the paper is arranged in a way where Section 2 shall explain the related work; Section 3 describes the details of the NEIHS dataset. Section 4 will be the methods used on the datasets, and detailed results with error analysis will be shown in section 5. Lastly, we conclude our work in section 6.

2. Related works. Here, we shall discuss existing hate speech detection research. To correctly and reliably identify hate speech from widely spoken languages like English, many works have been proposed [13, 14]. Traditional algorithms like SVM, NB, Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Gradient-Boosted Trees (GBT) have relied on manual feature engineering for quite some time. On the other hand, CNN or LSTM networks form the backbone of Deep Learning (DL) based techniques that learn multilayers of abstract features from raw texts. Since linear models have proven less accurate and scalable when dealing with billions of such texts, these methods pale compared to DLs. When extracting features, CNN excels, while LSTM excels at modelling neatly sequenced learning issues. For example, CNNs can extract n-grams, sequences of words or characters, and LSTM may learn the long-term relationships between words or characters in texts. Each network architecture has benefits, but various works have investigated what would happen if combined [15]. With Conv-LSTM, the class of a word sequence depends on previous word sequences, making it a powerful architecture that captures long-term correlations between characteristics retrieved by CNN. This architecture is more effective than structures exclusively based on CNN or LSTM. Variation of BERT (Bidirectional Encoder Representations from Transformers) [16], RoBERTa (Robustly optimized BERT) [17], ALBERT (A Lite BERT) [18], DistilBERT (Distilled version of BERT) [19] and their pre-trained models such as mBERT [16], MuRILBERT [20], RoBERTaHindi[6], Indic Bert [22], MahaBERT [23], MahaRoBERTa [24], XLM-RoBERTa [25], BanglaBert [26] etc. is used for hate speech detection experiments. This section mainly focuses on hate speech detection in Indian languages. Nowadays, mainly code-mixed languages like Hindi-English [27], Tamil-English [28] etc. are preferred for the research as social media users use complex code-mixed language, like their native language as well as English in between sentences, but we shall stick to the monolingual experiments for the discussion. HASOC (Hate Speech and Offensive Content Identification), a shared task organized by FIRE (Forum for Information Retrieval Evaluation)[7], which published hate datasets in Indian languages such as Hindi, Marathi, etc. HASOC offers four subtracks, one of which is relevant to us: **HASOC - English and Indo-Aryan Languages**. Datasets are distributed in tab-separated format. HASOC and most other collections require mechanisms to detect hateful content from the text of a post. In 2019, the HASOC-Hindi dataset offered three tasks [7]. The first task is binary classification, i.e., subtask A. The second task is to find whether the hate comment was profane or abusive (multiclass), i.e., subtask B. The third is to predict whether the hate comment is targeted or untargeted (multiclass), i.e., subtask C. In the Hindi language, ninety-three runs were submitted across three sub-tasks. Regarding the Hindi subtask A, the winner

[6]https://huggingface.co/flax-community/roberta-hindi (Access on 02.10.2023)
[7]http://fire.irsi.res.in/fire/2022/home (Access on 02.10.2023)

team, QutNocturnal [29], employed a CNN base technique with Word2vec embedding and got better Marco F1 and Weighted F1 values, 0.8149 and 0.8202, respectively. The second team LGI2P [30], trained a fastText model for the proposed Hindi language and later used BERT for classification. The system achieved 0.8111 Marco-F1 and 0.8116 Weighted-F1 values. For sub-task B on the Hindi Dataset, 3Idiots [31] scores 0.5812 and 0.7147 in Marco-F1 and Weighted-F1 utilizing BERT. Team A3-108 [32] achieves a high Marco-F1 score on sub-task C Hindi Dataset, which is 0.5754. According to them, Adaboost [33] was the best-performing classifier among the three classifiers, i.e., Adaboost or Adaptive Boosting (AB), RF, Linear SVM. They merge multiple weak classifiers to construct a robust prediction model, but an ensemble of SVM, RF, and AB with hard voting performed even better. This classifier used TF-IDF features of word unigrams and characters 2, 3, 4, and 5 grams with an additional feature of the length of every tweet. In HASOC 2020, two Hate Speech detection tasks [8], sub-task A (binary class) and sub-task B (multiclass) are proposed with another Hindi dataset in the research area. NSIT_ML_Geeks [34] outperforms other competing teams, scoring Marco-F1 0.5337 and 0.2667 in sub-task A and sub-task B, respectively, utilizing CNN and BiLSTM. In 2021, HASOC published a Hindi dataset [9] with sub-tasks A and B again. The best submission was achieved Macro F1 0.7825 in sub-task A with a fine-tuned Multilingual-BERT (20 epochs) with a classifier layer added at the final phase. The second team also fine-tuned Multilingual-BERT and scored Macro F1 0.7797. NeuralSpace [35] got Macro F1 0.5603 in sub-task B. They use an XLM-R transformer, vector representations for emojis using the system Emoji2Vec, and sentence embeddings for hashtags. After that, three resulting representations were concatenated before classification. In the paper [36], they used the pre-trained multilingual BERT (m-BERT) model for computing the input embedding on the Hostility Detection Dataset (Hindi) later SVM, RF, MultiLayer Perceptron (MLP), LR models are used as classifiers. In coarse-grained evaluation, SVM reported the best weighted-F1 score of 84%, whereas they obtained 84%, 83%, and 80% weighted-F1 scores for LR, MLP, and RF. In fine-grained evaluation, SVM has the most excellent F1 score for evaluating three hostile dimensions, namely Hate (47%), Offensive (42%), and Defamation (43%). LR beats the others in the Fake dimension with an F1 score of 68%. Authors [28] prepare a Tamil dataset to identify Homophobia and Transphobia from comments. They experiment with baseline models, machine learning, deep learning, and transformer-based models. We found a paper[8], authors prepare a large-scale monolingual Indic to hate speech dataset in 5 languages: Hindi (Hi), Tamil (Ta), Telugu (Te), Malayalam (Ml) and Kannada (Kn). A multilingual, monolingual transformer comparison of hate speech dataset on HASOC-2019(Hindi), HASOC-2019(Marathi), and Bangla hate dataset [37].

3. NEIHS Dataset.

3.1. Dataset collection. Our primary goal while creating the dataset was to create it with different varieties of data, so we chose different kinds of ™Facebook pages like political, entertainment, etc. We first identify some controversial posts, including recent events, politicians, and actors who are more likely to contain hate speech. Then, we go through comments on such posts and find comments that are written in at least 80-90% monolingual. Then, we manually check whether these include hate and categorize them accordingly. Figure 3.1 shows both datasets' word clouds. Bodo shares the Devnagari script, and Assamese shares the Assamese script, likely the Bengali-Assamese script. Figure 3.2 shows the most used hate words in the NEIHS dataset.

We have collected all the data from ™Facebook and ™YouTube comments for the NEIHS dataset. We followed a few Bodo ™Facebook pages, such as news, entertainment, celebrity, politics, etc., and selected some uploads to manage related comments and retrieved over a thousand comments with non-hate and hate comments from various posts published between January 2022 and September 2022 using open source scrapper tools[9]. The same we did with some Bodo local news, celebrity, politics, etc. ™Youtube channels. Most of the comments we collected are in English transliteration, and the sentences include unwanted symbols like `-`, `'`, `(`, `)`, etc. We cleaned and preprocessed sentences. Removed unnecessary symbols, structured sentences translated and rewrote sentences, and Bodo Script. Finally, the sentences were annotated as either 'hate' or 'non-hate' by native speakers. Sentences with 'hate' that include hate words are considered hate-offensive statements, while sentences that convey formal information, suggestions, or questions are considered non-hate sentences. We have

[8]https://openreview.net/forum?id=HCnb1TByvx7 (Access on 02.10.2023)
[9]https://github.com/kevinzg/facebook-scraper (Access on 02.10.2023)

(a) (b)

Fig. 3.1: Wordclouds of NEIHS - a) Assamese and b) Bodo datasets.

Hate words/ Assamese	Hate words/ Bodo
माकचुदु (Motherfucker), बनाৰি/बुनৰी/ৰेनादि (Prostitute), কেলা (It is a word that doesn't itself has a meaning but is used within sentences to express anger, irritation etc), চুদুৰভাই (Fucker), জহৰি (indicates women for tricking people), গেদা (indicates Bangladeshis mostly muslims)	फाग्ला (mad), गुनदा (Rapist), हारिनि सुथुर (Enemy for comminity), हारि बेफारि (community seller), जामबा (mad man), बुरबक (Dump), जोलैनि हांगार (betrayer)

Fig. 3.2: Some examples of Hate words in NEIHS - a) Assamese and b) Bodo datasets

collected the Assamese dataset in the same manner.

3.2. Dataset annotation. Two native speakers, young adults in the 19 to 24 age range, annotated the data for each language. These annotators are students at the Central Institute of Technology in Kokrajhar, Assam, India. The annotation team's task was to manually categorize comments from NEIHS as either containing hateful content or not, using binary labels. In cases where there was a disagreement between the two student annotators regarding the label assignment, a third student with expertise in social media research was consulted to make the final call. Hate speech is a highly subjective issue. As a result, defining what constitutes hate speech is difficult. As a result, we've established specific strict guidelines. These regulations are based on the community standards of ™Facebook[10] and ™YouTube[11]. Comments with the following aims should be marked as hate. (a) *Profanity*: Comments that contain profanity, cussing, or swear words are marked as hate. (b) *Sexual orientation*: Sexual attraction (or a combination of these) to people of the opposite sex or gender, to people of the same sex or gender, to both sexes, or to people of more than one gender. (c) *Personal*: remarks on clothing sense, content selection, language selection, etc. (d) *Gender chauvinism*: Comments in which people are targeted because of gender. (e) *Religious*: Comments in which the person is criticized for their religious beliefs and practices. For example, comments challenging the use of a turban or a burkha (the veil), (f) *Political*: Comments criticizing a person's political beliefs. For instance, bullying people for supporting a political party. (g) *Violent intention*: Comments containing a threat or call to violence.

Examples as in figure 3.3 are '' which roughly translates to 'What kind of man is he?', which contains a word usually identified as the hate word '' but the meaning does not contain any hate. ''. Now the sentence roughly translates to 'Total AAP supporters– 80% Miya(50% of them are fake accounts and another 30% is their real account), after 80% Miya's. The remaining ones are Badan (These are the supporters of Congress and Akhil).' We need to know the context 'Miya' is used as a word to describe illegal Bangladeshi immigrants, usually 'Muslims', and 'Badan' is considered a traitor because he invited the rulers of Myanmar to attack

[10]https://web.facebook.com/communitystandards/ (Access on 02.10.2023)
[11]https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ (Access on 02.10.2023)

Comments (Assamese/ English translation)	Hate category
চাল্লা কি মানুহ (What kind of man is he?)	non-hate
দেখাত জেহাদি জেহাদি লাগে (You look like a Jihadi.)	Personal Attack
মুঠ আপ সমৰ্থক--৮০% মিঞা(তাৰে ৫০% ফেক নামত কমেণ্ট দিয়ে বাকী ৩০% নিজ নামত) গ'ল ৮০% মিঞা। বাকী থাকিল ২০% বদন(এইকেইটা হ'ল কংগ্ৰেছ অখিল সমৰ্থক)" (Total AAP supporters-- 80% Miya(50% of them are fake accounts and the other 30% is their real accounts), after 80% Miya's. The remaining ones are Badan (These are the supporters of Congress and Akhil).)	Political Hate
99% গেদাই অসমত চোৰ ধৰ্ষন হত্যা কৰে ..। (99% of Rapes and Robberies are done by Gedas..)	Religious Hate

Comments (Bodo/ English translation)	Hate category
आंग सानो जाय बाराद्राय बखियो बे जेबो मावनो रोङा (Those who constantly abuse others will achieve nothing.)	non-hate
AASU नि आसाम Accord नि खोथाखौ सोरबा बरफोरा मोनथिगौब्ला गोसो खांबानो जाबाय (Those Bodos who are familiar with the AASU and the Assam Accord are requested to keep this in mind.)	Personal Attack
हनै मालाय हारसा हिनजावनि खिबु सुग्राफोरा मिनिसोदों (Look, Assamese women bum cleaner are laughing.)	Gender
सोरबा माबा मोनसे खामानि मावनो थांनायाव मानि हेंथा गिखफोर ? (Why was there always a barieer when they were going to work for the good ?)	non-hate

Fig. 3.3: Examples from NEIHS dataset. Hate examples and their translations in English with descriptions (Hate category) are boxed in red, whereas non-hate examples are in green.

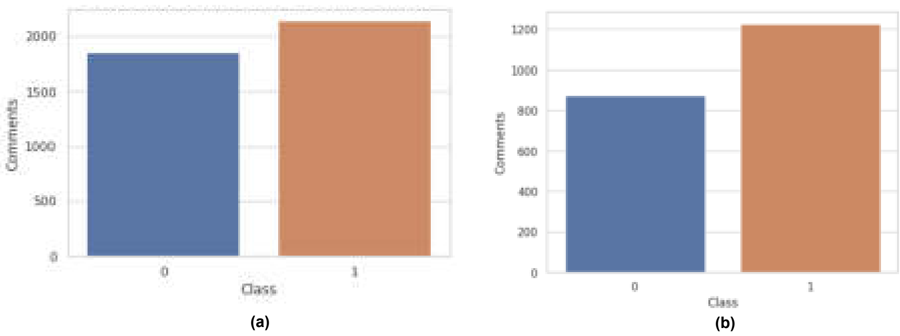


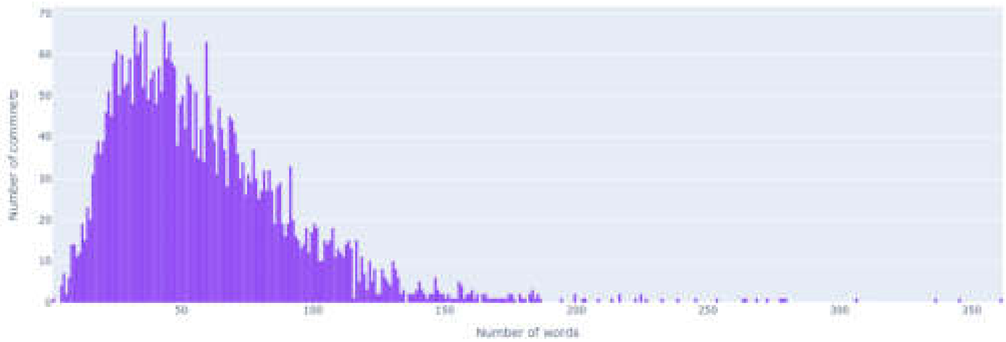
Fig. 3.4: Class distribution of NEIHS - a) Assamese and b) Bodo datasets where '0' is 'non-hate' and '1' is 'hate'.

Assam, and they did many unspeakable things and led the way to the end of Ahom rule and paved the way for the British.' '' . '' is a word used to describe illegal Muslim immigrants, kind of religious attack again. It is also used to mean criminally minded people. The sentence roughly translates to '99% of rapes and robberies are done by Gedas.'.

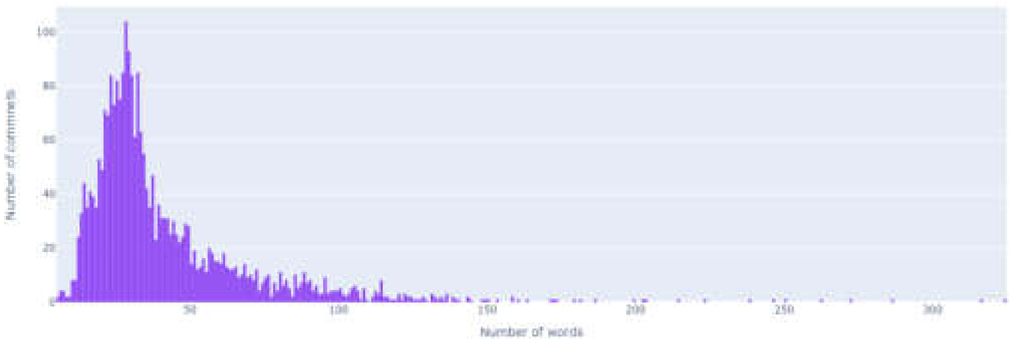
3.3. Dataset analysis. We summarize the key statistics of NEIHS in Table 3.1. For the Assamese dataset, 2,143 comments are hate out of 3,996. As a result, our data set is slightly skewed in favour of containing hate speech. Of 2,099 comments in our NEIHS - Bodo dataset, 1,225 contain hate speech. Figure 3.4 shows the details of class distribution. The length (number of words) of comments in each language is shown in figure 3.5. We split the dataset into a training set and a test set by 80:20. In NEIHS - Assamese, 1,705 comments are hate out of 3,196 in training data and 438 hate comments in test data out of 800. In the NEIHS - Bodo training set, 998 comments are hate out of 1,679 comments, and 227 are hate out of 420 in the test set.

4. Methods. Our input sample is x, consisting of m number of texts (each row) indicated as $x = \{r_1, r_2, r_3, .., r_i, ..., r_m\}$, where r_i is the i^{th} text or row, and m is equal to the total number of texts present in the set. Given a text or row r_i, the text has n sentences ($n <= 4$).

4.1. Preprocess. To improve performance, text data must be cleaned and noise-free before being used in the DL models. The languages spoken by those with few resources in India have several things in common. Researchers used identical preprocessing procedures for all datasets, even though they were written in various languages. Some tasks and datasets may call for slightly different text preparation processes. Raw comments



(a)



(b)

Fig. 3.5: The length (number of words) of comments in NEIHS a) Assamese and b) Bodo datasets.

Table 3.1: Class-wise distribution analysis for NEIHS dataset

Datasets	hate		non-hate		Total
	Train	**Test**	**Train**	**Test**	
NEIHS - Assamese	1,705	438	1,491	362	3,996
NEIHS - Bodo	998	227	681	193	2,099

with emojis, punctuation, and other undesired characters were present in some datasets. Commonly, the following procedures are used:

1. **Remove url, user**: We remove unwanted characters, url, URL occurrence with `xxurl`.

2. **Remove Punctuation**: Punctuation and numbers often don't add extra meaning to the text, hence being removed from the text.

3. **Remove stopwords**: Stopwords are the most commonly used words. Removed 15 Bodo stopwords from the NEIHS - Bodo dataset and 116 Assamese stopwords from the NEIHS - Assamese dataset.

4. **convert emojis to text**: We prepare an emoji dictionary to convert emojis to text. Only the Assamese dataset has emojis, and the Bodo dataset doesn't have emojis. So, examples are 😀 : '', 😂 : '

’, 😊 : ‘’.

5. **Stemming**: Stemming is a technique that reduces any inflected word to its root form. Here, lightweight stemming is used, adapted from Sarmah et al. [38]. Some examples of Assamese stemming like, ‘’, ‘’, ‘’ are the Assamese words, so we get the root words ‘’, ‘’, ‘’ truncating ‘’, ‘’, ‘’. We didn’t apply any stemming for the Bodo dataset, as any lightweight stemming does not exist.

6. **Tokenization**: We tokenize the sentences into words with the basic Keras’ Tokenizer.

7. **Label encoding**: The class is tagged as ‘non-hate’ and ‘hate’ in both datasets. We encode them into a unique number. Like ‘non-hate’ to ‘0’ and ‘hate’ to ‘1’.

After preprocessing we get $r_i = \{w_{i,1}, w_{i,2}, w_{i,3}, ..., w_{i,p}, .., w_{i,q}\}$, where $w_{i,p}$ denotes the p^{th} word in the i^{th} row and q is the number of words.

4.2. Machine Learning baseline models. Machine learning baselines such as NB, SVM are trained with TF-IDF-weighted character n-grams and word uni-grams.

4.3. Word embedding. A word’s vector representation is necessary for any neural network model. Thomas Mikolov’s Word2Vec models, described in detail at Google [39], are implemented to study word embeddings. Word2Vec’s learned vector [40] allows for the inference of word relationships and similarities. Here, we use the Skip-gram model. To create a Word2Vec model, we used the gensim[12] module to train on the NEIHS dataset.

Now, the Word2Vec skip-gram model is trained using the training dataset used for this study. To train word embedding, we fit the parameters as embedding dimension = 300, window = 10, and saved the trained Word2Vec model for the next step. We embed each word $w_{i,p}$ to our pre-trained (trained on NEIHS dataset) word vector after loading the model into memory i.e. each word in the text is converted into a d-dimension embedding vector, where $w^v_{i,p} \in \mathbb{R}^d$ is d-dimension embedding vector of p^{th} word. The word level embedding as $r^v_i = \{w^v_{i,1}, w^v_{i,2}, w^v_{i,3}, ..., w^v_{i,p}, ..., w^v_{i,q}\}$ and $H = H_1, H_2, H_3, ..H_n$ is a hidden layer. We represent a text r_i with q words as a matrix $r_i \in \mathbb{R}^{d \times q}$.

4.4. Deep neural networks.

1. **LSTM**: It is an extension of Recurrent Neural Network (RNN) [41], capable of learning long dependencies. The LSTM neural networks contain three gates and a cell memory state. Here, $\{w^v_{i,1}, w^v_{i,2}, w^v_{i,3}, ..., w^v_{i,p}, ..., w^v_{i,q}\}$ denotes the word vector, q is the length of a text, $\{h_1, h_2, h_3, .., h_o, .., h_q\}$ represents the hidden vector. Now, we give r^v_i as a input to LSTM for feature extraction, namely F^{LSTM}_i, in equation

$$F^{LSTM}_i = LSTM(r^v_i) \tag{4.1}$$

In the architecture $\{fh_1, fh_2, ..., fh_q\}$ is the forward hidden vector. The output layer is with the *sigmoid* activation function. Loss calculation is done by *binary_crossentropy*, and *adam* optimizer is used.

2. **BiLSTM**: Model works from both directions, where the equation will be

$$F^{BiLSTM}_i = BiLSTM(r^v_i) \tag{4.2}$$

In Bidirectional LSTM, sequence data is processed in both directions with forward LSTM and backward LSTM layers, and these two hidden layers are connected to the same output layer. $\{fh_1, fh_2, ..., fh_q\}$ and $\{bh_1, bh_2, ..., bh_q\}$ represent the forward and the backward hidden vector respectively. $\{h_1, h_2, h_3, .., h_o, .., h_q\}$ represents final hidden layer. The final hidden vector h_o of the BiLSTM is shown as the following equation:

$$h_t = [fh_o, bh_o] \tag{4.3}$$

Our model is enhanced with two BiLSTM layers. The fully connected layer with 256 neurons and *relu* activation is added. Because we have two classes, the output layer with *sigmoid* activation is introduced. Finally, use *binary_crossentropy* to compile the model.

[12]https://radimrehurek.com/gensim/models/word2vec.html (Access on 02.10.2023)

3. **CNN**: We transform a text r_i composed of q words into a matrix denoted as $r_i \in \mathbb{R}^{q\times d}$. Then, we perform a convolution operation on this matrix with a single stride. For each convolution operation, we utilize a filter denoted as $f_l \in \mathbb{R}^{a\times d}$, with a specific size of a, where d represents the dimension of the word vector. We apply 128 filters for $a \in \{3\}$, 256 filters for $a \in \{4\}$ and 512 filters for $a \in \{5\}$. $c(f_l, b) = relu(f_l \cdot b_{g:g+a-1})$ is the convolution function where f_l is the given filter, $b_{g:g+a-1}$ is the g^{th} vertical slice of the text matrix from g to $g : g + a - 1$ position. The result of this convolution is then passed through the *relu* activation function [42]. The function $c(f_l, b)$ generates a feature, denoted as c_g, which resembles n-grams for each slice g, resulting in a total of $q - a + 1$ features. To capture the most significant feature among these $q - a + 1$ features, we apply a max-pooling operation, i.e., $\hat{c}_l = max(c(f_l, b))$, which selects the maximum value. This max-pooling operation captures the most crucial aspect of each filter. In the proposed model, we have 896 filters (comprising 128+256+512), meaning we learn 896 of the most important features from the convolution layer. We pass these features to a *dense* layer with 256 perceptrons that use the *relu* activation function. Another *dense* layer with one perceptron is applied at the end with the *sigmoid* activation function.

4.5. Transfer learning. One challenge of the machine learning research known as 'transfer learning' is to find ways to generalize the insights obtained from addressing one problem to a new one that is conceptually similar. These two languages are so rare that no pre-trained transformer model is available. Google developed BERT, a transformer-based technique for NLP. BERT can generate contextualized embeddings. We insert [CLS] in the beginning and [SEP] between sentences or at the end. The sentence will then be as $r_i = \{[CLS]s_{i,1}[SEP]s_{i,2}[SEP]s_{i,3}....s_{i,j}[SEP]s_{i,j+1}....s_{i,n}[SEP]\}$, where $s_{i,j}$ is the j^{th}sentence of i^{th} row. Each sentence with p words. The text will need to be tokenized now. Tokenizing a text yields a dictionary containing the input ids, numerical representations of the tokens, and the attention mask, indicating whether the token is a [PAD]. $s_{i,j} = \{[CLS], t_{i,j,1}, t_{i,j,2}, ..., t_{i,j,k}, ..., t_{i,j,p}, [SEP]\}$, where $t_{i,j,k}$ is the k^{th} token in the j^{th} sentence of i^{th} row. The entire input looks like $r_i = \{[CLS], t_{i,1,1}, t_{i,1,2}, ..., t_{i,1,k}, ..., t_{i,1,r}, [SEP], t_{i,2,1}, t_{i,2,1}...., t_{i,2,s}, [SEP]\}$, where $t_{i,j,k}$ is k^{th} token of j^{th} sentence of i^{th} row. r and s are the number of tokens in sentences. Words and tokens should not be the same number as sentencepiece/ wordpiece tokenizers perform build-in stemming. Here, as transfer learning, we use pre-trained transformer-based BERT models.

1. **MuRILBERT**[13]: Multilingual Representations for Indian Languages (MuRIL) is a BERT model pre-trained on 17 Indian languages and their transliterated counterparts, i.e. monolingual segments, and parallel segments.
2. **mBERT (cased/ uncased)**[14][15]: It is pre-trained with the largest Wikipedia over 104 top languages worldwide, including Hindi, Bengali, and Marathi, using a masked language modelling (MLM) objective. Though Bodo and Assamese languages don't have capital letter issues, we applied both to observe the performance.
3. **RoBERTaHindi**[16]: This is a transformers model pre-trained on a large corpus of Hindi data (a combination of mc4, oscar and indic-nlp datasets).
4. **MahaRoBERTaMarathi**[17]: A Multilingual RoBERTa (xlm-roberta-base) model fine-tuned on publicly available Marathi monolingual datasets and L3Cube-MahaCorpus.
5. **BanglaBERT**[18]: Using mask language modeling, bangla-Bert-Base was pre-trained on data downloaded from OSCAR and Bengali Wikipedia Dump Dataset.

4.6. Experiments. Table 4.1 shows the preprocessing steps used in both languages before employing the data in the models. We kept 80% of the dataset as a train set and 20% as a test for both datasets. The architectural parameters used to train LSTM, BiLSTM, and CNN models are listed in Table 4.2, respectively. Due to memory and *GPU* issues, we did limited experiments with hyperparameters like batch size, epoch, learning rate, etc. In the first experiment, we use a TF-IDF vectorizer to transform words into features, and then we train the Multinomial NB model. We used the linear kernel and kept all other parameters at their default value. In the case of LSTM/BiLSTM/ CNN without Word2Vec, all the preprocessing steps are done

[13]https://huggingface.co/google/muril-base-cased (Access on 02.10.2023)
[14]https://huggingface.co/bert-base-multilingual-cased (Access on 02.10.2023)
[15]https://huggingface.co/bert-base-multilingual-uncased (Access on 02.10.2023)
[16]https://huggingface.co/flax-community/roberta-hindi (Access on 02.10.2023)
[17]https://huggingface.co/l3cube-pune/marathi-roberta (Access on 02.10.2023)
[18]https://huggingface.co/sagorsarker/bangla-bert-base (Access on 02.10.2023)

Table 4.1: Experiment wise preprocessing steps on NEIHS dataset

Preprocessings	**Assamese**		**Bodo**	
	Others	**BERT**	**Others**	**BERT**
Remove url, user	✓	✓	✓	✓
Remove Punctuation	✓	×	✓	×
Remove stopwords	✓	×	✓	×
Convert emojis to text	✓	✓	×	×
Stemming	✓	×	×	×
Tokenization	✓	✓	✓	✓
Label encoding	✓	✓	✓	✓

before employing the input to the first layer for the model; for embedding, the Keras embedding layer is applied. We set num_words to 5,142 for the Bodo dataset and 9,629 for the Assamese dataset. In transfer learning, we choose the advantage of MuRILBERT, which includes 17 Indian languages along with Assamese, and mBERT cased/uncased, which includes 104 top languages, BanglaBERT for only NEIHS-Assamese and RoBERTaHindi, MahaRoBERTaMarathi for the NEIHS-Bodo dataset, as Bodo shares Devnagari script and Assamese shares Assamese script likely Bengali-Assamese script. Besides this, we are using one pre-trained model, which is trained on the monolingual Hindi and Marathi language used for Bodo, and one Bangla pre-trained monolingual model for Assamese. However, we only performed lightweight preprocessing like removing URL, user name, emoji, and for the tokenization part, a sentencepiece/ wordpiece tokenizer is used, as previous studies have demonstrated that BERT-based models achieve higher classification accuracy on unclean texts. We use MuRILBERT, mBERT-uncased/cased for both NEIHS languages. Specifically, for transfer learning, we use $batchsize = 8$, $epochs = 20$, $learning\ rate = 2e-5$, and $epsilon = 1e-8$.

Below are all types of model combinations we tested on our dataset.

1. NB and SVM baseline evaluation
2. LSTM without Word2Vec
3. LSTM with Word2Vec
4. BiLSTM without Word2Vec
5. BiLSTM with Word2Vec
6. CNN without Word2Vec
7. CNN with Word2Vec
8. Transfer learning

5. Result. In this section, we discuss the precision, recall, weighted F1 score and accuracy obtained by training all the ML, DL, and transformer-based models on Assamese and Bodo datasets. Table 4.2 represents the results of all models trained on the datasets; we intentionally prefer a weighted F1 score over an accuracy score to evaluate the models because imbalanced class distribution exists in most classification problems. To evaluate our models, we use two class precisions ($P_{non-hate}$, P_{hate}), recalls ($R_{non-hate}$, R_{hate}), F1 scores ($F1_{non-hate}$, $F1_{hate}$) then calculate precision (W_P), recall (W_R), and F1 score (W_{F1}) here. At last, we calculate *Accuracy*.

$$P_{non-hate} = \frac{True_{non-hate}}{True_{non-hate} + False_{hate}} \tag{5.1}$$

$$P_{hate} = \frac{True_{hate}}{True_{hate} + False_{hate}} \tag{5.2}$$

$$R_{non-hate} = \frac{True_{non-hate}}{True_{non-hate} + False_{non-hate}} \tag{5.3}$$

$$R_{hate} = \frac{True_{hate}}{True_{hate} + False_{non-hate}} \tag{5.4}$$

$$F1_{non-hate} = 2 * \frac{P_{non-hate} * R_{non-hate}}{P_{non-hate} + R_{non-hate}} \tag{5.5}$$

$$F1_{hate} = 2 * \frac{P_{hate} * R_{hate}}{P_{hate} + R_{hate}} \tag{5.6}$$

Table 4.2: Combination of parameters for LSTM, BiLSTM and CNN models

Model name	**Parameter name**	**Parameter value**
LSTM	Epochs	20
	Batch size	8
	Embedding dimension	200, 300
	LSTM layer 1	128, *relu* activation
	LSTM layer 2	128, *relu* activation
	Dropout	0.1
	Dense layer 1	32, *relu* activation
	Dropout	0.1
	Dense layer	1, *sigmoid* activation
	Learning-rate	0.001
BiLSTM	Epochs	20
	Batch size	8
	Embedding dimension	200, 300
	BiLSTM layer 1	64
	BiLSTM layer 1	64
	Dense layer 1	256, *relu* activation
	Dense layer 1	1, *sigmoid* activation
	Learning-rate	0.001
CNN	Epochs	20
	Batch size	8
	Embedding dimension	200, 300
	Conv1D layer 1	128 filters, 3 kernel_size, *relu* activation, 1 stride
	Global MaxPooling1D 1	-
	Conv1D layer 2	256 filters, 4 kernel_size, *relu* activation, 1 stride
	Global MaxPooling1D 2	-
	Conv1D layer 3	512 filters, 5 kernel_size, *relu* activation, 1 stride
	Global MaxPooling1D 3	-
	Concatenate	Global MaxPooling1D 1, Global MaxPooling1D 2, Global MaxPooling1D 3
	Dense layer 1	256, *relu* activation
	Dense layer 2	1, *sigmoid* activation
	Learning-rate	0.001

$$W_P = \frac{P_{non-hate} * T_{non-hate} + P_{hate} * T_{hate}}{T_{non-hate} + T_{hate}} \tag{5.7}$$

$$W_R = \frac{R_{non-hate} * T_{non-hate} + R_{hate} * T_{hate}}{T_{non-hate} + T_{hate}} \tag{5.8}$$

$$W_{F1} = \frac{F1_{non-hate} * T_{non-hate} + F1_{hate} * T_{hate}}{T_{non-hate} + T_{hate}} \tag{5.9}$$

$$Accuracy = \frac{True_{non-hate} + True_{hate}}{T_{non-hate} + T_{hate}} \tag{5.10}$$

where $True_{non-hate}$ = True-negative (model predicted the texts as non-hate, and the actual value of the same is also non-hate), $True_{hate}$ = True-positive (model predicted the texts as hate, and the actual value of the same is also hate), $False_{non-hate}$ = False-negative (model predicted the texts as non-hate, but the true value of the same is hate), $False_{hate}$ = False-positive (model predicted the texts as hate, but the true value of the same is non-hate), $P_{non-hate}$ = Precision of non-hate class, P_{hate} = Precision of hate class, $R_{non-hate}$ = Recall of non-hate class, R_{hate} = Recall of hate class, $F1_{non-hate}$ = F1 score of non-hate class, $F1_{hate}$ = F1 score of hate class, $T_{non-hate}$ = The total number of non-hate class text present in the test set, T_{hate} = The total number of hate class text present in the test set.

5.1. Analysis of hate speech detection. We evaluated two ML models, three DLs with or without Word2Vec, five transformer-based models on the NEIHS - Bodo dataset, and four variants on the NEIHS - Assamese dataset. mBERT-cased, mBERT-cased, and MahaRoBERTaMarathi are the best performing and

Table 5.1: Precision, Recall, F1 score, and Accuracy of all the models applied on NEIHS-Assamese and Bodo dataset

	Precision			Recall			F1 score			Accuracy
(NEIHS-Bodo)	**non-hate**	**hate**	**weighted**	**non-hate**	**hate**	**weighted**	**non-hate**	**hate**	**weighted**	
NB	0.79	0.64	0.71	0.41	0.91	0.68	0.54	0.75	0.65	0.68
SVM	0.43	0.54	0.49	0.07	**0.93**	0.53	0.12	0.68	0.42	0.53
LSTM with Word2Vec	0.77	0.84	0.80	0.83	0.78	0.80	0.80	0.81	0.81	0.80
LSTM without Word2Vec	0.78	0.85	0.82	**0.84**	0.79	0.81	0.81	0.82	0.81	0.81
BiLSTM with Word2Vec	0.82	0.83	0.83	0.80	0.85	0.84	0.81	0.84	0.83	0.83
BiLSTM without Word2Vec	0.81	0.82	0.82	0.79	0.84	0.82	0.80	0.83	0.82	0.82
CNN with Word2Vec	0.82	0.84	0.83	0.82	0.84	0.83	0.82	0.84	0.83	0.83
CNN without Word2Vec	0.80	0.83	0.81	0.80	0.83	0.81	0.80	0.83	0.81	0.81
mBERT-cased	**0.86**	0.84	**0.85**	0.80	0.89	**0.85**	**0.83**	**0.87**	**0.85**	**0.85**
mBERT-uncased	0.85	0.85	**0.85**	0.82	0.88	**0.85**	**0.83**	0.86	**0.85**	**0.85**
MuRILBERT	0.80	**0.86**	0.83	**0.84**	0.82	0.83	0.82	0.84	0.83	0.83
RoBERTaHindi	0.82	0.84	0.83	0.81	0.85	0.83	0.81	0.84	0.83	0.83
MahaRoBERTaMarathi	**0.86**	0.83	**0.85**	0.79	0.89	**0.85**	0.82	0.86	0.84	**0.85**
(NEIHS-Assamese)										
NB	0.60	0.63	0.61	0.47	0.74	0.61	0.52	0.68	0.61	0.61
SVM	0.50	0.55	0.53	0.00	**1.00**	0.55	0.01	**0.71**	0.39	0.55
LSTM with Word2Vec	0.65	0.69	0.67	0.60	0.73	0.67	0.62	**0.71**	0.67	0.67
LSTM without Word2Vec	0.61	0.70	0.66	0.66	0.66	0.66	0.64	0.68	0.66	0.66
BiLSTM with Word2Vec	0.61	0.69	0.66	0.65	0.66	0.65	0.63	0.67	0.65	0.65
BiLSTM without Word2Vec	0.64	0.70	0.67	0.63	0.71	0.67	0.63	0.70	0.67	0.67
CNN with Word2Vec	0.64	0.69	0.67	0.60	0.72	0.67	0.62	**0.71**	0.67	0.66
CNN without Word2Vec	0.62	0.73	0.68	0.72	0.63	0.67	0.66	0.68	0.67	0.67
mBERT-cased	0.63	**0.76**	0.70	**0.76**	0.63	**0.69**	**0.69**	0.69	**0.69**	**0.69**
mBERT-uncased	0.64	0.73	0.69	0.69	0.68	0.68	0.66	0.70	0.68	0.68
MuRILBERT	**0.70**	0.73	**0.72**	0.63	0.66	0.65	0.67	0.70	**0.69**	**0.69**
BanglaBERT	0.61	0.73	0.67	0.71	0.63	0.67	0.66	0.67	0.67	0.67

best-fitted models on the Bodo dataset, giving the highest weighted F1-score of 85%, which is about 1% to 5% better than other transformer models. At the same time, BiLSTM with Word2Vec, CNN with Word2V, MuRILBERT, and RoBERTaHindi also performed moderately well, i.e., weighted F1-score of 83%. LSTM with without Word2Vec performs only 1% less. We notice that DL and transformer-based models perform almost the same. Both of the ML results are very poor compared to other models. mBERT-cased and MuRILBERT give the best result on the Assamese dataset, giving the highest weighted F1-score of 69%, which is about 1% to 8% better than other transformer models. mBERT-uncased achieved 68% weighted F1-score. At the same time, CNN without Word2Vec, LSTM with Word2Vec, BiLSTM without Word2Vec, and BanglaBERT scored 67%. CNN with Word2Vec, LSTM without Word2Vec and BiLSTM with Word2Vec also performed moderately well. NB and SVM perform very poorly. Figure 5.1 shows the confusion matrix of the best models on two NEIHS datasets separately. As best models predict almost the same, we provide only one confusion matrix for each dataset.

5.2. Error analysis. The results show that the model's performance is not up to the mark in the Assamese dataset but good in the Bodo dataset. However, as a pioneering work in the Assamese and Bodo, it needs an error analysis to identify the system's weakness for further improvement. Here, the error analysis is not performed at a granular level; rather, we have tried to identify the major source of error.

Errors in the preprocessing. A DL model performs very well on the Bodo dataset, but we need to experiment with this model by removing stopwords. We found only 15 stop words from native speakers (Students) but need to gather more if any others are left. Stop words are the most commonly used words. Commonly used words don't mean frequent words in the datasets, as both concepts are slightly different and depend on a particular dataset. Wordpiece tokenizer is checked on the Assamese dataset but found shocking results, and resultant tokens are not meaningful. This sentence ‘’ (English translation: He/ She looks like a Jihadi.) is giving accurate result like [‘’, ‘’, ‘’, ‘’, ‘’, ‘’] but at the same time ‘’ (English translation: They think life is theirs only) is giving [‘’, ‘##’, ‘##’, ‘##’, ‘[UNK]’, ‘’, ‘##’, ‘’, ‘##’, ‘[UNK]’, ‘’, ‘[UNK]’, ‘’], which is an error. This can be one reason for the worst result despite having more Assamese data than Bodo data. So,

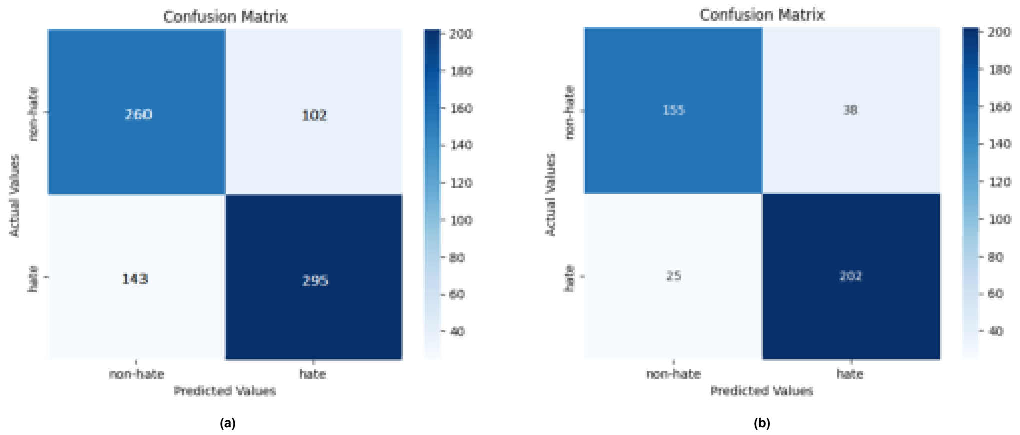


Fig. 5.1: Confusion matrix of the best models on two NEIHS datasets separately, (a) NEIHS-Assamese (best model: MuRILBERT) and (b) NEIHS-Bodo (best model: mBERT-cased)

tokenizer training for Assamese data is required. For both languages, an emoji-to-text Python library would help researchers. We noticed in other languages that converting emojis to text gives 2%-4% better results than the complete removal of emojis from the text. We need a proper lightweight stemming list if the lemmatization of words is complex or expensive. It could give more accurate results while training with DL.

Errors in the model. We perform fine-tuning on transformer-based models that have been pre-trained on either multilingual or monolingual datasets. However, these pre-trained models do not currently incorporate Bodo data. At the same time, Google Muril's dataset includes Assamese among its 17 languages. Due to insufficient raw datasets, we rely on word embedding for the same dataset rather than training a neural word embedding model from scratch.

Errors in the data set. In some instances, confusion surrounding annotation can adversely affect performance. To ensure optimal performance, annotators may need more contextual information. Moreover, more data is required.

Errors in the language phenomenon. Some inherent language problems are difficult to address at the computation level. In the context of hate speech, sometimes the text does not contain any hate-related words but still conveys a piece of hate information. On the other hand, though the text contains hate words, the text is not a hateful sentence. Sometimes, it needed the word sense disambiguation to identify the hateful information.

Errors in prediction. For the NEIHS-Assamese, mBERT-cased and MuRILBERT is the best model based on our analysis. We divide the error cases into the following categories[19]:

1. Implicit hate: Situations where there are no openly abusive words but express a complex thought. ‘’ (English translation: Come, we Assamese people should come together to beat up miyas.). There are no swear words here, but the message is meant to inspire intolerance against a particular religion.
2. Annotator confusion: Annotators get confused about whether the sentence is complement or slang. ‘ ’ (English translation: Your figure is sexy.), this comment is ambiguous. Here, the model prediction cannot be considered incorrect as the comment could be analysed in both ways, relying on cultural perceptiveness.
3. Contextual abuse: Some words are used more often in the abuse samples. Depending on the context, it changes the meaning. ‘’ (English translation: Godi media is buying all the news channels. They ended the freedom of speech in this country.). Here, ‘’ (Godi media) indicates media is supporting a particular person's opinion.

[19]https://openreview.net/forum?id=HCnb1TByvx7 (Access on 02.10.2023)

4. Spelling mistakes: It's common for profanity to have an incorrect spelling on social media, whether on purpose to get through the moderation system or accidentally owing to the relaxed nature of the medium. '' (English translation: We will throw you miya's out of assam.). '' is actually misspelled of ''.

6. Conclusion and future work. This paper describes our contribution, i.e., the NEIHS dataset, which includes Assamese and Bodo datasets, a binary classification task. We performed training NB, SVM, LSTM, BILSTM, and CNN on the NEIHS datasets. Transfer learning is also employed for the task. The best-performing model we found for NEIHS-Assamese is MuRILBERT and mBERT-cased, with a 69% weighted F1-score, where mBERT-cased and mBERT-uncased got 86% on NEIHS-Bodo dataset. Transformer-based models perform sufficiently but are burdened by longer computation times than traditional DL models, where DL models perform a little less in this aspect. As mentioned earlier, due to the lack of Bodo and Assamese tools in the NLP field, we failed to perform experiments like stemming and emoji to text in the Bodo language. In the case of transfer learning, we cannot use a monolingual transformer-based model. Nevertheless, the shortage of extensive datasets poses a significant obstacle to advancing automatic hate speech detection systems for Indic languages, particularly for North East Indian languages. To address this issue, we have contributed the human-annotated NEIHS dataset, which covers multiple languages from a popular social media platform. However, we intend to expand this dataset by collecting additional data with diverse categorical annotations, focusing on multiclass classification. In the future, we aim to incorporate more data from the same languages and other North-Eastern languages such as Meiteilon (spoken by 2 million native speakers), Manipuri (spoken by 1.8 million native speakers), and Mizo (spoken by 830,846 native speakers). Additionally, recognizing the complexity of code-mixed comments, where Indic languages are expressed using Roman characters, we also plan to enhance the NEIHS dataset by including such instances.

7. Acknowledgement. We thank Debarshi Sonowal, Abhilash Basumatary, and Bidisha Gogoi for their help in collecting and tagging the Assamese hate dataset.

REFERENCES

[1] G. W. Giumetti and R. M. Kowalski, *Cyberbullying via social media and well-being, Current Opinion in Psychology*, Volume 45, 2022, 101314, ISSN 2352-250X, https://doi.org/10.1016/j.copsyc.2022.101314.

[2] P. Sarkar, *Media And Politics. Pratishthaa Publishing House*, 2019, https://books.google.co.in/books?id=8HSPDwAAQBAJ.

[3] Z. Laub, *Hate speech on social media: Global comparisons. Council on foreign relations 7* (2019).

[4] M. Alhajji, S. Bass, and Dai, T, *Cyberbullying, Mental Health, and Violence in Adolescents and Associations With Sex and Race: Data From the 2015 Youth Risk Behavior Survey. Global Pediatric Health*, 2019, https://doi.org/10.1177/2333794X19868887.

[5] A. Guterres et al, *United Nations strategy and plan of action on hate speech.* Taken from: https://www. un. org/en/genocideprevention/documents/U 20Strategy (2019).

[6] R. Ramdev, S.D. Nambiar, and D. Bhattacharya, *Sentiment, Politics, Censorship: The State of Hurt. SAGE Publications*, 2015, https://books.google.co.in/books?id=i4GTCwAAQBAJ.

[7] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel, *Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '19). Association for Computing Machinery, New York, NY, USA*, 14–17. https://doi.org/10.1145/3368567.3368584.

[8] T. Mandl, S. Modha, A. K. Jaiswal, and B. R. Chakravarthi, *Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In Forum for Information Retrieval Evaluation (Hyderabad, India) (FIRE 2020). Association for Computing Machinery, New York, NY, USA*, 29–32. https://doi.org/10.1145/3441501.3441517.

[9] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, and M. Zampieri, *Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech. In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '21). Association for Computing Machinery, New York, NY, USA*, 1–3, 2022. https://doi.org/10.1145/3503162.3503176.

[10] R. Kalra and A. K. Dutt, *Exploring Linguistic Diversity in India: A Spatial Analysis. Handbook of the Changing World Language Map* (2019).

[11] M. Narzary, G. Muchahary, M. Brahma, S. Narzary, P. K. Singh, and A. Senapati, *Bodo Resources for NLP-An Overview of Existing Primary Resources for Bodo. AIJR Proceedings* (2021), 96–101.

[12] S. Narzary, M. Brahma, M. Narzary, G. Muchahary, P. Kumar Singh, A. Senapati, S. Nandi, and B. Som, *Generating Monolingual Dataset for Low Resource Language Bodo from old books using Google Keep. In Proceedings of the Thirteenth*

Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 6563–6570, 2022. https://aclanthology.org/2022.lrec-1.705.

[13] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, *Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. Proceedings of the International AAAI Conference on Web and Social Media 12, 1 (Jun. 2018).* https://doi.org/10.1609/icwsm.v12i1.15041.

[14] R. Izsák, *Hate speech and incitement to hatred against minorities in the media. UN Humans Rights Council. Technical Report*, 2015, A/HRC/28/64.

[15] J. Salminen, H. Almerekhi, M. Milenković, S. G. Jung, J. An, H. Kwak, and B. J. Jansen, *Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In Twelfth international AAAI conference on web and social media.*

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* CoRR abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* CoRR abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692.

[18] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.* CoRR abs/1909.11942 (2019). arXiv:1909.11942 http://arxiv.org/abs/1909.11942.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* CoRR abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108.

[20] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. P. Talukdar, *MuRIL: Multilingual Representations for Indian Languages.* CoRR abs/2103.10730 (2021). arXiv:2103.10730 https://arxiv.org/abs/2103.10730.

[21] K. Jain, A. Deshpande, K. Shridhar, F. Laumann, and A. Dash, *Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages*, 2020, arXiv:2011.02323 [cs.CL].

[22] D. Kakwani, A. Kunchukuttan, S. G. Gokul N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948–4961, Online. Association for Computational Linguistics. Online*, 4948–4961. https://doi.org/10.18653/v1/2020.findings-emnlp.445.

[23] R. Joshi, *L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources.* CoRR abs/2202.01159 (2022). arXiv:2202.01159 https://arxiv.org/abs/2202.01159.

[24] R. Joshi, *L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources. In Proceedings of The WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France*, 97–101, 2022.

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, *Unsupervised Cross-lingual Representation Learning at Scale.* CoRR abs/1911.02116 (2019). arXiv:1911.02116 http://arxiv.org/abs/1911.02116

[26] S. Sarker, *BanglaBERT: Bengali Mask Language Model for Bengali Language Understading*, 2020, https://github.com/sagorbrur/bangla-bert

[27] A. Sharma, A. Kabra, and M. Jain, *Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. Information Processing Management* 59, 1 (2022), 102760. https://doi.org/10.1016/j.ipm.2021.102760.

[28] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, and J. P. McCrae, *Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments.* CoRR abs/2109.00227 (2021). arXiv:2109.00227 https://arxiv.org/abs/2109.00227.

[29] Md. A. Bashar and R. Nayak, *QutNocturnal@HASOC'19: CNN for Hate Speech and Offensive Content Identification in Hindi Language. CoRR abs/2008.12448 (2020).* arXiv:2008.12448 https://arxiv.org/abs/2008.12448.

[30] J. C. Mensonides, P. A. Jean, A. Tchechmedjiev, and S. Harispe, *IMT mines ales at HASOC 2019: automatic hate speech detection. In FIRE 2019-11th Forum for Information Retrieval Evaluation, Vol.* 2517, p–279, 2019.

[31] S. Mishra and S. Mishra,*3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages. In FIRE (Working Notes).* 208–213, 2019.

[32] V. Mujadia, P. Mishra, and D. M. Sharma, *IIIT-Hyderabad at HASOC 2019: Hate Speech Detection.. In FIRE (Working Notes)*, 271–278, 2019.

[33] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences* 55, 1 (1997), 119–139.

[34] R. Raj, S. Srivastava, and S. Saumya. *"NSIT & IIITDWD@ HASOC 2020: Deep learning model for hate-speech identification in Indo-European languages." In FIRE (Working Notes), pp.* 161-167. 2020.

[35] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, and A. Dash, *One to rule them all: Towards Joint Indic Language Hate Speech Detection.* CoRR abs/2109.13711 (2021). arXiv:2109.13711 https://arxiv.org/abs/2109.13711.

[36] M. Bhardwaj, Md S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, *Hostility detection dataset in Hindi.* arXiv preprint arXiv:2011.03588 (2020).

[37] K. Ghosh and Dr. A. Senapati, *Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, pages 853–865, 2022, Manila, Philippines. De La Salle University.*https://www.paclic2022.net/papers.html.

[38] J. Sarmah, S. K. Sarma, and A. K. Barman, *Development of Assamese Rule based Stemmer using WordNet. In Proceedings of the 10th Global Wordnet Conference. Global Wordnet Association, Wroclaw, Poland*, 135–139, 2019.

https://aclanthology.org/2019.gwc-1.17.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates*, Inc.https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[40] A. Khatua, and E. Cambria,*A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. Information Processing Management*, 56, 1 (2019), 247–257. https://doi.org/10.1016/j.ipm.2018.10.010

[41] S. Hochreiter, and J. Schmidhuber; *Long Short-Term Memory. Neural Computation* 1997; 9 (8): 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735.

[42] V. Nair, and G.E. Hinton; *Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp.* 807–814 (2010)

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 31, 2023
Accepted: Oct 5, 2023

1.5em 0pt

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 957–969, DOI 10.12694/scpe.v24i4.2470

DETECTING ACADEMIC AFFECTIVE STATES OF LEARNERS IN ONLINE LEARNING ENVIRONMENTS USING DEEP TRANSFER LEARNING

PURUSHOTTAMA RAO K *AND B JANET †

Abstract. Online Learning Environments (OLEs) have become essential in global education, especially during and after the COVID-19 pandemic. However, OLEs face a challenge in recognizing student emotions, hindering educators' ability to provide effective support. To address this issue, researchers emphasize the importance of a balanced dataset and a precise model for academic emotion detection in OLEs. However, the widely-used DAiSEE dataset is imbalanced and contains videos captured in well-lit environments. However, real-time observations reveal students' diverse lighting conditions and proximity to cameras. Consequently, models trained on DAiSEE dataset exhibit poor accuracy. In response, this work suggests a customized DAiSEE dataset and proposes the Xception-based transfer learned model and AffectXception model. Our customization process involves selectively extracting single-label frames with intensity levels 2 or 3 from the original DAiSEE dataset. To enhance dataset diversity and tackle the issue of dataset imbalance, we meticulously apply data augmentation techniques on these extracted frames. This results in frames that showcase variations in lighting, both low and high, as well as diverse camera perspectives. As a result, the customized DAiSEE dataset is now well-balanced and exceptionally suitable for training deep learning models to detect academic emotions in online learners. Then we trained and tested both proposed models on this dataset. The AffectXception model outperforms existing models, achieving significant improvements. For Boredom, Engagement, Confusion, and Frustration, it attains accuracy rates of 77%, 79.28%, 83.76%, and 91.87%, respectively. Additionally, we evaluate the AffectXception model on the Online Learning Spontaneous Facial Expression Database (OL-SFED), obtaining competitive results across various emotion classes. This work empowers educators to adjust their content and delivery methods based on learners' emotional states, resulting in more effective and informative online sessions. As OLEs continue to play a crucial role in education, our approach enhances their capacity to address students' emotional needs.

Key words: Academic Affective States, Affective Computing, Deep Learning, Fine-tuning, Online Learning Environment, Transfer Learning

1. Introduction. Education is crucial for the overall development of an individual and plays a vital role in shaping up their future. It provides individuals with the necessary knowledge, skills, values, and attitudes to succeed in life [1, 2]. The classroom is a crucial setting for education as it provides a structured environment for learning and interaction with teachers and peers. It also plays a vital role in promoting critical thinking, problem-solving, and collaboration skills [3, 4]. So it is the responsibility of the learners to be engaged and attentive in the class time to gain more knowledge and to achieve course outcomes [5]. In the class time, students show positive expressions or emotions such as attention, engagement, and understanding to indicate that he or she is comprehending the material. They express Negative emotions such as confusion, frustration, or boredom to indicate that the student is struggling and may need additional support or clarification [6, 7]. Table 1.1 gives the cases when students deliver either positive or negative emotions. Therefore it concludes that emotions play a crucial role in the learning process and it is important for a teacher to understand the emotions of their students to make the class effective and to improve the academic performance of learners [8, 9]. A seasoned teacher is capable of identifying the understanding level of all students by observing their affective states during the class in traditional classroom environment and it is one of the main reasons for the success of these offline classroom environments [10]. Teachers can use this emotion feedback to make adjustments such as slowing down or speeding up the pace, repeating the subject, or changing their way of delivering the concept

*Department of Computer Applications, National Institute of Technology Tiruchirappalli, Tamilnadu - 620015, India. (k.purushottamaa@gmail.com). Questions, comments, or corrections to this document may be directed to that email address.

†Department of Computer Applications, National Institute of Technology Tiruchirappalli, Tamilnadu - 620015, India. (janet@nitt.edu).

by using innovative teaching methods and active based methods, which all assist to keep the session interesting and lively [4].

Table 1.1: Cases when students deliver either positive or negative emotions

Students express Positive emotions when they:	Students express Negative emotions when they:
Adhere to the lecture	Cannot keep up with the lecture
Satisfied with the lecture	Confused and want the instructor to repeat it
Capable of understanding the lecturer's thoughts	Try to get the lecturer's assistance
Would like to emphasize how well the lecture was received	Unable to keep up with the lecturer's pace

However, with the COVID-19 epidemic and also the enrichment and development of Internet technology the learning environment has shifted from purely traditional classroom to a hybrid learning environment (HLE) [11, 12]. Online learning environments (OLEs) have made it possible for students all around the world to pursue their education during and after the COVID-19 pandemic. Google Meet, WebEx, MS Teams, Zoom are some of the examples for online learning environments. OLEs are widely accessible, adaptive, and versatile due to their non-physical nature [12, 13]. Though these are having some benefits over offline classrooms, E-learning, unlike traditional classrooms, lacks the ability to capture students' emotions and dynamically take steps to improve their engagement. It can be challenging for teachers using an e-learning system to identify students' emotions and issues, particularly those brought on by uncertainty or apathy. With the help of prior research, it has been determined that e-learning systems typically ignore the Emotion, a crucial part of the learning process [14, 15]. As a result, even the important lectures or courses frequently results in the student's disinterest. Furthermore, during lectures, e-learning students have been shown to feel distinct negative emotional states such as apathy, indifference, drowsiness, and confusion [16]. As a result, the instructor's and instructional aids' efforts are rendered ineffective [17].

It is vital for online educators to accurately and efficiently detect their online learners' engagement status in order to provide personalized pedagogical support. However, teachers are finding it difficult to gauge pupils' engagement without having direct interaction or eye contact with them. As a result, a model that detects students' affective states in virtual classes is needed. The primary contributions of this paper are: (1) We create a customized DAiSEE dataset, by extracting single-label frames (frames that only belong to one class) from original DAiSEE dataset [18] and applied data augmentation techniques to make the new dataset balanced; (2) Using the customized DAiSEE dataset and the OL-SFED dataset [33, 34], which are more appropriate for academic affective states, we evaluate the suggested AffectXception model in detecting academic emotions; (3) and we compare the proposed model's performance against that of existing methods on original DAiSEE dataset. Our obtained results demonstrate that, compared to previous efforts, our technique classifies different academic emotional states more accurately.

The remaining portions of the paper are organized as follows: The background and a review of relevant existing literature are provided in Section 2. The details of the dataset modified and the model's training procedure for identifying students' academic affective states are presented in Section 3. Section 4 of the paper details the experiments and their results. The last section summarizes the findings and potential directions for future research.

2. Related Work. The works on (1) Correlation between Students' Emotions and their Engagement, (2) Students' Emotion Recognition in E-Learning Platforms, and (3) Academic Affective States are reviewed in this section.

According to existing studies, there is a positive correlation between learners' emotional status, their attention and their academic performance. According to the existing research, ensuring learner engagement is one of the most important aspects of quality online education [19]. As per [20], one of the parameters to determine students' engagement during lecture is their emotions. Emotions can have a significant impact on a student's engagement and academic achievement by influencing their interest and motivation in a lecture or course [21, 22]. It has been demonstrated that pleasant feelings such as happiness and neutrality can have favorable benefits on students, whereas unpleasant emotions such as sadness, anger, and boredom can have detrimental consequences [21]. Authors [23] demonstrated the importance of knowing emotional feedback to both students and teachers in another study. The authors also show that students are more motivated, engaged, and perform well in assessments while experiencing pleasant emotions. Students, on the other hand, are unmotivated, disengaged, and underperform on assessments with negative emotions.

Previous studies have proved that there is a need to recognize the students' emotions in online classrooms in order to find out the comprehension level of both individual student and the overall classroom. One of the primary means of recognizing and interpreting human emotions is facial expressions [24]. In order to detect the emotional states of students during lessons, majority of researchers have utilized seven basic emotions namely happy, anger, fear, sadness, joy, disgust, and surprise [5, 25]. The authors used deep learning-based algorithms to detect student involvement by detecting at their facial expressions. Students expressing 'Sad' emotions were found to be the least interested, while those displaying 'Happy' and 'Neutral' emotions were found to be the most involved [25]. Authors [26] developed a framework for evaluating students' emotions using facial expressions in an online lecture. They have presented the overall class feedback towards the lecture based on students' emotions. Authors have proposed a CNN based model to detect the emotions of students in E-learning platforms. Then they have used another module in their proposed architecture to send this emotional feedback of students during the class to the corresponding teacher that helps to engage the class in a more productive way. Authors have used single face images in their proposed work [15]. By combining global and local face features, [27] established a framework called Multi-region Attention Transformation to recognize facial expressions.

Few scholars have stated that the fundamental emotions for human being play a minimal role when we want to detect the emotional state of learners in educational settings [28]. Authors found that non-basic emotions including engagement, boredom, confusion, and frustration happened five times more frequently than basic emotions in classroom. Based on the results it was suggested that these non-basic affective states are more suitable for detecting learners' academic affective states [29]. In another work [30], authors have also stated that these affective academic states are more relevant than basic emotions to detect students' emotional state. In another work [20], authors proved that the predominant emotions experienced by students were curiosity, frustration, boredom, confusion, happiness, and anxiety, with contempt, anger, sadness, fear, disgust, eureka, and surprise being less frequently reported. Therefore, academic emotion is a term used to describe a variety of emotional reactions that students experience in academic activities like learning [32].

It is important to have a relevant dataset in order build a robust model that meets our requirements. But unfortunately there is a lack of academic affective states datasets. Then "Dataset for Affective States in E-Environments" (DAiSEE), a dataset created by [18], was released in 2015. It is the first multi-label video classification dataset and contains 9068 video clips from 112 users. It was created to identify the affective states of boredom, confusion, engagement, and frustration in users "in the wild." Each of the four affective states in the dataset has a label at one of four different levels: very low, low, high, and very high. They have used InceptionNet, C3D, C3D with fine tuning, LRCN and EmotionNet models to analyse and give benchmark results for DAiSEE dataset. Results have showed that there is a huge scope to get even better accuracies. An online learning spontaneous facial expression database (OL-SFED) [33, 34] for academic emotions is created in another study to address the lack of academic emotions datasets. This dataset contains 30,184 images and 1274 videos with 224 x 224 and 1280 x 720 resolutions, respectively. Authors have used different CNN models to give benchmark results on the created dataset.

Authors [35] mentioned the difficulties online teachers face in identifying the comprehension level of students during the lecture. He proposed a model to find the facial expressions including Enlightened, Confused and Bored expressions of students and a method to transfer this emotional feedback to the corresponding teacher. Another work uses a deep learning system to evaluate academic emotions of distance learners using data from online learning behavior. The multimodal weighted feature fusion algorithm is applied to extract the data, and an academic cognition motivation model and online learning emotion measurement framework are constructed. This study finds a positive correlation between distance learners' academic emotions and their learning outcomes [7]. To assist teachers in real-time monitoring of student engagement, authors suggested a framework for an academic emotions monitoring system. In order to train a CNN model, they built a dataset containing six academic emotional states from classroom lecture videos and incorporated relevant samples from the BAUM-1, DAiSEE, and YawDD databases in the dataset. Their approach analyzed student facial expressions captured

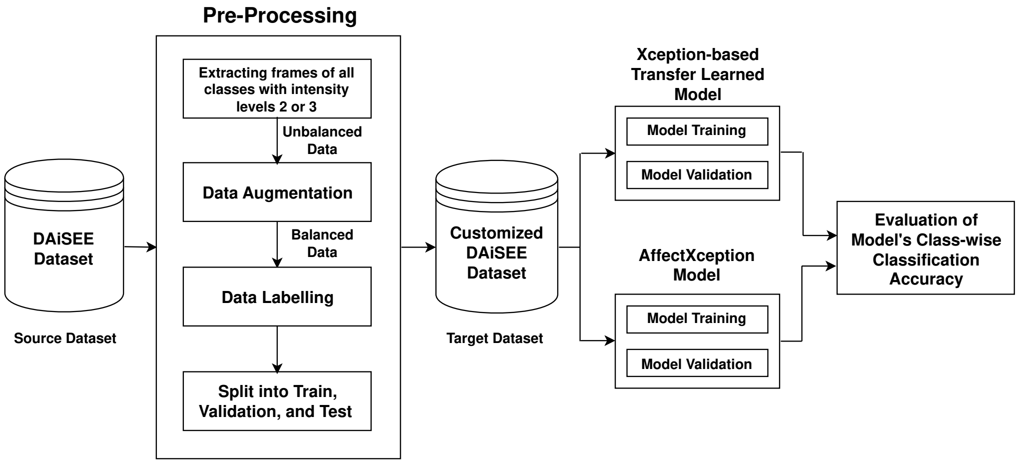


Fig. 3.1: Workflow of Proposed method

in lecture recordings to evaluate the level of participation of the entire class [36]. The authors [37] opted for academic affective states instead of basic emotions due to their significance in detecting students' involvement in videos in learning settings. They employed Conv3D, VGG16, ResNet50, and LSTM models, and trained them with the DAiSEE dataset. Results showed that Conv3D has achieved better results than the remaining models. The authors in [38] proposed a method for automatically estimating student engagement in online class by analyzing videos. This approach took into account facial expressions, head pose, and gaze movements to determine engagement levels. The authors suggested a model that predicts engagement through an LSTM network, trained using the EmotiW 2019 challenge dataset.

3. Proposed Work. In this section, we will discuss the issues with the existing dataset, creation of target dataset, and the workflow of our proposed method. Figure 3.1 shows the architecture diagram of the proposed work.

3.1. Dataset. In this work, we utilized the publicly accessible dataset named "Dataset for Affective States in E-Environments" (DAiSEE) [18] to train and evaluate our deep learning model. As its name implies, this dataset is well-suited for training models to detect affective states in e-learning environments. It includes 9068 labelled video clips of size 15 GB featuring 112 students participating in online lectures while seated in front of a webcam. The dataset encompasses four ideal classes for our work: Boredom, Engagement, Confusion, and Frustration, which are frequently expressed by learners during lectures. Each affective state is labelled with one of four intensity levels: 0 (very low), 1 (low), 2 (high), and 3 (very high), reflecting the observed level of emotion in the given video.

Based on the literature review, it has been discovered that the majority of researchers focusing on academic affective states utilized the DAiSEE dataset to train their models in order to determine student engagement in learning environments [5, 6, 8, 9, 12, 36, 37]. A commonly noted challenge among researchers working with the DAiSEE dataset is the issue of dataset imbalance [10, 14, 30, 37]. Furthermore, a common limitation observed in existing works that utilized the DAiSEE dataset is the accuracy of the models employed.

3.2. Pre-processing for Customized DAiSEE Dataset Creation. The pre-processing procedure seeks to enhance the data's quality and produce customised datasets from the original dataset in accordance with requirements in order to get better outcomes [39, 40, 41]. Our aim in this work is to use single-label classification method to improve the robustness of the deep learning model and attain higher accuracy. The dataset includes two categories of frames: single-label frames (frames that belong to only one class, such as [0, 2, 0, 0]), and multi-label frames (frames that belong to multiple classes, such as [2, 1, 0, 1]). See Fig. 3.2 for sample multi-label images in the dataset. To apply single-label classification method, we have only considered

Fig. 3.2: Sample multi-label images with annotations

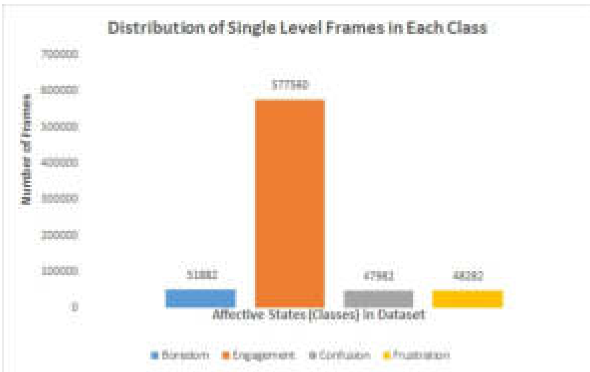


Fig. 3.3: Number of Single-label Frames in Each Class in DAiSEE Dataset

single-label frames that have high (2) or very high (3) intensity levels. Two diagrams Fig. 3.3 and Fig. 3.4 below present the number of single-label frames and single-label frames with high (2) or very high (3) intensity levels per class in the dataset.

It is evident from the information above that the DAiSEE dataset is significantly unbalanced. If a deep learning model is trained with the current unbalanced dataset, it can lead to biased results. This means that the model may perform well on the majority class but poorly on the minority class, as it has seen more examples of the majority class during training. This can result in poor accuracy or incorrect predictions for the minority class, and the overall performance of the model will be affected. Therefore, a balanced dataset is crucial for training an effective model, which requires presenting the model with roughly equal amounts of data samples for each class.

This work's important module is determining what kind of data must be extracted from the videos in order to have related and significant features for the detection of academic affective states in learning contexts. First we have extracted the frames with intensity level High (2) or Very High (3) of each class. We have noticed that most of the participants in online class videos are either in bright or dim environments, and are either up close or far away from the camera. See below Fig. 3.5 for reference.

So we have chosen and applied Brightness, Zoom, and Horizontal Flip data augmentation techniques on available images of classes Boredom, Confusion, and Frustration to make sure our new customized dataset

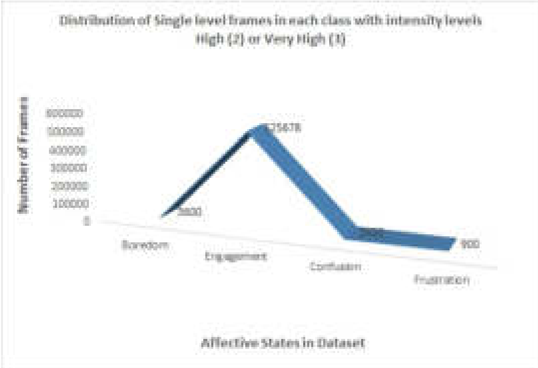


Fig. 3.4: Number of Single-label Frames in Each Class with intensity levels High (2) or Very High (3) in DAiSEE Dataset

Fig. 3.5: Samples to show online class participants in different conditions. (a) and (b) in high light environment; (c) and (d) shows in low light; (e) and (f) are far away from camera; and (g) and (h) are close to the camera

contains the images with above said conditions and to make the dataset balanced. And we have applied Under-sampling technique on Engagement class since it is having huge number data samples. After data augmentation we have partitioned data into Train, Validation, and Test with 70:20:10 ratio. The number of frames in each of the four classes after balancing the dataset is given in below Table 3.1. There are 40,320 images in the training data where each class contains 10,080 images, 11,520 images in the validation data, and 5,760 images in the test data. Then we relabeled the dataset with values 1 for the specific class and 0 for all remaining classes.

The algorithm 5 for creating customized DAiSEE dataset from the original DAiSEE dataset is given below:

3.3. Methodology. Here, authors used Xception [42], a deep learning model for image classification, to detect learners' academic affective states in online classrooms. The Xception model uses depth wise separable

Table 3.1: Class-wise distribution of customized DAiSEE dataset samples into train, validation, and test sets

Class	Train	Validation	Test	Total
Boredom	10080	2880	1440	14400
Engagement	10080	2880	1440	14400
Confusion	10080	2880	1440	14400
Frustration	10080	2880	1440	14400

Algorithm 5 Customized DAiSEE Dataset

dataset = load_DAiSEE_dataset() Load the DAiSEE dataset frames = convert_videos_to_frames(dataset) Convert the videos in the dataset into frames selected_frames = select_frames_classwise_with_levels_2_or_3(frames) Select class wise frames with levels 2 or 3 Class_distribution = count_classwise_frames(selected_frames) Find the class distribution from selected_frames augmented_frames = apply_data_augmentation(frames, techniques=[”brightness”, ”zoom”, ”horizontal_flip”]) Apply data augmentation techniques to address data imbalance train_data, validation_data, test_data = split_dataset(augmented_frames, ratio=(0.7, 0.2, 0.1)) Split the selected frames into train, validation, and test sets with an 70:20:10 ratio train_labels = label_frames(train_data) Label the frames in the train set validation_labels = label_frames(validation_data) Label the frames in the validation set test_labels = label_frames(test_data) Label the frames in the test set

convolutions to learn complex representations of input data while reducing computational cost by applying a single filter to each channel of the input feature map. The model also includes residual connections, global average pooling, and fully connected layers. The Xception model consists of an input layer, entry flow (Fig. 3.6), middle flow (Fig. 3.7), exit flow (Fig. 3.8), and output layer, with each flow containing a series of depth wise separable convolutions, activation functions, and residual connections. The output layer is a fully connected layer that outputs the predictions for the task at hand using the Softmax activation function.

4. Experimental Results. In order to assess Xception-based transfer learning model and AffectXception model performance on a customized DAiSEE dataset, an experimental study of the suggested approach is described in this section.

4.1. Xception-Based Transfer Learned Model. We used the four-class customized dataset mentioned above to train the Xception deep learning model to predict learners' affective states. Figure 4.1 shows the workflow of Xception based transfer learned model. There are 40,320 images in the training data where each class contains 10,080 images. The input dimension was 299 × 299 × 3, where 299 × 299 represents the image resolution, and 3 represents the three channels of RGB color image. First we have trained the lower layer of the model by freezing the remaining layers with the specified parameters. The model is trained with 10 epochs by setting a learning rate of 0.0001 to learn the parameters. To train the model authors have used the loss function by using Eq. 4.1:

$$Loss = \Sigma_1^C log(f(s)_i) \tag{4.1}$$

where C denotes the total number of classes, y_i represents ground truth value, and $f(s)_i$ indicates predicted value.

A model is generated following the training procedure and is predicted to be able to categorize images of test data. The metric we have used to assess the model performance is Accuracy. Refer Eq. 4.2. Accuracy is defined as the number of correct predictions made by the model as a percentage of the total number of predictions.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{4.2}$$

Class wise results obtained for Xception-based transfer learning model are given in Table 4.1 below. Please refer Table 4.3 for a comparison of these results with prior works.

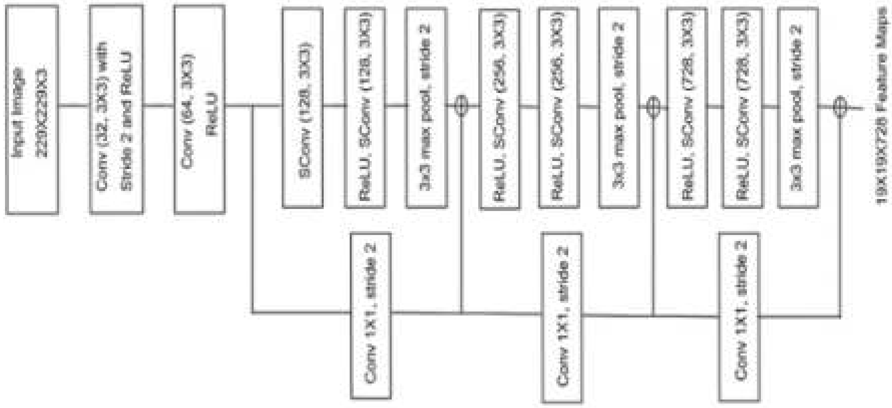


Fig. 3.6: Architecture diagram of Entry Flow

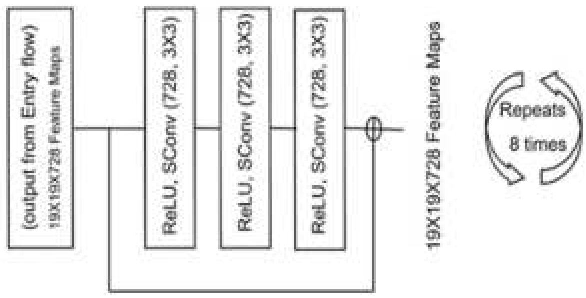


Fig. 3.7: Architecture diagram of Middle Flow

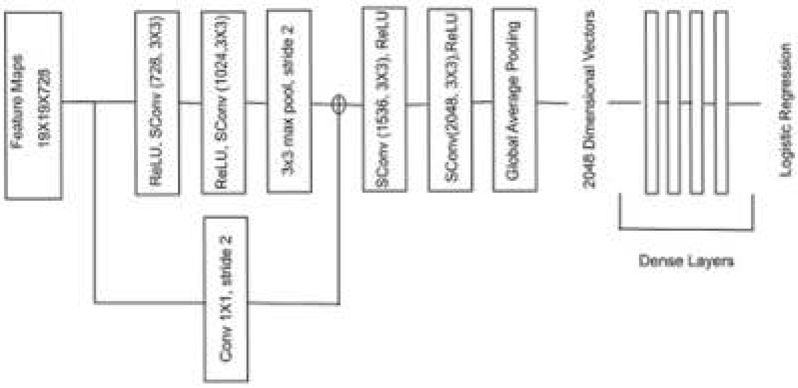


Fig. 3.8: Architecture diagram of Modified Exit Flow

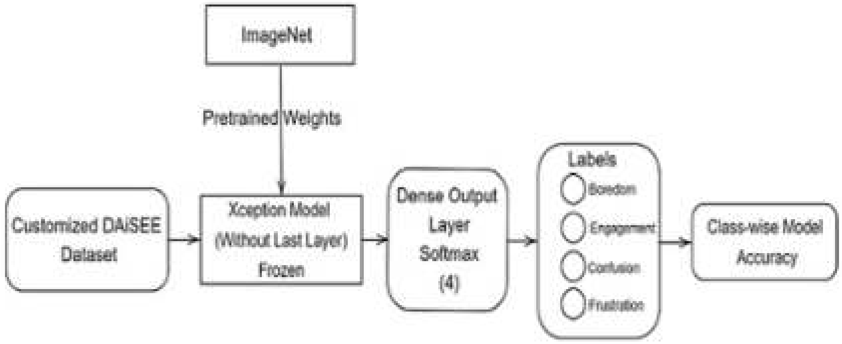


Fig. 4.1: Workflow of Xception-based transfer learned model

Table 4.1: Class-wise accuracies of Xception-based Transfer Learned Model

Class Name	Obtained Class-wise Accuracy
Boredom	72%
Engagement	75.6%
Confusion	75%
Frustration	68.38%

4.2. AffectXception Model. By seeing the above unimpressive results of the Xception-based transfer learning model, we thought that while it is possible to employ a pre-trained deep learning model like Xception trained on the ImageNet dataset for face emotion recognition, doing so may not be the best option. The reason is that the ImageNet dataset is designed for object recognition, and the categories it contains (such as dogs, cats, and cars) may not necessarily be relevant for academic emotion detection. Therefore, a model trained on ImageNet may not be optimized to detect affective states of a person or learner. Affective computing requires a model that is trained specifically on facial images datasets like fer-2013, AffectNet datasets for basic emotions, and DAiSEE, OL-SFED datasets for academic affective states. Training a deep learning model on datasets like DAiSEE, OL-SFED will result in a more accurate and specialized academic affective states detection model as it will have learned to recognize the specific features that are indicative of different emotions of learners during lectures. So, we chose fine tuning as a solution to this problem and to achieve better results. In this instance, we unfreeze the entire model and retrain it using the new customized DAiSEE dataset with a learning rate of 0.00001 and 20 epochs rather than freezing all the layers except last one. Results obtained with the AffectXception model are given below. Table 4.2 gives the obtained class-wise accuracies of AffectXception model. Figure 4.2 shows the comparison of both the models' results.

4.3. Comparison with Existing Methods. With the AffectXception model trained on our customized DAiSEE dataset, we got the better results over the existing methods to detect academic affective states in E-learning platforms on DAiSEE dataset. The results and comparison of the proposed AffectXception model trained on customized DAiSEE dataset with the most recent state-of-the-art approaches are shown in the Table 4.3 below. Here, we utilized Accuracy as a metric for comparing our results to prior studies.

4.4. Performance Analysis of OL-SFED Dataset. The proposed AffectXception model's performance is also assessed on Online Learning Spontaneous Facial Expression Database (OL-SFED) [33, 34]. A total of 30,184 facial expression images with a resolution of 224X224 and 1274 videos with a resolution of 1280X720 are included in the OL-SFED. There were 82 healthy students who voluntarily took part in the experiment, including 53 female and 29 male students. Five significant academic emotions have been chosen in the database

Table 4.2: Class-wise accuracies of AffectXception model

Class Name	Obtained Class-wise Accuracy
Boredom	77%
Engagement	79.28%
Confusion	83.76%
Frustration	91.87%

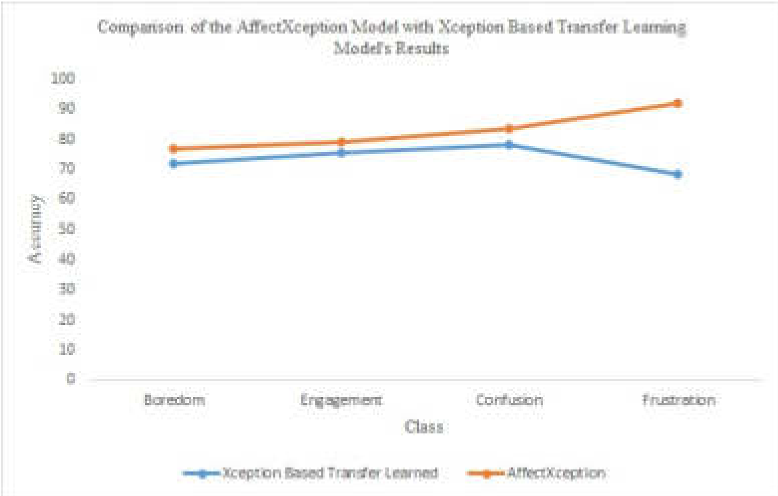


Fig. 4.2: Comparison of Xception-Based Transfer learned and AffectXception models results

as labels for facial expressions: Confusion, Distraction, Happy, Neutral, and Tired. The details of the number of images that each class contains is given in the below Fig. 4.3.

The above figure makes it clear that the OL-SFED dataset is considerably imbalanced. But it is required to have a balanced dataset if we train an efficient model to detect academic affective states of students' in online learning environments. And as seen in Fig. 4, most participants in online class recordings are either in bright or dark settings, and they are either close to or far from the camera. To make sure our dataset comprises of the images with the aforementioned conditions and to ensure the dataset is balanced, we have chosen and used Brightness, Zoom In, Zoom Out, and Horizontal Flip data augmentation techniques on available photographs of each class. The number of images after applying data augmentation techniques on OL-SFED dataset are given in Table 4.4.

Then we have trained the AffectXception model with a learning rate of 0.00001, batch size of 32, and with 20 epochs. Class-wise results obtained by AffectXception model on OL-SFED dataset are given Table 4.5 below.

5. Conclusion. Because teachers can analyze students' facial expressions to determine how attentive they are and change their teaching style appropriately, traditional classrooms continue to be a successful educational setting. The accessibility, affordability, and adaptability of online learning, particularly in the wake of the Covid-19 outbreak, have increased its popularity. One significant distinction between traditional classroom learning and online learning which is hybrid now is that during virtual sessions, teachers are unable to gauge participants' levels of participation in real-time. As a result, it could be more challenging to tell whether or not students are paying attention.

In this work, we have worked on Dataset for Affective States in E-Environments (DAiSEE) to build an effective model that can detect the academic affective states of online learning learners. We have customized the dataset by extracting the frames those represents a single emotion strongly. Then we have applied Brightness,

Table 4.3: Performance comparison with existing methods. Best results are given in bold

Method	Academic Affective States			
	Boredom	Engagement	Confusion	Frustration
DenseAttNet [5]	54.27%	63.59%	69.22%	78.58%
GaussianNB Classifier [30]	69%	**84%**	81%	84%
InceptionNet Frame Level [18]	36.5%	47.1%	70.3%	78.3%
InceptionNet Video Level [18]	32.3%	46.4%	66.3%	77.3%
C3D [18]	47.2%	48.6%	67.9%	78.3%
C3D Fine Tuning [18]	45.2%	56.1%	66.3%	79.1%
LRCN [18]	53.7%	57.9%	72.3%	73.5%
Multi-level Classification Approach [43]	63.82%	77.28%	81.95%	86.6%
Xception-based Transfer Learned Model (Ours)	72%	75.6%	75%	68.38%
AffectXception (Ours)	**77%**	79.28%	**83.76%**	**91.87%**

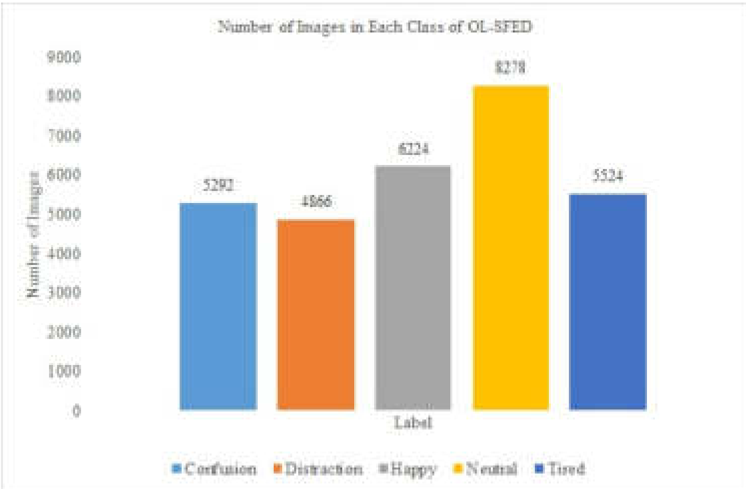


Fig. 4.3: Number of Images in each academic emotion label of OL-SFED

Table 4.4: Class-wise distribution of balanced OL-SFED images into train, validation, and test sets

Class	Train	Validation	Test	Total
Confusion	10500	3000	1500	15000
Distraction	10500	3000	1500	15000
Happy	10500	3000	1500	15000
Neutral	10500	3000	1500	15000
Tired	10500	3000	1500	15000

Table 4.5: Class-wise accuracies of AffectXception model on OL-SFED

Class Name	Obtained Class-wise Accuracy
Confusion	80.01%
Distraction	80%
Happy	80.24%
Neutral	93.01%
Frustration	70.2%

Zoom In, Zoom Out, and Horizontal Flip data augmentation techniques on extracted frames to make the dataset balanced and more suitable for academic emotion detection. To identify the academic affective states of learners in online learning environments, we trained both the Xception based transfer learned model and the AffectXception model on a bespoke DAiSEE dataset. In terms of class-wise performance, the AffectXception model outperformed the Xception-based transfer learnt model and cutting-edge techniques. Findings from the AffectXception model for boredom, engagement, confusion, and frustration are 77%, 79.28%, 83.76%, and 91.87%, respectively. The performance of the AffectXception model on the online learning spontaneous facial expression database was then examined (OL-SFED). For the classes of Confusion, Distraction, Happy, Neutral, and Tired, the model has scored 80.01%, 80%, 80.24%, 93.01%, and 70.2%, respectively.

Though the obtained results are better than existing works on DAiSEE dataset, there is still a chance to improve the results. Future work includes developing comprehensive academic emotions dataset, detecting individual student academic affective states during class, and applying our learners' emotion detection models with edge devices in real-time classroom scenarios.

REFERENCES

[1] Raghupathi, Viju, and Wullianallur Raghupathi, *The influence of education on health: an empirical assessment of OECD countries for the period 1995–2015*, Archives of Public Health, vol. 78, no. 1, pp. 1-18, 2020.

[2] Idris, Fazilah, Zaharah Hassan, Azizah Ya'acob, Saran Kaur Gill, and Noor Aziah Mohd Awal, *The role of education in shaping youth's national identity*, Procedia-Social and Behavioral Sciences, vol. 59, pp. 443-450, 2012.

[3] Rao, K. Purushottama, and B. Janet, *Teacher Assistance System to Detect Distracted Students in Online Classroom Environment*, 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2022, pp. 1743-1749.

[4] Umarale, Dhaval, Shyam Sodhani, Amodh Akhelikar, and Reeta Koshy, *Attention Detection of Participants during Digital Learning Sessions using Edge Computing*, International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020), 2020.

[5] Mehta, Naval Kishore, Shyam Sunder Prasad, Sumeet Saurav, Ravi Saini, and Sanjay Singh, *Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement*, Applied Intelligence, vol. 52, no. 12, pp. 13803-13823, 2022.

[6] Thiruthuvanathan, M., Balachandran Krishnan, and M. A. D. Rangaswamy, *Engagement detection through facial emotional recognition using a shallow residual convolutional neural networks*, International Journal of Intelligent Engineering and Systems, vol. 14, no. 2, 2021.

[7] Sun, Yang, and Shaoze Wang, *Study on the Distance Learners' Academic Emotions Using Online Learning Behavior Data*, Mobile Information Systems, vol. 2022, 2022.

[8] Hu, Yan, Zeting Jiang, and Kaicheng Zhu, *An Optimized CNN Model for Engagement Recognition in an E-Learning Environment*, Applied Sciences, vol. 12, no. 16, pp. 8007, 2022.

[9] Batra, Sarthak, Hewei Wang, Avishek Nag, Philippe Brodeur, Marianne Checkley, Annette Klinkert, and Soumyabrata Dev, *DMCNet: Diversified model combination network for understanding engagement from video screengrabs*, Systems and Soft Computing, vol. 4, pp. 200039, 2022.

[10] Ai, Xusheng, Victor S. Sheng, and Chunhua Li, *Class-attention Video Transformer for Engagement Intensity Prediction*, 2022, arXiv preprint arXiv:2208.07216.

[11] Samuel Kolawole Ajiboye, David Obafemi Adebayo, Saidu Abubakar, *Teachers' Assessment of Inattentive Classroom Behavior Among Primary School Students in Ilorin City, Kwara State, Nigeria*, Mimbar Sekolah Dasar, vol. 7, no. 2, pp. 172-183, 2020, doi: 10.17509/ mimbar-sd.v7i2.24101

[12] Ma, Xiaoyang, Min Xu, Yao Dong, and Zhong Sun, *Automatic student engagement in online learning environment based on neural turing machine*, International Journal of Information and Education Technology, vol. 11, no. 3, pp. 107-111, 2021.

[13] Abedi, Ali, Chinchu Thomas, Dinesh Babu Jayagopi, and Shehroz S. Khan, *Bag of States: A Non-sequential Approach to Video-based Engagement Measurement*, 2023, arXiv preprint arXiv:2301.06730.

[14] Abedi, Ali, and Shehroz S. Khan, *Detecting disengagement in virtual learning as an anomaly*, 2022, arXiv preprint arXiv:2211.06870.

[15] Sun, Ai, Ying-Jian Li, Yueh-Min Huang, and Qiong Li, *Using facial expression to detect emotion in e-learning system: A deep learning method*, Emerging Technologies for Education: Second International Symposium, SETE 2017, Held in Conjunction with ICWL 2017, Cape Town, South Africa, September 20–22, 2017, Revised Selected Papers 2, pp. 446-455. Springer International Publishing, 2017.

[16] M. Ali Akber Dewan, Mahbub Murshed and Fuhua Lin, *Engagement detection in online learning: a review*, Smart Learning Environments, vol. 6, no. 1, pp. 1-20, 2019, https://doi.org/10.1186/s40561-018-0080-z.

[17] Krithika.L.B, Lakshmi Priya GG, *Student Emotion Recognition System (SERS) for e-learning improvement based on learner concentration metric*, International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science, vol. 85, pp. 767-776, 2016.

[18] Gupta, A., A. D'Cunha, K. Awasthi, and V. Balasubramanian, *Daisee: Towards user engagement recognition in the wild*, 2016, arXiv preprint arXiv:1609.01885.

[19] Yousra Banoor Rajabalee, Mohammad Issack Santally, Frank Rennie, *A study of the relationship between students' engagement and their academic performances in an eLearning environment*, E-Learning and Digital Media, vol. 17, no. 1, pp. 1–20, 2020, doi: 10.1177/2042753019882567.

[20] Dixson, Marcia D., *Creating effective student engagement in online courses: What do students find engaging?*, Journal of the Scholarship of Teaching and Learning, pp. 1-13, 2010.

[21] Pekrun, R., and Linnenbrink-Garcia, L., *Academic Emotions and Student Engagement*, Handbook of Research on Student Engagement, pp. 259–282, 2012.

[22] Khawlah Altuwairqi, Salma Kammoun Jarraya, Arwa Allinjawi, Mohamed Hammami, *A new emotion–based affective model to detect student's engagement*, Journal of King Saud University – Computer and Information Sciences, vol. 33, no. 1, pp. 99-109, 2021.

[23] Marta Arguedas, Thanasis Daradoumis and Fatos Xhafa, *Analyzing How Emotion Awareness Influences Students' Motivation, Engagement, Self-Regulation and Learning Outcome*, Educational Technology & Society, vol. 19, no. 2, pp. 87–103, 2016, ISSN 1436-4522 (online) and 1176-3647 (print).

[24] Ram Krishn Mishra, Siddhaling Urolagin, J. Angel Arul Jothi, Pramod Gaur, *Deep hybrid learning for facial expression binary classifications and predictions*, Image and Vision Computing, vol. 128, pp. 104573, 2022, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2022.104573.

[25] Prakhar Bhardwaj, P.K. Gupta, Harsh Panwar, Mohammad Khubeb Siddiqui,Ruben Morales-Menendez, Anubha Bhaik, *Application of Deep Learning on Student Engagement in e-learning environments*, Computers and Electrical Engineering, vol. 93, pp. 107277, 2021.

[26] Weiqing Wang, Kunliang Xu, Hongli Niu, and Xiangrong Miao, *Emotion Recognition of Students Based on Facial Expressions in Online Education Based on the Perspective of Computer Simulation*, Complexity, vol. 2020, pp. 1-9, 2020, https://doi.org/10.1155/2020/4065207.

[27] Yifei Guo, Jian Huang, Mingfu Xiong, Zhongyuan Wang, Xinrong Hu, Jihong Wang, Mohammad Hijji, *Facial expressions recognition with multi-region divided attention networks for smart education cloud applications*, Neurocomputing, vol. 493, pp. 119-128, 2022, https://doi.org/10.1016/j.neucom.2022.04.052.

[28] Wei, Qinglan, Bo Sun, Jun He, and Lejun Yu, *BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels*, Signal Processing: Image Communication, vol. 59, pp. 168-181, 2017.

[29] D'Mello, Sidney, and Rafael A. Calvo, *Beyond the basic emotions: what should affective computing compute?*, in CHI'13 extended abstracts on human factors in computing systems, 2013, pp. 2287-2294.

[30] Joshi, Nidhi, Niklas Beecken, Hawa Bah, Frank Steinicke, and Juliane Degner, *Advanced Emotion Analytics of Virtual Group Meetings involving Intelligent Virtual Agents*, 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), IEEE, 2022, pp. 344-350.

[31] D'Mello, Sidney K., Blair Lehman, and Natalie Person, *Monitoring affect states during effortful problem solving activities*, International Journal of Artificial Intelligence in Education, vol. 20, no. 4, pp. 361-389, 2010.

[32] Pekrun, R., Goetz, T., Titz, W.; Perry, R.P., *Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research*, Educational Psychologist, vol. 37, no. 2, pp. 91–105, 2002.

[33] Bian, Cunling, Ya Zhang, Deliang Wang, Yi Liang, Bo Wu, and Weigang Lu, *An academic emotion database and the baseline evaluation*, 13th International Conference on Computer Science & Education (ICCSE), IEEE, 2018, pp. 1-6.

[34] Bian, Cunling, Ya Zhang, Fei Yang, Wei Bi, and Weigang Lu, *Spontaneous facial expression database for academic emotion inference in online learning*, IET Computer Vision, vol. 13, no. 3, pp. 329-337, 2019.

[35] Haobang Wu, *Real Time Facial Expression Recognition for Online Lecture*, Wireless Communications and Mobile Computing, vol 2022, Article ID 9684264, 2022.doi: https://doi.org/10.1155/2022/9684264

[36] Pabba, Chakradhar, and Praveen Kumar, *An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition*, Expert Systems, vol. 39, no. 1, pp. e12839, 2022.

[37] Werlang, Pablo S., and Patrícia A. Jaques, *Student Engagement Recognition from Videos: A Comparison Between Deep Learning Neural Network Architectures*, IEEE Computer Society Technical Community on Learning Technology, vol. 21, no. 3, pp. 7-12, 2021.

[38] Buono, Paolo, Berardina De Carolis, Francesca D'Errico, Nicola Macchiarulo, and Giuseppe Palestra, *Assessing student engagement from facial behavior in on-line learning*, Multimedia Tools and Applications, pp. 1-19, 2022.

[39] Mishra, Ram Krishn and Raj, Harshit and Urolagin, Siddhaling and Jothi, J Angel Arul and Nawaz, Nishad, *Cluster-based knowledge graph and entity-relation representation on tourism economical sentiments*, Applied Sciences, vol. 12, no. 16, pp. 8105, 2022, https://doi.org/10.3390/app12168105.

[40] Mishra, Ram Krishn and Jothi, J Angel Arul and Urolagin, Siddhaling and Irani, Kayan, *Knowledge based topic retrieval for recommendations and tourism promotions*, International Journal of Information Management Data Insights, vol. 3, no. 1, pp. 100145, 2023, https://doi.org/10.1016/j.jjimei.2022.100145.

[41] Ram Krishn Mishra, *Deep Learning Model for Multiclass Classification of Diabetic Retinal Fundus Images Using Gradient Descent Optimization*, Advances in Signal Processing and Communication Engineering: Select Proceedings of ICASPACE 2021, pp. 27-35, 2022.

[42] Chollet, François, *"Xception: Deep learning with depthwise separable convolutions*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[43] Solanki, Naman, and Souraj Mandal, *Engagement Analysis Using DAiSEE Dataset*, 17th International Conference on Control, Automation, Robotics and Vision (ICARCV), IEEE, 2022, pp. 223-228.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 31, 2023
Accepted: Sep 26, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 971–984, DOI 10.12694/scpe.v24i4.2471

A STATISTICAL ANALYSIS OF SENTIMENT OVER DIFFERENT SOCIAL PLATFORMS ON DRUG USAGE ACROSS HIGH, MIDDLE AND LOW-INCOME COUNTRIES

ANUREET CHHABRA,* AKASH SHARMA,† KUNAL CHHABRA,‡ AND AMIT CHHABRA§

Abstract. Social media serves as a platform for sharing information and connecting with others on various subjects, including healthcare and drugs. Analyzing drug sentiment from various social media like Twitter, reddit, quora etc are crucial for monitoring drug safety, identifying adverse reactions, and providing an early warning system for potential safety concerns, benefiting healthcare organizations and governments. This paper aims to study the opinions of people from all over the world by analyzing their messages, posts on social media. For this study 39,069 drug related message corpus were fetched making a comparison between the high income countries, and middle income and low income countries on the basis of drug consumption from 2021 to 2023. The dataset used for the study consisted of 41.63% text-corpus from high income countries out of which on an average from 2021 to 2023, 40.2% was found to have a positive sentiment. Whereas 34.65% of text-corpus are from middle-income countries out of which on an average 26.4% of was of a positive sentiment and 23.70% text-corpus are from low income countries with 23.6% having a positive sentiment. Furthermore, the primary factor for having such differences from people's sentiment on drug consumption from high to low income countries includes Cultural and Social Norms, Legalization, availability etc therefore, In high-income countries, drug use is more socially accepted than in other regions. This proposed study gives an insight into people's opinion on various drugs from different countries and regions. The results of this study attempted to understand how the public is responding to different types of information and to identify potential misinformation which can be used to formulate policies for existing and future drug prevention campaigns in order to improve public health and promote public education.

Key words: Sentiment Analysis, Drug consumption, drug overdose

1. Introduction. One of the finest places in the world right now for individuals to share their opinions on a given topic are Twitter, Reddit, Quora etc [1]. Social media platforms are widely used by individuals, businesses, and organizations to share news, opinions, and updates with their followers, and is a popular platform for real-time news and discussions on a wide range of topics. Also these platforms provide access to its API (Application Programming Interface), which is a set of tools and protocols that developers can use to access and interact with data and functionality [2]. Due to APIs of these platforms, they have become a significant source of real-time data that can be utilized in studies to analyze sentiments of the people towards social and political issues. Currently, an increased consumption of drugs and alcohol by youth is a growing concern among parents, educators, and healthcare professionals. Drugs and alcohol can have severe negative effects on individuals and society as a whole. They are highly addictive and can lead to physical and psychological dependence [3]. The increased consumption of alcohol by teenagers and youth can have serious negative consequences, including increased risk of accidents and injuries, impaired judgment and decision-making, and increased risk of alcohol-related health problems [4]. The worldwide deaths due to drug overdose are exponentially increasing as shown in figure 1.1 and deaths due consumption of different drugs in the USA is shown in figure 1.2. The rate of smoking and tobacco is very high as per the latest surveys as shown in figure 1.3 and figure 1.4 respectively. Excessive drug consumption can lead to addiction, legal problems and social problems. Narcotics drug abuse can also lead to overdose and death. In addition to the harm caused to individuals, narcotics drug abuse can also have significant social and economic costs. It can lead to increased crime and reduced productivity, and can put a strain on healthcare and criminal justice systems [5].

Sentiment analysis on drug-related messages, posts etc can be used to develop targeted interventions and prevention strategies, such as educational campaigns or treatment programs. For example, if sentiment analysis

*Chandigarh College Of Engineering and Technology, Panjab University. (anureet0409@gmail.com)
†Chandigarh College Of Engineering and Technology, Panjab University (sharmaaakash080@gmail.com).
‡Vellore Institute of Technology (kunalch7677@gmail.com).
§Chandigarh College Of Engineering and Technology, Panjab University. (amitchhabra@ccet.ac.in).

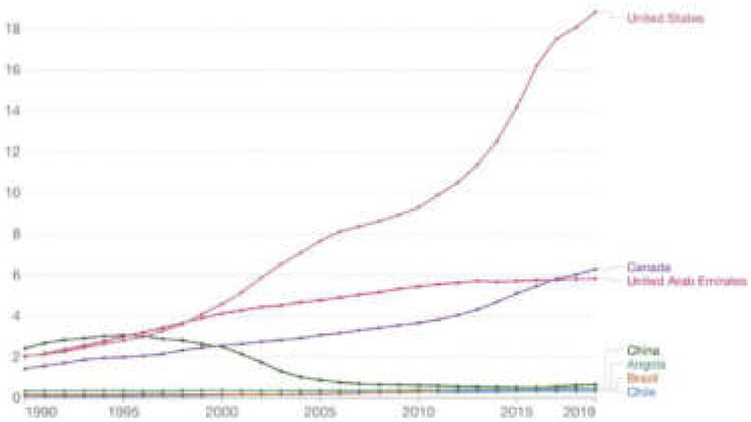


Fig. 1.1: Country wise Drug Use disorder death rate (Per 100,000) on Y›-axis, Year (1980 to 2019) on X-axis[6].

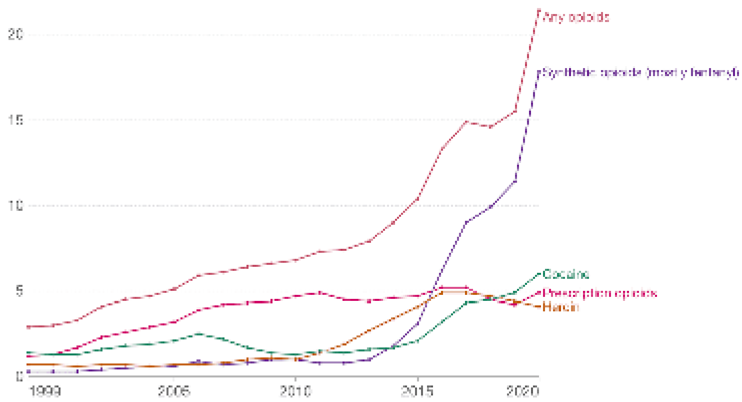


Fig. 1.2: Deaths due to drug overdose in the USA (Per 100,000) on Y-axis from the year 1999 - 2020 (X-axis) [6].

reveals an increase in the use of a particular drug, targeted educational campaigns could be developed to raise awareness of the risks associated with that drug and discourage its use. In addition, sentiment analysis can be used to monitor the effectiveness of prevention strategies and treatment programs over time. Moreover, through the analysis of drug-related messages, researchers can identify various concerns such as driving under the influence and underage drinking, substance abuse, and drug intoxication. Which can be valuable in developing effective public health policies [7].

The aim of this study is to present a detailed sentiment analysis of drug-related posts, messages at a country and also at a global level. The methodology involved in this paper utilizes several libraries, including Snscrape, Tweepy, NLTK, TextBlob, and API's of Twitter, Quora and reddit to generate the desired result. NLTK is an open-source library which provides a wide range of tools and resources for natural language processing, including tools for sentiment analysis. The further sections are divided as follows. Section 2 contains the review of the existing work on drug related sentiment analysis. Section 3 discusses the societal impact of this study. Section 4 contains the methodology followed to collect the data and perform sentiment analysis on it. Section 5 contains the results and analysis for both global and country wise segmented data and discussion about the results obtained are present in section 6. Section 7 contains the limitations and the future directions of this study.

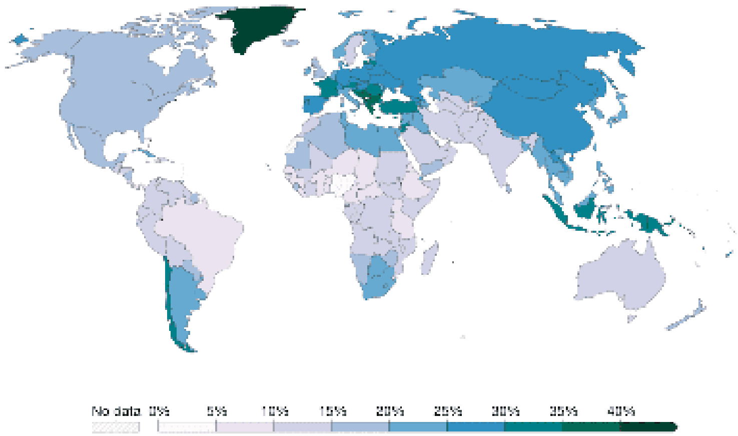


Fig. 1.3: Country-wise Prevalence of Global daily smoking populations (people of age 15 years and older) for the year 2021 [6].

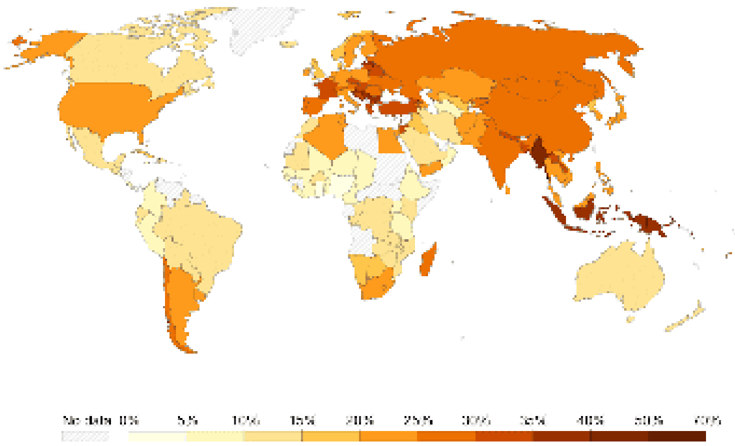


Fig. 1.4: Global use of tobacco (age 15 years and older) for the year 2020 [6].

2. Literature Review. The research [8] explores the use of sentiment analysis of tweets to extract information about adverse drug events (ADEs). The study uses a combination of natural language processing and machine learning techniques to analyze tweets and to identify ADEs related to specific drugs. The study found that sentiment analysis can be an efficient and effective way to extract information about ADEs from tweets, and that the proposed method was able to identify ADEs with high precision and recall. The research shows that sentiment analysis of tweets can be a useful tool for monitoring ADEs and for identifying potential safety issues related to specific drugs. In [9] the study looks at the impact of sentiment analysis on extracting adverse drug reactions (ADRs) from tweets and forum posts. The study found that sentiment analysis can be used to identify specific ADRs that are not reported in traditional sources such as clinical trials. The research [10] looks at the use of sentiment analysis on tweets to understand public opinion of the application of drugs in different countries but no categorisation has been made. The study uses textblob for identifying the sentiment of tweets in different languages and different countries.

In [11] the use of sentiment analysis and transfer learning to improve the classification of adverse drug reactions (ADRs) has been used. The study found that the use of transfer learning from a pre-trained model can be effective in handling the class imbalance present in the ADR data. The study [12] explores the use of sentiment analysis on Twitter data to extract information about adverse drug reactions (ADRs). The study found that sentiment analysis can be an effective way to extract ADRs from tweets and that the proposed method was able to identify ADRs with high precision and recall. The study also found that the use of sentiment analysis can be useful for monitoring ADRs and identifying potential safety issues related to specific drugs. The study concludes that sentiment analysis of tweets can be a useful tool for discovering ADRs and for identifying potential safety issues related to specific drugs. Study [13] looks at public opinion of the application of drugs in different countries by analyzing tweets. The study uses VADER Sentiment Analysis, a pre-trained lexicon-based model, to classify tweets according to sentiment (positive, negative, or neutral) and to determine the viewpoint of the tweets. The study found that the sentiment of tweets varies depending on the country and the drug being discussed, and that VADER can effectively classify the sentiment of tweets in different languages and different countries. The study also found that the use of VADER can provide a quick and efficient way to understand public opinion of the application of drugs in different countries.

In study [14] presents a comprehensive review and benchmark evaluation of the state-of-the-art techniques for sentiment analysis of tweets is given. The study provides a benchmark evaluation of the performance of various techniques on a dataset of tweets. The study found that the most effective techniques for sentiment analysis of tweets are deep learning approaches such as convolutional neural networks and recurrent neural networks. The study also found that these techniques performed better than traditional machine learning approaches such as support vector machines and decision trees. The study [15] is a survey and comparative study of different semi-supervised learning techniques for sentiment analysis of tweets. The aim of the study is to evaluate the performance of these techniques and to compare their performance. The study found that the semi-supervised learning techniques performed well on the sentiment analysis task, with the best performance achieved by the proposed method, which combined a combination of feature extraction, feature selection, and classification.

The study [16] presents a Sentiment analysis on gabapentinoid (pregabalin and gabapen- tin) has been performed. Some keywords taken are: Pregabalin, Lyrica, gabapentin, Neurontin. From 8 March to 7 May 2021. They have extracted 2931 pregabalin-related tweets out of which 30% are found to be having a positive sentiment, 21% are found to be negative and 2736 gabapentin-related tweets out of which 22.7 are found positive, 9.1% are found to be negative. Sentiment analysis on drugs used for therapy of COVID-19 using VADER is performed in paper [17] and a dictionary based approach. Keywords taken for in study are Hydroxychloroquine etc. on which 12.8%, 13.33% and 13.5% of tweets are found to be of positive sentiment in India, USA, China. Out of many countries African countries were found to be having maximum number of positive tweets with 18.3% In Study [18] authors have improved precision and recall of sentiment analysis on data from DailyStrength forum and tweets with a total messages of 8051. Keywords taken for drug related sentiment analysis are Trazodone, Cymbalta, Quetiapine, Seroquel. Proposed approaches by the authors have increased F-measure to 69.16% and 80.14% from twitter and DailyStrength respectively. The study [19] presents a total 267,215 tweets from Nov-2014 to Feb-2015 have been extracted using 250 drug-related tweets keywords such as adderall, xanax, Aspirin etc. Although the study has not done sentiment analysis but rather the authors focuses on ADE in which they found the maximum tetra-gram language model scores to be 1 and minimum score to be 0.21.

The above mentioned literature survey does not provide a comprehensive analysis based on income division and population of countries which indirectly affect education in the country. Furthermore, the available studies don't provide sentiments of people on drug related keywords that are taken in this study from social media platforms such as Twitter, Reddit, Quora etc. To address the limitations in the existing literature, our research aims to conduct a comparative sentiment analysis of drug-related message corpus like posts across low, middle, and high-income countries. By analyzing sentiments from diverse income-level nations, our work seeks to provide a more comprehensive and representative view of public opinion on drug usage, efficacy, and societal acceptance. Additionally the study also provides a global sentiment analysis on a wide range drug related keywords like "heroin", "cocaine", "nicotine", "smoking" etc

3. Societal Impact of the Study. Drug sentiment analysis on social media messages, tweets, posts can help society in several ways. By analyzing messages related to drugs, researchers in the current study provides insight into public opinion on various drugs and the potential side effects of these drugs [20]. This

Table 4.1: Sample Dataset generated using APIs of Twitter, Reddit, Quora

Date	Content	Location	Sentiment
2022-12-08 05:32:24+00:00	I can't wait to go back to university and commit Zina, listen to music, do drugs, go to parties, etc. and most importantly freemix	USA	Positive
2022-12-03 14:40:05+00:00	Mumbai: Man stabs friend to death during quarrel over paying for alcohol, held	India	Negative
2022-12-08 14:24:13+00:00	This should have never happened. Seems like the attacker has a problem with alcohol	USA	Negative
2022-04-10 03:19:06+00:00	Best Alcohol Drink in Euphoria??	USA	Positive
2022-12-08 14:34:43+00:00	It's Thursday, which means it's almost Friday, which means that the alcohol delivery is set to arrive any minute. I'm hyped	USA	Positive

information can be used to inform public health policy and to identify potential safety concerns related to drugs. This study can be used by pharmaceutical companies to monitor the public's perception of their drugs and to identify potential issues related to drug development and marketing [21]. The Sentiment analysis in this study aims to provide most common misconceptions and misunderstandings about drugs, which can then be addressed through public education campaigns.

Furthermore, Through this study early signs of negative public sentiment towards a drug can be identified, which can be used as an early warning system to allow pharmaceutical companies and regulatory agencies to take action before a safety concern becomes a major public health issue [22,23]. Sentiment analysis can be used to monitor the impact of drug-related news and events on public sentiment. This can be used to understand how the public is responding to different types of information and to identify potential misinformation [24]. Sentiment analysis on drugs related messages can provide valuable information to help improve public health, inform drug development and marketing, and promote public education. Misinformation about drugs on social media can be identified using sentiment analysis, and this information can be used to inform public education campaigns [25].

4. Methodology.

4.1. Data Collection. The dataset consists of 39,069 message corpus extracted or scraped from reddit, twitter and quora using their APIs. The dataset's properties include date-time, content, location and sentiment. The dataset comprised of 40% tweets from twitter and 60% of the data from reddit and quora. The dataset has been categorized to represent negative sentiment, neutral sentiment, and positive emotion which is due to the combined result of cultural, social, accessibility, availability and other multiple factors as further discussed in section 6. Negative sentiment within the dataset reflects the disapproval and condemnation of drugs by individuals, while positive emotion corresponds to instances where people express happiness and approval of drug consumption, often seen in celebratory contexts and similar situations. The neutral sentiment category, on the other hand, captures instances where drug-related sentiments neither lean towards negativity or positivity. A sample of the records has been shown in table 4.1. The messages are sourced from India, Mexico, USA, Russia, Australia, Bangladesh, Sri Lanka which have been segmented into low, middle and high income and population wise as shown in Table 4.2 to facilitate a comprehensive analysis and obtain a generalized demographic overview.

4.2. Data Preprocessing. Initially, messages such as posts related to drugs and alcohol are gathered from twitter, reddit and quora. The gathered messages corpus are cleaned, and prepared. We utilized SNSCRAPE and TWEEPY, Reddit and Quora's API to search for text concerning the drug such as Morphine, Ganja, Alcohol etc. The option "near" is set to the country on which we are doing analysis to get a message with geolocation coordinates inside or close to that country. The language option "Lang" to "en" to get texts/messages in English. Apart from the text collected from social media platforms, the data associated with each text is also collected—such as the text's location, links to pertinent media, or user mentions—represents a sizable quantity of information. The overall methodology has been shown in figure 4.1.

Table 4.2: Countries selected for the case study based on Population and Income as a factor.

Country	**Population Level**	**Income**
India	High	Middle
Mexico	Middle	Middle
USA	High	High
Russia	High	High
Australia	Low	High
Bangladesh	High	Low
Sri Lanka	Low	Low

Data cleaning is necessary prior to sentiment analysis. Techniques for data preparation have evolved across a wide range. The Python Natural Language Toolkit is one of several tools for minimizing text format complexity [26]. Retrieve raw data, operate with HTML elements, and parse text using regular expressions . It also does text stemming, tokenization, and text similarity calculations [27]. Exclude terms and phrases that give no information as part of our data preprocessing, such as (is, the, a, etc). NLTK is used for data preparation in our study. Text similarity as a preprocessing approach minimizes the amount of data collected. The value of the data set is increased by carefully removing duplicate texts with a lot of similarities using a text similarity measure. To preprocess the data links, emojis and punctuation are removed from the text. Then the preprocessed text is converted to ASCII values to pass it to the VADER tool for sentiment analysis.

4.3. Tokenization of Text. Tokenization is a rapid method for converting raw input into a usable data string. Even though tokenization is most typically utilized in the building of NFTs and cybersecurity, it is still an important stage in the NLP process [28]. Tokenization is a NLP approach that breaks down phrases and paragraphs into more manageable, language-assignable bits. The first steps in the NLP process involve obtaining information (a phrase) and breaking it down into digestible chunks [29].

In this study we used VADER (Valence Aware Dictionary for Sentiment Reasoning) to do sentiment analysis and categorize retrieved texts as positive, negative, or neutral as shown in algorithm 1. VADER takes into account both emotional polarity (positive/negative/neutral) and intensity (strong). Using the wisdom of the crowd and human raters to validate the text sentiment analysis, VADER text sentiment analysis was introduced in 2014 [30]. Based on the amount of exclamation points and question marks at the conclusion, VADER sentiment analysis increases the sentiment score of the phrase. The VADER algorithm accounts for word capitalization by increasing or decreasing its emotion score by 0.733 based on the term's positive or negative meaning.

4.4. Generation of Word Clouds. Simple APIs are provided to enable common NLP tasks, including part-of-speech tagging, noun phrase extraction, sentiment analysis, classifying, and translating [31]. It provides straightforward APIs for several NLP applications, including sentiment analysis, noun phrase extraction, parts-of-speech tagging, translation, and classification. To determine the polarity of the messages, TextBlob has been used. Tools for sentiment analysis, pattern recognition, and natural language processing are available in TextBlob. 2,918 terms make up TextBlob's emotion vocabulary [32]. When using TextBlob's sentiment analysis tool, it is possible to determine whether a text contains true information or someone's opinion by looking at its polarity and subjectivity scores.

5. Result and Analysis. In this case study, the analysis has been segmented into three distinct categories based on the countries' income and population levels: high, middle, and low. In order to present a comprehensive understanding of how drug sentiment varies across different economies and countries over the years. This drug sentiment analysis is provided to gain valuable insights into the prevailing attitudes and perceptions towards drugs in each group and how the change over time. Moreover, to present an overview, This study has also been extended to global study that encapsulates the trends and patterns in drug sentiment across the entire spectrum of countries. As per the word clouds generated, the most frequent words obtained for each sentiment has been summarized in table 5.1.

ALGORITHM 1 Texts/messages Scraping from APIs and Sentiment Analysis using VADER

Input: Extracted Drug Related message using APIs

Output: Predicted Emotion of the messages as either positive/negative or neutral

Begin:

```
1. Import Libraries
2. Message_Corpus ← Initialize Empty List
3. While i < Number of Messages to be scraped
4.       em← Extracted messages using APIs with keywords "Ganja" , "alcohol","drugs"...
5.       Message_Corpus .append(em);
6. End While
7. Remove redundant messages in the dataset
8. Preprocess the messages by removing links and emojis
9. message df = Convert Message Corpus to Dataframe with Attributes "message",
   "date-time", "Location","Sentiment"
10. For i in message_df
11.         IF Sentiment_polarity_score(message_df["message"][i]) >= 0.05 then
12.                 message_df["Sentiment"][i] = Positive
13.         Else
14.                 IF Sentiment Polarity Score (message_df ["message"][i]) <= -0.05 then
15.                         message_df ["Sentiment"][i] = Negative
16.                 ELSE
17.                         message df ["Sentiment"][i] = Neutral
18.                 END IF
19.         END IF
20. END FOR
```

Table 5.1: Most frequent words in each sentiment for Global analysis

Negative	Positive	Neutral
Suicide	Recreational	Legalization
Depress	Nice	Food
Addiction	Sweet	Morning
Abuse	Enjoy	Medication
Dangerous	Celebrate	Treatment
Overdose	Appreciate	Dosage
Deadly	Party	Government

5.1. Global Analysis. For the global data analysis of drug related text sentiment analysis in this paper keywords such as "alcohol", "Ganja", "nicotine" have been used in the APIs. The word cloud of each keyword has been visualized from figure 5.1 to figure 5.4. The opioid crisis has been a significant public health issue in many countries, particularly the United States along with other countries as seen in figure 1.3. Opioids, including prescription painkillers and illegal drugs like heroin and fentanyl, have led to a surge in overdose deaths and addiction cases. Although the smoking rates have been declining in many countries, especially in developed nations, the numbers are still significant. This decline is due to public health campaigns, increased awareness of the health risks associated with smoking, and the implementation of tobacco control policies such as higher taxes, smoking bans, and graphic warning labels which can be enhanced further. Excess of Tobacco consumption is a major challenge in many countries as shown in figure 1.4. Tobacco use remains a leading cause of preventable deaths worldwide. It is estimated that millions of people die each year due to tobacco-related

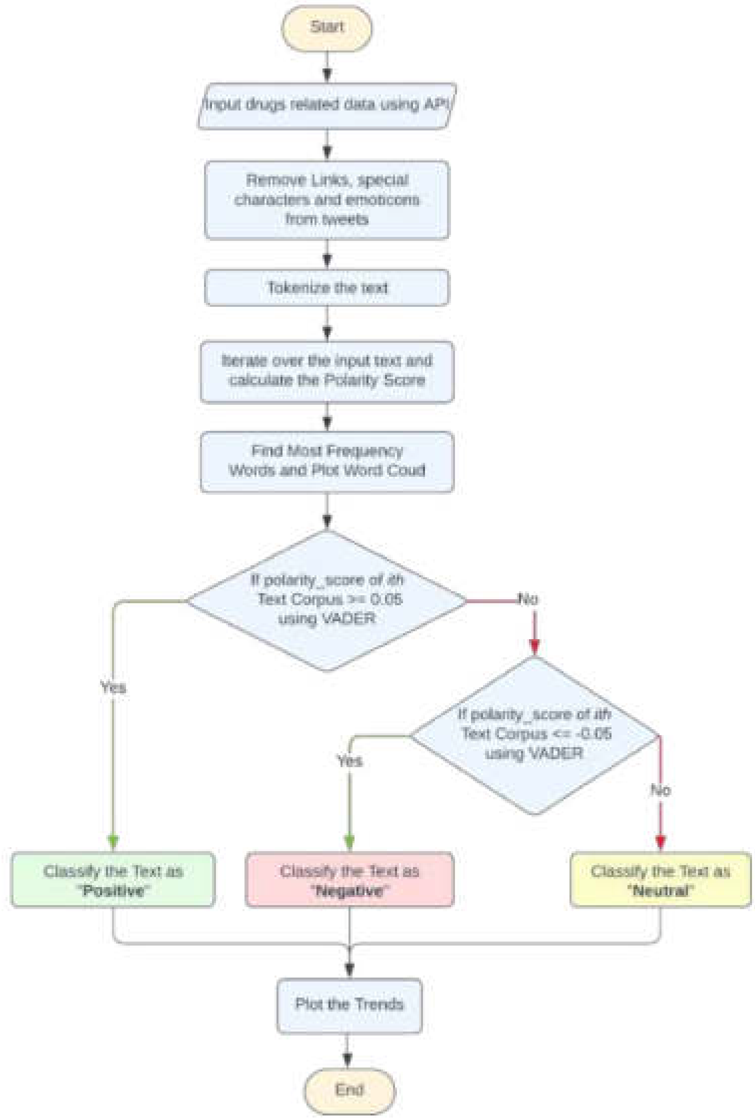


Fig. 4.1: Methodology followed for drug related posts or text Sentiment Analysis.

Fig. 5.1: Word-cloud of keyword “ganja” (Left) and “alcohol” (Right) on the global dataset.

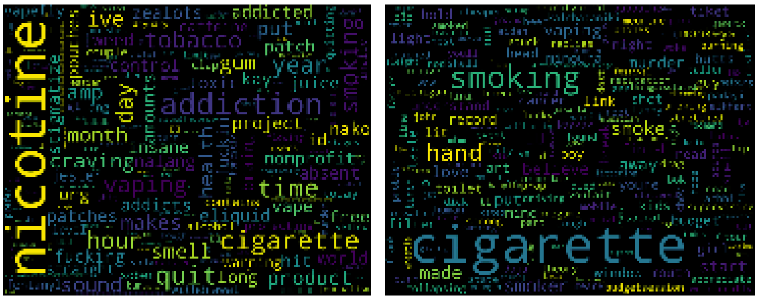


Fig. 5.2: Word-cloud of keyword "nicotine" (Left) and "cigarette" (Right) on the global dataset

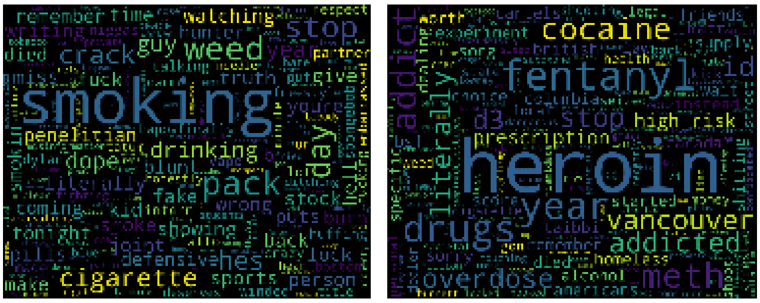


Fig. 5.3: Word-cloud of keyword "smoking" (Left) and "Heroin" (Right) on the global dataset

illnesses, including lung cancer, heart disease, respiratory problems, and other health issues. Furthermore in table 6.1 global drug sentiments has been shown for each keyword and the word-wise sentiments in the texts is presented in table 5.1. Furthermore, the analysis for the data of global messages or posts are present as follows in table 5.2. Various keywords are taken such as ganja, cocaine, smoking nicotine etc.

5.2. Income as a factor. High-income countries, also known as developed countries, are nations with a high gross national income (GNI) per capita. These countries typically have a well-developed infrastructure, a high standard of living, and a strong economy. Some examples of high-income countries include: United States, Russia, Australia etc. Low-income countries, also known as developing countries, are nations with a low gross national income (GNI) per capita [33]. These countries typically have a less developed infrastructure, lower standard of living, and a weaker economy. Some examples of low-income countries include: Bangladesh, Sri Lanka etc.

In high-income countries, drug consumption is higher than in lower and middle-income countries. The use of opioids, such as heroin, is a major concern in these countries. According to the World Health Organization (WHO), opioid use disorders affect an estimated 14 million people globally, with the highest prevalence in North America and Western Europe. The use of other types of drugs such as cocaine and amphetamines is also commonly consumed in high-income countries [34]. The dataset generated for the study consisted of 41.63% (16,268) text from high income countries (USA, Russia, Australia) whereas the rest 34.65% (13,538) text are from middle-income countries like (India, Mexico) and 23.70% (9,263) text are from low income countries (Bangladesh, Sri Lanka).

In the generated dataset we observed that the average as follows: 35.60% of the extracted text from high income countries (USA, Australia, Russia) were classified as neutral and 40.02% as positive from the year 2021-2023 which is a matter of concern. As seen in the dataset, in high income countries, drugs and alcohol is a major part in every celebration and in their daily lives. Whereas In middle-income countries, drug and alcohol consumption tends to be lower than in high-income countries but higher than in low-income countries. The

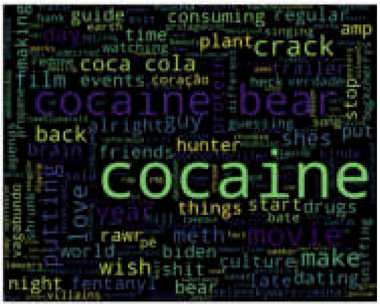


Fig. 5.4: Word-cloud of keyword “cocaine” in the global dataset

Table 5.2: Most frequent words in each sentiment for Global analysis

Sr. no.	Keyword	Positive Tweets	Negative Tweets	Neutral Tweets
1	Ganja	30.49 %	9.88%	59.63%
2.	Alcohol	33.41%	14.86%	51.73%
3.	Nicotine	15.0%	15.0%	70.0%
4.	Cigarette	29.83%	23.91%	46.26%
5.	Smoking	29.837%	16.599%	53.564%
6.	Heroin	17.018%	23.376%	59.606%
7.	cocaine	20.158%	18.346%	61.496%

use of opioids and cannabis is relatively common in some countries. The non-medical use of prescription drugs is also a concern. Through detailed analysis we have observed that on an average of the three years, 44.10% of the text were neutral with 26.40% of text turned out to be negative in middle income countries . In Low income countries the average sentiments are summarized as follows, 35% of text are Positive and 23.6% of text are Neutral. The pie charts summarizing and comparing the results of low, middle and high income countries average from 2021 to 2023 are shown in figure 5.5.

In low-income countries, drug consumption tends to be lower than in high-income and middle-income countries. However, the production and trafficking of narcotics drugs is a major concern in many of these countries like Afghanistan. Many low-income countries are major producers of opium and coca, the raw materials used to make heroin and cocaine, respectively. The illegal trade in these drugs generates huge profits, fueling corruption and destabilizing governments [35]. Drug consumption in low-income countries is also a concern, especially in urban areas.

6. Discussion about the Results. The drug-related sentiment of people from the year 2021 to 2023 from countries with income (high, middle, low) as a factor has been summarized in table 6.1 and year wise analysis has been shown in figures 6.1 - 6.3. Which shows that the although consuming drug creates addiction and is morally negative but still there is a lot of positive and neutral sentiment of people towards it. In this study various such factors are studied to understand the sentiment of people towards drugs. The primary factor that impacts the drug-use of the people in a nation includes Socio-economic, Cultural and Social Norms, Legalization and Decriminalization etc as discussed below in this section.

Due to socio-economic factors in populated and more developed nations, income disparities and social inequalities can lead to marginalization, where drug use might be more prevalent as a coping mechanism. Through the results of this study it has been found that in high income and middle income countries the percentage of positive and neutral sentiment is increasing throughout the years whereas in low income countries positive and neutral sentiment is decreasing. The projected values till 2023 are discussed as follows. In high income countries positive sentiment increased from 38.52% to 41.46% and negative sentiment 26.51% to 21.97% from 2021 to 2023. In middle income countries, the positive sentiments increased from 25.38% to 27.27%

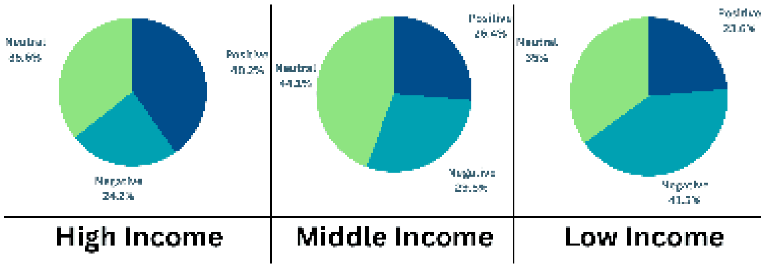


Fig. 5.5: Pie charts for results of of High, Middle and Low income countries average from year 2021 to 2023

Table 6.1: Drug related text sentiments of countries with income as a factor from year 2021 to 2023

Classification	Year	Positive Tweets	Negative Tweets	Neutral Tweets	Total Tweets
High Income	2021	38.52%	26.51%	34.97%	16,268
	2022	40.65%	24.07%	35.28%	
	2023	41.46%	21.97%	36.57%	
Middle Income	2021	25.38%	30.8%	43.74%	13,538
	2022	26.40%	29.48%	44.12%	
	2023	27.27%	28.15%	44.58%	
Low Income	2021	24.28%	40.33%	35.39%	9,263
	2022	23.91%	40.65%	35.24%	
	2023	22.37%	43.35%	34.28%	

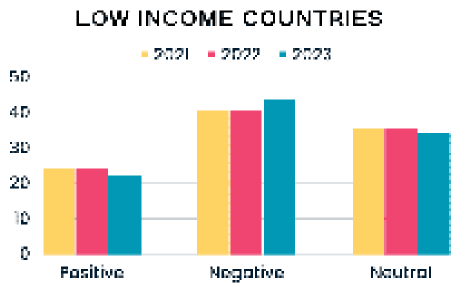


Fig. 6.1: Year wise sentiment analysis of Low Income countries

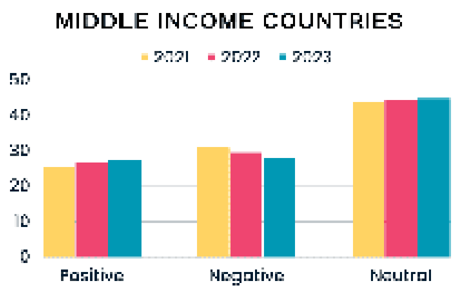


Fig. 6.2: Year wise sentiment analysis of Middle Income countries

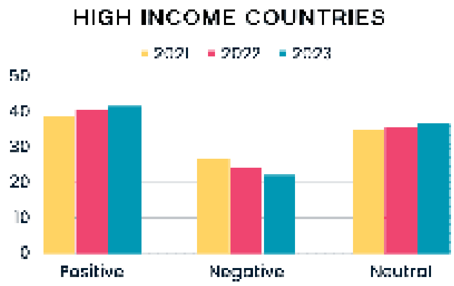


Fig. 6.3: Year wise sentiment analysis of High Income countries

and the negative sentiment decreased from 30.8% to 28.15% from the year 2021 to 2023. On the other hand, in smaller and medium-sized countries, the distribution of wealth and access to resources is more balanced, potentially resulting in lower overall drug consumption rates. Furthermore, Different societies and cultures have varying attitudes and norms towards drug use. In high-income countries, drug use is getting more socially acceptable with time or perceived differently than in other regions which is perhaps the reason for increasing positive sentiment on drug consumption. In low income countries positive sentiment decreased from 24.28% to 22.37% and the negative sentiment is increasing from 40.33% to 43.35% from the year 2021 to 2023. This trend signifies that people are discouraging drug usage and not accepting the western culture or norms from high and middle income countries. High-educated countries have deeply ingrained cultural norms that discourage drug use, leading to lower consumption rates regardless of population size. Conversely, in countries where drug use is more normalized, higher consumption rates may be observed. Understanding cultural factors and social norms is crucial for designing effective drug prevention and harm reduction strategies.

In some high-income countries, the people consume more drugs as compared to other countries due to legalized or decriminalization of specific drugs. This can lead to a shift in public perception and a more open discussion about drug use, which is reflected in positive sentiments of the input text message. Moreover, The availability and accessibility of drug treatment and support services can play a pivotal role in mitigating the impact of drug consumption. Smaller and under-developing countries face challenges in establishing comprehensive treatment programs due to limited resources, whereas larger countries may struggle to reach all affected individuals effectively. Understanding the relationship between population size, the prevalence of treatment services, and their effectiveness in reducing drug abuse is crucial for improving public health interventions. Smaller countries are more vulnerable to drug trafficking due to their limited resources for border security and law enforcement . The influx of illicit drugs contributes to higher rates of drug consumption and related problems. On the other hand, larger countries have more extensive border control measures but can still face significant challenges in combating drug trafficking.

7. Limitation and Future Scope. The current research has certain limitations that should be acknowledged. Firstly, the retrieved messages might not be exhaustive, as some drug-related data could have been removed by the respective social media platforms in accordance with their community guidelines. Moreover, the analysis relies on Text Mining techniques, inadvertently disregarding special characters, emojis, and images that may carry significant contextual information or emotional expressions related to drug use. However, it's not always possible to accurately interpret the sentiment of a message, and there may be biases in the data that affect the analysis. Additionally, sentiment analysis can only provide a snapshot of public perception at a particular point in time, and the sentiment of social media data may change over time. Despite these limitations, drug sentiment analysis can be a valuable tool for healthcare professionals and researchers who are looking to stay informed about public perception of drugs and develop effective strategies for drug development, marketing, and safety.

8. Conclusion. Drug sentiment analysis on data collected from various social media like twitter, reddit and quora provides valuable insights into public drug perception, monitoring safety, and enabling effective strategies for government and non-profit organizations to take necessary actions. This study examined drug-related sentiments of people over various social media platforms in different countries. It was found that in high income countries 40.2%, and in low income countries 23.6% of the average input text corpus from the year 2021-2023 were found to be positive, Which shows that people on the internet are getting more inclined towards a positive sentiment for drugs and alcohol yearly in high and middle income countries. Whereas a decreasing trend is seen in low income countries which signifies that people are not promoting drug usage. Drugs are a major part of celebrations and parties in the current world. Various factors, including socio-economic disparities, cultural norms, drug legalization, and treatment availability, influence drug consumption patterns. Smaller countries may face challenges in establishing comprehensive treatment programs and combating drug trafficking, whereas some developed countries have legalized various drugs and alcohol making it accessible to common people. Analyzing these trends is essential in designing targeted drug prevention and harm reduction strategies for different regions, promoting public health, and ensuring the well-being of communities worldwide.

REFERENCES

[1] Singh, D. K. S., Nithya, N., Rahunathan, L., Sanghavi, P., Vaghela, R. S., Manoharan, P., ... & Tunze, G. B. (2022). Social network analysis for precise friend suggestions for twitter by associating multiple networks using ml. International Journal of Information Technology and Web Engineering (IJITWE), 17(1), 1-11.

[2] Biswas, S., & Ghosh, S. (2022). Drug Usage Analysis by VADER Sentiment Analysis on Leading Countries. Mapana Journal of Sciences, 21(3).

[3] Saraswat, M., Dubey, A. K., & Gupta, R. Sentiment Analysis of Drugs related Post Impacting Society Healthcare.

[4] Abbasi, N., Vahdat-Nejad, H., Mansoor, W., & Salmani, F. (2022, December). Extracting Drug-Related Tweets in COVID-19 Pandemic. In 2022 5th International Conference on Signal Processing and Information Security (ICSPIS) (pp. 103-106). IEEE.

[5] Singh, M., Dhillon, H. K., Ichhpujani, P., Iyengar, S., & Kaur, R. (2022). Twitter sentiment analysis for COVID-19 associated mucormycosis. Indian Journal of Ophthalmology, 70(5), 1773-1779.

[6] Hannah Ritchie, Pablo Arriagada and Max Roser (2022) - "Opioids, cocaine, cannabis and other illicit drugs". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/illicit-drug-use' [Online Resource]

[7] Panda, B., Panigrahi, C. R., & Pati, B. (2022). Exploratory data analysis and sentiment analysis of drug reviews. Computación y Sistemas, 26(3), 1191-1199.

[8] Peng, Y., Moh, M., & Moh, T. S. (2016, August). Efficient adverse drug event extraction using Twitter sentiment analysis. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1011-1018). IEEE.

[9] Garg, S. (2021, January). Drug recommendation system based on sentiment analysis of drug reviews using machine learning. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 175-181). IEEE.

[10] Vijayaraghavan, S., & Basu, D. (2020). Sentiment analysis in drug reviews using supervised machine learning algorithms. arXiv preprint arXiv:2003.11643.

[11] Alhuzali, H., & Ananiadou, S. (2019, August). Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In Proceedings of the 18th BioNLP Workshop and Shared Task (pp. 339-347).

[12] Bose, D., Aithal, P. S., & Roy, S. (2021). Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries. International Journal of Management, Technology, and Social Sciences (IJMTS), 6(1), 110-127.

[13] Moh, M., Moh, T. S., Peng, Y., & Wu, L. (2017). On adverse drug event extractions using twitter sentiment analysis. Network Modeling Analysis in Health Informatics and Bioinformatics, 6(1), 1-12.

[14] Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. ACM Transactions on Management Information Systems (TMIS), 9(2), 1-29.

[15] Mathieson S, O'Keeffe M, Traeger AC, Ferreira GE, Abdel Shaheed C. Content and sentiment analysis of gabapentinoid-related tweets: An infodemiology study. Drug Alcohol Rev. 2022 Dec 20. doi: 10.1111/dar.13590. Epub ahead of print. PMID: 36539307.

[16] Bose, Rajesh & Aithal, Sreeramana & Roy, Sandip. (2021). Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries. International Journal of Management, Technology, and Social Sciences. 10.47992/IJMTS.2581.6012.0132.

[17] Korkontzelos, Ioannis & Nikfarjam, Azadeh & Shardlow, Matthew & Sarker, Abeed & Ananiadou, Sophia & Gonzalez, Graciela. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of Biomedical Informatics. 62. 10.1016/j.jbi.2016.06.007.

[18] Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. Data Brief. 2016 Nov 23;10:122-131. doi: 10.1016/j.dib.2016.11.056. PMID: 27981203; PMCID: PMC5144647.

[19] Silva, N. F. F. D., Coletta, L. F., & Hruschka, E. R. (2016). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Computing Surveys (CSUR), 49(1), 1-26. Chicago

[20] Ribeiro, L. A., Cinalli, D., & Garcia, A. C. B. (2021, May). Discovering adverse drug reactions from twitter: A sentiment analysis perspective. In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design

(CSCWD) (pp. 1172-1177). IEEE.

[21] Asderis, G. A. (2022). Sentiment Analysis on Twitter Data, a Detailed Comparison of TextBlob and VADER.

[22] Verma, R., Chhabra, A. & Gupta, A. A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective. Soc. Netw. Anal. Min. 13, 12 (2023). https://doi.org/10.1007/s13278-022-01015-2

[23] Najafizada, M., Rahman, A., Donnan, J., Dong, Z., & Bishop, L. (2022). Analyzing sentiments and themes on cannabis in Canada using 2018 to 2020 Twitter data. Journal of Cannabis Research, 4(1), 1-14.

[24] Talahaturuson, E., Gumelar, A. B., Sooai, A. G., Sueb, S., Suprihatien, S., Altway, H. A., ... & Sulistiyani, S. (2022, September). Exploring Indonesian Netizen's Emotional Behavior Through Investment Sentiment Analysis Using TextBlob-NLTK (Natural Language Toolkit). In 2022 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 244-248). IEEE.

[25] Shahzad, M., & Alhoori, H. (2022). Public Reaction to Scientific Research via Twitter Sentiment Prediction. Journal of Data and Information Science, 7(1), 97-124.

[26] Abednego, L., Nugraheni, C. E., & Fedora, A. (2022). Forex Sentiment Analysis with Python. International Journal of Advanced Research in Economics and Finance, 4(1), 46-55.

[27] Yin, H., Song, X., Yang, S., & Li, J. (2022). Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web, 25(3), 1067-1083.

[28] Qian, C., Mathur, N., Zakaria, N. H., Arora, R., Gupta, V., & Ali, M. (2022). Understanding public opinions on social media for financial sentiment analysis using AI-based techniques. Information Processing & Management, 59(6), 103098.

[29] Benmelech, E., Eberly, J., Papanikolaou, D., & Krieger, J. (2021, May). Private and social returns to R&D: drug development and demographics. In AEA Papers and Proceedings (Vol. 111, pp. 336-340). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

[30] Panda, B., Panigrahi, C. R., & Pati, B. (2022, September). Exploratory and Sentiment Analysis of Classified Drug Groups. In Proceedings of the 6th International Conference on Advance Computing and Intelligent Engineering: ICACIE 2021 (pp. 229-239). Singapore: Springer Nature Singapore.

[31] Sunitha, D., Patra, R. K., Babu, N. V., Suresh, A., & Gupta, S. C. (2022). Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries. Pattern Recognition Letters, 158, 164-170.

[32] Schlegelmilch, B. B., Sharma, K., & Garg, S. (2022). Employing machine learning for capturing COVID-19 consumer sentiments from six countries: a methodological illustration. International Marketing Review.

[33] Cameron, D., Smith, G. A., Daniulaityte, R., Sheth, A. P., Dave, D., Chen, L., ... & Falck, R. (2013). PREDOSE: a semantic web platform for drug abuse epidemiology using social media. Journal of biomedical informatics, 46(6), 985-997.

[34] Phan, N., Chun, S. A., Bhole, M., & Geller, J. (2017, April). Enabling real-time drug abuse detection in tweets. In 2017 IEEE 33rd international conference on data engineering (ICDE) (pp. 1510-1514). IEEE.

[35] Daniulaityte, R., Chen, L., Lamy, F. R., Carlson, R. G., Thirunarayan, K., & Sheth, A. (2016). "When 'bad'is 'good'": identifying personal communication and sentiment in drug-related tweets. JMIR public health and surveillance, 2(2), e6327.

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 31, 2023
Accepted: Sep 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 985–997, DOI 10.12694/scpe.v24i4.2475

A DEEP COMMUNITY DETECTION APPROACH IN REAL TIME NETWORKS

DEEPJYOTI CHOUDHURY* AND TAPODHIR ACHARJEE†

Abstract. Community detection in real time networks is one of the important aspect of social network analysis. Deep learning has been applied successfully in a variety of research fields in recent years. Proximity matrix is frequently used as the representation of the network structure. However, there are issues with the proximity matrix's insufficient spatial contiguity information. As a result, this research provides a deep learning applied community identification approach that combines the reorganization of the matrices, spatial attribute uprooting, and community identification. For obtaining a spatial proximity matrix, the primary proximity matrices in a real time graph is recreated using the highest weight and adjacent users. The dimensional proximity matrix can obtain a subdomain of the network, allowing the convolutional neural network (CNN) to draw out dimensional localization more easily and fast. Ten different real time datasets of social networks are used in tests to examine our proposed approach. Our results show that the proposed community identification approach has higher compatibility than existing deep learning-based strategies. As a result, the proposed deep community identification approach is capable of detecting the excellent clusters in real time networks.

Key words: real-time network, deep learning, community detection, social network, proximity

1. Introduction. It is conventional that social networks have been extensively deliberated to analyze behaviors of human considering a number of layouts, including information extraction, domination analysis, community detection, individual profile details, social data privacy etc [1]. Community identification in real-time social networks is a well-known aspect of networked systems in biology, economics, politics, and computer science.

Deep learning (DL) has showed excellent performance in a wide number of research domains, including real-time networks for analysing user structural information [2]. DL-applied network embedding can be executed in both ways: with random walk [3] and without random walk [4][5]. As we know that proximity matrix is used to reserve the information of the connected nodes in the network. However, the adjacency matrix has inadequate spatial proximity information.Several Auto-Encoder (AE)-based network entrenching studies improved the input vector by performing divergent pre-processing on proximity matrices to improve the elicitation correctness of spatial characteristics extraction to overcome this complication and boost the correctness of feature uprooting [6][7][8]. Convolutional neural network (CNN) is used in network embedding because it is an effective technique for extracting spatial localisation [9][10]. The convolution operation can be simulated on the graph. As a result, the difficulty is how to enhance the proximity matrix such that it can store spatial closeness among vertices.

1.1. Contribution. Given the resilience and efficacy of AE and CNN-applied network embedding, this paper integrates 'AutoEnc+CNN' to increase the aspect of feature uprooting from the nodes. As a result, this research offers a deep community detection approach that combines (a) matrix reorganization, (b) 'AutoEnc+CNN'-applied spatial feature uprooting, and (c) community identification. Furthermore, this paper provides a spatial characteristics uprooting strategy based on AutoEnc and CNN to uproot the spatial features of the reorganised proximity matrix.

In recent decades, efforts have been considerably made to build efficient models to identify communities in social networks. The main objectives of this paper are mentioned below:

1. To obtain the dimensional proximity matrices in real-time networks, a matrix reorganisation strategy based on a unique structure reorganisation approach is proposed, which can aid CNN in quickly and

*Department of Computer Science and Engineering, Assam University, Silchar, Assam, India 788011 (deepjyotichoudhury05@gmail.com).
†Department of Computer Science and Engineering, Assam University, Silchar, Assam, India 788011 (tapacharjee@gmail.com).

easily determining geographical localization.

2. To obtain a dimensional features uprooting strategy based on autoencoder and CNN, which is proposed to derive spatial eigenvectors and successfully uproot the spatial properties of real-time networks.
3. To explore the topographic architecture of real-time networks based on the matrix reorganisation and dimensional characteristics uprooting approach in order to improve social community detection.

After a brief introduction in Section 1, Section 2 presents important related investigations performed in recent years. In Section 3, proposed approach is discussed in detailed manner. Execution & Results Analysis is elaborated in Section 4. We have compared some previous works with our proposed approach in this section. Section 5 highlights the conclusion, challenges and future work.

2. Related Works. It has been observed in recent years that the applications of Deep Neural Network (DNN) can embed networks using arbitrary walks while maintaining network properties. DeepWalk is one of the most well-known deep learning method along with the enhancement of arbitrary walk. Without arbitrary walk, the same method applies deep approaches over the entire network. AE and CNN are two prominent deep learning models used in network embedding that do not employ random walks. Resler et al. [11] presented in their paper that the use of a metric learning enhanced deep CNN on an archaeological dataset. The associations between sites were determined by them using a community detection algorithm based on the confusion matrix data. A deep learning-based weighted network community discovery technique was also presented [12] based on a deep sparse autoencoder. A solution to problem of overlapping community structure for large graph was presented using autoencoder [13]. Ferraro et al. [14] presented a different technique to resolve community detection problem based on DL and they proposed a hypergraph-based data model for representing all forms of user connections within an MSN, which were frequently mediated by multimedia data.

Again, a parallel deep learning-applied community identification strategy using particle swarm optimization in massive composite networks was proposed [15]. The findings demonstrated how well the suggested deep learning with hybrid optimisation works for identifying communities in large networks. Essaid et al. [16] presented a method for detecting communities inside the Bitcoin network that uses a deep feature representation algorithm and Deep Feedforward Autoencoders. Their findings demonstrated that, compared to a random P2P network, the Bitcoin network has a stronger clustering coefficient and community structure. Sun et al. [17] developed a system that combined CNN and Transformer, used wavelet and inverse wavelet transforms for encoding and decoding, and employed wavelet transform and inverse wavelet transform for learning. A dual graph autoencoder (DGAE) was suggested by Zhang et al. [18] to develop discriminative representations for hyperspectral images. In order to characterise the geometric structures of hyperspectral pictures, DGAE first built the super pixel-based similarity graph with spatial information and band-based similarity graph using the relationships of pair-wise pixels within homogeneous regions and pair-wise spectral bands. More discriminative feature representations were learned from the hidden layer via the encoder-decoder structure of DGAE using the newly created dual graph convolution.

Self-Supervised Contrastive Graph Clustering (SCGC) was suggested by Kulatilleke et al. [19], which enforced graph structure using contrastive loss signals to acquire discriminative node representations and iteratively revised soft cluster labels. A thorough analysis of AE-based industrial applications was published by Qian et al. [20]. It was primarily divided into two sections: AE-based representation learning and monitoring techniques, which showed how AE-based monitoring methods are designed from start to finish. Second, a thorough analysis of AE-based representation learning from the viewpoints of industrial data characteristics was conducted. A study was carried out by Lim et al. [21] recently in order to offer a thorough overview and to investigate potential future possibilities for the best reinforcement learning-based virtual network embedding solutions.

Proximity matrix was well elaborated by Goel et al. [22] and by utilising the user's own behaviours as well as those of other users in their social network, they suggested a novel way to build a strong User Interest Profile (UIP) in community detection. The same authors proposed a methodology [23] that focused on UIP augmentation using multiple strategies, as well as a novel approach to handle outlier tags that caused ambiguity in the collective Resource Illustration Profile (RIP). The fuzzy satisfaction requirement-based novel mapping functions were designed to measure query relevance score and user interest relevance score for a web resource. A novel approach [24] was proposed that employed Cohen's k as a similarity measure for each pair of nodes;

the values were then clustered to discover communities. A new community detection approach was presented by Wu et al. [25] that proved the spatial proximity matrix could obtain a subdomain of the graph, allowing the convolution neural network to extract spatial localization more easily and fast. An auto-encoder based on a convolution neural network can extract the spatial eigenvector of the reconstructed adjacency matrix to improve modularity.

With the aid of some recent works [26][27][28][29][30], we also want to lay the groundwork for how researchers have implemented deep learning models in the network embedding technique:

- **AE-applied Graph Embedding Approach:** The AE algorithm is a powerful facts condense technology. AutoEnc-applied graph embedding approaches frequently modify the parameters of the input and converts it into a new depiction. To absorb the composition of the network, SDNE was proposed to design a mini-supervised model based on DNN that enhanced the inputs by combining 1^{st} and 2^{nd} order adjacency values.
- **CNN-applied Graph Embedding Approach:** CNN and its variations have found widespread application in network embedding. CNN-applied graph embedding employs the original CNN model, which was built for both Euclidean and non-Euclidean domains.

2.1. Motivation. The following are the primary observations that motivated us to present the proposed work provided in this paper:

1. 'Matrix reorganisation' is needed because it is a strategy technique used to improve efficiency, flexibility, and overall performance to identify communities in a large network. The network structure is restructured from a typical hierarchical model to a matrix form that contains features of both functional and project-based structures. Matrix reorganisation can convert network data into a more appropriate structure, such as an adjacency matrix or a modularity matrix, making it easier to discover communities or groups of nodes within the network. Matrix reorganisation can help increase the scalability of community detection algorithms in large-scale social networks. It decreases computing complexity and enables more effective analysis of massive datasets, both of which are critical in comprehending and managing online communities with millions of members. Matrix reorganization can assist in identifying influential nodes or key community members within social networks. This is why we have adopted this approach in our proposed method.
2. 'Spatial feature uprooting' method is required because of its ability to select the most discriminating features, where a deep learning based method can ingest more readily. Communities in many real-world circumstances are defined not only by social connections but also by geographical proximity. The inclusion of geographic information into spatial feature uprooting allows community detection algorithms to incorporate both social and spatial dimensions. Different regions within a network may have varying degrees of community structure. In order to measure spatial heterogeneity and pinpoint areas with distinct community boundaries where communities converge, spatial feature uprooting is used. When spatial linkages and geographic context are important, spatial feature uprooting is crucial for community detection. It makes it possible to analyse networks more thoroughly by taking into account both social and spatial dimensions.
3. 'Community identification' in real time networks is the final goal based on matrix reorganisation and spatial feature uprooting method. Communities frequently display recognisable behavioural patterns. Researchers can learn more about the habits, pursuits, and pursuits of various groups inside the network by identifying communities. In social networks, community identification is essential for a variety of purposes, from boosting user experiences and marketing tactics to upholding safety and comprehending the social dynamics of online communities. It offers insightful information that aids in decision-making and promotes a better comprehension of the intricate linkages and behaviours seen in social networks.

3. Proposed Approach. As mentioned in the Section 1, our proposed approach is the combination of matrix reorganization approach, spatial feature uprooting and finally community identification. Let's elaborate these approaches in the subsections below.

3.1. Matrix Reorganization Approach. Workflow of the proposed approach has been shown in Fig. 3.1. The proposed approach comprises of three different sub-approaches: Highest Weight User Selection, Adjacent

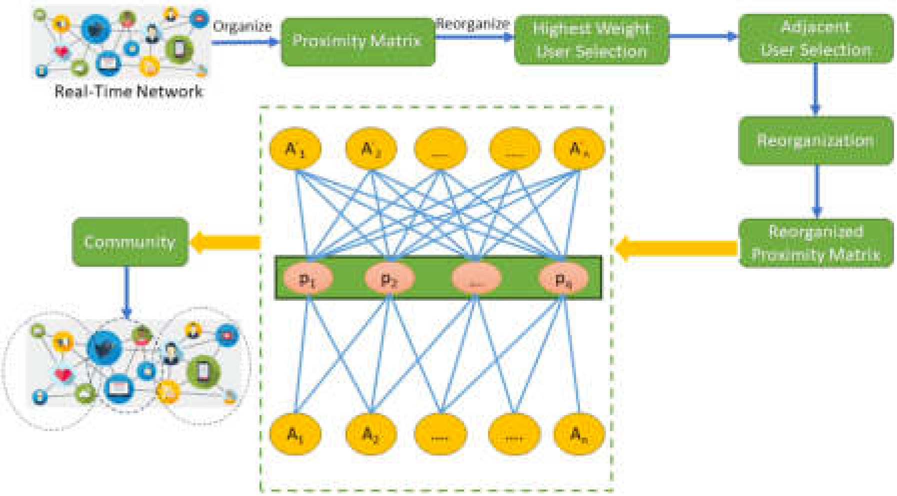


Fig. 3.1: Workflow of the Proposed Matrix Reorganization Approach

User Selection and Reorganization of Proximity Matrix. Let's discuss these sub-approaches below.

3.1.1. Highest Weight User Selection. The user is a key member of a group who may influence the views of other members using this strategy. In real-time social networks, numerous users can follow and make friend with the most connected person in a community. As a result, the Highest Weight User Selection technique proposed here would assess each node's influence on the next most significant node in the proximity matrix to identify a starting node for matrix reconstruction. The pseudo codes of this method is presented in Algorithm 6.

Algorithm 6 Proposed Highest Weight User Selection Algorithm

Require: a proximity matrix with nodes and the number of repetitions
Ensure: the user with the highest weight

Organize the proximity matrix A_{nxn}
while $p = 1, ..., n$, $q = 1, ..., n$ **do**
    Calculate the weight of the connections from user a_p to a_q based on the number of interconnections of a_p
end while
for $p = 1, ..., n$ **do**
    Initialize the weight of user a_p in the starting moment as 1.
end for
while (the amount of repetitions is less than the user having highest weight) **do**
    for $p = 1, ..., n$ **do**
        Initialize the weight of user a_p in the next repetition as 0.
        for $q = 1, ..., n$ **do**
            Compute the weight of user a_p grown by the connection weight from user a_q to user a_p as the merit of substitute.
            Initialize the weight of user a_p in the next repetition along with the merit of substitute.
        end for
    end for
end while
Return the user has the maximum weight.

3.1.2. Adjacent User Selection. After identifying the person with the highest weight (superior), an adjacent user selection method is presented to identify the adjacent user who is most relevant to the superior user.

Based on Equation 3.2, the superior user having the highest weight is represented as user a_p. Here, we have applied Euclidean Distance (ED), $z(p,q)$ to compute the distance between user a_p and user a_q. According to Equation 3.3, the user a_q with the least distance based on the nearest neighbor is initiated.

$$z(p,q) = \sqrt{\sum_{l=1}^{n} d(p,q,l)^2} \tag{3.1}$$

where

$$d(p,q,l) = a_{p,l} - a_{q,l} \tag{3.2}$$

$$AdjacentNeighbor = least_{z(p,q)} \tag{3.3}$$

where $q \neq p, 1 \leq q \leq n$

3.1.3. Reorganization of Proximity Matrix. The proposed highest weight user selection and adjacent user selection methods may be used to determine the order of users in the proximity matrix A. The superior user may be selected as the starting user, and the adjacent of the superior user can be selected as the next user; then, using Equations 3.2 and 3.3, the adjacent neighbor of the second user can be selected, and so on. Algorithm 7 can rebuild the proximity matrix A as matrix Z.

Algorithm 7 Reorganization of Proximity Matrix

Require: a proximity matrix A with n nodes
Ensure: the reorganized proximity matrix Z

Organize the proximity matrix A_{nxn}
Construct the directory R_n
Construct the directory S_n
for $p = 1, ..., n$ **do**
    Enter the user a_p into the directory R_n.
end for
Initialize the superior leader from the highest weight user selection approach as user a_p.
while Length of $R_n > 0$ **do**
    Extract user a_p from R_n
    for $q = 1, ..., n$ **do**
        Determine the Euclidean distance from user a_p to user a_q
    end for
    Detect as well as initialize adjacent user as user a_p
    Extract user a_p from S_n
end while
Build the proximity y matrix Z_{nxn} based on S_n
Return the reorganized proximity matrix Z.

3.2. Spatial Feature Uprooting. In a basic instance, both the input as well as output layers have 4 number of neurons, that means 4 users present in the network. The first hidden layer (the convolutional layer) has a filter size of 1 by 3, therefore two neurons are instructed in the concealed layer. Figure 3.2 depicts the architecture of a basic CNN-based auto-encoder.

In a normal scenario, the rebuilt proximity matrix may be divided into n records, with each record having a dimension of 1 by n. Both the input and output layers have n number of neurons. The first concealed layer has q neurons.

Figure 3.3 depicts the general case structure of a CNN-based auto-encoder. The loss function considers mean squared deviation (MSD). During the execution and performance stages, spatial information may be retrieved based on neuron values.

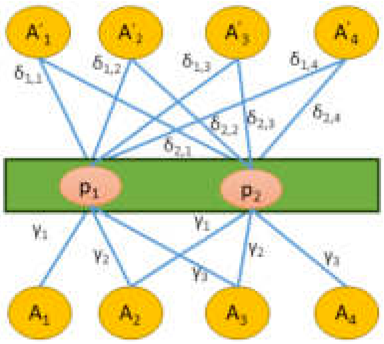


Fig. 3.2: A basic example of a CNN-applied auto-encoder

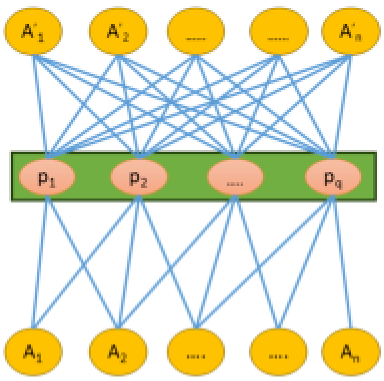


Fig. 3.3: A general instance of a CNN-applied auto-encoder

3.3. Community Identification. After extracting spatial information, the dimension of each record is represented as 1xn and used in the K-means method.

$$P_q = [p_{q,1}, p_{q,2}, ..., p_{q,n}] \tag{3.4}$$

The proposed community detection approach consists of three phases, which are as follows:

1. The k records are chosen at random from the n records to serve as k cluster centers.

$$R_p = [r_{p,1}, r_{p,2}, ..., r_{p,n}] \tag{3.5}$$

2. Based on Equation 3.1, ED is used to calculate the separation space from the q^{th} position to the p^{th} position of the cluster. Data n are divided into k number of clusters depending on their distance, also the core of every community is re-evaluated established on the data in the community.
3. If no modifications are made to any of the cluster centers, the community identification process is completed. Otherwise, Point 1 & 2 must be repeated.

4. Execution & Results Analysis. This section comprises of dataset description, evaluation matrices, followed by results analysis.

4.1. Dataset. We have considered here 10 different real time datasets to experiment using our proposed approach. The details is shown in Table 4.1.

4.2. Evaluation Matrices. We have used here Q-modularity, Normalized Mutual Information, Mean Reciprocal Rank, and Mean Average Precision as the evaluation matrices.

Table 4.1: Dataset [31]

Sl No.	Dataset	Nodes	Edges
1	Karate	34	78
2	Football	115	613
3	Dolphins	62	159
4	Polbooks	105	441
5	Cora	2,708	5,429
6	Facebook	4,039	81,800
7	Artists	50,515	819,306
8	CiteSeer	3,312	4,732
9	Polblogs	1,490	16,718
10	School	68	220

- Q-Modularity [32]: This measure is calculated as:

$$Q = \frac{1}{2m}\sum_{pq}(A_{pq} - \frac{l_p l_q}{2m})\delta(t_p, t_q) \tag{4.1}$$

  where, A represents an adjacency matrix, m represents the quantity of edge, l_p represents the degrees of the p-th node.

- Normalized Mutual Information [33]: NMI is defined as:

$$NMI(P,Q) = \frac{-2\sum_{l=1}^{c_P}\sum_{m=1}^{c_Q} R_{lm}\log(\frac{R_{lm}.N}{R_{l.}R_{.m}})}{\sum_{l=1}^{c_P} R_{l.}\log(\frac{R_{l.}}{N}) + \sum_{m=1}^{c_Q} R_{.m}\log(\frac{R_{.m}}{N})}, \tag{4.2}$$

  where, c_P & c_Q are considered as the numeral of the clusters in the partition, $P(Q)$. The total number of the users in the error matrix are depicted as $R_{l.}(R_{.m})$. N is considered here as the entire users in the network.

- Mean Reciprocal Rank [34]: MRR is calculated as:

$$\frac{1}{|p|}\sum_{a=1}^{|p|}\frac{1}{q_a} \tag{4.3}$$

  where $|p|$ is a total number of queries, q_a is the rank position of a first relevant community among all the communities retrieved for the a^{th} query. A value of MRR ranges between 0 and 1.
- Mean Average Precision: MAP can be defined as:

$$\frac{1}{|p|}\sum_{a=1}^{p} AvgP_a \tag{4.4}$$

  where $|p|$ is a total number of queries, MAP is the mean of Average Precision (AvgP) of each query in a query. A value of MAP ranges between 0 and 1.

4.3. Results Analysis. Four scenarios were created to assess the modularity of different approach combinations in order to evaluate the proposed approach.

1. To find communities in a social network, the auto-encoder approach is used to extract features from the original proximity matrix. Case (1)'s label is written as 'AutoEnc'.

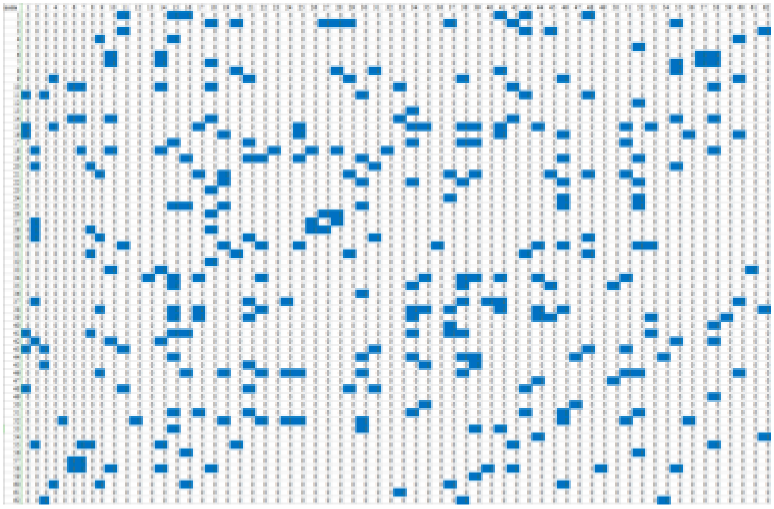


Fig. 4.1: Original Proximity Matrix of Dolphin Network

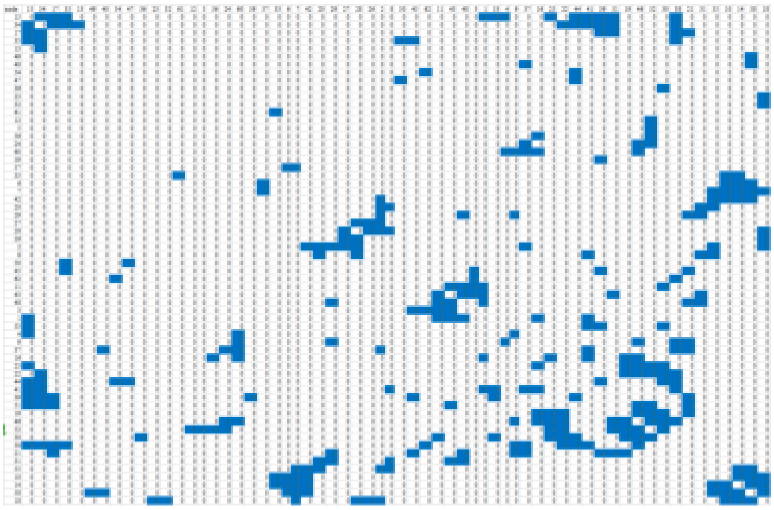


Fig. 4.2: Reorganized Proximity Matrix of Dolphin Network

2. To discover communities in a social network, the auto-encoder approach is used to extract characteristics from the rebuilt proximity matrix. Case (2)'s label is written as 'ReMat+AutoEnc'.
3. To bring out the characteristics of the primary proximity matrix in a real-time network to recognize community, the CNN-based auto-encoder approach is used. Case (3)'s label is written as 'CNN+AutoEnc'.
4. To extract the characteristics of a rebuilt proximity matrix in a real-time network to recognize community, an CNN-based auto-encoder approach is used. Case (4) is labeled as 'AutoEnc+ReMat+CNN'.

4.3.1. Execution of Reorganized Proximity Matrices. The process of the reorganization of proximity matrices is already elaborated in section 3. For the visualisation of the reorganized proximity matrices, we have considered two datasets as the samples. Figures 4.2 & 4.4 exhibit the reconstructed proximity matrices of dolphin and karate club network.

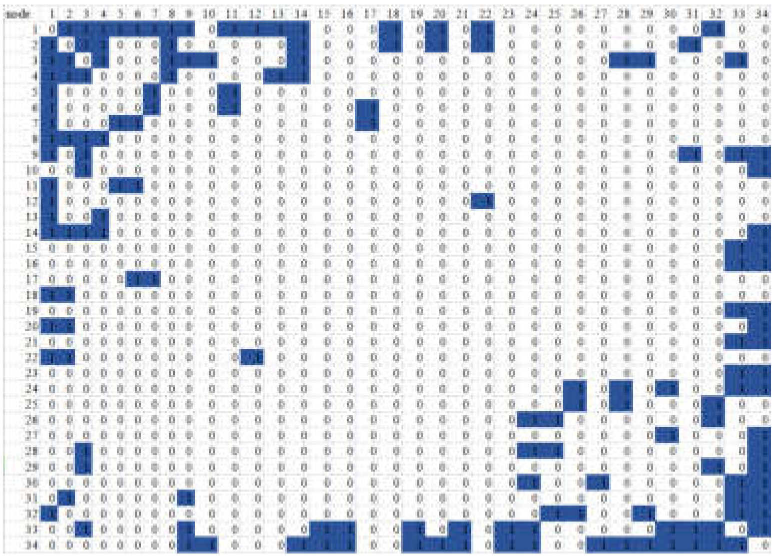


Fig. 4.3: Original Proximity Matrix of Karate Network

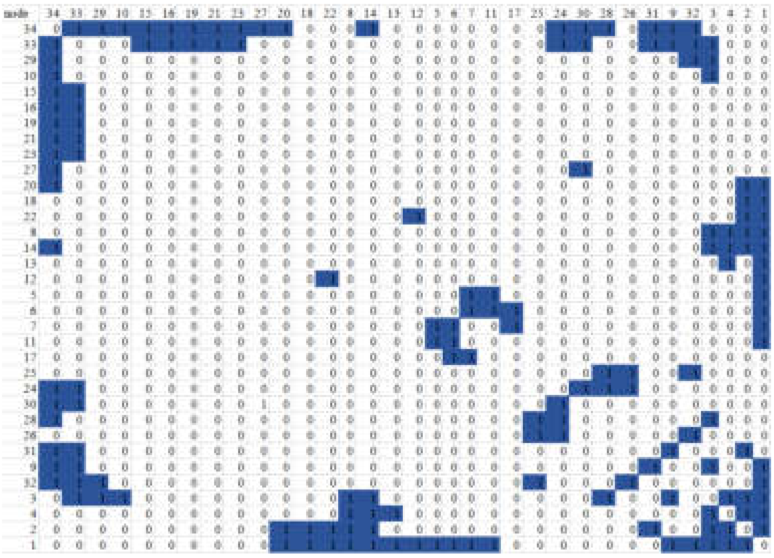


Fig. 4.4: Reorganized Proximity Matrix of Karate Network

4.3.2. Execution of Evaluation Matrices. This subsection provides the modularity score of all the four cases as shown in Table 4.2. Among all the four scenarios, ‘AutoEnc+ReMat+CNN’ has attained the highest modularity score. That is why we have considered the fourth case for comparing the modularity score with other existing algorithms. Here, we have considered other three popular algorithms for comparison, namely, ‘Kmeans+NetRA’,‘Kmeans+Node2Vec’, and ‘Kmeans+SDNE’. For all the 10 number of real time datasets, our proposed approach has gained the highest modularity score for community identification. The rebuilt proximity matrix is essentially a representation of the network’s structure learned by the auto-encoder. The quality of the proximity matrix has a significant impact on how accurately communities are detected. A common technique for dimensionality reduction is auto-encoders. The reconstructed matrix’s properties should

Table 4.2: Modularity Score of 4 Scenarios

Sl No.	Dataset	AutoEnc	ReMat+ AutoEnc	CNN+ AutoEnc	AutoEnc+ ReMat+ CNN
1	Karate	0.158	0.272	0.292	0.327
2	Football	0.539	0.629	0.730	0.768
3	Dolphins	0.484	0.539	0.549	0.625
4	Polbooks	0.544	0.593	0.637	0.683
5	Cora	0.358	0.397	0.478	0.483
6	Facebook	0.628	0.728	0.794	0.835
7	Artists	0.472	0.493	0.528	0.632
8	CiteSeer	0.469	0.528	0.573	0.624
9	Polblogs	0.372	0.448	0.472	0.527
10	School	0.573	0.576	0.638	0.735

Table 4.3: Modularity Score Compared with Existing Algorithms

Sl No.	Dataset	Kmeans+ NetRA	Kmeans+ Node2Vec	Kmeans+ SDNE	AutoEnc+ ReMat+ CNN
1	Karate	0.273	0.284	0.264	0.327
2	Football	0.528	0.618	0.593	0.768
3	Dolphins	0.528	0.492	0.519	0.625
4	Polbooks	0.492	0.439	0.542	0.683
5	Cora	0.293	0.346	0.274	0.483
6	Facebook	0.639	0.629	0.737	0.835
7	Artists	0.583	0.428	0.484	0.632
8	CiteSeer	0.529	0.553	0.514	0.624
9	Polblogs	0.384	0.418	0.474	0.527
10	School	0.618	0.531	0.683	0.735

preserve pertinent data while becoming less dimensional. Since the network is real-time, the characteristics of the rebuilt proximity matrix is generated quickly and efficiently. Real-time networks require low-latency processing, so the auto-encoder approach should be designed to produce the matrix in a timely manner. A rebuilt proximity matrix generated by a CNN-based auto-encoder is required for accurate, efficient, and flexible community detection in real-time networks. These qualities influence the approach's quality of community detection, scalability, tolerance to noise and changes, and overall efficacy in real-time applications. These are significance that our proposed approach (Case 4) has improved the performance among all the mentioned existing methods in essence.

The result is displayed in Table 4.3. Our proposed one, 'AutoEnc+ReMat+CNN' has attained the highest modularity score in Facebook network with 83.5% and achieved lowest in Karate club network with 32.7%. While 'Kmeans+NetRA' generates its better modularity score in Facebook and School network with 63.9% and 61.8% consecutively, 'Kmeans+Node2Vec' method provides its best modularity score in Football and Facebook network with 61.8% and 62.9%. 'Kmeans+SDNE' method attained its best modularity score in Facebook network with 73.7%.

Table 4.4 represents the NMI score comparison with the existing algorithms. In this experiment, our proposed 'AutoEnc+ReMat+CNN' approach has out-beats the other algorithms. Our method has achieved the highest NMI score in Karate network with 100% follwed by Dolphins network with 95%, Football network with 93% etc. 'Kmeans+SDNE' has also generated the better values of NMI than the other two existing

Table 4.4: NMI Score Compared with Existing Algorithms

Sl No.	Dataset	Kmeans+ NetRA	Kmeans+ Node2Vec	Kmeans+ SDNE	AutoEnc+ ReMat+ CNN
1	Karate	0.63	0.81	0.88	1
2	Football	0.59	0.72	0.83	0.93
3	Dolphins	0.65	0.75	0.87	0.95
4	Polbooks	0.62	0.68	0.72	0.78
5	Cora	0.53	0.59	0.60	0.62
6	Facebook	0.70	0.74	0.78	0.84
7	Artists	0.58	0.64	0.68	0.73
8	CiteSeer	0.61	0.67	0.68	0.69
9	Polblogs	0.52	0.63	0.66	0.71
10	School	0.71	0.83	0.85	0.91

Table 4.5: MRR Score Compared with Existing Algorithms

Sl No.	Dataset	Kmeans+ NetRA	Kmeans+ Node2Vec	Kmeans+ SDNE	AutoEnc+ ReMAT+ CNN
1	Karate	0.62	0.69	0.78	0.82
2	Football	0.66	0.73	0.81	0.84
3	Dolphins	0.71	0.79	0.83	0.88
4	Polbooks	0.68	0.74	0.79	0.83
5	Cora	0.63	0.71	0.75	0.81
6	Facebook	0.72	0.78	0.84	0.90
7	Artists	0.65	0.74	0.79	0.85
8	CiteSeer	0.63	0.69	0.74	0.80
9	Polblogs	0.59	0.66	0.71	0.78
10	School	0.75	0.80	0.84	0.92

algorithms. Karate club network has achieved 88% followed by Dolphins network with 87%, Footbal network with 83% NMI score in ‘Kmeans+SDNE’ method. ‘Kmeans+Node2Vec’ method provides its best NMI score in Karate club network with 81%, followed by Dolphins with 75% and Facebook with 74%. ‘Kmeans+NetRA’ generates its best NMI score in School network with 71%, followed by Facebook with 70%.

MRR score is also considered as one of the evaluation measures which displayed in Table 4.5. After the experiments, our proposed approach has attained the best MRR score in School network with 92%, followed by Facebook with 90%. Rest of the datasets have also performed well in our proposed approach. On the other hand, ‘Kmeans+NetRA’ generates its best MRR score in School network with 75%, followed by Facebook with 72% and Dolphins with 71%. ‘Kmeans+Node2Vec’ method provides its best MRR score in School network with 80%, followed by Dolphins network with 79% and Facebook with 78%. ‘Kmeans+SDNE’ has also generated the better values of MRR than the other two existing algorithms. It has achieved 84% in both Facebook and School network, followed by Dolphins with 83% and Football with 81%.

Table 4.6 represents the MAP score comparison with the existing algorithms. Our proposed method has outperformed among all the existing methods. It has attained the highest MAP score in School network with 88%, followed by Facebook with 87% and Artists with 83%. ‘Kmeans+NetRA’ generates its best MAP score in School network with 73%. School network has performed well in ‘Kmeans+Node2Vec’ also with 78% MAP score. But ‘Kmeans+SDNE’ has outperformed the other two methods and generate its best MAP score in Facebook network with 82%, followed by School with 81%.

Table 4.6: MAP Score Compared with Existing Algorithms

Sl No.	Dataset	Kmeans+ NetRA	Kmeans+ Node2Vec	Kmeans+ SDNE	AutoEnc+ ReMAT+ CNN
1	Karate	0.58	0.66	0.76	0.80
2	Football	0.63	0.71	0.78	0.82
3	Dolphins	0.65	0.72	0.75	0.81
4	Polbooks	0.64	0.70	0.76	0.80
5	Cora	0.58	0.68	0.72	0.78
6	Facebook	0.69	0.75	0.82	0.87
7	Artists	0.61	0.71	0.77	0.83
8	CiteSeer	0.60	0.66	0.71	0.78
9	Polblogs	0.57	0.64	0.69	0.75
10	School	0.73	0.78	0.81	0.88

5. Conclusion and Future Work. This research paper proposes a combined auto-encoder and CNN-based deep community identification approach for real time networks. During our experiments, we have first evaluated the modularity score on selected datasets using four different cases. Our proposed combined CNN and auto-encoder based method provides the prominent results on all the datasets. That is why we have considered our combined approach for evaluation and compare it with other existing approaches. To gather spatial adjacency matrices and reorganized proximity matrices, a novel matrix reorganization approach is proposed here. The matrix extends the standard proximity matrix with spatial closeness, obtaining obvious subspace features, and making convolutional processes simple and rapid to extract network spatial localisation. In this paper, the 'AutoEnc+ReMat+CNN' based approach is designed to obtain spatial eigenvectors, which spontaneously bring out the graph spatial properties and improve the modularity score. The combined model of the 'AutoEnc+ReMat+CNN'based community identification approach serves as the basis for community identification in a dynamic environment of the network.

The total amount of neurons in the input and output layers remains constant once the DL-based method is applied in spite of our approach 'AutoEnc+ReMat+CNN', which can be a useful investigation of the network embedding method. The interactions among the users in a real time community may change dynamically. Therefore, enhanced time-sequence approaches are necessary to draw out spatio-temporal properties for arbitrary real-time networks as future approach.

REFERENCES

[1] Yang, D., Liao, X., Wei, J., Chen, G. & Cheng, X. Modeling information diffusion with the external environment in social networks. *Journal Of Internet Technology.* **20**, 369-377 (2019)

[2] Perozzi, B., Al-Rfou, R. & Skiena, S. Deepwalk: Online learning of social representations. *Proceedings Of The 20th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining.* pp. 701-710 (2014)

[3] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781.* (2013)

[4] Tao, Z., Liu, H., Li, J., Wang, Z. & Fu, Y. Adversarial graph embedding for ensemble clustering. *International Joint Conferences On Artificial Intelligence Organization.* (2019)

[5] Wang, C., Wang, C., Wang, Z., Ye, X., Yu, J. & Wang, B. DeepDirect: Learning directions of social ties with edge-based network embedding. *IEEE Transactions On Knowledge And Data Engineering.* **31**, 2277-2291 (2018)

[6] Cao, J., Jin, D., Yang, L. & Dang, J. Incorporating network structure with node contents for community detection on large networks using deep learning. *Neurocomputing.* **297** pp. 71-81 (2018)

[7] Cavallari, S., Zheng, V., Cai, H., Chang, K. & Cambria, E. Learning community embedding with community detection and node embedding on graphs. *Proceedings Of The 2017 ACM On Conference On Information And Knowledge Management.* pp. 377-386 (2017)

[8] Yu, W., Zheng, C., Cheng, W., Aggarwal, C., Song, D., Zong, B., Chen, H. & Wang, W. Learning deep network representations with adversarially regularized autoencoders. *Proceedings Of The 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining.* pp. 2663-2671 (2018)

[9] Xu, Y., Chi, Y. & Tian, Y. Deep convolutional neural networks for feature extraction of images generated from complex

networks topologies. *Wireless Personal Communications.* **103** pp. 327-338 (2018)
[10] Xu, J., Li, M., Jiang, J., Ge, B. & Cai, M. Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network. *IEEE Access.* **7** pp. 92248-92258 (2019)
[11] Resler, A., Yeshurun, R., Natalio, F. & Giryes, R. A deep-learning model for predictive archaeology and archaeological community detection. *Humanities And Social Sciences Communications.* **8** (2021)
[12] Li, S., Jiang, L., Wu, X., Han, W., Zhao, D. & Wang, Z. A weighted network community detection algorithm based on deep learning. *Applied Mathematics And Computation.* **401** pp. 126012 (2021)
[13] Bhatia, V. & Rani, R. A distributed overlapping community detection model for large graphs using autoencoder. *Future Generation Computer Systems.* **94** pp. 16-26 (2019)
[14] Ferraro, A., Moscato, V. & Sperli, G. Deep learning-based community detection approach on multimedia social networks. *Applied Sciences.* **11**, 11447 (2021)
[15] Al-Andoli, M., Tan, S. & Cheah, W. Distributed parallel deep learning with a hybrid backpropagation-particle swarm optimization for community detection in large complex networks. *Information Sciences.* **600** pp. 94-117 (2022)
[16] Essaid, M. & Ju, H. Deep Learning-Based Community Detection Approach on Bitcoin Network. *Systems.* **10**, 203 (2022)
[17] Sun, K., Meng, F. & Tian, Y. Multi-level wavelet-based network embedded with edge enhancement information for underwater image enhancement. *Journal Of Marine Science And Engineering.* **10**, 884 (2022)
[18] Zhang, Y., Wang, Y., Chen, X., Jiang, X. & Zhou, Y. Spectral–spatial feature extraction with dual graph autoencoder for hyperspectral image clustering. *IEEE Transactions On Circuits And Systems For Video Technology.* **32**, 8500-8511 (2022)
[19] Kulatilleke, G., Portmann, M. & Chandra, S. SCGC: Self-supervised contrastive graph clustering. *ArXiv Preprint ArXiv:2204.12656.* (2022)
[20] Qian, J., Song, Z., Yao, Y., Zhu, Z. & Zhang, X. A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes. *Chemometrics And Intelligent Laboratory Systems.* pp. 104711 (2022)
[21] Lim, H., Ullah, I., Han, Y. & Kim, S. Reinforcement learning-based virtual network embedding: A comprehensive survey. *ICT Express.* (2023)
[22] Goel, S. & Kumar, R. Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels. *Neurocomputing.* **315** pp. 425-438 (2018)
[23] Goel, S. & Kumar, R. Collaboratively augmented UIP–Filtered RIP with relevancy mapping for personalization of web search. *Information Sciences.* **547** pp. 163-186 (2021)
[24] Hoffman, M., Steinley, D., Gates, K., Prinstein, M. & Brusco, M. Detecting clusters/communities in social networks. *Multivariate Behavioral Research.* **53**, 57-73 (2018)
[25] Wu, L., Zhang, Q., Chen, C., Guo, K. & Wang, D. Deep learning techniques for community detection in social networks. *IEEE Access.* **8** pp. 96016-96026 (2020)
[26] Zhao, S., Peng, R., Hu, P. & Tan, L. Heterogeneous Network Embedding: A Survey. *CMES-Computer Modeling In Engineering & Sciences.* **137** (2023)
[27] Pham, P., Nguyen, L., Nguyen, N., Kozma, R. & Vo, B. A hierarchical fused fuzzy deep neural network with heterogeneous network embedding for recommendation. *Information Sciences.* **620** pp. 105-124 (2023)
[28] Abbas, K., Abbasi, A., Dong, S., Niu, L., Chen, L. & Chen, B. A Novel Temporal Network-Embedding Algorithm for Link Prediction in Dynamic Networks. *Entropy.* **25**, 257 (2023)
[29] Ding, Y., Zhai, Y., Hu, M. & Zhao, J. Deep Forest Auto-Encoder for Resource-Centric Attributes Graph Embedding. *Pattern Recognition.* pp. 109747 (2023)
[30] Vo, T. An integrated topic modeling and auto-encoder for semantic-rich network embedding and news recommendation. *Neural Computing And Applications.* pp. 1-16 (2023)
[31] Al-Andoli, M., Tan, S., Cheah, W. & Tan, S. A review on community detection in large complex networks from conventional to deep learning methods: A call for the use of parallel meta-heuristic algorithms. *IEEE Access.* **9** pp. 96501-96527 (2021)
[32] Lázár, A., Abel, D. & Vicsek, T. Modularity measure of networks with overlapping communities. *Europhysics Letters.* **90**, 18001 (2010)
[33] Kv, T. On normalized mutual information: measure derivations and properties. *Entropy.* **19**, 631 (2017)
[34] Goel, S., Kumar, R., Kumar, M. & Chopra, V. An efficient page ranking approach based on vector norms using sNorm (p) algorithm. *Information Processing & Management.* **56**, 1053-1066 (2019)

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 31, 2023
Accepted: Oct 5, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 999–1010, DOI 10.12694/scpe.v24i4.2479

USER SENTIMENT ANALYSIS METHODS FOR ELDERLY SOCIAL MEDIA NETWORKS

SHUAIZHI SHEN*

Abstract. For the analysis of user sentiment in social media networks for the elderly population, emotional sentences are first extracted to classify movie reviews. Afterwards, social network data of the elderly population based on user search behavior is analyzed. The movie reviews of elderly social media users are analyzed for rating prediction. The research results indicate that the accuracy of sentiment classification results is in descending order of Dirichlet, maximum entropy, and support vector machine. The highest classification accuracy of the three algorithms is 87.1%, 86.9%, and 86.5%, respectively. The classification accuracy of the first level classifiers of Dirichlet, maximum entropy, and support vector machine are 90.7%, 88.7%, and 87.4%, respectively. The classification accuracy of the second level classifier is 86.7%, 83.7%, and 80.4%, respectively. The predictive analysis results of the research method are superior to those generated by using Slope One. The method proposed in the study can promote emotional analysis of film review texts, improving the analysis accuracy.

Key words: Elderly population; Social media networks; Emotional analysis; Film reviews; Personalized recommendations

1. Introduction. In recent years, the rapid development of the Internet has made communication between people more and more convenient. With the emergence of various new multimedia social platforms, communication around these platforms is also increasing day by day. Every day, tens of thousands of users express their opinions on these platforms [1]. In textual information, there are abundant vocabulary with emotional tendencies. These words can well reflect the user's emotional state at a certain moment [2]. The comment text of movies belongs to the common comment information on multimedia social media platforms, which contains the emotions and opinions of social media network users. Moreover, by analyzing the same specific movie, different users can better understand their emotional tendencies and analyze their emotions. At the same time, analyzing the emotions in movie reviews will help clarify everyone's overall view of the movie, thereby promoting adjustments in movie promotion and scheduling. There is currently a lot of research on film criticism, but in the traditional field of criticism, it is difficult to break through emotional analysis [3]. Personalized movie recommendations are generally based on proactive recommendations, which use software to conduct a series of data on users who have operated on the platform, including useful historical footprints, personal information, and operations, in order to infer user preferences. After organizing user data, it is then analyzed and combined with user preferences, Furthermore, targeted movie recommendations can be made to users. Due to the fact that on multimedia social media platforms, movie review information not only affects audience choices and viewing decisions to a certain extent, but also helps producers obtain feedback information from audiences in a timely manner after watching a movie, allowing for targeted promotion. In order to analyze the emotions of the elderly towards movies, make targeted recommendations based on emotions, and expand the film market for the elderly, this study intends to use comment text information from multimedia social platforms to conduct research on film recommendation and rating prediction methods. The article is mainly divided into five parts. The first part is an introduction, which mainly introduces the research background and purpose. The second part is a literature review, mainly summarizing the current research situation of different scholars at home and abroad. The third part is the research method, the fourth part is the result analysis, and the fifth part is the conclusion.

2. Related Work. Multimedia social platforms are gradually becoming text databases for viewpoint and sentiment analysis. Higher requirements have also been put forward for the mining and analysis of public opinion information. Chakraborty K and other scholars proposed a detection method based on social network user similarity selection to analyze the emotions in social media data. Community-based user data technology and

*School of Media and Art Design, Guilin University of Aerospace Technology, Guilin, 541004, China (PHDshen2023@163.com)

emotional evaluation methods have been studied and classified. The research results are conducive to analyzing the emotions of social media data and promoting accurate understanding of social media expression content [4]. Abid et al. designed an emotional analysis method based on bidirectional recursive convolutional neural network to solve the low accuracy of emotional data analysis in multi-source social media. The research results show that compared with conventional sentiment analysis methods, the accuracy of this method is greatly improved, reaching 89.67% [5]. The traditional Convolutional neural network cannot accurately capture the semantic features in emotion analysis. Therefore, researchers such as Alam proposed a domain specific distributed word representation method based on social media text resources. Then, the convolutional neural network is expanded. The results show that this method can reduce the convolution dimension and expand the size of Receptive field. It can combine different inflation rates to obtain long-term contextual semantic information and reduce computational costs [6]. Ducange et al. designed a decision support system based on machine learning text classification to improve the accuracy of online social media sentiment analysis. By combining sentiment analysis engines, user emotions in comments are analyzed across sources. The research results indicate that the accuracy of sentiment analysis using this method is as high as 90%, which is beneficial for analyzing users' emotions and making improvements [7]. Vasishtha S and other scholars proposed an unsupervised analysis system based on fuzzy rules for sentiment analysis of posts on social media. Fuzzy rules are used to calculate and analyze the emotions in posts. The research results indicate that this method can be applied to all datasets. Compared to traditional methods, it has higher performance [8].

Social media has become a new channel for product opportunities. Therefore, Jeong et al. proposed an opportunity recognition method based on social media data. Through theme modeling and sentiment analysis, product opportunities are mined and identified. Emotional analysis is used to evaluate product satisfaction. The research results indicate that this method can promote users' understanding of product emotions and recommend products based on customer needs [9]. The mixing of multiple languages increases the difficulty of recognition. Therefore, Bansal N and other scholars designed a language recognition system with mixed data sets and logistic regression classifiers. The results show that the accuracy of Logistic regression is the highest, 86.63. This method helps to improve the recognition accuracy of mixed languages [10].

Based on the above research, it can be seen that machine learning methods have certain effects in emotional analysis of social media data. However, there is limited analysis of user emotions in social media networks for the elderly. Given the increasing popularity of movie reviews on social media, in order to analyze the emotions of elderly users, strengthen movie recommendations for elderly users, and expand the video market for the elderly, research will combine machine learning and other methods to analyze the user emotions of movie reviews.

3. User sentiment analysis of social media networks targeting the elderly population. In order to analyze the emotional analysis of elderly social media network users towards movies, we aim to expand the film market for the elderly population through emotional analysis. The study first analyzes social network data based on user search behavior, then extracts emotional sentences from comments of the elderly population, classifies movie comments, and finally analyzes movie comments of elderly social media users before conducting rating predictions.

3.1. Analysis of Social Network Data for the Elderly Based on User Search Behavior. Currently, with the rapid development of social media, a large amount of information has emerged on social media platforms. Analyzing these data can provide more valuable information. So, research based on cloud technology has become a very important way to obtain information. By exploring different search patterns, different search results are revealed from different perspectives. Then the retrieval quality and retrieval behavior are analyzed [11]. An algorithm that combines user search behavior has been proposed. Based on the user's search logs, their relevant behaviors are extracted. Analyzing user behavior patterns can increase the volume of search information services for users and find keywords. This operation can provide user behavior data for future rating predictions, improving rating accuracy.

As a stable distributed computing architecture, Hadoop's core is HDFS and MapRecurce. The most important aspect of the Hadoop architecture is the implementation of distributed storage for the underlying framework through HDFS. Distributed programming support for parallel tasks can be achieved through MapReduce. q and C represent search behavior and click through volume, respectively. Users generally ignore the return

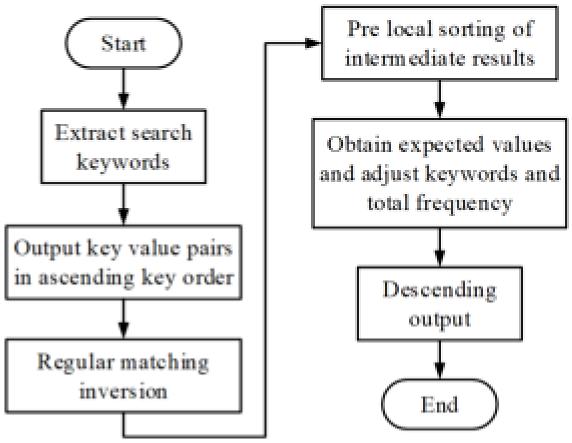


Fig. 3.1: The statistical process of users' search keywords

result URL, which greatly affects their search behavior. This defect can be remedied through equation 3.1.

$$U_q = \sum_{i=1}^{n} c(A,q) * \text{click}(A,q) \tag{3.1}$$

In equation 3.1, $c(A,q)$ represents the click through volume of the balance factor click(A,q) of webpage A. The higher the U_q value, the more popular the page is. When a user performs a search activity, if they believe there are similarities in the search activity, they will stay for a period of time. This behavior does not affect users' satisfaction with search activities. For this purpose, equation 3.2 is used to represent the weight of the search time.

$$\text{Time}(A,q) = \frac{t_i}{\sum_{i=1}^{n} t_i} \tag{3.2}$$

In equation 3.2, t_i is the set of words q that users find when spending time browsing websites. In the cloud computing search process, there is a correlation between pages i and j, but there is a significant difference in their weights [12]. When performing N iterations, in $[0,t]$, the page construction matrix clicked by the user is C_{NxN}. $C_{i,j}$ represents the number of clicks on i and j. If $C_{i,j}$ and $C_{j,k}$ are greater than 0, there is a relationship between I, j, and k, as shown in equation 3.3.

$$K(A,T_i) = \lambda(\text{ID}_A, \text{ID}_{T_i}) \tag{3.3}$$

In equation 3.3, $K(A,T_i)$ represents the correlation between A and T_i. $\lambda(\text{ID}_A$ is used to describe the relevant values found for two-page IDs. When users search for their desired content, their search behavior characteristics and browsed information will be recorded in the log directory. Therefore, the main research content of the article is the data analysis of logs. User log analysis includes searching for key words and user time periods. The specific implementation steps are shown in Figure 3.1.

From Figure 3.1, it can be seen that the input data of the Map function is a segment in the log text. The first MapReduce calculation process analysis found that the text content in the first column is a string. The second column is the number of times this text is presented. During the second MapReduce calculation, a set interface is used to pass the parameters to the user. Then the task is controlled by the main control program [13]. However, when conducting searches, users choose web pages based on a large number of similar topics. The imbalance is inevitable. Therefore, in addition to simple connections, an implicit connection also needs to be considered. Therefore, the PR expression for calculating the conventional page ranking and page

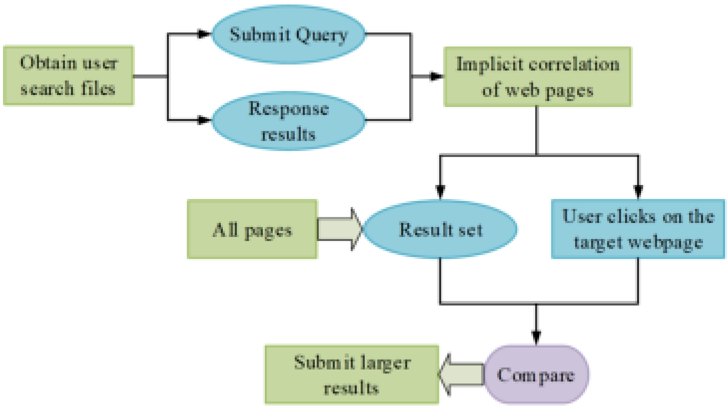


Fig. 3.2: Detailed comparison process for real-time feedback

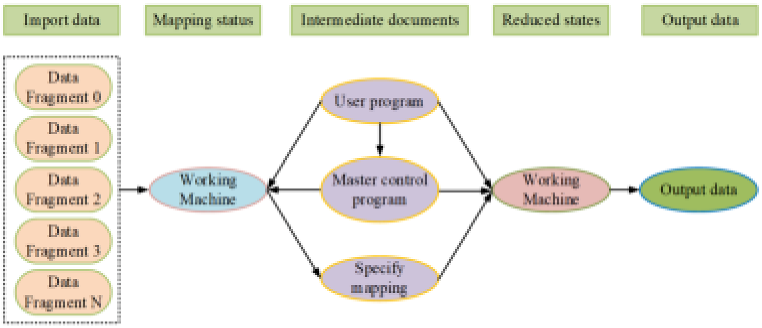


Fig. 3.3: Map/Reduce model structure

X is shown in equation 3.4

$$PR(X) = \sum_{(X,T_i)\in R} \left(\frac{PR(T_i)\left(\delta_1 f(X,T_i) + \delta_2 T(X,q) + \delta_3 k(X,T_i)\right)}{\sum_{k=1}^{M} click(T_i,X)} \right) \tag{3.4}$$

In equation 3.4, the parameters δ_1, δ_2 and δ_3 all represent the influencing factors. $\delta_1 + \delta_2 + \delta_3 = 1$. E is used to describe the total number of web pages. d represents the damping factor. (T_i, X) represents the number of clicks on web pages T_i and X. The higher the click through volume, the greater its relevance. During user retrieval using cloud computing, a specific example of real-time feedback comparison can be obtained. This process is shown in Figure 3.2. From Figure 3.2, it can be seen that the user selects the target webpage after receiving a set of results from the search engine. The relevant ID number can be obtained. In addition, based on the implicit value of association, the result set is compared with the implicit degree of association. The close correlation of web pages is fed back to users as new search results [14]. The Hadoop distributed computing platform can analyze and mine data from multiple aspects to infer user search preferences. Based on this, search results are correlated with content to infer users' emotional tendencies. The Map/Reduce model is shown in Figure 3.3.

From Figure 3.3, it can be seen that the Map/Reduce model mainly consists of five parts. They are input data, mapping state, intermediate document, reduced state, and output data. Hadoop cloud computing is used to count user query records. Due to the large number of logs queried, Mapreduce needs to be calculated in parallel. Through this operation, the number of keyword queries, the ranking of URL feedback results, and

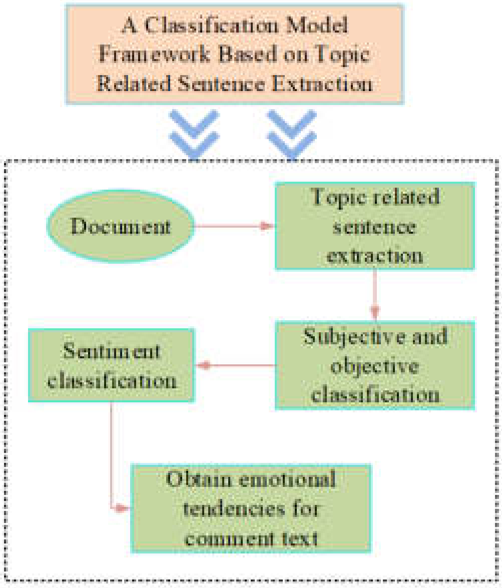


Fig. 3.4: A Classification Model Framework Based on Topic Related Sentence Extraction

the number of users can be calculated. Based on the user's query patterns and preferences, a summary of association algorithms is conducted. User search topics can improve user search efficiency.

3.2. Classification of movie reviews based on sentiment sentence extraction. In movie reviews, text information often includes multiple movie themes. In addition, other movie theme reviews also include some emotional words [15]. Therefore, when classifying words based on emotions, words and sentences that are not related to the movie theme should be eliminated. Given the divergent nature of film review themes, a method for emotion classification is proposed based on this research. That is to extract emotional theme related sentences from movie review texts. The model framework for sentiment classification is shown in Figure 3.4. From Figure 3.4, it can be seen that the extraction method is mainly divided into three stages. Firstly, sentences related to the topic are found in the text. Then, sentences related to the theme are analyzed to determine the subjectivity and objectivity of the text. Finally, the theme sentences related to subjectivity can be found in the text, which is the theme sentiment sentences. The sentiment analysis of text is an sentiment classifier constructed based on machine learning methods. Then the entire comment text's emotions are classified. Finally, the emotional orientation of the text is obtained [16]. Sentences related to the topic are divided into subjective and objective categories. Objectivity sentences are eliminated. Afterwards, sentiment classification is performed on the topic sentiment sentences. Machine learning methods are used to conduct emotional analysis on comment texts and assess their emotional tendencies. Common machine learning methods include support vector machine (SVM), maximum entropy, Latent Dirichlet Allocation (LDA) and other methods. SVM is a supervised learning model, which can analyze data, recognize high-dimensional patterns, classify emotions, Logistic regression, etc. The SVM model performs well in classification performance. The samples are divided into two types. A sample set is a collection of sample points $\{(X_1, C_1), (X_2, C_2), \ldots, (X_n, C_n)\}$. In the sample point set, C_j represents the category, with a value range of [-1,1]. A high-dimensional space vector is represented by a number X_i. The mathematical expression of the classification hyperplane is shown in equation 3.5.

$$w \cdot x - b = 0 \tag{3.5}$$

The two are parallel to the optimal hyperplane and closest to the support vector. SVM modeling method is used to find the optimal hyperplane, so as to transform the problem into the Quadratic programming

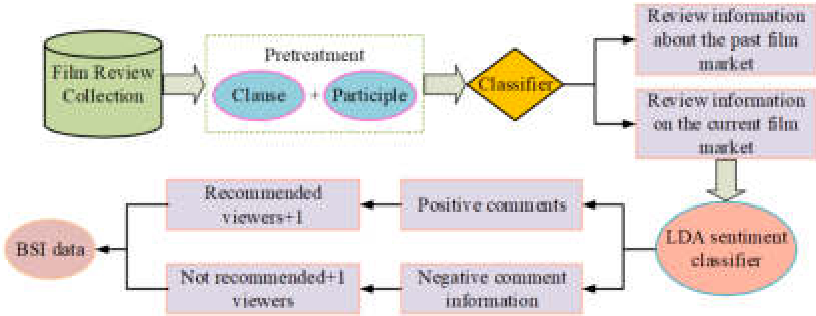


Fig. 3.5: Step diagram of user sentiment prediction

optimization [17]. Based on Lagrand's Mean value theorem, the hyperplane can be described by equation 3.6.

$$f(x) = \sum_{i=1}^{n}(a_i c_1 x_i^T) + b \tag{3.6}$$

The maximum entropy can predict the results of probability distribution. That is to say, the prediction result is to try to choose a uniformly distributed result. x represents the feature vector of the sample. $y$represents the category of the sample. The eigenvector values of sample x in $p(y|x)$ represent probability. The expression for maximum entropy is shown in equation 3.7

$$H(p) = -\sum_{(x,y)} p(y \mid x)\log(y \mid x) \tag{3.7}$$

Combined with adjustment function, known information can be presented relatively easily. So, the classification based on the maximum entropy model evolves into the optimization of feature functions. This feature function has constraints. The expression for obtaining the maximum entropy probability is shown in equation 3.8.

$$f_i(x,y) = \begin{cases} 1 & \text{if "meet the conditions"} \\ x & \text{otherwise} \end{cases} \tag{3.8}$$

The optimized maximum entropy expression is shown in equation 3.9.

$$P(y \mid x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)}{Z(x)} \tag{3.9}$$

In equation 3.9, $Z(x)$represents the factor. λ is used to describe the weight value of feature i. The maximum entropy model does not need to consider whether each feature value is related, nor does it require independent feature assumptions. However, based on attribute vectors, different attributes are randomly selected according to their differences [18]. Therefore, when studying multiple categories, the maximum entropy model is the most effective. On the basis of early emotional analysis, a prediction model that can automatically identify users' future emotional tendencies towards movies has been constructed. In movie reviews, comments on future movie emotions also include current movie sentiment reviews. Therefore, when conducting emotional analysis of the text, this factor should be included. Based on the previous emotional analysis, a new emotional prediction model has been constructed. The implementation steps for user sentiment prediction are described in Figure 3.5.

From Figure 3.5, the model has two classifiers. Therefore, it can be divided into two stages to analyze the emotions of words. The sentences recognized by the first level classifier are sentiment classified by the second level classifier. This classification will generate two types of surprises and disappointments. Among them, BSI data is a type of data in statistical economics. The expression is shown in equation 3.10.

$$\begin{cases} BSI(1) = \frac{TJ}{TJ+BTJ} \times 100\% \\ BSI(2) = \frac{BTJ}{TJ+BTJ} \times 100\% \end{cases} \tag{3.10}$$

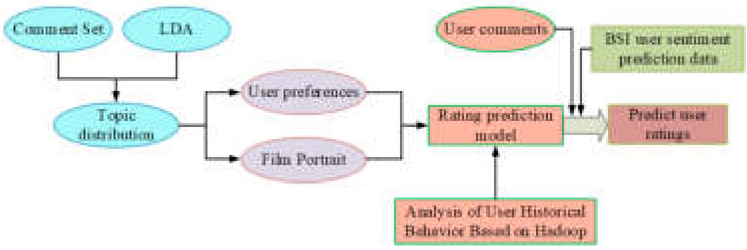


Fig. 3.6: Rating prediction process based on user comments

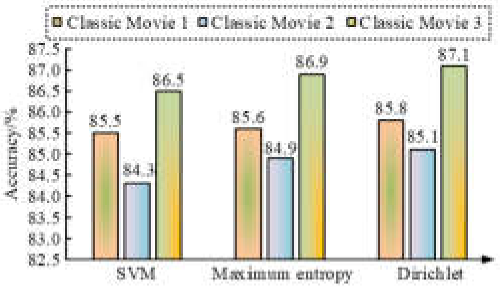
In equation 3.10, TJ and BTJ represent the number of recommended users and the number of non-recommended users, respectively. The first classifier is a classifier based on temporal order. The second classifier is the potential Dirichlet classifier. The accuracy of both classifiers has a significant impact on the accuracy of BSI. A secondary classifier is constructed in the study. It is used as an influential factor in Sentiment analysis.

3.3. Rating prediction based on film review analysis. With the development of the Internet, recommendation technology is increasingly widely used in multimedia social media. The recommendation system can accurately predict users' preferences and help them find the most suitable product for themselves. The current research focus is on how to accurately judge user preferences and product categories, thereby improving the accuracy of recommendations. A new emotion prediction model is constructed by utilizing words and symbols with emotional meanings such as comments' text information, user ratings, and emoticons. On this basis, a sentiment prediction model based on evaluation is proposed. This model not only considers the application of evaluation methods, but also needs to refine the evaluation text to mine information such as user preferences and movie classification [19]. First, through the topic model, the potential topic distribution in the comment information is determined. Then, through the Logistic regression model, each potential word that has an impact on the generation of comment information and scoring is searched out. The internal correlation between possible topics and actual comment information and ratings is determined. After understanding the preferences and search behavior of elderly users, a new method is proposed to accurately grade and predict text information. Then, the scores of the scoring prediction results can be sorted in descending order. Traditional recommendation models focus on the impact of evaluation results on elderly users. However, traditional models have cold start issues [20, 15]. After extracting the scoring features in the comment text, feature analysis can be performed to associate user preferences with product categories. This can alleviate the cold start issue during the recommendation process. In addition, the information in the comments can also be presented to elderly users by selecting the most representative comment information based on the potential topic distribution. The specific implementation process of a rating prediction model based on user comments is shown in Figure 3.6.

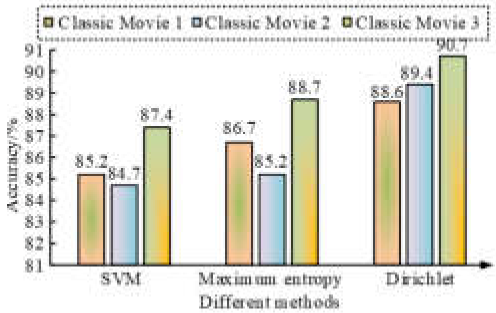
From Figure 3.6, the movie portrait module and rating prediction module are the two functional modules that the model must have. User preferences, movie portraits, and Hadoop based user behavior analysis are organically combined. Based on the regression model, a user rating prediction model is constructed. After training the model, the trained model is used to predict ratings for movies that have not been rated by users.

4. User emotions analysis in social media networks. The movie reviews of elderly social media network users are combined with machine learning, first level classifiers, and second level classifiers for sentiment analysis. According to movie review analysis, rating prediction is achieved.

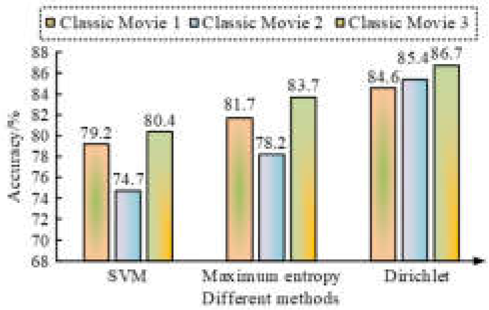
4.1. Emotional Analysis Results of Film Review Text. Based on machine learning for sentiment analysis, various methods are used to analyze emotions in various movies, including SVMs, maximum entropy, and Dirichlet. Then, based on this, a new model is established to analyze the text to determine the classification accuracy of different machine learning methods. In the experiment, three methods are still used, including SVM, maximum entropy and Dirichlet. The first level classifier is based on time series. The differentiated feature values are used as evaluation indicators for the classifier. In the experiment, the 5-fold crossover method is used to validate in the first level classifier and the second level classifier, respectively. The classification accuracy



(a) The accuracy of emotional classification in movies



(b) The classification results of the first level classifier on different feature items



(c) The classification results of a two-level classifier on different feature items

Fig. 4.2: Classification accuracy and sentiment analysis results of the first and second level classifiers

and sentiment analysis results of the first and second level classifiers are shown in Figure 4.2.

From Figure 4.4a, under machine learning, the accuracy of emotion classification results is in descending order of Dirichlet, maximum entropy, and support vector machine. The highest classification accuracy of the three algorithms is 87.1%, 86.9%, and 86.5%, respectively. From Figures 4.4b and 4.1c, in the sentiment analysis of the first and second level classifiers, the three classifiers have good classification performance for three different types of features. Especially when both unary and binary grammars are used simultaneously, their classification performance is better than the first two. The classification accuracy of the first level classifiers of Dirichlet, maximum entropy, and support vector machine are 90.7%, 88.7%, and 87.4%, respectively. The classification accuracy of the second level classifier is 86.7%, 83.7%, and 80.4%, respectively. From the classification results of the three classifiers, Dirichlet has better classification performance than the first two classifiers. Therefore, the method of analyzing the text of Dirichlet's film reviews will be integrated into the research method for the next step of emotional analysis.In the first and second level classifiers, the lack of distinction between emotions results in the presence of both positive and negative emotion words in the comment text. In addition, the mixing of different types of movies increases the difficulty of emotion classification. Thus, the classification results are not satisfactory. Therefore, the emotional analysis of the film review text continues. In the experiment, 5 PC servers are used to build a Hadoop distributed computing platform. Figure 4.2 shows the comparison results of nodes and accuracy based on different data volumes.

From Figure 4.3, the overall effectiveness of the research method is superior to the other two traditional methods. Hadoop is used to analyze the search behavior of elderly users, which can obtain relevant information needed by elderly users and be used for subsequent emotional analysis. In addition, cloud computing models will help better analyze user data. It can also effectively compensate for the shortcomings in the existing

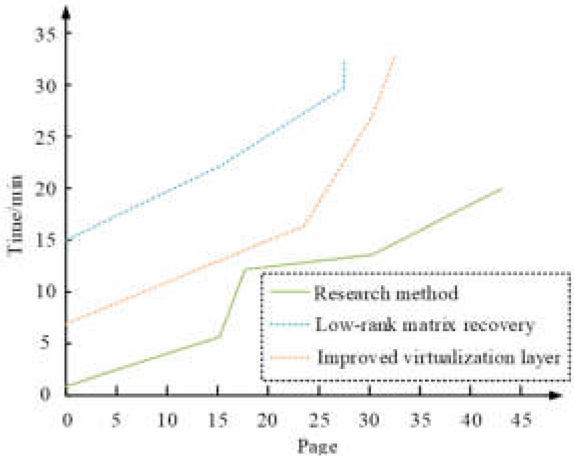


Fig. 4.3: Comparison results of three algorithms based on different page views

Table 4.1: Statistical results of user data on social media networks for the elderly population

Number	Film Category	Number of Users/People	Number of Movies/Films	Comment/Article
(1)	Action	13,171	229	101,719
(2)	Comedy	17,018	119	170,175
(3)	Affectional film	12,843	151	58,419
(4)	Science fiction film	8,657	120	86,562
(5)	Suspense film	2,353	61	23,522
(6)	Horror film	5,452	70	54,519
(7)	Total	59,494	750	494,916

Hadoop distributed computing architecture.

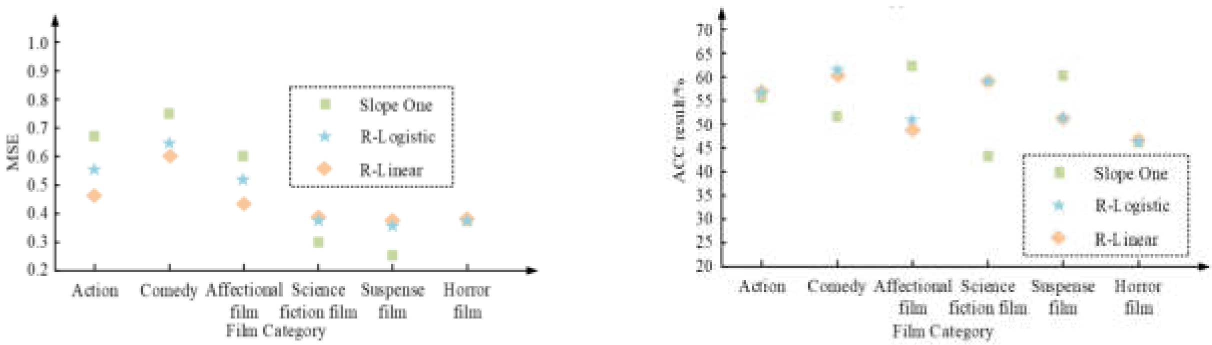
4.2. Scoring prediction results based on movie review analysis. The experiment collected a large amount of comment information from middle-aged and elderly social media network users in the data center, including 59494 elderly users, 750 movies, and 494916 comments. Among them, movies mainly include six types of domestic and foreign films: action films, comedy films, romance films, science fiction films, suspense films, and horror films. The user data statistics results are shown in Table 4.1.

The dataset of movie reviews was randomly divided into a training set and a testing set in an 8:2 ratio. The theme content and ratings in movie reviews are linked by HFT. The HFT source code is downloaded from the website. The collaborative filtering (CF) method means that users with the same preferences can choose the same movie. The subjective decision-making of traditional CF associated users is used to filter comment text. Slope One is a widely used commodity Collaborative filtering method. This method has the characteristics of simplicity and high efficiency, which is implemented by the open-source tool My Media Lite3.10. In this chapter's experiment, four methods, Slope One, HFT, R-Linear, and R-Logistic, are used as test systems. 5 tests are conducted to demonstrate the effectiveness of the system. The Mean squared error (MSE) and accuracy (ACC) results of the four methods are shown in Figure 4.5.

From Figure 4.5, the results of MSE are a comparison of the prediction accuracy of the scores. In actual scores, the size of the score is generally between 1 and 5. However, the scores obtained through Slope One, HFT, and linear regression are all decimal. Therefore, the evaluation values containing decimals were rounded to meet the calculation requirements of ACC. Among them, random allocation and temporal allocation have the highest MSE results on HFT, with 0.78 and 1.32, respectively. Random allocation and temporal allocation

(a) MSE Results (b) ACC Results

Fig. 4.5: MSE and ACC results of four methods



(a) MSE Results (b) ACC Results

Fig. 4.7: Prediction Results of Six Types of Movies Using Slope One, R-Linear, and R-Logistic Methods From Figure 4.7, the linear regression method has achieved good results in all six movie categories.

achieved the highest ACC results on Slope One, with 56.5% and 55.4%, respectively. Slope One, R-Linear, and R-Logistic methods are used to predict action, comedy, romance, science fiction, suspense, and horror films. The results are shown in Figure 4.7.

However, when the data suddenly changes, the results obtained by using the Slope One method to predict six categories have significant fluctuations. From the experimental results, it can be seen that the predictive analysis results of the research method are superior to the results generated by using Slope One. In the case of sparse data, the research method can still have good stability and small fluctuations, with good MSE and ACC results. When the recommendation system recommends a movie to users, they will first browse the review information of the movie before they can understand the general content of the movie. The selection of user representative comments is predicted. The results are shown in Figure 4.8.

From Figure 4.8, a video recommendation model has been established through the evaluation of viewers and text analysis of their comments. Finding the comment topic in the movie's comment text to obtain the topic distribution of the comment text can project both user information and movie information into the same space. The above process can obtain user preferences and movie portraits to establish a regression model. The theme distribution and prediction are linked to recommend movies to users. Combining predictive methods, representative evaluations were selected as support for this model. From the experimental data, the scoring prediction model has shown good performance and special adaptability to data sparsity.

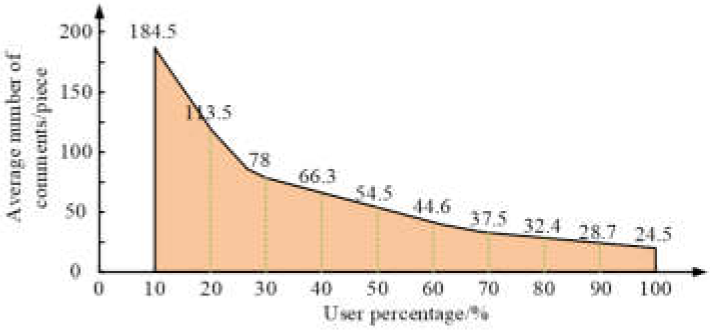


Fig. 4.8: Prediction of user representative comments selection

5. Conclusion. With the development of the Internet, there are more and more comment texts that integrate user emotions. There are more emotional commentary texts in the film field. To strengthen the user sentiment analysis of social media networks for the elderly population, movie reviews are analyzed through user search behavior. The results indicate that by analyzing the logs of elderly users, the behavioral patterns of users can be discovered. The accuracy of sentiment classification results is ranked from high to low by Dirichlet, maximum entropy, and SVM. The highest classification accuracy of the three algorithms is 87.1%, 86.9%, and 86.5%, respectively. The overall effectiveness of the research method is superior to the other two traditional methods. The predictive analysis results of the research method are superior to those generated by using Slope One. In the case of sparse data, the research method can still have good stability and small fluctuations, with good MSE and ACC results. Therefore, this study will have certain guiding significance and practical value. In the future, research will be conducted on automatic classification of subjective and objective texts.

Funding. The research is supported by: 2021 Guangxi Philosophy and Social Science Planning Research Project. Approval No:21FXW013.

REFERENCES

[1] Derakhshan, A. & Beigy, H. Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications Of Artificial Intelligence.* **85** pp. 569-578 (2019)

[2] Babvey, P., Borrelli, D., Lipizzi, C. & Galaxies, R. Digital Fingerprints of Discussions on Social Media. *IEEE Transactions On Computational Social Systems.* **8**, 294-307 (2021)

[3] Liu, H., Chatterjee, I., Zhou, M., Lu, X. & Abusorrah A. Aspect-Based Sentiment Analysis:, A. Survey of Deep Learning Methods. *IEEE Transactions On Computational Social Systems.* **7**, 1358-1375 (2020)

[4] Chakraborty, K., Bhattacharyya, S. & Bag, R. Survey of Sentiment Analysis from Social Media Data. *IEEE Transactions On Computational Social Systems.* **7**, 450-464 (2020)

[5] Abid, F., Li, C. & Alam, M. Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks. *Computer Communications.* **157**, 102-115 (2020)

[6] Alam, M., Abid, F., Cong, G. & Yunrong, L. Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications. *Computer Communications.* **154**, 129-137 (2020)

[7] Ducange, P., Fazzolari, M., Petrocchi, M. & Vecchio, M. An effective Decision Support System for social media listening based on cross-source sentiment analysis models. *Engineering Applications Of Artificial Intelligence.* **78** pp. 71-85 (2019)

[8] Vashishtha, S. & Susan, S. Fuzzy Rule based Unsupervised Sentiment Analysis from Social Media Posts. *Expert Systems With Applications.* **138**, 1-11283 (2019)

[9] Jeong, B., Yoon, J. & Lee, J. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal Of Information Management.* **48** pp. 280-290 (2019)

[10] Bansal, N., Goyal, V. & Rani, S. Experimenting Language Identification for Sentiment Analysis of English Punjabi Code Mixed Social Media Text. *International Journal Of E-Adoption (IJEA).* **12**, 52-62 (2020)

[11] Rathore, P., Soni, J., Prabakar, N., Palaniswami, M. & Santi, P. Identifying Groups of Fake Reviewers Using a Semisupervised Approach. *IEEE Transactions On Computational Social Systems.* **8**, 1369-1378 (2021)

[12] Amina, B. & Azim T. Scancpeclens:, A. FrameWork for Automatic Lexicon Generation and Sentiment Analysis of MicroBlogging Data on China Pakistan Economic Corridor. *IEEE Access.* **7**, 33876-13388 (2019)

[13] Liu, H., Burnap, P., Alorainy, W. & Williams, M. Fuzzy Approach to Text Classification with Two-Stage Training for Ambiguous Instances. *IEEE Transactions On Computational Social Systems*. **6**, 227-240 (2019)

[14] Lu, R., Tsao, H., Lin, H., Ma, Y. & Chuang, C. Sentiment Analysis of Brand Personality Positioning Through Text Mining. *Journal Of Information Technology Research*. **12**, 93-103 (2019)

[15] Nazir, A., Rao, Y., Wu, L. & And, S. and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. *IEEE Transactions On Affective Computing*. **13**, 845-863 (2022)

[16] Ibrahim, N. & Wang, X. Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Computers In Human Behavior*. **96** pp. 32-45 (2019)

[17] Shi, L., Liu, L., Wu, Y., Jiang, L. & Panneerselvam, J. Human-Centric Cyber Social Computing Model for Hot-Event Detection and Propagation. *IEEE Transactions On Computational Social Systems*. **6**, 1042-1050 (2019)

[18] Wang, B., Wang, E., Zhu, Z., Sun, Y., Tao, Y. & Wang, W. An explainable sentiment prediction model based on the portraits of users sharing representative opinions in social sensors. *International Journal Of Distributed Sensor Networks*. **17**, 3323-3330 (2021)

[19] Liu, Y., Jiang, C. & Zhao, H. Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems*. **3079**, 1-11307 (2019)

[20] Khaled, A., Ouchani, S. & Chohra, C. Recommendations-based on semantic analysis of social networks in learning environments. *Computers In Human Behavior*. **101** pp. 435-449 (2019)

[21] Nimrah, S. & Saifullah, S. Context-Free Word Importance Scores for Attacking Neural Networks. *Journal Of Computational And Cognitive Engineering*. **1**, 187-192 (2022)

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Aug 1, 2023
Accepted: Oct 16, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1011–1019, DOI 10.12694/scpe.v24i4.2105

A COMPETITIVE SEGMENTATION OF THE HUMAN BRAIN USING ARTIFICIAL NEURAL NETWORK APPROACH TOWARDS MRI

S. VENI[*], C. BALAJI [†], AND T. GENISH[‡]

Abstract. The diagnosis of brain abnormalities, prognosis monitoring, and treatment evaluation all rely heavily on the magnetic resonance scan of brain tissue segmentation. Although numerous automated or semi-automatic methods have been proposed in the literature to reduce the need for human intervention, the degree of accuracy is frequently still significantly lower than that of manual segmentation. We give a clever technique for fragmenting the cerebrum utilizing a managed counterfeit brain organization system called artificial neural network (ANN) and volumetric shape models. In the beginning, in addition to the usual spatial-based and intensity-based image features, a level-set oriented brain boundary fitting technique is used to accomplish this. This is controlled by the picture intensity. The ANN is then informed of the number of important structures. Additionally, rather than directly applying standard guidelines to local appearances, this ANN learns local adaptive feature classification conditions. The outcomes demonstrate that the proposed strategy achieves competitive results in a relatively shorter time spent training.

Key words: Magnetic Resonance, supervised, Artificial Neural Network, level-set, classification.

1. Introduction . In the last twenty years, non-invasive brain imaging technologies have advanced rapidly, providing fresh avenues for studying the brain's structure and function. Magnetic resonance imaging (MRI) has made significant strides [4] in exploring brain anatomy and detecting brain injuries. Moreover, the quality of brain MR imaging has steadily improved, leading to a considerable increase in data.

Extracting vital data from vast and intricate MRI datasets is now a laborious and demanding task for clinicians. This manual analysis is time-consuming and error-prone due to the considerable variation in interoperation among studies. However, the advent of computerized diagnostic and testing techniques has resolved these challenges in interpreting brain MRI data. In various clinical applications, accurately segmenting the brain in MRI is critical since it impacts the overall investigation's outcome. This is because precise segmentation[13, 5, 2] of anatomical regions is essential for various processing steps.

MRI segmentation is a commonly used technique for measuring and visualizing various brain structures, studying brain development, detecting lesions, planning surgeries, and guiding interventions. The diverse applications of image processing have led to the development of numerous segmentation techniques with varying levels of complexity and accuracy. Automated MRI brain segmentation is frequently requested to obtain quantitative measurements of different brain regions and provide context data for lesion diagnosis and quantification. These numerical measurements play a vital role in evaluating brain atrophy, monitoring the prognosis of multiple sclerosis patients, and investigating the development of the brain across different ages. Furthermore, the structural data obtained during segmentation is a valuable visual aid for image-guided surgeries. Despite the numerous automatic or semi-automatic brain segmentation methods proposed in the literature, the current cutting-edge methods still lack sufficient performance in clinical practice.

The motivation behind research on brain tumor segmentation is to improve the accuracy and efficiency of diagnosing and treating brain tumors. Brain tumors are a severe medical condition that can have a significant impact on a patient's quality of life and survival. MRI is a commonly used imaging technique for detecting brain tumors, but manual segmentation can be time-consuming and error-prone, leading to inconsistent results and delays in treatment. Automated or semi-automated brain tumor segmentation methods have been proposed to overcome these challenges, allowing faster and more accurate diagnosis, treatment planning, and monitoring. By

[*]Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore
[†]Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore., email: cbalaji22@outlook.com
[‡]School of Computing Science, KPR College of Arts Science and Research, Coimbatore.

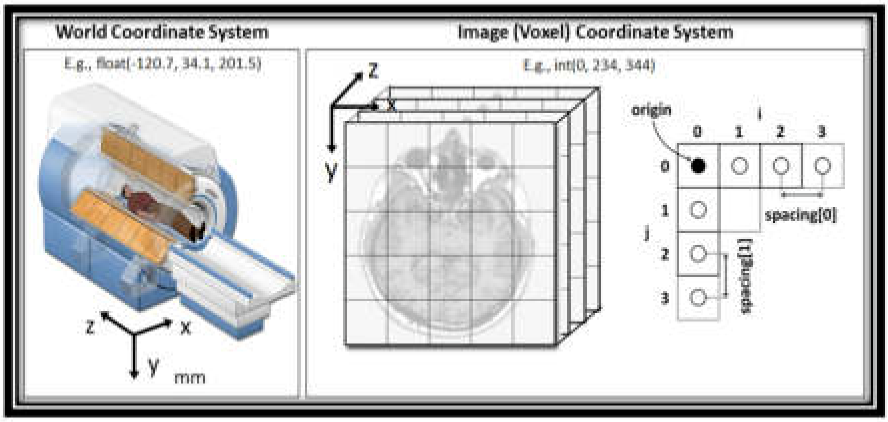


Fig. 2.1: Coordinate system of MRI scan

accurately segmenting the tumor region from normal brain tissue, clinicians can obtain vital information about the tumor's location, size, and shape, which can be used to guide surgery, radiation therapy, and chemotherapy.

The contribution of this research to segment the brain tumor is:

1. level-set oriented brain boundary fitting technique is used to extract and learn image features;
2. studied features are classified using ANN for accurate classification.

In this paper, section 2 gives a basic image concept regarding MRI and image segmentation. Section 3 reviews the works of literature with state-of-the-art methods. The proposed methodology is given in section 4. Results and their detailed discussion is compiled in section 5, and the conclusion is presented in section 6.

2. Contextual MR Image Concepts. Magnetic resonance imaging (MRI) [1] is frequently the preferred technique for structural brain investigation because of its high spatial resolution, strong contrast for soft tissues, and absence of known health hazards. In the quantitative study of the brain, MR images are common. Segmentation plays a significant role in quantitative analysis. Manual segmentation is the gold standard for in vivo pictures. To introduce the reader to the intricacy of the brain MRI segmentation problem and discuss its challenges, we start by outlining the fundamental concepts of image and its segmentation. This covers the definition of two dimensions and three-dimension images, MRI intensity distributions of the brain tissue, an issue with image segmentation, and image features. Fig 2.1 depicts the coordinate system of MRI slices. In the brain MRI, the grey numbers 0 to 255 are commonly used to represent the intensity values that make up the values of the functions Img(x,y) and Img(x,y,z). Each image comprises a fixed number of elements in voxels in three-dimensional space and pixels in two-dimensional space. Each image element has a distinct intensity value and corresponding coordinates x,y for pixels and x,y,z for voxels.

Image features are a representation of the distinctive properties of a segmented image or object. Examples of features that employ numerical measures to differentiate between the structures of interest and their surroundings are quantitative descriptors of visual appearance and shape. The success of picture segmentation depends on choosing the most pertinent characteristics and accurately extracting those features. The first and second-order statistics of an image's gray-level intensities (Fig. 2.2) serve as the foundation for the statistical features, typically extracted and classified in MRI using a statistical approach. The pattern or texture is defined by a collection of statistically extracted features represented as a vector in a multidimensional feature space.

Each grey intensity is dependent on a subset of the intensities that are nearby in the spatial interaction models (Fig. 2.3). Markov Random Field models (MRF) are the most widely used models for capturing local spatial interactions between pixel/voxel intensities [7]. Markov random field (MRF) theory provides a framework for modeling an image's local properties, in which the global image properties follow the local interactions. To reduce misclassification errors caused by image noise, MRF models have been successfully incorporated into a variety of brain MRI segmentation algorithms [9].

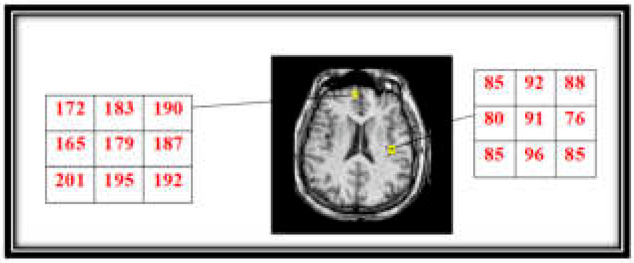


Fig. 2.2: Gray value by the intensity in Brain Image

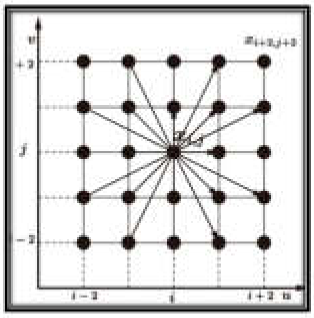


Fig. 2.3: Pixel x and Neighbouring pixels

3. Artificial Neural Network. One of the most well-known machine learning models is artificial neural networks (ANNs) (Fig. 3.1). The computational units known as interconnected neurons in a neural network typically span several layers. The neural network has an input layer where data enters, one or more hidden layers that transform the data as it moves through, and an output layer where predictions are made. An objective function is used to compare the network's outputs to the actual labels so that it can be trained to make accurate predictions from a set of labeled training data.

During training, the network's parameters of each neuron's strength are changed until the patterns the network discovers reliably predict the training data. Once the patterns have been learned, the network can generalize to new data by making predictions based on never-before-seen data. Despite their ability to model and solve complex problems, ANNs are well recognized as challenging to train and resource-intensive. This has reduced the practical value of other machine learning models, which were previously the focus of attention. However, artificial neural networks are among the most popular and well-researched machine learning techniques.

Artificial neural networks are a good example of supervised learning. An artificial neural network learned the information as an interconnected network unit. Humans find it difficult to separate this information. This element served as the inspiration for the data mining categorization rule. The basis for the categorization procedure is the dataset. The data set is divided into training and test samples. The test sample is used to evaluate the classifier's accuracy while the training sample is used to train the network. Some techniques for segmenting a data set include the hold-out method, cross-validation, and random sampling. Typical neural network learning procedures include, The network, which has a set number of nodes in the input, output, and hidden layers, uses an algorithm to learn its topology. Artificial intelligence can benefit from neural networks since they can alter the network's structure and learn by varying the weights.

4. Review of Literature. MRI segmentation is typically tricky because acquired MR images are erratic and frequently affected by noise and other defects. The extensive range of applications for image processing has led to the emergence of numerous techniques for image segmentation. This is because no single strategy,

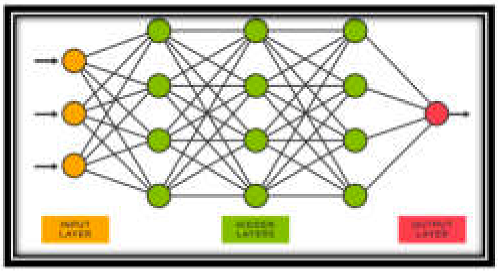


Fig. 3.1: Artificial Neural Network

nor are all approaches, equally successful for all sorts of photos. Some approaches, for example, merely use the grey-level histogram [12], while others incorporate spatial picture data to deal with noisy settings. Images can only be segmented using individual pixel or voxel intensities when the backdrop and the object of interest have distinctly different powers. By simply comparing the intensity values to the threshold, the intensity value that distinguishes the object from the background, the entire object—or the majority of its pixels or voxels—can be distinguished from the background. The image's overall intensity distribution determines the threshold. For image segmentation, edges are one of the most frequently used features. An object's edges are places on its surface where the intensities dramatically change. Typically, these changes are detected by thresholding. However, image smoothing is frequently required as a pre-processing step because this edge detection method is susceptible to image noise. Currently, available methods for segmenting MR images of the brain include the threshold method, the region method, the random field method, the clustering method, and Neural Network.

The process of manually segmenting and labeling an image by a human operator is known as manual segmentation. This segmentation is frequently carried out slice by slice for three-dimensional volumetric images. Because it is difficult to reliably and accurately delineate structures in medical imaging, the manual method is regarded as the most accurate. Artifacts and poor image quality are to blame for the difficulties with segmentation.

However, manual segmentation is required to quantitatively evaluate automated segmentation methods and establish the "ground truth," a substitute for actual delineation. Additionally, one essential component of atlas-based segmentation techniques, which is how the brain atlas was created, is the manual segmentation of various brain regions.

1. The major research gap of the literature section,
2. Firstly, many existing segmentation techniques rely on manually selected features or thresholds, which may not apply to all patients or tumor types. More robust and adaptive methods that can account for variations in image quality and tumor characteristics are needed.
3. Secondly, many existing techniques rely on 2D images, which may not capture the full extent of the tumor's 3D shape and location. Improvements in 3D segmentation techniques could improve the accuracy of tumor localization and volume measurement.
4. Thirdly, there is a need for more comprehensive evaluation metrics to compare and validate different segmentation techniques. The lack of standardized metrics can make it challenging to compare the performance of different methods and hinder their translation to clinical practice.

5. Data and Methodology. The proposed methodology encompasses pre-processing, feature extraction, level set-based segmentation, ANN classifier

5.1. Data Sets Used. Five volumes of T2-Weighted MRI data are collected from IBSR. We also used five volumes of MRI T2 data sets from SBC scans performed in Dindigul, India [10]. Additionally, five volumes of T2 - w MR images from the KGS Scan and Diagnostic Centre in Madurai, India. All these data sets have ground truth brain-segmented images done by experts.

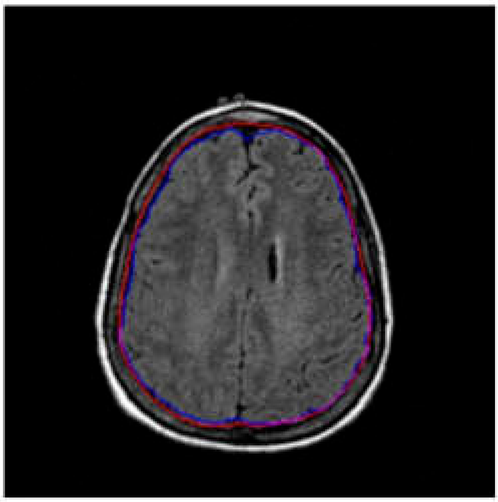


Fig. 5.1: Brain boundary detection by Level Set (Blue: Ground Truth, Red: Proposed Method

5.2. Pre-processing. Pre-processing is done to ensure that the data is correctly ready for the classifier. The following pre-processing processes were carried out in this work before submitting the data to the classifiers.

A. Correction of Bias field in MR Input Image

Bias field correction is adjusting for fluctuations in image contrast brought on by inhomogeneity in the magnetic field. The most often applied method is N4 bias field correction in medical images.

B. Non-brain (Skull) Tissue Removal

Skull stripping [8, 6, 14] involves the removal of the skull from pictures to concentrate on cerebral tissues. Removing non-brain structures like the skull, which have a big impact on the outcomes, is necessary to increase segmentation accuracy. The level set (Fig. 5.1) method was applied to this process.

5.3. Feature Extraction. In the following process, feature extraction of the normalized image is obtained. Feature extraction helps detect the brain region from MRI. In this work, we have extracted four features from the enhanced image I_E: correlation, contrast, entropy, and homogeneity.

Correlation is useful to locate the image featur by calculating spatial dependency between the pixels.

$$I_{crr} = \sum_{i,j=0}^{n-1} I_{E(i,j)} \frac{(i-\mu)(j-\mu)}{\sigma^2} \tag{5.1}$$

Contrast calculates the intensity of contrast between a pixel and its neighbor pixel for the enhanced image.

$$I_{Cnt} = \sum_{i,j=0}^{n-1} I_{E(i,j)} (i-j)^2 \tag{5.2}$$

Entropy is used to calculate randomness of the image

$$I_{Ent} = \sum_{i,j=0}^{n-1} -\ln(I_{E(i,j)}) I_{E(i,j)} \tag{5.3}$$

Homogeneity is used to calculate the homogeneous of image pixels in the given image.

$$I_{Hom} = \sum_{i,j=0}^{n-1} \frac{I_{E(i,j)}}{1+(i-j)^2} \tag{5.4}$$

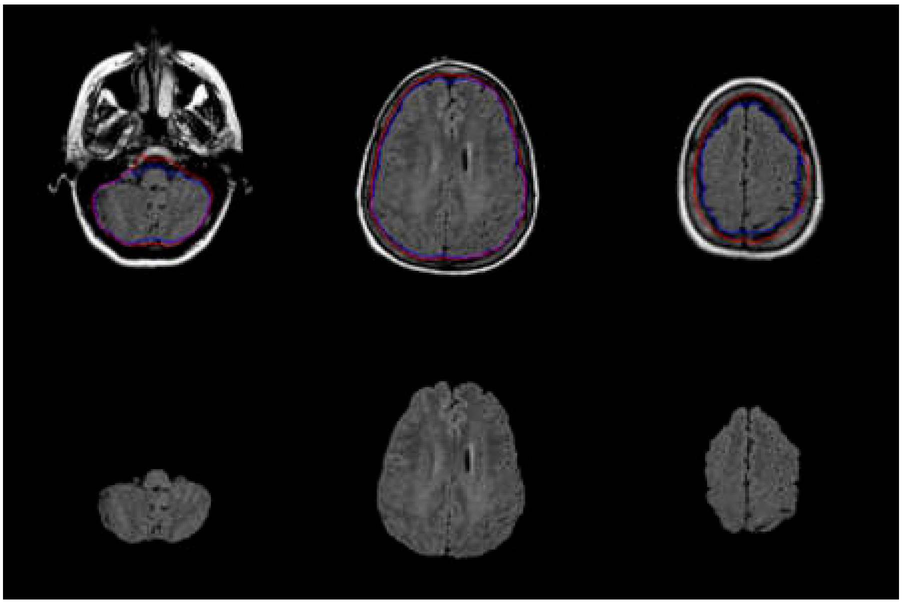


Fig. 6.1: Brain boundary detection and brain extraction

5.4. ANN Classifier. The following element of the system is neural network control. The neural network is trained to recognise brain. The solution that is suggested makes use of feed-forward neural networks. The network lacks feedback links or loops due to its directed acyclic structure. It has an input layer, an output layer, and hidden layers, as shown in Figure 3.1. Each node in the layer contains a neuron, which is referred to as the basic processing unit of a neural network. Each neuron calculates the weighted sum of its inputs and then applies an activation function to normalise the sum. The activation function functions as a decision-making body at a neuron's output.

Training:. Several factors are taken into account when building a network, including the number of layers, the neurons per layer, the connections between layers, and the training function. There are also inputs and outputs. The attributes of each pixel are provided as input, and the expert conclusions regarding each pixel are delivered as output vectors. Following the network's processing of the inputs, a comparison between the desired and produced outputs is made.

5.5. Testing. During the segmentation of the test dataset, the trained classifier, output the probabilities of each pixel belonging to each of the characteristics that were retrieved earlier. The pixel was given to the class with the highest probability. Classified values were incorporated into the results since they were used in the networks' training.

6. Results and Discussion. Skull stripping has been the subject of a significant number of published research, but none of the algorithms or techniques that come to mind specifically offer quantitative performance assessments for a variety of image sequence configurations. As different images convey diverse information about the brain, MR images are frequently used in brain research, including brain segmentation. This section highlights the results of the proposed approach on the test datasets. In this case, the network was trained using all training datasets and compared the performance of the proposed approach for stripping skulls with the manual process.

Row 1 shows brain boundary detection at slice 5, slice 21, and slice 26. Row 2 shows the segmented brain of slice 5, slice 21, and slice 26.

Table 6.1 presented the performance of the skull removal phase concerning mean and standard deviation. Our algorithm offers an average dice [3] of 0.972±0.027 based on overall testing cases.

The table 6.1 presents the performance evaluation of the skull removal phase using five different metrics,

Table 6.1: Performance evaluation of skull removal phase

	Dice	Precision	Recall	FPR	FNR
Mean	0.9721	0.9872	0.9239	0.0015	0.0306
Standard Deviation	0.0272	0.0088	0.0265	0.0018	0.0359

Table 6.2: Performance comparison of the proposed method with the existing method

No of Volumes	Dice	FPR	FNR	Precision	Recall
Proposed Method	0.9721	0.0015	0.0306	0.9872	0.9239
BET[11]	0.8213	0.0075	0.0105	0.6489	0.9782

namely Dice, Precision, Recall, False Positive Rate (FPR), and False Negative Rate (FNR). These metrics are commonly used to evaluate the accuracy and quality of segmentation results. The mean and standard deviation values for each metric are reported in the table. The mean value represents the average performance of the skull removal phase across all test cases, while the standard deviation indicates the variability of the results.

The Dice coefficient measures the similarity between the segmented image and the ground truth. It ranges from 0 to 1, where 1 indicates a perfect match between the two images. In this case, the mean Dice coefficient is 0.9721, indicating a high degree of overlap between the segmented image and the ground truth. Precision measures the proportion of true positive results among all positive results. It indicates how often the algorithm correctly identifies pixels that belong to the skull. The mean precision value is 0.9872, indicating a high degree of accuracy in identifying skull pixels.

Recall measures the proportion of true positive results among all actual positive cases. It indicates how well the algorithm detects all pixels that belong to the skull. The mean recall value is 0.9239, indicating that the algorithm has detected most of the skull pixels.False Positive Rate (FPR) measures the proportion of false positive results among all negative cases. It indicates how often the algorithm incorrectly identifies non-skull pixels as belonging to the skull. The mean FPR value is very low at 0.0015, indicating a high degree of specificity in identifying skull pixels.

False Negative Rate (FNR) measures the proportion of false negative results among all positive cases. It indicates how often the algorithm incorrectly fails to identify skull pixels. The mean FNR value is 0.0306, indicating that the algorithm has missed some skull pixels in a few cases.

Table 6.2 and Figure 6.2 prove that the sum of all images yields the best performance since the model employs the testing data throughout the testing phase.

Table 6.2 compares the performance of the proposed method with an existing method (BET) on the basis of various evaluation metrics. The table includes the number of volumes, as well as the Dice coefficient, false positive rate (FPR), false negative rate (FNR), precision, and recall for both methods. The proposed method outperforms the existing method in all metrics, with a higher Dice coefficient of 0.9721 compared to 0.8213 for BET, indicating better similarity between the predicted and ground truth masks. The FPR and FNR values are also lower for the proposed method, indicating fewer false positives and false negatives in the segmentation results. The precision and recall values are also higher for the proposed method, indicating a better balance between the true positives and false positives. Overall, the results suggest that the proposed method is superior to the existing method in skull removal phase segmentation.

Figure 6.3 shows execution time and they are exceptionally proficient. In less than three seconds, 2D input images are converted into a segmented image. Hence, the training takes place the majority of the automation time.

7. Conclusion. In this research, a fully automatic model is proposed using a level set method and an ANN-based learning algorithm. The findings demonstrate that the suggested method was reasonably accurate, had an acceptable standard deviation, and required much less time to train and test images. The existing approach needs more investigation in order to increase accuracy. Further investigation is necessary to see

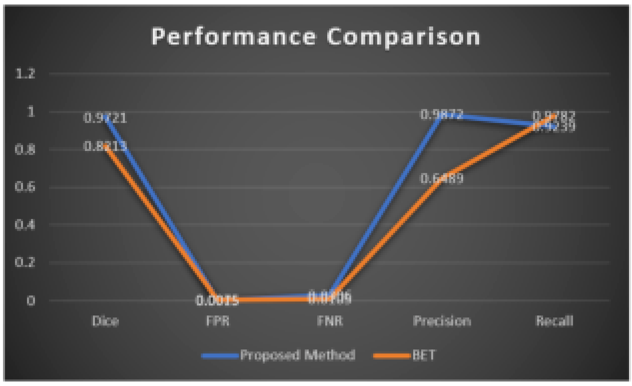


Fig. 6.2: Comparability of the proposed method with the prevailing method

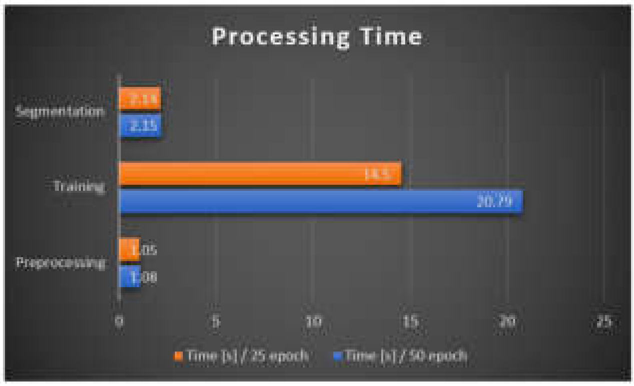


Fig. 6.3: Execution time for each phase of the proposed method

whether the proposed technique can be used for segmenting medical images in other contexts.

In this research, a novel method for skull removal phase segmentation in MRI brain images has been proposed. The proposed method combines the use of deep learning techniques with morphological operations and edge detection algorithms to achieve high accuracy in skull segmentation. The performance of the proposed method was evaluated using several metrics such as Dice coefficient, FPR, FNR, precision, and recall. The results showed that the proposed method outperformed an existing method (BET) in all the evaluation metrics, with a higher Dice coefficient of 0.9721 compared to 0.8213 for BET. The FPR and FNR values were also lower for the proposed method, indicating fewer false positives and false negatives in the segmentation results. The precision and recall values were also higher for the proposed method, indicating a better balance between the true positives and false positives.

The proposed method has demonstrated high accuracy and efficiency in skull removal phase segmentation. The results suggest that the proposed method can be a valuable tool in various clinical applications that require accurate skull removal phase segmentation, such as brain tumor detection and analysis, and neuroimaging studies. The proposed method's superior performance can help reduce the need for manual intervention, which can save time and increase efficiency in the segmentation process. Additionally, the proposed method's high accuracy can help improve the diagnosis and treatment of brain disorders, leading to better patient outcomes.

The limitation of the proposed method is that it uses a small number of training and test instances and hence it is impossible to anticipate how well the approach will perform, despite a large amount of training data.

REFERENCES

[1] M. A. Balafar, A. R. Ramli, M. I. Saripan, and S. Mashohor, *Review of brain mri image segmentation methods*, Artificial Intelligence Review, 33 (2010), pp. 261–274.
[2] I. Bankman, *Handbook of medical imaging: processing and analysis management*, academic press, 2000.
[3] L. R. Dice, *Measures of the amount of ecologic association between species*, Ecology, 26 (1945), pp. 297–302.
[4] S. Gayathri, R. Siva Shankar, and K. Somasundaram, *Fetal brain segmentation using improved maximum entropy threshold*, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN, (2020), pp. 2278–3075.
[5] T. Genish, K. Prathapchandran, and S. Gayathri, *An approach to segment the hippocampus from t 2-weighted mri of human head scans for the diagnosis of alzheimer's disease using fuzzy c-means clustering*, in Advances in Algebra and Analysis: International Conference on Advances in Mathematical Sciences, Vellore, India, December 2017-Volume I, Springer, 2018, pp. 333–342.
[6] J. Hwang, Y. Han, and H. Park, *Skull-stripping method for brain mri using a 3d level set with a speedup operator*, Journal of Magnetic Resonance Imaging, 34 (2011), pp. 445–456.
[7] S. Z. Li, *Markov random field models in computer vision*, in Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings, Volume II 3, Springer, 1994, pp. 361–370.
[8] L. Pei, A. Murat, N. Tahon, S. Zenkin, S. Karawi, A. Kamal, M. Yilmaz, L. Chen, M. Er, and R. Colen, *A generative skull stripping of multiparametric brain mris using 3d convolutional neural network*, (2022).
[9] M. Rajchl, J. S. Baxter, A. J. McLeod, J. Yuan, W. Qiu, T. M. Peters, J. A. White, and A. R. Khan, *Map-based brain tissue segmentation using manifold learning and hierarchical max-flow regularization*, (2014).
[10] S. Scans, *Dindigul*, State of Tamilnadu, India, (2020).
[11] S. M. Smith, *Fast robust automated brain extraction*, Human brain mapping, 17 (2002), pp. 143–155.
[12] Somasundaram, Shankar, and Siva, *Automated skull stripping method using clustering and histogram analysis for mri human head scans*, International Journal of Advanced Research in Computer Science & Technology, 2 (2014).
[13] K. Somasundaram, S. Gayathri, R. Rajeswaran, and D. Manjiri, *Fetal brain extraction from magnetic resonance image (mri) of human fetus*, The Imaging Science Journal, 66 (2018), pp. 133–138.
[14] K. Somasundaram and R. S. Shankar, *A novel skull stripping method for t1 coronal and t2 axial magnetic resonance images of human head scans based on resonance principle*, in Proceedings of the 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV, vol. 2012, 2012, pp. 29–35.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Feb 20, 2023
Accepted: May 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 1021–1030, DOI 10.12694/scpe.v24i4.2137

INTELLIGENT DEEP LEARNING AND SOFTMAX ROUTING FOR ENERGY-EFFICIENT WIRELESS SENSOR NETWORKS IN PUBLIC SPACE DESIGN

MENGMENG CHANG*

Abstract. The increasing usage of several nodes to transfer the massive volume of data to the remotes in wireless sensor networks is a challenging task to reduce the loss. The high volumes of data transmission in wireless sensor networks (WSN) can surpass their capacity, resulting in congestion, latency issues, and packet loss. However, computational intelligence (CI) models can aid in managing and creating intelligent networks in WSN. The WSN congestion issues result in information loss and increased energy usage. CI-based models have been used to resolve this issue, reducing the latency. This paper proposes SoftMax Routing with Deep Neural Network (SRDNN) for efficient routing in WSN. This will route the data packets by choosing the high energy and lower load. It consists of two parts, such as the construction of the routing path, which determines the residual energy of the node. It is analyzed using SoftMax routing to decide whether the node is efficient in energy. The route request and reply established various paths between the source and destination. The path with minimum buffer space and maximum bandwidth is chosen in the optimal routing. The simulation results under the metrics such as energy consumption, data loss rate, throughput, and delay show the proposed model performance.

Key words: Wireless Sensor Networks, routing, SoftMax, Deep neural network, Deep learning, energy efficiency.

1. Introduction. WSNs are a significant technological improvement enabling sensors to gather data from various environmental sensing devices[14, 2]. It has been used for intelligent data processing and decision-making [34]. In WSN, a more significant number of Sensor nodes (SN) since the data is processed in a self-organizing manner with the sink [13]. The SN transfers the gathered data to the sink, which will integrate, process and upload the data to the respective server [7]. The WSN has various advantages, including high reliability, deployment, and reduced power consumption [16]. Due to these reasons, it has been widely used in medical care, environmental monitoring, and other fields [29, 9].

However, the constrained power and the capacity to process the SNs reduce the WSN lifetime[18]. In general, WSN sensors are equipped with limited battery power that is not changeable during the deployment. Hence, energy is the primary concern in WSN, and energy-efficient approaches can prolong the network lifetime[31]. Identification of the load balancing model that efficiently utilizes the limited resources and increases the network lifetime. Specifically, the energy-efficient routing methods significantly reduce the energy of the WSN and increase its survival rate[4].

Traditional routing approaches focused on the clustering-based model for data flow control to extend the WSN lifetime[5]. The low energy adaptive clustering hierarchy (LEACH) routing[10] categorizes the SN into various clusters and uses the hierarchical cluster head to process the sensed data. In [8], forwarding the data is separated from data transmission to assist the node in avoiding premature death of the nodes. These approaches have some issues, such as difficulty in formulating it, which consumes large amounts of energy and needs to be adaptable to various network structures. Therefore, adopting new strategies to solve these issues is necessary.

In [3], The network on chip (NoC) structure was developed for efficient data transmission. It performs on-chip antenna usage among the long-distance nodes to reduce the latency. However, the EE still needed to be improved. In [20], opportunistic routing (OR) was introduced for multiple input multiple output WANs. It also fails to reduce the latency. In [30], Softmax Regressed Tanimoto Reweight Boost Classification (SRTRBC) method was developed to reduce energy utilization with reduced latency. This model identifies the underloaded EE nodes for data transmission. To solve the issues such as increased packet loss, reduced packet delivery, increased latency, and high computational cost, this paper contributes the following:

*Department of Art Design Taiyuan Normal University, Jinzhong, Shanxi, 030619 China (mengmengchang1@outlook.com)

1. This paper developed an energy-efficient congestion-aware routing scheme for WSN by identifying underloaded energy-efficient nodes for data transmission.
2. The Model comprises two parts: routing path construction using Softmax regression routing and congestion-aware routing using DNN.
3. The softmax routing identifies the residual energy (RE) of the node by examining whether it is an efficient or inefficient node through multiple paths from source to destination.
4. Based on the buffer space and the bandwidth, DNN performs the congestion-aware classification of the node with increased bandwidth and reduced buffer space for optimal routing path selection.
5. The proposed SRDNN experiments in the simulation environment and the congestion-aware scheme reduce packet loss, energy, and latency.

The remaining section of this paper is as follows: Section 2 discusses the related literature on energy-efficient approaches. Section 3 introduced the proposed network model with materials for energy-efficient routing. Section 4 addressed the simulation results and comparison of numerical outcomes. Section 5 concludes the proposed model's merits and future direction.

2. Related work. The article[12] highlights the importance of big data in surveillance applications and introduces a novel object-tracking approach using graph-based modeling and multilevel fusion. The performance evaluation results indicate the effectiveness of the proposed method [12]. The wireless multimedia sensor node (WMSN) [11] with video and audio capabilities that fuses object recognition results to improve performance and reduce data transmission size in a WMSN. The performance test results indicate that the proposed approach can significantly enhance object recognition capability while maintaining low overhead for auditory data processing. The article[26] presents a new congestion-aware routing algorithm that combines global and local routing algorithms to reduce average packet latency without increasing system power. The experimental results demonstrate better performance than DyAD, CATRA, and ERCA algorithms under various traffic conditions.

The paper [17] proposes a new method for layer-wise training of deep residual-like networks with statistical guarantees on multi-class classification tasks. The proposed method uses functional gradient boosting and shows global convergence under a standard margin assumption, eliminating a worse dependence on network depth in a generalization bound. The paper[21] proposes an enhanced routing mechanism for sensor networks that considers congestion and uses fuzzy rule sets to make decisions. Fuzzy rule sets may only sometimes provide the most accurate or optimal results, as they rely on imprecise linguistic descriptions rather than precise mathematical models. The computational complexity of fuzzy rule [1] sets can be high, making it difficult to scale to large sensor networks with many nodes. Wireless Sensor Networks (WSNs) can suffer from congestion due to many nodes and limited network resources. Technique[6] congestion and clustering control mechanism is used to mitigate congestion. Also, the article uses a hybrid multi-objective approach to improve the network performance in congestion [23]. By using bidirectional communication between[22] the ground control station and the drones, the protocol can dynamically adjust the data rate and prioritize data transmission based on network conditions. This helps to mitigate congestion and improve the overall performance and reliability of the Internet of drones.

The paper [33] employs a centralized controller to manage network congestion using multiple metrics, allowing for optimized routing decisions that minimize packet loss and energy consumption in low-power and lossy networks. A mechanism that improves congestion control in CoAP (Constrained Application Protocol) [25] observes group communication by dynamically adjusting the data transmission rate and considering the reliability of network paths. The approach in [19] utilizes fuzzy logic and sliding mode control to regulate congestion in wireless sensor networks by jointly considering network layer and physical layer parameters to optimize data transmission while minimizing energy consumption. The author in [24] uses a protocol that combines rate-based congestion control and energy-aware routing to reduce energy consumption in wireless sensor networks by adjusting data rates and routing paths based on the energy levels of individual nodes.

To optimize congestion control in wireless sensor networks by considering traffic patterns and energy consumption, protocol [32] dynamically adjusts data rates and routes based on network traffic. The energy levels of individual nodes are supposed to minimize congestion and energy consumption. The protocol in the paper [15] employs clustering techniques to group sensor nodes and uses a congestion control mechanism to regulate data transmission while considering the sinks' mobility. By optimizing the routing of data and reducing the

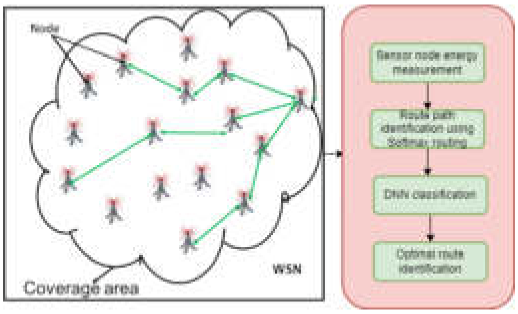


Fig. 3.1: Proposed system model

energy consumption of individual nodes, this protocol can extend the lifetime of the WSN and improve overall network performance.

The above literature review shows more congestion control and energy management techniques in WSN. However, deep learning techniques are very limited in this process. This research presents an attractive deep-learning model for conserving the energy in the WSN model.

3. Proposed Model. WSN consists of sensor nodes that sense the data and transfer it into the sink node. Fig 3.1 illustrates the proposed optimal routing path identification model that performs the routing identification that enhances data delivery and reduces the latency. It computes each distributed sensor node energy. Next, the route path between the source and destination is identified using the request and reply message. The minimum route between the sender and receiver is selected for transmission using the proposed Softmax routing with the DNN-based routing model. This congestion-awake routing scheme reduces the network latency and congestion.

The proposed system network model consists of the sensor nodes called S_i, i = 1,2,3,. . .,N distributed in the N*N area within the transmission range T. The sensor nodes are independently located and gather the data from the environment. The gathered data packets called P_i, i=1,2,3,..M are forwarded to the destination node d through the congestion awake nodes called C_i, i=1,2,3,..N to increase the WSN lifetime with the optimal routing.

3.1. Construction of Route path using Softmax routing. The SRDNN has been used to find the optimal route path classification among the sensor nodes using the softmax routing analysis. It is an ML model that finds the relationship among the dependent and independent variables. The SR observes the energy of the SN. For each SN, the energy is computed, and at the initial stage, all SN have the same energy and is reduced during the sensing procedure. The energy of SN is calculated as the multiplication of time and power as shown in Eqn 3.1. The RE of the node is computed using Eqn 3.2

$$E_S(Joule) = Power\,(watts) \times Time\;(seconds) \tag{3.1}$$

$$RE_{E_S} = Total_{E_S} - Consumed_{E_S} \tag{3.2}$$

where, RE denotes the residual energy of SN, Total declares the SN total energy and consumed energy of SN is declared as $Consumed_{E_S}$. The SR examines the predicted SN energy and finds the energy-efficient node. This activation function can categorize the SN over the predicted output as formulated in Eqn 3.3

$$SR = \frac{e^{RE_i}}{\sum_{k=1}^{p} e^{RE_k}} \tag{3.3}$$

where, RE_i is the residual energy of SN i that ranges from 0 to p. The SR output ranges from 0 to 1. While SR is analyzed with regression then the values ranges categorized as follows

$$= \begin{cases} Non\ energy - efficeint\ node & if\ 0 \le SR < 0.5 \\ Energy\ efficeint\ node & if\ 0.5 \le SR \le 1 \end{cases} \tag{3.4}$$

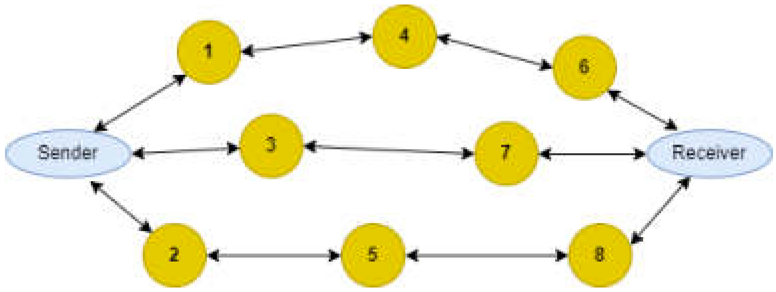


Fig. 3.2: Construction of Routing path

Once the nodes are categorized, the SN with a higher SR is chosen to construct the routing path. The sender distributes the request to the receiver through the intermediate nodes. The path between the sender and receiver is constructed using these two control messages, such as route requestA_{req} and route replyA_{rep}.

$$Sender\xrightarrow{A_{req}}\sum_{j=1}^{d}(I_j)\xrightarrow{A_{req}}Receiver \tag{3.5}$$

The sender node transfers the request for a route to the sink node via intermediate node I. Once, the request messages are collected, the receiver replies to the sender as follows

$$Sender\xleftarrow{A_{rep}}\sum_{j=1}^{d}(I_j)\xleftarrow{A_{rep}}Receiver \tag{3.6}$$

The multiple paths among the sender and receiver using SR are demonstrated in Fig 3.2. Once the routing path is constructed, the possible route from sender to receiver is [S,1,4,6,R], [S,3,7,R] and [S,2,5,8,R]. These three routes with the intermediate nodes are the possible paths from the sender to the receiver.

3.2. DNN based classification . The DNN structure is illustrated in Fig 3.3 which consists of input, output, and various hidden layers. The output layer classifies the routing path as a congestion-aware and non-congestion-aware path. DNN is a kind of feed-forward network with multiple layers that increase the network capability [28]. Let X be the input with the selected routing path and Y is the output layer that classifies the path into two categories which range the value as 0 to 1. The output computation of each hidden layer h is denoted in Eqn 3.7

$$h_i\left(X\right)=\sigma(\mathrm{W_i^T}(\mathrm{X})+b_\mathrm{i}) \tag{3.7}$$

Where σ is the activation function called ReLU used in hidden layer and Softmax function used in output layer, W is the weight and b is the bias. The activation functions are denoted din Eqn 3.8 and 3.9.

$$\sigma_{ReLU}\left(X\right)=\max(0,X)\in[0,\infty] \tag{3.8}$$

$$\sigma_{sigmoid}\left(\mathrm{X}\right)=\frac{1}{1+e^{-\mathrm{X}}}\in(0,1) \tag{3.9}$$

This paper considers the DNN with the input of three layers to represent the selected routing path. The dense layers with $2^{10}, 2^{8}$ *and* $2^{6}$with the sigmoid function represent the output as congestion aware route or not. Once it is classified, the weak learners are grouped to form the strong one as in Eqn 3.10

$$WL=\sum_{k=1}^{m}\omega_k \tag{3.10}$$

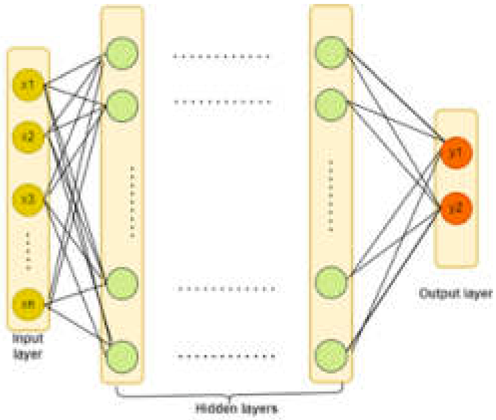


Fig. 3.3: DNN

Table 4.1: Energy Efficiency (%) comparison

No. of SN	NoC structure-based routing	Opportunistic routing (OR)	RL-AR	Proposed SRDNN
200	88.7	91.3	96.2	98.4
400	90.2	92.4	94.1	97.6
600	88.9	91.3	97.8	98.9
800	88.3	90.8	96.8	98.8
1000	84.5	93.4	94.6	97.4
Avg performance	88.12	91.84	95.9	98.22

where WL denotes the weak classifier output and ω_k declares the number of weak learners. In order to reduce the network loss constraint, the binary cross entropy loss function [27] is used as in Eqn 3.11

$$Loss{=}L\left(\omega ,b\right)+L_{const} \tag{3.11}$$

$$L\left(\omega ,b\right)=\frac{1}{n}\sum_{i=1}^{n}-[\mathrm{Y}\left(i\right)\ln(\mathrm{Y}_L\left(i\right)+(1-Y\left(i\right)\ln\left(1-Y_L\left(i\right)\right)\ ] \tag{3.12}$$

Where $Y\left(i\right)$ denotes the actual output and $\mathrm{Y}_L\left(i\right)$ denotes the DNN output.

4. Simulation results. The simulation of proposed SRDNN is experimented using MATLAB. The simulation environment consists of 1000 sensor nodes and it is experimented for 10 runs. The proposed and existing approaches such as network on chip structure [3], opportunistic routing [20] and RL based adaptive routing [30] are compared. It is evaluated using the metrics such as energy, data delivery and data loss rate, latency and throughput.

4.1. Analysis in terms of Energy Efficiency (EE). EE: It is the proportion of output and input energy which is determined as in Eqn 4.1. The performance comparison of EE of the proposed and existing methods is shown in Table 4.2.

$$EE\ (\%) = \frac{E_{output}}{E_{Input}} \times 100 \tag{4.1}$$

From table 4.1, the developed routing and existing approaches EE are compared. Among the state of the art approaches, the proposed SRDNN model outperforms with the increased EE. For instance, 600 number of SNs,

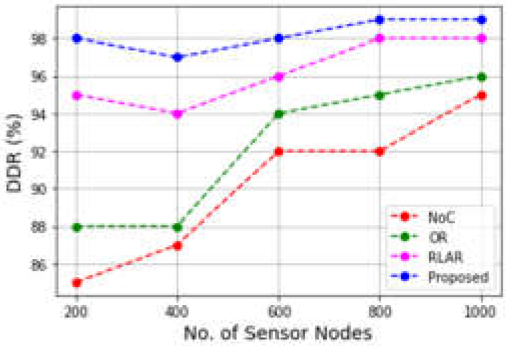


Fig. 4.1: DDR comparison

the EE of proposed model is 98.9%. Whereas, the other approaches such as NoC secured 88.9%, OR obtained 91.3% and RL-AR secured 97.8%. Due to the transmission process, the EE of SNs are reduced while increasing the number of SNs. As an average, the performance of the proposed model is superior to other approaches. Comparatively, the SRDNN is improved by 10.1% than NoC, 7.6% than OR and 2.32% than RL-AR because of the efficient routing and classification model.

4.2. Analysis in terms of Data delivery rate (DDR): . It is computed as the ration between the correctly delivered number of packets and total count of sent packets which is formulated as in Eqn 4.2

$$DDR = \frac{No.of\ correctly\ delivered\ packets}{Total\ no.of\ packets\ sent} * 100 \tag{4.2}$$

Fig 4.1 illustrates the DDR comparison among the proposed and existing approaches. The X axis denotes the No of sensor nodes transfer the data packets. The Y axis denotes the DDR of four approaches in terms of %. The illustration of this graph promotes the performance of the proposed model with increased DDR than other approaches. Due to the implementation of the softmax routing to find the residual energy, the node that is congestion aware is identified efficiently using DNN. For instance, 1000 number of SNs, the DDR of proposed model is 99% which is superior to other models such as NoC (95%), OR (96%) and RLAR (98%). The average performance of the proposed model is increased by 8% than NoC, 6% than OR and 2% than RLAR.

4.3. Analysis in terms of Latency. :
It is the variation between the expected and actual arrival time of the data packets and it is denoted in Eqn 4.3

$$L\ (ms) = Actual_{AT} - Expected_{AT} \tag{4.3}$$

Fig 4.2 illustrates the latency comparison among the proposed and existing approaches. The X axis denotes the No of sensor nodes transfer the data packets. The Y axis denotes the latency of four approaches in terms of milliseconds. The illustration of this graph promotes the performance of the proposed model with reduced latency than other approaches. Due to the implementation of the softmax routing to find the residual energy, the node that is congestion aware is identified efficiently using DNN. For instance, 600 number of SNs, the latency of proposed model is 24ms which is lesser than other models such as NoC (47ms), OR (38ms) and RLAR (35ms). The average performance of the proposed model is reduced by 21.2% than NoC, 13.2%% than OR and 8.4% than RLAR.

4.4. Analysis in terms of Data loss rate (DLR). It is the ratio between the count of data packets that are correctly delivered and total number of transmitted packets which is denoted in Eqn 4.4

$$DLR = \frac{No\ of\ data\ packet\ delivered}{No\ of\ data\ packet\ sent} * 100 \tag{4.4}$$

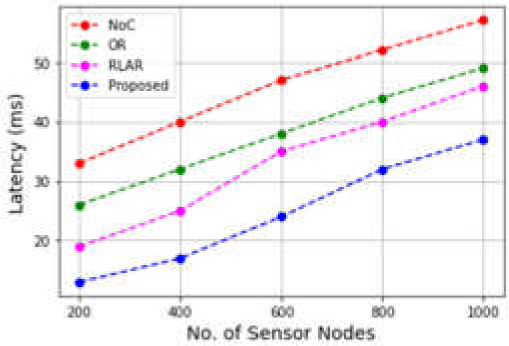


Fig. 4.2: Latency comparison

Table 4.2: DLR (%) comparison

No. of SN	NoC structure based routing	Opportunistic routing (OR)	RL-AR	Proposed SRDNN
200	15	12	10	4
400	13	12	8	3
600	8	6	5	3
800	8	5	5	2
1000	5	4	4	2
Avg performance	9.8	7.8	6.4	2.8

Table 4.2 represents the DLR comparison of proposed SRDNN model and existing approaches such as NoC, OR and RL-AR. The results show that the performance of the proposed model with reduced DLR than other approaches. Due to the implementation of the softmax routing to find the residual energy, the node that is congestion aware is identified efficiently using DNN. As an average, the proposed model secured the reduced DLR of 2.8% than other approaches such as NoC (9.8%), OR (7.8%) and RLAR (6.4%). The average performance of the proposed model is reduced by 7% than NoC, 5%% than OR and 3.6% than RLAR.

4.5. Analysis in terms of Throughput. It is the amount of data packets that are broadcasted from the SN at a particular time interval which is measured using Eqn 4.5

$$Throughput\ (bps) = \frac{No.of\ transmitted\ packets}{Time\ interval} \tag{4.5}$$

Fig 4.3 illustrates the Throughput comparison among the proposed and existing approaches. The X-axis denotes the No of sensor nodes to transfer the data packets. The Y axis denotes the throughput of four approaches in terms of bits per second (bps). The illustration of this graph promotes the performance of the proposed model with reduced throughput than other approaches. Due to the implementation of the softmax routing to find the residual energy, the node that is congestion aware is identified efficiently using DNN. For instance, 1000 SNs, the throughput of the proposed model is 890bps which is lesser than other models such as NoC (924bps), OR (954bps), and RLAR (978bps). The average performance of the proposed model is reduced by 4.8% than NoC, 7.11% than OR and 14.9% than RLAR.

The computational complexity of the proposed SRDNN model can be considered to be normal when compared with existing efficiency. Therefore, the computational complexity of the proposed SRDNN model for each run can be estimated to be in the order of O(n∧2 * r), where r is the number of iterations required for convergence during the training process. Assuming a moderate number of iterations required for convergence, the computational complexity of the proposed SRDNN model for 10 runs can be estimated to be in the order of O(n∧2 * r * 10).

The limitations of the research are as follows,

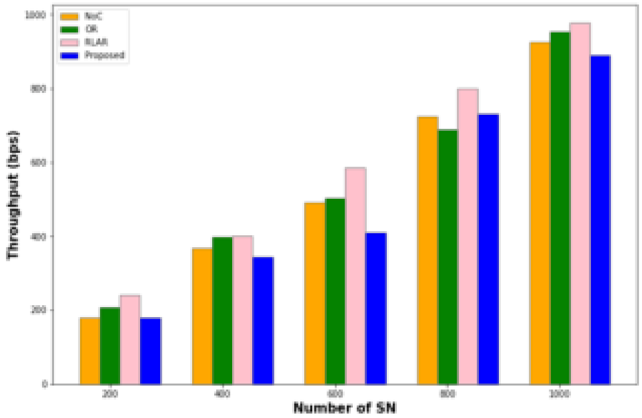


Fig. 4.3: Throughput comparison

- The study does not consider the effects of external interference or signal jamming, which can successfully test the performance of the proposed model. so it can be considered in future work.
- The proposed model relies heavily on the energy levels of the nodes in static, which may not be accurate in dynamic network conditions where node energy levels fluctuate frequently.

Applications

The proposed SoftMax Routing with Deep Neural Network (SRDNN) method for efficient routing in wireless sensor networks (WSN) has practical applications in various fields, including environmental monitoring, healthcare, and industrial control.

1. In environmental monitoring, the SRDNN method can be used to route data from sensors monitoring parameters such as temperature, humidity, and air quality. The proposed method can ensure efficient routing of data while reducing energy consumption, minimizing packet loss, and increasing throughput. This can improve the accuracy and reliability of environmental monitoring data, allowing for better decision-making regarding resource management, urban planning, and pollution control.
2. In healthcare, the SRDNN method can be applied to WSNs used for patient monitoring, such as monitoring vital signs, medication adherence, and activity levels. The proposed method can ensure timely and efficient delivery of data while minimizing data loss and reducing energy consumption. This can improve patient outcomes by enabling better decision-making by healthcare providers, leading to improved patient care and reduced hospital readmissions.
3. In industrial control, the SRDNN method can be used to route data from sensors used to monitor and control industrial processes such as manufacturing, oil and gas production, and power generation. The proposed method can ensure efficient and reliable delivery of data while minimizing energy consumption and reducing packet loss. This can lead to improved process control and optimization, resulting in increased efficiency and reduced costs.

The potential impact of the proposed method in these fields is significant, as it can lead to improved data collection, analysis, and decision-making. This can improve efficiency, reduce costs, and enhance safety in various industries. Additionally, the proposed method can enable the deployment of larger and more complex WSNs, leading to the development of new applications and use cases.

5. Conclusion. This paper introduced the efficient routing and classification scheme using SRDNN for congestion-aware data transmission over WSN. This model constructs the routing path using the SN residual energy. This residual energy identifies the EE and non-EE nodes. The multiple path from sender to receiver is constructed using the intermediate nodes. The DNN classification performs the classification of congestion awake routing with reduced latency and data loss during the data transfer through WSN. The simulation and numerical results show that the proposed SRDNN outperforms other existing approaches in terms of EE, data delivery and data loss rate, latency, and throughput. The SRDNN secured improved delivery and EE rates and reduced latency, data loss and throughput. In the future, the proposed model will be updated with Meta-heuristic algorithms to optimize the usage of energy and increase the data delivery rate for a dynamic environment. Also, network interference will be considered in future system design.

REFERENCES

[1] P. Aimtongkham, P. Horkaew, and C. So-In, *Multistage fuzzy logic congestion-aware routing using dual-stage notification and the relative barring distance in wireless sensor networks*, Wireless Networks, 27 (2021), pp. 1287–1308.
[2] I. Butun, P. Österberg, and H. Song, *Security of the internet of things: Vulnerabilities, attacks, and countermeasures*, IEEE Communications Surveys & Tutorials, 22 (2019), pp. 616–644.
[3] M. Devanathan, V. Ranganathan, and P. Sivakumar, *Congestion-aware wireless network-on-chip for high-speed communication*, Automatika, 61 (2020), pp. 92–98.
[4] U. M. Durairaj and S. Selvaraj, *Two-level clustering and routing algorithms to prolong the lifetime of wind farm-based wsn*, IEEE Sensors Journal, 21 (2020), pp. 857–867.
[5] 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.
[6] M. Farsi, M. Badawy, M. Moustafa, H. A. Ali, and Y. Abdulazeem, *A congestion-aware clustering and routing (ccr) protocol for mitigating congestion in wsn*, IEEE Access, 7 (2019), pp. 105402–105419.
[7] G. Han, H. Wang, X. Miao, L. Liu, J. Jiang, and Y. Peng, *A dynamic multipath scheme for protecting source-location privacy using multiple sinks in wsns intended for iiot*, IEEE Transactions on Industrial Informatics, 16 (2019), pp. 5527–5538.
[8] A. A.-H. Hassan, W. M. Shah, A.-H. H. Habeb, M. F. I. Othman, and M. N. Al-Mhiqani, *An improved energy-efficient clustering protocol to prolong the lifetime of the wsn-based iot*, Ieee Access, 8 (2020), pp. 200500–200517.
[9] X. Jiang, H. Zhang, E. A. B. Yi, N. Raghunathan, C. Mousoulis, S. Chaterji, D. Peroulis, A. Shakouri, and S. Bagchi, *Hybrid low-power wide-area mesh network for iot applications*, IEEE Internet of Things Journal, 8 (2020), pp. 901–915.
[10] F. A. Khan, M. Khan, M. Asif, A. Khalid, and I. U. Haq, *Hybrid and multi-hop advanced zonal-stable election protocol for wireless sensor networks*, IEEE Access, 7 (2019), pp. 25334–25346.
[11] M. Koyuncu, A. Yazici, M. Civelek, A. Cosar, and M. Sert, *Visual and auditory data fusion for energy-efficient and improved object recognition in wireless multimedia sensor networks*, IEEE Sensors Journal, 19 (2018), pp. 1839–1849.
[12] C. Küçükkeçeci and A. Yazici, *Multilevel object tracking in wireless multimedia sensor networks for surveillance applications using graph-based big data*, IEEE Access, 7 (2019), pp. 67818–67832.
[13] X. Liu, T. Qiu, X. Zhou, T. Wang, L. Yang, and V. Chang, *Latency-aware path planning for disconnected sensor networks with mobile sinks*, IEEE Transactions on Industrial Informatics, 16 (2019), pp. 350–361.
[14] D. Martín-Lammerding, A. Córdoba, J. J. Astrain, P. Medrano, and J. Villadangos, *An ontology-based system to collect wsn-uas data effectively*, IEEE Internet of Things Journal, 8 (2020), pp. 3636–3652.
[15] M. Masdari, *Energy efficient clustering and congestion control in wsns with mobile sinks*, Wireless Personal Communications, 111 (2020), pp. 611–642.
[16] A. Mehmood, Z. Lv, J. Lloret, and M. M. Umar, *Eldc: An artificial neural network based energy-efficient and robust routing scheme for pollution monitoring in wsns*, IEEE Transactions on Emerging Topics in Computing, 8 (2017), pp. 106–114.
[17] A. Nitanda and T. Suzuki, *Functional gradient boosting for learning residual-like networks with statistical guarantees*, in Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics, vol. 108, 2020, pp. 1–10.
[18] A. Pal and A. Jolfaei, *On the lifetime of asynchronous software-defined wireless sensor networks*, IEEE Internet of Things Journal, 7 (2020), pp. 6069–6077.
[19] S. Qu, L. Zhao, and Z. Xiong, *Cross-layer congestion control of wireless sensor networks based on fuzzy sliding mode control*, Neural Computing and Applications, 32 (2020), pp. 13505–13520.
[20] Y. Richter and I. Bergel, *Opportunistic routing based on partial csi in mimo random ad-hoc networks*, EURASIP Journal on Wireless Communications and Networking, 2021 (2021), pp. 1–26.
[21] G. Sangeetha, M. Vijayalakshmi, S. Ganapathy, and A. Kannan, *An improved congestion-aware routing mechanism in sensor networks using fuzzy rule sets*, Peer-to-Peer Networking and Applications, 13 (2020), pp. 890–904.
[22] B. Sharma, G. Srivastava, and J. C.-W. Lin, *A bidirectional congestion control transport protocol for the internet of drones*, Computer Communications, 153 (2020), pp. 102–116.
[23] K. Singh, K. Singh, A. Aziz, et al., *Congestion control in wireless sensor networks by hybrid multi-objective optimization algorithm*, Computer Networks, 138 (2018), pp. 90–107.
[24] V. Srivastava, S. Tripathi, K. Singh, and L. H. Son, *Energy efficient optimized rate based congestion control routing in wireless sensor network*, Journal of Ambient Intelligence and Humanized Computing, 11 (2020), pp. 1325–1338.
[25] C. Suwannapong and C. Khunboa, *Encoco-red: Enhanced congestion control mechanism for coap observe group communication*, Ad Hoc Networks, 112 (2021), p. 102377.
[26] N. Taherkhani, R. Akbar, F. Safaei, and M. Moudi, *A congestion-aware routing algorithm for mesh-based platform networks-on-chip*, Microelectronics Journal, 114 (2021), p. 105145.

[27] S.-M. Tseng and S.-C. Kao, *User selection approach in multiantenna beamforming noma video communication systems*, Symmetry, 13 (2021), p. 1737.
[28] M. Usman, Z. Ahmad, and M. Wajid, *Dataset of raw and pre-processed speech signals, mel frequency cepstral coefficients of speech and heart rate measurements*, in 2019 5th International Conference on Signal Processing, Computing and Control (ISPCC), IEEE, 2019, pp. 376–379.
[29] T. Van Nguyen, T.-N. Do, V. N. Q. Bao, D. B. da Costa, and B. An, *On the performance of multihop cognitive wireless powered d2d communications in wsns*, IEEE Transactions on Vehicular Technology, 69 (2020), pp. 2684–2699.
[30] D. Wang, J. Liu, D. Yao, and I. Member, *An energy-efficient distributed adaptive cooperative routing based on reinforcement learning in wireless multimedia sensor networks*, Computer Networks, 178 (2020), p. 107313.
[31] Y. Xiong, G. Chen, M. Lu, X. Wan, M. Wu, and J. She, *A two-phase lifetime-enhancing method for hybrid energy-harvesting wireless sensor network*, IEEE Sensors Journal, 20 (2019), pp. 1934–1946.
[32] S. L. Yadav, R. Ujjwal, S. Kumar, O. Kaiwartya, M. Kumar, and P. K. Kashyap, *Traffic and energy aware optimization for congestion control in next generation wireless sensor networks*, Journal of Sensors, 2021 (2021), pp. 1–16.
[33] Y. Yukun, L. Jiangbing, X. Dongliang, R. Zhi, and H. Qing, *Centralized congestion control routing protocol based on multi-metrics for low power and lossy networks*, The Journal of China Universities of Posts and Telecommunications, 24 (2017), pp. 35–43.
[34] L. Zhou, C. Ge, S. Hu, and C. Su, *Energy-efficient and privacy-preserving data aggregation algorithm for wireless sensor networks*, IEEE Internet of Things Journal, 7 (2019), pp. 3948–3957.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Mar 9, 2023
Accepted: Apr 13, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1031–1040, DOI 10.12694/scpe.v24i4.2139

AN EFFICIENT COOPERATIVE ROUTING WITH ML BASED ENERGY EFFICIENCY MODEL FOR DISTRIBUTED UNDERWATER WSN ELECTRICITY METER WARNING SYSTEM

ZHENGANG SHI[*], TAO PENG[†], LINHAO ZHANG[‡], BO GAO[§], AND HONGXI WANG[¶]

Abstract. Underwater wireless sensor network that operates underwater, typically in oceans, lakes, and rivers. UWSNs are composed of a large number of small sensor nodes that are equipped with various sensing and communication capabilities. These nodes are deployed in the underwater environment to collect and transmit data, which can be used for a variety of applications such as environmental monitoring, oceanography, and marine biology. The Underwater WSN (UWSN) consists of sensor nodes to sense the data and transmit it to the sink node. These sensor nodes (SN) are equipped with limited batteries, which is the central issue. Therefore, the routing protocols were developed for researchers to save energy. However, the increment of network lifetime remains an open challenge. Forwarding the data to the nearest SN to the sink will reduce the network reliability and stability, draining SN's energy early. To overcome these issues, this paper focused on developing an efficient Cooperative based routing (CR) with a machine learning (ML) model to improve the network's lifetime. The cooperative routing discovers the route path from the sender to the destination. The best possible way from the sender to the receiver has been selected using the ML approach called the Self-organizing network (SON). By identifying congestion-free multi-hop transmission using CRSON, the data packet is transmitted from sender to receiver with reduced energy, increasing the network's lifetime and reliability. This model is simulated and experimented with energy efficiency, packet delivery, loss rate, latency, and throughput metrics.

Key words: Underwater WSN, cooperative routing, machine learning, Self-organizing network, energy efficiency

1. Introduction. Underwater Wireless Sensor Networks (UWSNs) are a type of network that is used to collect data and monitor the environment in underwater settings. These networks typically consist of a large number of small sensor nodes that are deployed in the water and are capable of gathering data on various environmental factors such as temperature, pressure, and water quality [4, 18]. UWSNs are used in a variety of applications, including oceanography, environmental monitoring, and underwater exploration. Because of the unique challenges posed by the underwater environment, such as limited communication range and high levels of interference, designing and deploying effective UWSNs is a complex task. However, advances in technology have made it possible to create sophisticated and reliable UWSNs that are capable of gathering and transmitting data in even the most challenging underwater environments.

UWSNs are the most significant area that supports monitoring the environment and surveillance of the military. In UWSN, the sensor nodes were connected to sink nodes, surface stations, and other nodes in the respective area [2]. Instead of radio signals, acoustic signals are used to transmit the data from the sender to the receiver since the salt water interrupts the radio signals. A de-centralized UWSN leads to a low-cost solution that deploys the sensors rapidly for parameter measurement, which will harm the marine system. The UWSN transmission while monitoring has been interrupted by a list of issues, including limited bit error rate (BER), bandwidth, high energy consumption, and propagation delay [5]. Due to the node dying of insufficient power in UWSN, the network lifetime also gets reduced. To overcome this, efficient routing protocols are needed. The primary factor in determining the routing scheme is the selection of a relay node depending on distance, number of hops, and residual energy.

The UWSNs node with the constrained battery backup and replacement are restricted in the environment. While developing a routing scheme, the node with limited battery power must be considered. The sink node,

[*]State Grid HeBei Marketing Service Center, Shijizhuang Hebei 050000, China (zhengangresar1@outlook.com)
[†]State Grid HeBei Marketing Service Center, Shijizhuang Hebei 050000, China
[‡]State Grid HeBei Marketing Service Center, Shijizhuang Hebei 050000, China
[§]State Grid HeBei Marketing Service Center, Shijizhuang Hebei 050000, China
[¶]State Grid HeBei Marketing Service Center, Shijizhuang Hebei 050000, China

Sensor node, and other respective equipment are appropriately deployed to improve the network's lifetime[3]. However, the network's topology is also significant in reducing the energy of USWNs. The well-designed topology can consume less energy, and the node will die early if not adequately designed. Cooperative communication is the best solution for reliable transmission from sender to receiver, reducing data loss. Proper routing (CR) transmits the data through various routes to improve the receiving data's possible to reach its destination [17]. Some of the key challenges that need to be addressed in order to create effective UWSNs:

1. Radio waves do not propagate well underwater, which means that the communication range between UWSN nodes is limited. This makes it challenging to maintain reliable communication between the nodes and requires careful planning of the network topology.
2. The underwater environment is full of various sources of interference, such as noise from other marine animals or equipment. This can make it difficult for UWSN nodes to communicate with each other and can lead to data loss or corruption.
3. UWSN nodes are typically battery-powered, and it is often difficult or impossible to limit the communication range between UWSN nodes place or recharge the batteries once they have been deployed. This means that UWSN nodes must be designed to be highly energy-efficient and that the network must be carefully managed to conserve power.
4. The underwater environment is harsh, with high pressures, corrosive saltwater, and low visibility. This can make it difficult to design UWSN nodes that can withstand these conditions and operate reliably over long periods of time.
5. UWSNs typically consist of a large number of nodes, which can make it challenging to manage and scale the network. Designing an effective network topology and routing protocol that can handle large numbers of nodes is a key challenge in UWSN design.

In this research controlling energy usage and routing process is proposed to address few challenges of above points.

The CR methods are divided into the fixed relay node and the incremental relay node. The selected relay approaches improve data reliability through total collaboration. The relay node boosts the data before it is forwarded to the sink. The CR approaches can improve data delivery and reduce data loss. The energy and time consumption of this can make it difficult in the case of acoustic waves with limited energy. Machine learning (ML) approaches are recently the solution to address the green routing issues in WSNs[11]. The ML models provide flexibility and versatility to deal with complex data transmission with efficient routing approaches[16, 8]. To improve the network lifetime with reduced energy consumption and latency, this paper contributes the following:

1. An efficient ML-based routing model has been proposed to reduce the network's energy consumption with reduced latency.
2. This approach consists of two processes: constructing a route path using cooperating routing and optimizing relay node selection using SON.
3. Rather than directly transmitting the packets from sender to receiver, the boxes are transferred through the relay nodes using the CRSON model, which reduces the end-to-end data delivery delay and ensures network reliability.
4. The model's performance is simulated and compared with the existing models regarding energy efficiency, packet delivery rate, packet loss rate, latency, and throughput.

This paper is organized as follows: Section 2 discussed the related research of routing approaches in WSN and UWSN. Section 3 introduced the proposed system model and routing scheme. Section 4 simulates and compares the proposed model with other existing routing approaches. Section 5 concludes the proposed model with its future directions.

2. Related work. This section discusses the related literature on routing approaches. Ahmad et al., [1] developed cooperative energy-efficient routing (CEER) for UWSN to improve the network's reliability. The authors utilized sink mobility to reduce the power by removing the hotspot challenge. Wang et al.,[15] developed distributed adaptive routing with a reinforcement learning-based routing scheme for wireless multimedia sensor networks (WMSN). Based on the knowledge of the relay node and reliability, the quality of service and energy consumption of the network is improved. Sridhar et al., [12] Softmax Regressed Tanimoto Reweight Boost

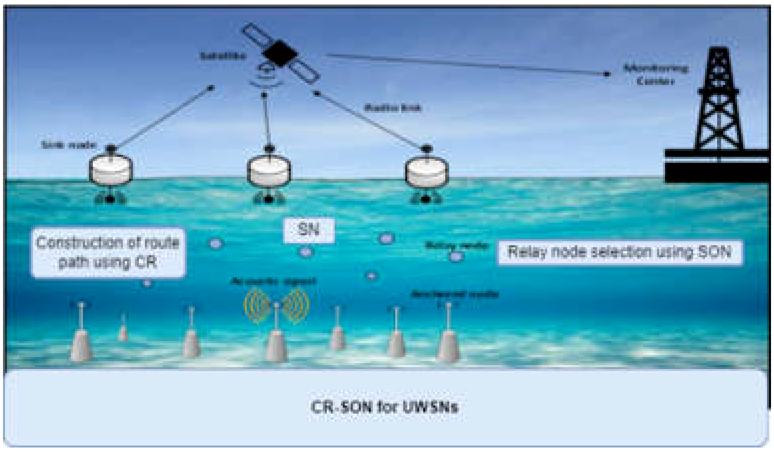


Fig. 3.1: Overview of CR-SON-based EE model

Classification (SRTRBC) method was developed to reduce energy utilization with reduced latency. This model identifies the underloaded EE nodes for data transmission.

Ullah et al., [14] developed single and multiple route path selection using minimum BER, distance to the sink node with increased residual energy. The authors obtained improved energy usage and reliable transfer of data. Due to the cooperation of nodes, the delay is raised as the disadvantage of this work. Dang et al., [13] utilized the multi-hop model to gather the data and transmission. The authors transmit the data in two ways, such as a forwarded node and relay node, using maximal radio and received signal strength models. Yum et al., [6] developed a multi-media and multi-band-based strategy for routing. Manhattan distance has been used to find the closeness of the two methods. The RSSI model has been utilized to find the distance between the sensor and the surface station. The simulation results prove that the model secured improved bandwidth and reduced delay. Qadir et al., [10] developed a based noise-aware method to decrease the latency and increase the throughput using sink mobility. The energy of the system has maximized the disadvantage. Latif et al., [7] proposed a delay-intolerant energy-efficient model using joint sink mobility with the forward mechanism. The PDR is increased with collision avoidance. The energy consumed a lot is a significant drawback. Liang et al., [9] proposed a dependence-based communication protocol for MIMO to extend the C1G2 scheme. The experimented results prove that the proposed model is improved by 40% regarding communication ratio.

Major concerns on literature discussion are as follows. It did not provide any information about the performance of the model in terms of reliability, scalability, or energy efficiency. To estimate the distance between the sensor and the surface station, which can be affected by interference and noise and may not be accurate enough for some applications. Fee papers optimized the system's energy consumption by sacrificing its performance in terms of throughput. Their approach may not be suitable for applications that require both high throughput and low energy consumption. By focusing on delay-intolerant applications and not considering applications that require low latency. Also, their mechanism consumes a lot of energy, which can limit its scalability and practicality. With existing MIMO protocols, the improvement in communication ratio they reported may not be significant compared to other existing protocols.

3. Proposed System model . The overview of the proposed network model is shown in Fig 3.1. It consists of CR-based route path discovery and SON-based relay node identification for packet transmission from sender to receiver with reduced latency and energy.

3.1. Network model. The UWSNs consist of three-dimensional is that consist of equally distributed sensors in the surveillance area. It consists of N sensor nodes (SN) as S_i, i=1,2,3,..N with the transmission range T. The data packets are denoted as P_i, i=1,2,3,..M which is forwarded to the receiver through the relay nodes called R_i, i=1,2,3,..N. The 3D UWSN is denoted as a graph of G={V,E} with N nodes. Each SN has its location information. The underwater systems have bottom-mounted nodes with its location and the anchor nodes are not deployed on the seafloor. For distribution localization, an autonomous underwater vehicle (AUV) is positioned as a reference node.

3.2. Energy model. The 3D Euclidean distance among the nodes in the UWSN is declared as a function $\gamma(a, b)$ as the distance among the node 'a' to 'b' that is stated as in Eqn 3.1

$$\gamma : N * N?\delta \ : \gamma(a, b) \tag{3.1}$$

Each node in the UWSN has the sensors to gather the data from the exterior place, which is transferred to the sink node through single or multiple hops. The sink node produces the collected data to the receiver in the communication range from R_{min} denotes the minimum transmission radius to the maximum transmit radius called R_{max}. The distance between node a and node b is bound while the distance among them with the constraint $a\left(h\right) = \gamma\left(a, b\right) = b(h)$. The two SNs have an equal minimum hop distance called h. Due to this the network density 'ρ' has an impact on the quality of the boundary. For the case of h>0 secured,

$$\lim b\left(h\right) \ - a\left(h\right) = R_{min} \tag{3.2}$$

where R_{min} is the range of SN lowest communication. The sensing model of UWSN is denoted in Eqn 3.3

$$d\left(a, b\right) = \frac{\beta}{d\left(a, b\right) k} \tag{3.3}$$

Where, $d\left(a, b\right)$ is the distance between the SN a and b and k is the parameter and β is the positive constant[4]. Assume all the SN has less battery power and that is not able to recharge after the implementation process. The network lifetime is the time when the first SN dies out of energy.

The considered simulation area is categorized into four equal sizes: upper right, upper left, bottom right, and bottom left. The sink nodes can move in the three corned paths and collect the data from the SN of each part. The SN that is randomly installed can sense the packet and transmit it to the sink node. The acoustic signal transmission can differ between shallow and deep water. The parameters such as energy, distance, and the bit error rate are considered for the selection of the receiver. The packets are directly transferred to the sink node from the neighbor node. Otherwise, the packet is sent through multiple hop. The receiver selection parameters used for packet forwarding are listed as in Eqn 3.4

$$P = \frac{Residual\ Energy}{Distance * BER} \tag{3.4}$$

The SN with residual energy and less BER is the first destination. If the BER falls below the threshold, the corresponding node has been selected as a relay node, and the sender transmits the packet to the receiver once it is formed.

3.3. Proposed CR-SON routing . This section of proposed routing scheme consists of two sub-phases: route path discovery and relay node selection.

3.3.1. Cooperative routing-based route path discovery. The discovery of route path of the candidate nodes are illustrated in Fig 3.2.

The sink node does not need that much energy since it is not move and only broadcasts its location during the startup of the transmission. The SN has the data packets with the location, relay node, source and destination, and sink node data. The S is the sender node and the remaining nodes are the receiver nodes called r1, r2, r3, r4, and r5. The packets are sent from S to sink with the minimum radius. The cosine of S and r1 is computed by r1 since r1 is located nearer to the transmitter. The forwarding packet with radius rd_i is transmitted by r2 if its cosine value is larger than 0. The radius rd_i is computed as in Eqn 3.5

$$rd_i = MIN\left\{\left(1 + \frac{\varepsilon_i^{res-ene}}{\varepsilon_i^{max}}\right).R_{min},\ R_{max}\right\} \tag{3.5}$$

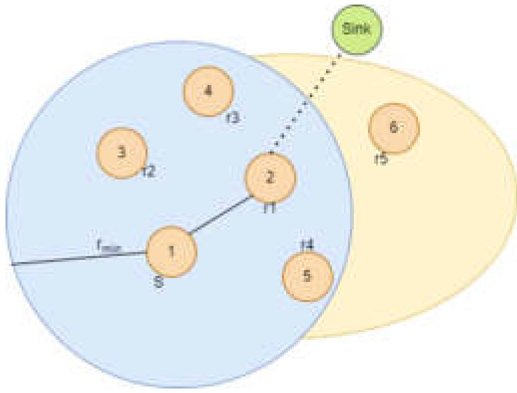


Fig. 3.2: Route discovery using CR

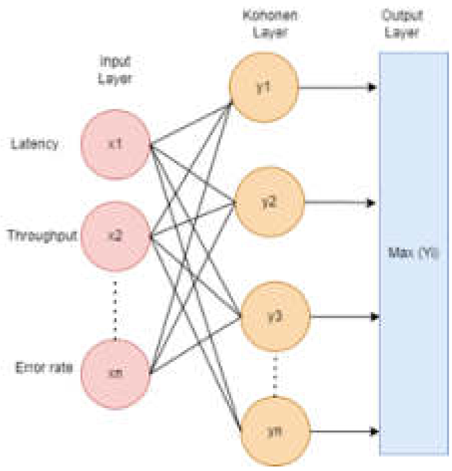


Fig. 3.3: Neural network structure of SONN

where ε_i^{max} is the SN improved energy and $\varepsilon_i^{res-ene}$ is the SN starting energy. The value of rd_i is ranged from $R_{min,}$ to $2R_{min}$. While the Forwarded packet with radius $R_{min,}$ reached S location the remaining energy of r1 is consumed. When the sink node gathers r1 packet, it extracts the residual energy and packet position. The node r3 is going to sleep mode and it does not transfer any packet for power saving since the residual energy of r3 is less than r2. The data packets are also received by r4 since it is also a receiver node. Since the radius from S to r4 and r4 to sink node is zero, it is also going to sleep mode. The node r3 is also in sleep mode since it does not receive the packets from S. Doing this transmission, the node which is in sleep mode can wake up and find its path in the communication range while the forwarding packets transfer.

3.3.2. Relay node identification using SON. SON is an unsupervised feed forward neural network proposed by Kohonen and it differs from standard neural networks with competitive learning called backpropagation. Using the geometric relationship, the SON converts the statistical data into low-dimensional space. Fig 3.3 illustrates the structure of SON which consists of an input layer to send the input data to the next layer. The second layer is the competitive layer which acts as the output node. In this layer, each neuron is connected to other neurons through inhibitory connections. To the immediate neuron, it is connected with excitatory connections. The Kohonen layer is the winner that takes all the layers. For the given input set, the output of this layer is 1 which does not need any training vector. The SON does not need an activation function or threshold, rather the output neuron is chosen as winner based on the given input pattern. That winning

Table 4.1: Parameters used for simulation

Parameters	Values
Simulation environment (W*D*H)	1000*1000*1000 m
No. of sensor nodes	500
No. of Sink nodes	20
Transmission range	550m
No. of runs	5
Packet time to live	20s
Bit rate	10kbps
Min and max speed	0 and 3 m/s

neurons are given the output. The SON enhances the understanding of data through effective visualization.

Initially, the weights of the neurons are assigned randomly. Input layer neurons are connected to all the neurons in the network. The neighbor neuron in the region is eligible to update the weight. While sending the data sample to the input layer, the Euclidean distance to all the weights is computed. The neuron weight with matching input is the best match which is adjusted toward the input. The process of SON is listed as follows:

Step 1: Initialization of network: Initialize $W_{ij}\left(t\right)$ $\left(0 = i = N-1\right)$ as the weight from the input node i to node j at time t. N is the total number of SN assigned with the weights. The radius of the neighborhood around jth node is declared as $N_j(0)$.

Step 2: Input $X_0\left(t\right), X_1\left(t\right),\ X_2\left(t\right), \ldots X_{n-1}(t)$ is initialized where $X_i\left(t\right)$ is the input from ith node at time t.

Step 3: The distance D is computed between the ith input and jth output using Eqn 3.6

$$D_j = \sum_{i=1}^{N-1} \left(X_i(t) - W_{ij}(t)\right)^2 \tag{3.6}$$

The node with minimum distance is chosen and it is the output node j called relay node for data transmission.

Step 4: The weight of j output node and its neighbor are updated using Eqn 3.7

$$W_{ij}\left(t+1\right) = W_{ij}\left(t\right) + \eta\left(t\right).X_i\left(t\right) - W_{ij}(t) \tag{3.7}$$

where, η is the learning rate in the range 0 to 1.

Step 5: Repeat the process until maximum iteration is reached.

4. Simulation and Analysis. The efficiency of the proposed CR-SON routing scheme is experimented with the simulation environment using MATLAB. The analysis is carried out using the simulation parameters listed in Table 4.1.

4.1. Evaluation Metrics. The performance of the proposed model is evaluated in terms of network lifetime, energy efficiency, latency, throughput, packet delivery rate, and packet loss rate.

1. **Network Lifetime:** It is the total time spent by the network to complete the operation.
2. **Energy efficiency (EE):** It is the proportion of output and input energy which is determined as in Eqn 4.1.

$$EE\ (\%) = \frac{E_{output}}{E_{Input}} \times 100 \tag{4.1}$$

3. **DDR**: It is computed as the ratio between the correctly delivered number of packets and the total count of sent packets which is formulated as in Eqn 4.2

$$DDR = \frac{No.of\ correctly\ delivered\ packets}{Total\ no.of\ packets\ sent} * 100 \tag{4.2}$$

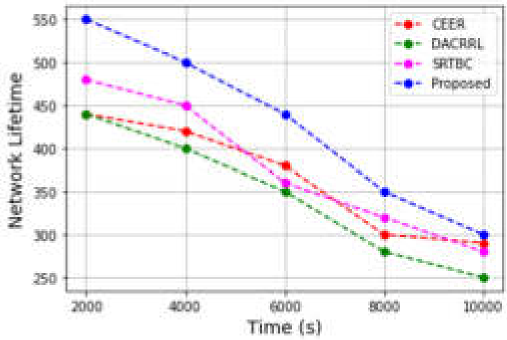


Fig. 4.1: Network lifetime comparison

4. **Latency**: It is the variation between the expected and actual arrival time of the data packets and it is denoted in Eqn 4.3

$$L\ (ms) = Actual_{AT} - Expected_{AT} \tag{4.3}$$

5. **DLR**: It is the ratio between the count of data packets that are correctly delivered and total number of transmitted packets which is denoted in Eqn 4.4

$$DLR = \frac{No\ of\ data\ packet\ delivered}{No\ of\ data\ packet\ sent} * 100 \tag{4.4}$$

6. **Throughput**: It is the amount of data packets that are broadcasted from the SN at a particular time interval which is measured using Eqn 4.5

$$Throughput\ (bps) = \frac{No.of\ transmitted\ packets}{Time\ interval} \tag{4.5}$$

4.2. Performance Analysis. The performance of the proposed model is evaluated and compared with existing routing approaches such as Cooperative energy efficient routing (CEER) [1], Distributed adaptive Cooperative routing with RL (DACR-RL)[15] and softmax regression with Tanimoto-Reweight-Boost-Classification (SRTBC) [12] routing schemes.

4.2.1. Impact on Network lifetime. Fig 4.1 illustrates the Network lifetime comparison of proposed and existing approaches in terms of time variation. It has been observed that proposed CRSON is better than other state-of-the-art approaches. The initial node of CRSON dies at 2000s which is approximately 1000s longer than other approaches. Where, the initial node of CEER, DACRRL and SRTBC die at 1000s respectively. The proposed model secured the more extended network lifetime which proves the system stability.

4.2.2. Impact on EE. The EE comparison is shown in Table 4.2. The overall EE of the proposed model is more efficient than existing approaches. As an average, the proposed model is efficient with 98.24% which is superior to other approaches such as CEER (95.54%), DACRRL (93.58%), and SRTBC (95.08%). Due to the implementation of cooperative routing and SOM, the efficiency of transmitting the packet from sender to receiver is effectively managed with reduced energy which improves energy utilization. As an average, the proposed model is 2.7% better than CEER, 4.6% better than DACRRL, and 3.16% better than SRTBC.

4.2.3. Impact on DDR. The illustration of DDR comparison between the proposed and existing approaches is shown in Fig 4.2. The X axis denotes Time in seconds and Y axis denotes the DDR in %. The observation from this fig shows that the improved performance of proposed CRSON with increased DDR than existing approaches. For the instance of 6000seconds, the DDR of proposed model is 98.3% which is better than other approaches such as CEER (96.5%), DACRRL (94.2%) and SRTBC (92.1%). The average performance of the proposed model is increased by 2.04% than CEER, 6.02% than DACRRL and 8.08% than SRTBC.

Table 4.2: Energy Efficiency (%) comparison

Time (s)	CEER	DACRRL	SRTBC	Proposed CRSON
2000	95.3	92.4	92.3	98.6
4000	94.3	93.1	95.2	98.1
6000	96.7	94.3	96.7	98.5
8000	96.9	93.6	94.5	98.2
10000	94.5	94.5	96.7	97.8
Avg performance	95.54	93.58	95.08	98.24

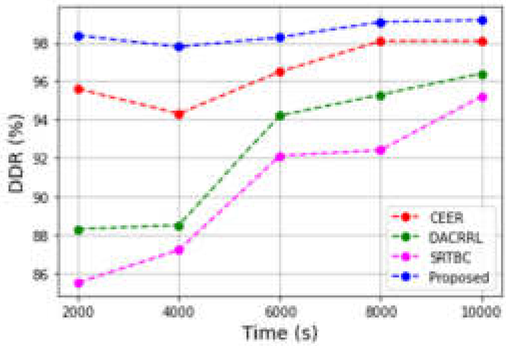


Fig. 4.2: DDR comparison

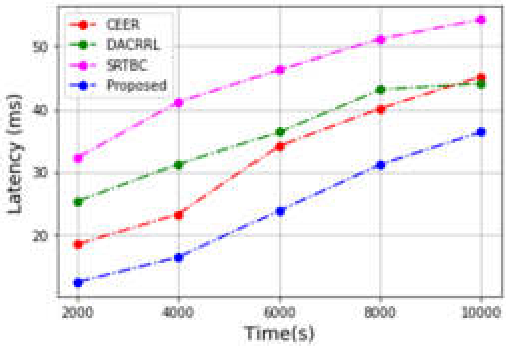


Fig. 4.3: Latency comparison

4.2.4. Impact on Latency. The illustration of latency comparison between the proposed and existing approaches is shown in Fig 4.3. The X axis denotes Time in seconds and the Y axis denotes the latency in ms. The observation from this fig shows the improved performance of the proposed CRSON with reduced latency than existing approaches. For the instance of 6000seconds, the latency of the proposed model is 23.8 ms which is reduced than other approaches such as CEER (34.2ms), DACRRL (36.4ms) and SRTBC (46.4ms). The average performance of the proposed model is reduced by 25% than CEER, 33% than DACRRL and 46% than SRTBC.

4.2.5. Impact on DLR. The illustration of DLR comparison between the proposed and existing approaches is shown in Fig 4.4. The X axis denotes Time in seconds and Y axis denotes the DLR in %. The observation from this fig shows that the improved performance of proposed CRSON with reduced DLR than existing approaches. For the instance of 6000seconds, the DLR of proposed model is 3% which is reduced than

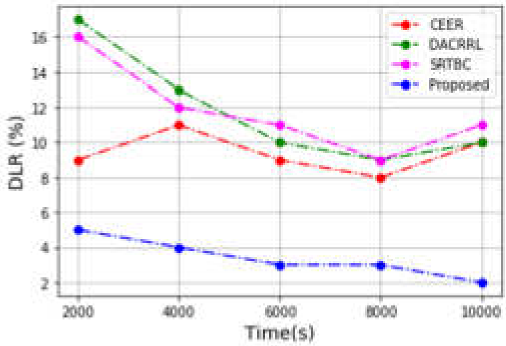


Fig. 4.4: DLR comparison

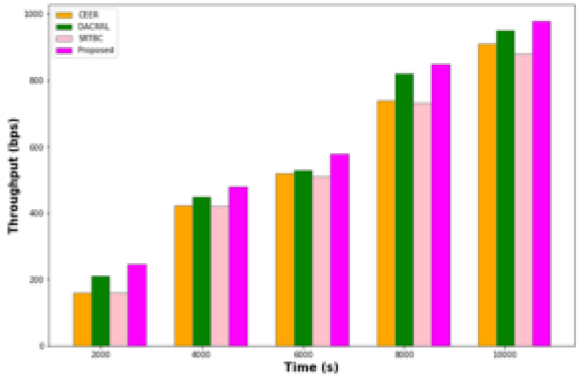


Fig. 4.5: Throughput comparison

other approaches such as CEER (9%), DACRRL (10%) and SRTBC (11%). The average performance of the proposed model is reduced by 8.4% than CEER, 8.4% than DACRRL and 8% than SRTBC.

4.2.6. Impact on throughput. The illustration of throughput comparison between the proposed and existing approaches is shown in Fig 4.5. The X axis denotes Time in seconds and Y axis denotes the throughput in bps. The observation from this fig shows that the improved performance of proposed CRSON with reduced DLR than existing approaches. For the instance of 6000seconds, the throughput of proposed model is 580bps which is improved than other approaches such as CEER (520bps), DACRRL (530bps) and SRTBC (510bps). The average performance of the proposed model is improved by 1.8% than CEER, 9.3% than DACRRL and 11.8% than SRTBC.

5. Conclusion. In this paper, presents a novel approach for routing in UWSNs that combines cooperative routing and a self-organizing network-based scheme. This proposed method utilizes a collaborative routing model to discover the optimal path for packet transmission, while also employing SON to select the most suitable relay node between the sender and receiver. It represents a significant advancement in the field of UWSN routing and has the potential to greatly improve the performance and efficiency of underwater communication networks. The simulation results and analysis in terms of energy efficiency, latency, throughput, data delivery and data loss rate, and network lifetime show an effective performance of proposed model. Compare to the existing approaches, the proposed model secured improved performance for network lifetime, energy efficiency and DDR with reduced latency and increased throughput and DLR. The efficiency of the proposed model is improved with 98.24% of energy efficiency which enhances the network stability and reliability. The proposed model was evaluated through simulation only, and its performance in a real-world deployment scenario will be checked in future research. The proposed model depends on the availability of relay nodes and the network's ability to self-organize. If there are not enough relay nodes or if the network's self-organizing capability is limited, the proposed model's performance may be affected. In future, the proposed model is enhanced with optimization-based model to enhance the congestion aware path for transmission.

REFERENCES

[1] I. Ahmad, T. Rahman, A. Zeb, I. Khan, M. T. B. Othman, and H. Hamam, *Cooperative energy-efficient routing protocol for underwater wireless sensor networks*, Sensors, 22 (2022), p. 6945.

[2] I. Ahmad, T. Rahman, A. Zeb, I. Khan, I. Ullah, H. Hamam, and O. Cheikhrouhou, *Analysis of security attacks and taxonomy in underwater wireless sensor networks*, Wireless Communications and Mobile Computing, 2021 (2021), pp. 1–15.

[3] F. Ahmed, Z. Wadud, N. Javaid, N. Alrajeh, M. S. Alabed, and U. Qasim, *Mobile sinks assisted geographic and opportunistic routing based interference avoidance for underwater wireless sensor network*, Sensors, 18 (2018), p. 1062.

[4] I. Butun, P. Österberg, and H. Song, *Security of the internet of things: Vulnerabilities, attacks, and countermeasures*, IEEE Communications Surveys & Tutorials, 22 (2019), pp. 616–644.

[5] S. Fattah, A. Gani, I. Ahmedy, M. Y. I. Idris, and I. A. Targio Hashem, *A survey on underwater wireless sensor networks: Requirements, taxonomy, recent advances, and open research challenges*, Sensors, 20 (2020), p. 5393.

[6] D. R. KM, S.-H. Yum, E. Ko, S.-Y. Shin, J.-I. Namgung, and S.-H. Park, *Multi-media and multi-band based adaptation layer techniques for underwater sensor networks*, Applied Sciences, 9 (2019), p. 3187.

[7] K. Latif, N. Javaid, I. Ullah, Z. Kaleem, Z. Abbas Malik, and L. D. Nguyen, *Dieer: delay-intolerant energy-efficient routing with sink mobility in underwater wireless sensor networks*, Sensors, 20 (2020), p. 3467.

[8] T. Liang, Y. Lin, L. Shi, J. Li, Y. Zhang, and Y. Qian, *Distributed vehicle tracking in wireless sensor network: A fully decentralized multiagent reinforcement learning approach*, IEEE Sensors Letters, 5 (2020), pp. 1–4.

[9] X. Liang, H. Zhou, W. Zhou, X. Cui, Z. Liu, and X.-Y. Li, *Onion: Dependency-aware reliable communication protocol for magnetic mimo wpt system*, in 2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS), IEEE, 2021, pp. 458–465.

[10] J. Qadir, A. Khan, M. Zareei, and C. Vargas-Rosales, *Energy balanced localization-free cooperative noise-aware routing protocols for underwater wireless sensor networks*, Energies, 12 (2019), p. 4263.

[11] S. Redhu and R. M. Hegde, *Cooperative network model for joint mobile sink scheduling and dynamic buffer management using q-learning*, IEEE Transactions on Network and Service Management, 17 (2020), pp. 1853–1864.

[12] V. Sridhar, K. Ranga Rao, V. Vinay Kumar, M. Mukred, S. S. Ullah, and H. AlSalman, *A machine learning-based intelligence approach for multiple-input/multiple-output routing in wireless sensor networks*, Mathematical Problems in Engineering, 2022 (2022), pp. 1–13.

[13] H. Tran-Dang and D.-S. Kim, *Channel-aware energy-efficient two-hop cooperative routing protocol for underwater acoustic sensor networks*, IEEE Access, 7 (2019), pp. 63181–63194.

[14] U. Ullah, A. Khan, M. Zareei, I. Ali, H. A. Khattak, and I. U. Din, *Energy-effective cooperative and reliable delivery routing protocols for underwater wireless sensor networks*, Energies, 12 (2019), p. 2630.

[15] D. Wang, J. Liu, D. Yao, and I. Member, *An energy-efficient distributed adaptive cooperative routing based on reinforcement learning in wireless multimedia sensor networks*, Computer Networks, 178 (2020), p. 107313.

[16] K. Wang, C.-M. Yu, and L.-C. Wang, *Dora: a destination-oriented routing algorithm for energy-balanced wireless sensor networks*, IEEE Internet of Things Journal, 8 (2020), pp. 2080–2081.

[17] X. Wang, D. Qin, M. Zhao, R. Guo, and T. M. Berhane, *Uwsns positioning technology based on iterative optimization and data position correction*, EURASIP Journal on Wireless Communications and Networking, 2020 (2020), pp. 1–19.

[18] L. Zhou, C. Ge, S. Hu, and C. Su, *Energy-efficient and privacy-preserving data aggregation algorithm for wireless sensor networks*, IEEE Internet of Things Journal, 7 (2019), pp. 3948–3957.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Mar 9, 2023
Accepted: Apr 13, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1041–1051, DOI 10.12694/scpe.v24i4.2140

VIRTUAL REALITY OPERA SPACE: ENERGY-EFFICIENT CLUSTERING PROTOCOL BASED ON CLUSTER CENTERS FOR WIRELESS SENSOR NETWORKS

ZHIYONG YANG*

Abstract. Energy efficiency presents a notable challenge in wireless sensor networks (WSNs) as the sensor nodes, which are responsible for collecting and transmitting data, operate under limited energy capacity. The sensor nodes must generate as much power as possible to make the network last longer. Generally, every sensor node in a WSN has batteries with limited capacities. It can be challenging to replace this little battery at times due to its vast number and numerous environmental issues. Therefore, it is believed that energy-efficient communication is essential for prolonging the lifespan of a sensor node. This work suggests a brand-new routing protocol clustering method that is energy efficient. Initially, the bamboo growth optimizer novel metaheuristic method was built on differential equations of bamboo growth and the Gaussian mixture model (BFGO). Secondly, a Cluster Centered Cluster Head Selection Algorithm (C3HA) of a routing protocol (BFGO-C3HA) is presented, with the encoding method and fitness function being revised. This algorithm is based on the BFGO technique. It can reduce transmission distance and increase energy efficiency. The results of simulations show that applying BFGO-C3HA can efficiently lessen the sensor network's energy consumption while increasing the amount of information transmitted to extend the network's lifetime.

Key words: wireless sensor networks; energy-efficient clustering mechanism; bamboo forest growth optimizer, Cluster Centered Cluster Head Selection

1. Introduction. Wireless sensor networks are made up of numerous sensor nodes placed throughout an area and collect data from the observed region. It connects the physical and digital worlds, making it a crucial component of the Internet of Things. WSNs are currently used in various applications, including monitoring the ecological environment, military operations, medical care, travel, and urban land use [23, 4, 21]. WSNs are widely used, which has drawn the interest of numerous academics in recent years. In WSNs, the sensor nodes cooperatively perceive, transmit, and gather data. There will be some energy used in this procedure.

Nevertheless, the battery-powered sensor nodes could only carry a certain quantity of power. The distribution of collected information will be somewhat hampered once the batteries run out and the power source cannot be replaced or augmented in time; even the entire sensor network may become paralyzed [8, 3]. Hence, a bottleneck issue in real-world sensor network applications is the need to be as energy efficient as appropriate in a situation of limited power to prolong the lifespan of the complete sensor network.

In real-time applications, batteries with a reduced capability are used to power the wireless sensor nodes. The wireless sensor network is installed in unsupervised outdoor or hazardous environments with a competitive field. The observation area is unregulated and powered by a charge, which makes it difficult to deliver power to or change the batteries in the sensor nodes. The most efficient method for extending a sensor network's lifespan is to reduce sensor nodes' energy consumption [1]. This will ensure that WSNs can operate continuously. To some extent, numerous specialists and academics have carried out studies to optimize the energy consumption of sensor nodes [25, 22, 17, 7, 25].

In addition, the WSN sensor node's sleep and idle listening phases resemble game mechanics. Sensor nodes will employ various tactics, resulting in multiple functionalities of sensor networks with varying advantages and results. The energy consumption of sensor nodes can be reduced by allowing them to sleep [14], although they will still use energy when changing between the active and sleep states [26]. During the state, the sensor nodes' energy use will grow if the energy is used because of the excessive transition. Either way, sensor nodes will use some power when just sitting around and observing. Long listening periods when idle sensor nodes result in significant energy waste. In all other terms, the phases of idle listening and sleep interact with one another and together define the lifespan of the wireless sensor network.

*Department of Art Design Taiyuan Normal University, Jinzhong, Shanxi, 030619 China (yangzy@tynu.edu.cn)

The main Contribution of the proposed method is given below:

1. The novelty of this work is that the proposed method works in Heterogenous Wireless Sensor Network (WSN) using a Gaussian mixture model (BFGO) based Cluster Centered Cluster Head Selection Algorithm (C3HA).
2. In a WSN, a few heterogeneous nodes can typically be utilized to increase network longevity and stability, and various node types have variable starting energy requirements and energy usage rates.
3. In the two-level heterogeneous WSN under study, energy heterogeneity is considered, and the sensor nodes are separated into advanced and ordinary nodes.
4. The proposed method minimizes energy usage and reduces the node communication transmission distance.

1.1. Motivation. The motivation behind this research is to address the significant challenge of energy efficiency in wireless sensor networks (WSNs). WSNs rely on sensor nodes to gather and transmit data, but the limited energy capacity of these nodes poses a significant obstacle to network longevity. Given that every sensor node in a WSN has batteries with limited capacities, it can be difficult to replace them due to their vast number and various environmental issues.

The researchers aim to demonstrate that applying BFGO-C3HA can efficiently reduce the sensor network's energy consumption while increasing the amount of information transmitted, thereby extending the network's lifetime. Ultimately, the motivation behind this research is to improve the energy efficiency of WSNs, which can help overcome the challenge of limited battery life and enhance the overall performance and longevity of these networks.

The rest of our research article is written as follows: Section 2 discusses the related work on Wireless Sensor Networks, the Internet of Things, and Clustering Methods. Section 3 shows the proposed work's algorithm process and general working methodology. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related works. WSNs are utilized in many industries because of the Internet of Things' fast growth. Numerous specialists and academics have undertaken multiple studies on energy efficiency because of the growing attention given to the energy issue in WSNs. Maximizing the network's lifespan is becoming a major topic in current conversation because of the constrained energy of sensor nodes. In [6], they employed two levels—corresponding to two steps, i.e., the creation and functioning of the cluster—to optimize the formation of groups and the selection of cluster heads (CHs). A new clustering optimization technique was put forth to accommodate wireless sensor networks with multilayer power variability.

Modern mathematics has a new subfield called game theory. It focuses mainly on the advantages and tactics of competitors, and it researches their optimization tactics. Currently, game theory is increasingly being applied in WSNs, primarily in data collection, voltage regulation, power efficiency, and security mechanisms [6, 2, 5, 12]. The use of game theory in recent years to improve WSN energy efficiency has produced some impressive outcomes. For instance, a power management game framework was established to balance energy usage and data packet signal quality [16]. To balance the consequences of its actions and network lifetime, this paper's first step was to define the trade-off using a multivariable optimization model.

No precise scientific hypothesis supports the algorithm's computational formula, and it's not intimately tied to the fundamentals of everything. We looked for formula derivations concerning the biological development concept of bamboo forests to create a meta-heuristic optimization with great results and a tight relationship to the reality of things. This research suggests a new meta-heuristic optimization technique called the bamboo forest growth optimizer (BFGO) and shows how the efficiency of the method's optimization ability is proved on the CEC test sets and through combinatorial optimization. The technique is built around the differential model of bamboo growth and the Gaussian mixture models [20].

The base station (sink) and the CH should be adequately installed for proper communication in clustered WSNs [9, 11, 15, 18]. Spatial variations enable efficient data flow on the network and a longer network lifespan. The current state-of-the-art problem of determining the correct placements of the sensors in the network has enabled the development of several algorithms in the literature. These studies generally concentrate on optimization methods [24, 19, 10, 13]. Based on this research, CHs were chosen randomly and may be close to one

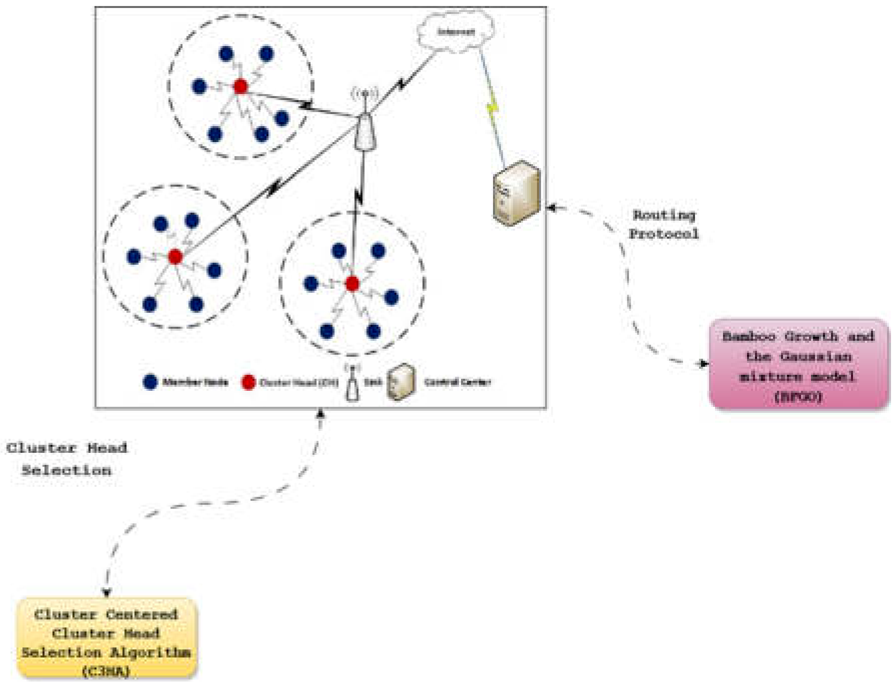


Fig. 3.1: Architecture of Proposed Method

another. This problem results in a rise in power usage. The selected CH in these clustering techniques may be far from the cluster.

The problem that was identified in previous study is the lack of standardized and universally accepted algorithms for optimizing network lifespan and energy efficiency. While many studies have been conducted on these topics, there is still a need for further research to develop more efficient and reliable algorithms. Additionally, some of the optimization techniques proposed in the literature may be highly dependent on specific network conditions, making them difficult to generalize across different WSN applications.

3. Proposed Bamboo-gaussian Cluster Head Selection Methodology. The Energy-Efficient clustering protocol minimizes the energy consumption taken during transmission sensor nodes in WSN. The use of bamboo growth is a novel approach to optimize the network performance, while the Gaussian mixture model is used to cluster the sensor nodes effectively. Additionally, the proposed method also includes a comprehensive system, network, and energy consumption model that can be used to evaluate the performance of the network accurately. Finally, the architecture of the proposed method, as shown in Figure 3.1, is also unique and can provide an efficient framework for implementing the proposed algorithm in real-world applications.

3.1. System Model. There are both advanced and standard nodes in the heterogeneous network, which has two levels [4]. This means that there are n nodes, m advanced nodes make up a certain percentage of all nodes, and each advanced node has an energy (EN) four times greater than a typical node. If EN_0 is assumed to be the starting energy for normal nodes, then EN_0 $(1 + \alpha)$ represents the starting energy for advanced nodes. Hence, the total HWSN's total node energy is:

$$EN_{total} = n * (1 - m) * EN_0 + n * m * (1 + \alpha) * EN_0 = n * EN_0 * (1 + \alpha * m) \tag{3.1}$$

To gather data, N nodes randomly and uniformly install a WSN. In the simulated environment, the following fundamental presumptions are created:

1. Nodes are fixed. The network's base station node is distinctive and strategically placed.
2. Every node has a distinct identification number.

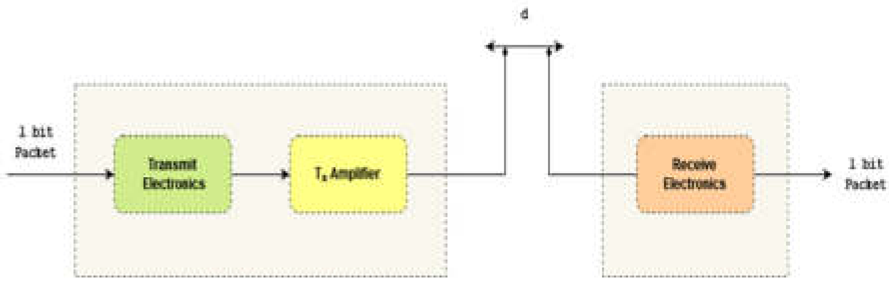


Fig. 3.2: Energy Consumption Model

3. Data fusion is carried out by the CH node, which also sends the merged information to the base station.
4. Sensor nodes' energy supply is constrained. They can no longer use the network after they pass away.
5. Nodes may compute, store information, calculate their remaining energy, and determine their location from other nodes.
6. The only node heterogeneity considered is their energy heterogeneity; additional node heterogeneity features are not considered.
7. The sink node has a reliable power source and does not go out.

3.2. Network Model. A WSN is a network assembled randomly from a specific number of sensor nodes, and each one forms a cluster with a cluster head node and several cluster members (CMs). Each CM evaluates its membership in a group considering its matching location. Also, each CM has a specific sensing range that each node can use to gather data from the monitored object and deliver it to the associated CH, which further transmits the information to the sink node.

3.3. Energy Consumption Model. The network's primary source of energy use is wireless transmission among nodes. The energy usage of nodes in this research is calculated using the energy loss formula in accordance with the first-order radio concept. Figure 3.2 displays the model diagram. The components of the sensor node that are responsible for sending and receiving information are identified and described. Specifically, there are two main components involved: a transmitter and a receiver. The transmitter is situated on the left module of the sensor node, while the receiver is located on the right module.

Furthermore, the amount of data that the sensor node can send or receive is measured in bits and denoted as "l". This means that the sensor node has a fixed capacity for transmitting or receiving data, which is determined by the number of bits it can handle at a time.

Lastly, the distance between the transmitter and receiver is referred to as the "separation" and is measured in units of length (e.g., meters, feet, etc.). This distance is critical because it determines how far the signal can travel before it weakens or becomes distorted, which can affect the accuracy of the data that is being transmitted or received.

The transmitting electronics provide the data to the amplifier, as seen in Figure 3.2, through a signal. They are wirelessly transmitted d metres to the receiving electronics. The node's total energy consumption during data transmission comes from the energy used by the amplifier circuit as well as the electrons that are transferring. The following equation depicts the routing protocol's estimation of transmitting volume based on BFGO-C3HA 3.2.

$$EN_{tx}(l,d) = \begin{cases} EN_{elec} \times l + \epsilon_{fs} \times l \times d^2, d < d_0 \\ EN_{elec} \times l + \epsilon_{mp} \times l \times d^2,\ else \end{cases} \tag{3.2}$$

Here, ϵ_{fs} and ϵ_{mp} are the power amplifier factors for the free spaces and multi-path fading models, respectively, and EN_{elec} stands for the energy used by the receiver and transmitter. Equation 3 illustrates that d_0 is the distance threshold between the transmitter and receiver.

$$d_0 = \sqrt[2]{\frac{\epsilon_{fs}}{\epsilon_{mp}}} \tag{3.3}$$

A multipath fading model is present if d > d_0; otherwise, a free space approach is present. Equation displays the received values for energy consumption, EN_{rx} 3.4.

$$EN_{rx}(l) = EN_{elec} \times l \tag{3.4}$$

In Equation 5 illustrates the data fusion's energy requirements.

$$EN_m(l) = EN_{DA} \times l \times (1+n) \tag{3.5}$$

Here EN_{DA} stands for the data fusion's unit energy usage.

3.4. Proposed Bamboo Growth and the Gaussian mixture model (BFGO) based Cluster Centered Cluster Head Selection Algorithm (C3HA). The proposed method works in the Heterogenous Wireless Sensor Network environment for energy efficient data transmission between sensor nodes. It uses Cluster Centered Cluster Head Selection Algorithm (C3HA) for CH selection. Then the bamboo growth and the Gaussian mixture model (BFGO) protocol is used for providing an optimal routing between sensor nodes.

3.4.1. Bamboo Growth and the Gaussian mixture model (BFGO). An herbaceous plant called bamboo can grow quickly to the length of a tree. Its shoot stage is when it experiences this quick growth. According to the "bamboo rule," bamboo only grows 3 cm in the first four years and then 30 cm each day from the fifth year on, growing to 15 m in under six weeks. The bamboo shoot is a brief phase of rapid development that takes place while the bamboo expands its roots several hundred metres into the soil. As a result, there are two phases to the establishment of a bamboo forest: (a) the underground extension of the bamboo whips, and (b) the development of the bamboo shoot.

While the meta-heuristic optimization method looks for possibilities, the two stages of bamboo forest growth may relate to global exploration, local exploitation, etc. As a result, by combining the mathematical problem of bamboo forest growth, a bamboo forest growth optimizer (BFGO) method can be created.

Three variables affect the trajectory of the underground bamboo whip expansion: the directives for the group cognitive items, the bamboo whip memory, and the bamboo forest center. This indicates that the worldwide optimal solution, the intra-group optimum, and the position of the central solution all influence the course of solution exploration. Equations 3.6–3.8 provide the formula for the development path.

$$cos_\alpha = \frac{\overrightarrow{X}_t . \overrightarrow{X}_G}{|\overrightarrow{X}_t| \times |\overrightarrow{X}_G|} \tag{3.6}$$

$$cos_\beta = \frac{\overrightarrow{X}_t . \overrightarrow{X}_{P(k)}}{|\overrightarrow{X}_t| \times |\overrightarrow{X}_{P(k)}|} \tag{3.7}$$

$$cos_\gamma = \frac{\overrightarrow{X}_t . \overrightarrow{C}_{(k)}}{|\overrightarrow{X}_t| \times |\overrightarrow{C}_{(k)}|} \tag{3.8}$$

Here $\overrightarrow{X}_t$ is the location of the current method, and $\overrightarrow{X}_G$ is the location of the person who is performing at the highest level globally. On the k-th bamboo whip, $\overrightarrow{X}_{P(k)}$ and $\overrightarrow{C}_{(k)}$ represent the central solution and intra-group optimal solution, respectively. The current person's expansion directions on $\overrightarrow{X}_G$, $\overrightarrow{X}_{P(k)}$, and $\overrightarrow{C}_{(k)}$ are represented by α, β and γ.

In the initialization step, the Gaussian mixture model leads people to a globally optimal outcome. The probability of the method entering local optimization grows with the number of iterations. Make it more probable that individuals will gravitate towards a centralised approach for preventing the algorithm from getting stuck in local optima. In doing so, the method's ability to locate the best solution is improved, and the distribution of options during the iteration phase is more diversified.

3.4.2. Cluster Centered–Cluster Head Selection Algorithm (C3HA). As a fresh clustering approach for WSNs, the C3HA was suggested. C3HA was created for sensor clusters. The K-means technique was used to identify the network's clusters. A new CH selection (CHS) method called the suggested C3HA was created to increase network longevity and optimise energy usage. This method requires two-fold clustering, in contrast to the methods described in the literature. Thus, a specific subset of the nodes that make up a cluster is initially identified for every cluster after the network has been grouped using k-means.

The nodes in this unique subset, known as CC (Cluster Centered), are given precedence during CH selection. OCC (Out of Cluster Centered) refers to the set of nodes outside of CC, also known as the complement of CC. Nodes inside this set are able to join CH once all the nodes inside CC have died. Next, we'll go into the specifics of how CC and OCC are determined. By giving the centre nodes more weight using the new method C3HA has proposed, CH selection will be more effective. Let's break down the suggested approach in terms of math Eq 9-12.

$$Cl_k = \frac{1}{n_i}\sum_{j=1}^{n_j} x_i \tag{3.9}$$

$$d_l^k = \sqrt{(x_r^k - x_c^k)^2 + (y_r^k - y_c^k)^2} \tag{3.10}$$

$$r_k = \frac{1}{n_i}\sum_{i=1}^{n_j} d_i \tag{3.11}$$

$$OCC_k = CC_k^{'} = (x_i, y_i) \in U_k|(x_i, y_i)| \notin CC_k \tag{3.12}$$

The CC border is a circle with radius r k that corresponds to the mean Euclidean distance from all sensors to the centre of the kth cluster, as provided in Eq 3.12. The final piece of set CC$_k$ is called OCC$_k$. Included in it are the sensors from the KTH cluster that are not listed in CC$_k$ according to Eq 3.12. The routing protocol uses the following as its implementation of BFGO-C3HA pseudo-code for Algorithm 1.

Algorithm 1 BFGO-C3HA

Input: CHs, dead normal nodes N$_a$, survival state N$_s$, maximum rounds R$_m$, current running rounds R$_c$

Output: Optimal path

Step 1: Initialize the deployment of the nodes in the WSN, the energy of the nodes in the HWSN, and the node information.

Step 2: while (R$_c$ $\leq$ R$_m$) do

Step 3: if N$_s$==false then

Step 4: end

Step 5: end if

Step 6: for j = 1; j $\leq$ N ; j + + do

Step 7: if energy $\leq$ 0 then

Step 8: N$_a$++;

Step 9: N$_n$++;

Step 10: end if

Step 11: end for

Step 12: Cluster Head Selection CH

Step 13: Initialize the nodes

Step 14: Random distributions of node to area

Step 15: Select Cluster Head using eq 9-12

Step 16: End

Step 17: Estimate the number of cluster heads (N$_{ch}$)

Step 18: for i = 1; j $\leq$ N ; j + + do
Step 19: if Node(i)==cluster head; then
Step 20: Transmitting Data
Step 21: Evaluate consumed energy, remaining energy, and transfer volume using.
Equations 2-5
Step 22: else Send data to cluster head
Step 23: end if
Step 24: end for
Step 25: end while

The algorithm is for optimizing energy consumption in a wireless sensor network while maintaining network connectivity. It involves initializing the network nodes, selecting cluster heads, and transmitting data to the base station or nearest cluster head. The algorithm iterates over the network nodes, checking their energy levels, and counting the number of dead nodes. The number of cluster heads is estimated, and data is transmitted to the base station or nearest cluster head. The output of the algorithm is the optimal path that minimizes energy consumption while maintaining network connectivity.

The algorithm starts by initializing the network nodes, their deployment, energy levels, and other relevant information. It then sets up a loop that will continue iterating until the current running rounds are less than or equal to the maximum rounds. Inside the loop, the algorithm checks the survival state of the network to see if it is still able to operate. If the network is not able to operate, the loop terminates. Otherwise, the algorithm iterates over all the nodes in the network, checking their energy levels and counting the number of dead nodes.

The next step is to select the cluster heads, which are responsible for collecting data from the sensor nodes and transmitting it to the base station. The algorithm randomly distributes the nodes across the network area and selects cluster heads based on a formula that takes into account the distance between nodes and their remaining energy levels.

Once the cluster heads are selected, the algorithm initializes the nodes and estimates the number of cluster heads in the network. The algorithm then iterates over all the nodes in the network and checks if a node is a cluster head. If a node is a cluster head, it transmits the data collected from its associated sensor nodes to the base station. The algorithm evaluates the consumed energy, remaining energy, and transfer volume using equations 2-5.

If a node is not a cluster head, it sends its data to the nearest cluster head. The algorithm continues iterating over all the nodes in the network, evaluating their energy levels and transferring data to the base station or nearest cluster head. Finally, when the current running rounds are greater than the maximum rounds, the algorithm outputs the optimal path that minimizes energy consumption while maintaining network connectivity.

3.5. Result Analysis. In this work, MATLAB was employed to evaluate the proposed method (BFGO-C3HA) and compare it with the low-energy adaptive clustering hierarchy (LEACH) [23], the distributed energy efficient clustering (DEEC) [23], and the enhanced LEACH-centralized (LEACH-C) [23]. The benefits of heterogenous WSN sensor nodes are then assessed regarding node survival state, node longevity, and data transfer. In this study, 100 sensor nodes were placed at random inside a 100-square-foot space, and the sink node was set to (50*50). Energy was no longer provided after the sensor nodes were set up. The precise parameter parameters used during the simulation are shown in Table 3.1.

The proposed method evaluates the lifetime of the network, remaining energy, volume of data transmission and throughput. The proposed method BFGO-C3HA is compared with the existing methods such as LEACH, DEEC and LEACH-C.

The lifespan of the network is measured by the number of rounds until the final node dies. Figure 3.3 displays the modifications in the remaining nodes of the LEACH, DEEC, LEACH-C, and BFGO-C3HA protocols together with the total number of rounds. Figure 3.3 shows the survival trend for the four clustering procedures (LEACH, DEEC, and LEACH-C) as well as the protocol based on BFGO-C3HA. The power used by nodes for transmission of data and processing increases as the network operates. Many nodes run out of power and die after several operation rounds. The graph shows that the network survives slightly longer and the nodes of the BFGO-C3HA method die slightly slower than the previous three protocols.

Table 3.1: Simulation Parameter

Parameters Used	**Values**
Network Area	100*100 m^2
Total Number of Nodes Used N	100
Position of BS	(50,50)
Packet Size (J)	4000 bits
Initial Energy of Node	0.5 J
Total Number of Iterations	20
Fitness function Weight	4
Transmitter Amplifier	0.0013 nJ/bit/m^4
Energy Cost	5 nJ/bit

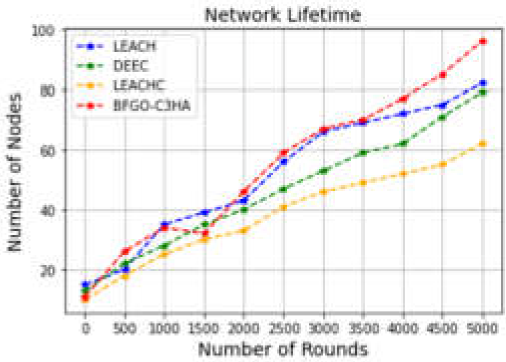


Fig. 3.3: Network Lifetime

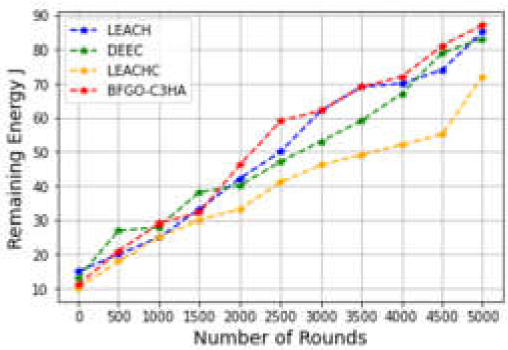


Fig. 3.4: Remaining Energy

The amount of energy used while a network is in use reflects how well it is working. Less energy usage and improved network performance are shown by more energy that remains. Figure 3.4 displays the fluctuation of remaining energy with the total number of rounds for the LEACH, DEEC, LEACH-C, and BFGO-C3HA protocol operating in the networks, while Figure 3.4's pattern in energy consumption reveals that the BFGO-C3HA treatment has more energy left over than the LEACH, DEEC, and LEACH-C protocols.

Each round, the network's surviving node transmits an incoming packet to the CH nodes, who subsequently send it to the base station, which records the amount of data packets received. The entire number of packets sent makes up the data transmission rate, which measures the network's throughput. Figure 3.5 displays the

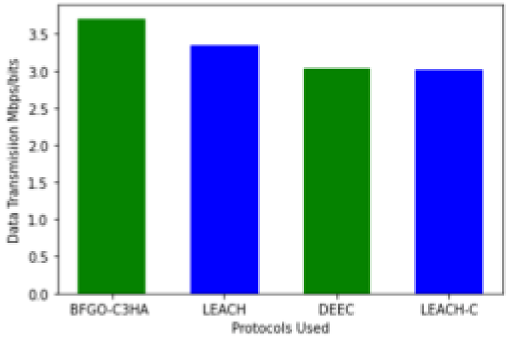


Fig. 3.5: Data Transmission between CH nodes

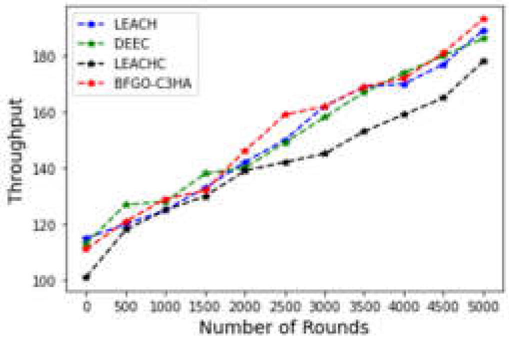


Fig. 3.6: Throughput

data transmission volume for the four protocols.

In Figure 3.6 depicts the variation in throughput that occurs when varying numbers of sensor nodes are implemented across a single network cross-section 3.10. The BFGO-C3HA algorithm used in this work greatly exceeded the other four algorithms as the number of deployed sensor nodes, or throughput, increased.

3.6. Summary. This work finds the optimal path for wireless sensor networks (WSN) by deploying nodes and selecting cluster heads (CHs) based on energy and survival state. The algorithm takes inputs of CHs, dead normal nodes Na, survival state Ns, maximum rounds Rm, and current running rounds Rc, and outputs the optimal path. The algorithm initializes the deployment of nodes, energy of nodes, and node information. It selects CHs using equations 9-12 and estimates the number of CHs (Nch). The algorithm then transmits data and evaluates consumed energy, remaining energy, and transfer volume using equations 2-5. The process is repeated until Rc reaches Rm. At each round data transmission rate, energy consumption, throughputs are measured and output are drawn. From all the outcome it is proved that proposed obtains better performance.

4. Conclusion. In this study, we present a BFGO-C3HA-based clustering technique for the routing protocol for heterogeneous WSNs. Using the BFGO-C3HA method's optimising capabilities, it aims to pick the best cluster heads, identify the best CH nodes, ensure the fairness of cluster allocation, and maximise network performance. An intelligent bionic optimization technique is initially developed for the optimization problem based on the development characteristics of a bamboo forest. This paper proposes a brand-new energy-efficient clustering technique for the routing protocol. At first, a brand-new metaheuristic technique called the bamboo growth optimizer was developed using the Gaussian mixture model and difference equations for bamboo development (BFGO). Second, a routing protocol (BFGO-C3HA) with a cluster-centered cluster head selection algorithm (C3HA) is described, with updates to the fitness function and encoding scheme. The BFGO method is the foundation of this algorithm. Both the transmission distance and energy efficiency can be decreased. The simulation results show that the BFGO-C3HA application can effectively reduce the energy consumption of the sensor network while increasing the amount of information transmitted, thereby extending the network lifetime. The proposed method efficiently overcomes the existing method and works in Heterogenous environment and enhances the data transmission between the nodes.

Funding.

1. This work was supported by the Shanxi Philosophy and Social Science Foundation Project (Research on the Inheritance and Innovation of Shanxi Opera Helmet Making Technique in the Context of Intangible Cultural Heritage, No. 2019B356);
2. The fourth batch of projects funded by young academic leaders of taiyuan normal University (Research on the Inheritance and Innovation of Shanxi Opera Helmet Making Technique in the Context of Intangible Cultural Heritage).
3. Shanxi Art Science Planning Project "Research on the Innovation and Development Strategy of Shanxi Opera Art Intangible Cultural Heritage Creative Products"

REFERENCES

[1] R. Ashween, B. Ramakrishnan, and M. Milton Joe, *Energy efficient data gathering technique based on optimal mobile sink node selection for improved network life time in wireless sensor network (wsn)*, Wireless Personal Communications, 113 (2020), pp. 2107–2126.

[2] D. Bai, W. Yao, S. Wang, and J. Wang, *Multiscale weighted permutation entropy analysis of schizophrenia magnetoencephalograms*, Entropy, 24 (2022), p. 314.

[3] S. Chamanian, S. Baghaee, H. Uluşan, Ö. Zorlu, E. Uysal-Biyikoglu, and H. Külah, *Implementation of energy-neutral operation on vibration energy harvesting wsn*, IEEE sensors journal, 19 (2019), pp. 3092–3099.

[4] Q. Feng, S.-C. Chu, J.-S. Pan, J. Wu, and T.-S. Pan, *Energy-efficient clustering mechanism of routing protocol for heterogeneous wireless sensor network based on bamboo forest growth optimizer*, Entropy, 24 (2022), p. 980.

[5] E. Harirchian and T. Lahmer, *Developing a hierarchical type-2 fuzzy logic model to improve rapid evaluation of earthquake hazard safety of existing buildings*, in Structures, vol. 28, Elsevier, 2020, pp. 1384–1399.

[6] S. Helali, A. Masmoudi, and Y. Slaoui, *Semi-parametric estimation using bernstein polynomial and a finite gaussian mixture model*, Entropy, 24 (2022), p. 315.

[7] S. Hu and X. Wang, *Game theory on power control in wireless sensor networks based on successive interference cancellation*, Wireless Personal Communications, 111 (2020), pp. 33–45.

[8] X. Huan, K. S. Kim, S. Lee, E. G. Lim, and A. Marshall, *A beaconless asymmetric energy-efficient time synchronization scheme for resource-constrained multi-hop wireless sensor networks*, IEEE Transactions on Communications, 68 (2019), pp. 1716–1730.

[9] H. H. Ibrahim, M. S. Singh, S. S. Al-Bawri, and M. T. Islam, *Synthesis, characterization and development of energy harvesting techniques incorporated with antennas: A review study*, Sensors, 20 (2020), p. 2772.

[10] M. Khademi Nori and S. Sharifian, *Edmara2: A hierarchical routing protocol for eh-wsns*, Wireless Networks, 26 (2020), pp. 4303–4317.

[11] R. La Rosa, P. Livreri, C. Trigona, L. Di Donato, and G. Sorbello, *Strategies and techniques for powering wireless sensor nodes through energy harvesting and wireless power transfer*, Sensors, 19 (2019), p. 2660.

[12] À. Nebot and F. Mugica, *Energy performance forecasting of residential buildings using fuzzy approaches*, Applied Sciences, 10 (2020), p. 720.

[13] U. N. Nisha and A. M. Basha, *Triangular fuzzy-based spectral clustering for energy-efficient routing in wireless sensor network*, The Journal of Supercomputing, 76 (2020), pp. 4302–4327.

[14] J.-S. Pan, X.-X. Sun, S.-C. Chu, A. Abraham, and B. Yan, *Digital watermarking with improved sms applied for qr code*, Engineering Applications of Artificial Intelligence, 97 (2021), p. 104049.

[15] F. Pereira, R. Correia, P. Pinho, S. I. Lopes, and N. B. Carvalho, *Challenges in resource-constrained iot devices: Energy and communication as critical success factors for future iot deployment*, Sensors, 20 (2020), p. 6420.

[16] E. Radwan, M. Nour, E. Awada, and A. Baniyounes, *Fuzzy logic control for low-voltage ride-through single-phase grid-connected pv inverter*, Energies, 12 (2019), p. 4796.

[17] P. P. Raj, A. M. Khedr, and Z. A. Aghbari, *Data gathering via mobile sink in wsns using game theory and enhanced ant colony optimization*, Wireless Networks, 26 (2020), pp. 2983–2998.

[18] R. S. Rathore, S. Sangwan, S. Prakash, K. Adhikari, R. Kharel, and Y. Cao, *Hybrid wgwo: whale grey wolf optimization-based novel energy-efficient clustering for eh-wsns*, EURASIP Journal on Wireless Communications and Networking, 2020 (2020), pp. 1–28.

[19] D. K. Sah and T. Amgoth, *A novel efficient clustering protocol for energy harvesting in wireless sensor networks*, Wireless Networks, 26 (2020), pp. 4723–4737.

[20] M. Shadoul, H. Yousef, R. Al Abri, and A. Al-Hinai, *Adaptive fuzzy approximation control of pv grid-connected inverters*, Energies, 14 (2021), p. 942.

[21] M. Tay and A. Senturk, *A new energy-aware cluster head selection algorithm for wireless sensor networks*, Wireless Personal Communications, 122 (2022), pp. 2235–2251.
[22] P. Thandapani, M. Arunachalam, and D. Sundarraj, *An energy-efficient clustering and multipath routing for mobile wireless sensor network using game theory*, International Journal of Communication Systems, 33 (2020), p. e4336.
[23] X. Yan, C. Huang, J. Gan, and X. Wu, *Game theory-based energy-efficient clustering algorithm for wireless sensor networks*, Sensors, 22 (2022), p. 478.
[24] S. Zeadally, F. K. Shaikh, A. Talpur, and Q. Z. Sheng, *Design architectures for energy harvesting in the internet of things*, Renewable and Sustainable Energy Reviews, 128 (2020), p. 109901.
[25] J. Zhang, J. Yin, T. Xu, Z. Gao, H. Qi, and H. Yin, *The optimal game model of energy consumption for nodes cooperation in wsn*, Journal of Ambient Intelligence and Humanized Computing, 11 (2020), pp. 589–599.
[26] L. Zhao, M. Gai, and Y. Jia, *Classification of multiple power quality disturbances based on pso-svm of hybrid kernel function.*, J. Inf. Hiding Multim. Signal Process., 10 (2019), pp. 138–146.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Mar 10, 2023
Accepted: May 2, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1053–1063, DOI 10.12694/scpe.v24i4.2141

EFFICIENT CLUSTERING OF BRAIN TUMOR SEGMENTS USING LEVEL-SET HYBRID MACHINE LEARNING ALGORITHMS

HE LVLIANG* AND WU HUA†

Abstract. Cluster computing is an essential technology in distributed environments for practical data analysis in complex datasets like tumor segmentation, disease classification etc. Today real-world applications like medicine and transport are needed for big data analytics environments. This research article considers complex image data environments like brain tumor segmentation based on advanced clustering techniques for effective tumor prediction. An a-state-of-art analysis used Hierarchical clustering to extract initial tumor segments from the image. The next segment is further refined using novel Noise detection-based level-setting techniques. The unsupervised Fuzzy C-means and k- means clustering is used to segment the diseases affected region to enhance noise detection used in the level set. Effective features are extracted using gray level co-occurrence matrix and redundant discrete wavelet transform. Finally, classifying malignant and benign brain tumor images is done using deep probabilistic neural networks. Publicly available datasets are used to validate the proposed algorithms. Experimental results prove that proposed pipeline techniques have effective performance in tumor segmentation and classification model.

Key words: Cluster computing, Distributed environment, brain tumor segmentation, deep learning classification, Hierarchical clustering, Fuzzy C-means, and k- means clustering.

1. Introduction. Recently cluster computing has emerged as a high performance in complex data operation. Normally, image segmentation [22] allows a vast number of data for efficient processing. This segmentation algorithm usually requires numerous pixel data to give better results. It needs more time to address more effective output. By using clustering techniques, the processing splits into multiple clusters (or) nodes and makes it possible quickly. This results in entirely satisfactory solutions with high-resolution images.[19, 2]. Clustering technique is essential in medical diagnosis like tumor segmentation, disease segmentation etc.

Brain tumor segmentation is a critical component of medical diagnosis and treatment planning. Still, it is a complex task due to the variability in tumor size, shape, and location and the similarity in intensity values between healthy and tumor tissue. To overcome these challenges, machine learning and deep learning-based methods have emerged as powerful tools that have significantly improved segmentation accuracy.

Deep learning, in particular, has demonstrated impressive results in overcoming segmentation errors and enhancing accuracy. One key advantage of deep understanding is that it can automatically learn features from the input data, allowing it to identify complex patterns that may be difficult to capture using traditional segmentation techniques. Furthermore, deep learning can perform feature extraction and segmentation in a single end-to-end process, eliminating the need for hand-crafted feature extraction. These advancements in deep learning-based techniques have greatly enhanced the accuracy and flexibility of brain tumor segmentation in clinical applications. With improved segmentation, medical professionals can make more informed diagnoses and treatment plans, monitor the tumor's progression and evaluate the treatment's effectiveness over time.

Medical imaging plays an essential role in diagnosing what the actual problem is and also use to treat diseases. Image segmentation using the clustering model plays a vital role in this process. The past day's image segmentation process mainly produced unsatisfied results, which is normal, because image segmentation includes a large amount of noised pixel data. Here we discuss Brain tumor image segmentation using a pipeline of clustering models for efficient tumor image segmentation[20]. Especially in this type of segmentation, the

*Electrical Engineering Department GuiZhou Vocational Technology College of Electronics & Information Kaili Guizhou, China, (helvliang1@outlook.com)

†Mechanical and electrical information Department computer major SongTao Miao Autonomous County Secondary Occupation School Tongren Guizhou, China

image can vary in shape and size. To meet such demand, we suggest combining clustering techniques for better results. So here we recommend using the below-mentioned techniques to reach an accurate destination.

The initial step in the segmentation process involves Noise Detection, which is crucial for removing any unwanted fluctuations in the intensity values of the input image. This pre-processing step is essential to ensure accurate results. Subsequently, we applied the Fuzzy C Means and K Means algorithms to handle noisy data[4]. The Fuzzy C Means algorithm uses a membership function to determine the similarity between data points and cluster centers, whereas the K Means algorithm uses the Euclidean distance[21]. The Fuzzy C Means approach offers a more flexible clustering technique, while K Means is faster and simpler to produce satisfactory results [25].

Brain tumor prediction and classification is regarded as a challenging research issue. This research considers A-State-of-art techniques and chooses the best strategy to process the brain tumor MRI image. The image is always composed of enormous noises. Hierarchical clustering is widely used in various image segmentation or tumor segmentation process. The algorithms for noise detection in this approach leverage their strengths to provide robust and accurate segmentation results. This noise detection and correction method with the hybrid Fuzzy C Means and K Means algorithms is highly effective for brain tumor segmentation in medical imaging. It helps in accurate diagnosis and proper treatment planning in clinical applications. We recommend using this method for optimal results.[12, 7].

The proposed methodology contribution to image segmentation is as follows:

1. Levelling the feature by removing noises appropriately using fuzzy c & K-means clustering methods.
2. Involves (GLCM) Gray Level Co-occurrence Matrix for future (feature) extraction for calculating co-occurrence of pixel intensities in the image and Involves (RDWT) Redundant Discrete Wavelet Transform for decomposing the image.
3. Deep learning for the final classification process is used with probability techniques.

The paper is further organized as follows: The related study is discussed in section 2 of the article. Section 3 has proposed algorithms and their working details. Next, 4^{th} section details the evaluated results, and finally, in section 5 whole research work is concluded.

2. Related Work. The authors[22] proposed a method for detecting brain tumors through MRI using image thresholding and K means clustering. They used an adaptive median filter to increase image clarity in the pre-processing stage. The results showed their method produced a high accuracy rate in detecting brain tumors. In paper [19], researchers discussed fuzzy c means optimization for brain tumor segmentation. They emphasized the importance of using membership to handle complex data to achieve better results. This work [2] presented an adaptive kernel fuzzy mean and clustering algorithm for brain tumor segmentation. The results showed that the proposed method effectively improved the accuracy of brain tumor segmentation. The paper [20] proposed a method for brain tumor detection using template-based K-means clustering. They found that using the TK template-based k means algorithm produced high-accuracy results for brain tumor diagnosis and treatment planning. The author[4] discussed using common fuzzy c means clustering techniques and concepts for MRI image segmentation. They used a fluctuation tracker to select images for segmentation and showed that using SVM improved the accuracy of brain tumor segmentation.

The proposed approach for liver tumor segmentation [21] utilized a 2D slice-based U-net for liver localization, a 3D patch-based network. For segmentation refinement, and unsupervised Fuzzy c means clustering for tumor estimation. In contrast, [25] proposed using Fuzzy c hyper means clustering to overcome traditional FCM disadvantages for breast tumor segmentation. The methods[24] applied Fuzzy c means and integrated K means clustering techniques for brain tumor classification using Fusion net HFCMIK segmentation. Based on these related works, we propose to use clustering techniques for brain tumor segmentation to improve the accuracy and robustness of the output. Authors in[12] proposed using the Superpixel fuzzy clustering and lattice Boltzmann method to obtain better results for Brain tumor segmentation.

According to authors [7] brain tumor classification is performed under a hybrid deep autoencoder with the Bayesian fuzzy clustering concept. For the preprocessing stage, using a mean filter for noise detection, the Bayesian method is used for brain segmentation. Article [11] motivation presents the deep learning approach to classify brain MRI images to assist medical practitioners. Also, it includes the k means clustering method for preprocessing. The paper[5, 14, 6, 23, 8] analyzes tumor localization by performing operations using k-

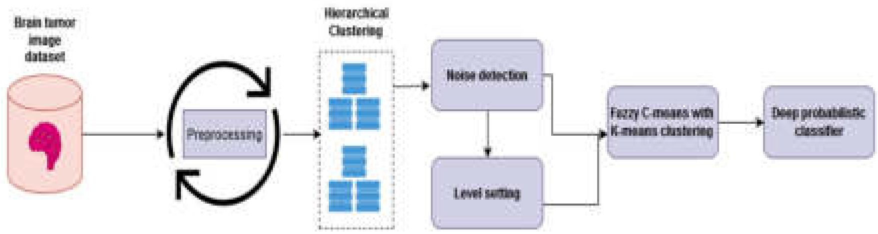


Fig. 3.1: Proposed HFCMKM Techniques for Brain tumor segmentation

means clustering techniques. The proposed method of [15] tumor segmentation is classified under Linear contrast enhancement. Deep learning feature extraction. Based on this strategy, they obtained the segmented output. Machine learning techniques for efficient brain tumor detection[17] Hybrid clustering techniques for brain tumor segmentation[9]. Categorization of brain tumors under MRI images under CNN and SVM[1]. Brain tumor segmentation under machine learning approach[10, 13].Brain tumor classification tumor detection: a long short-term memory (LSTM)-based learning model[18, 16].

3. Proposed system model. The research overview on accurate tumor segmentation and classification model is presented in this section. Figure 3.1 describes the architecture of the proposed work. Medical results based on computer-aided diagnosis have proven most accurate by physicians, and more technological perspective is needed to improve the detection and classification. The grayscale MRI image is considered in this research. The idea requires an effective preprocessing model to remove unwanted noises in the dataset. This research uses median filters to remove noise from the dataset. Then the tumor groups are recognized using hierarchical clustering. Further, our proposed noise-based level set function ideology is computed with hybrid machine learning techniques called Fuzzy C-means and k-means clustering (HFCMKM). In this process, tumor features are accurately segmented from the input image before final classification. The proposed HFCMKM model for MRI image segmentation involves a series of steps to enhance the quality of the input image and extract relevant features for accurate segmentation.

Medical imaging heavily relies on accurate image preprocessing. In the case of brain tumor image segmentation, the focus is on achieving high accuracy. The primary objective of image segmentation is to eliminate the noise present in the image, enhancing its quality and accuracy. This is achieved through three stages: changing the image mode, improving image quality, and analyzing intensity distribution. Image preprocessing includes various sub-stages:

(1) Converting the input image into grayscale.

(2) Smoothing the images to provide noise-free data. This can be achieved by extracting the brain's surface and de noising the images using Median filtering.

Median filtering works by replacing the intensity of each pixel with the median value of powers in its nearest pixel. The size of the filter kernel can be specified, and the entire image is evaluated using the filter kernel. The resulting image is then used to enhance the quality of the image and identify the noise-affected regions, improving the accuracy of the other segmentation process.

This research model employs a noise level set-based hybrid approach with Fuzzy C-means and K-means clustering algorithms to improve the accuracy of segmentation results. This approach overcomes the limitations of traditional fuzzy clustering and produces satisfactory output images. Finally, a post-processing step is performed to eliminate small and isolated regions in the segmented image, resulting in a more accurate and precise image for analysis.

3.1. Hierarchical clustering . The preprocessed image is further processed using hierarchical clustering to segment the image. This clustering method groups similar data points into clusters based on a similarity metric and is commonly used in brain tumor segmentation. Hierarchical clustering is a machine-learning technique that can be used for brain tumor segmentation. Brain tumor segmentation is identifying and delineating the boundaries of tumor tissue within medical images, such as Magnetic Resonance Imaging (MRI) scans. Hierarchical clustering is an unsupervised learning algorithm that groups similar data points in a hierarchical tree-like structure.

In brain tumor segmentation, hierarchical clustering can group pixels or voxels within the MRI scan with similar intensity values. The algorithm starts by considering each pixel or voxel as a separate cluster and then iteratively merges them based on their similarity. The similarity between two groups is determined using a distance metric, such as Euclidean or correlation distance. The algorithm continues to merge the clusters until a stopping criterion is met, such as a desired number of clusters or a threshold value for the distance metric.

Once the hierarchical tree-like structure is created, a threshold can be applied to the distance metric to cut the tree at a particular level, creating a set number of clusters. These clusters can then be labeled as tumor or non-tumor tissue based on their location within the MRI scan.

The process of hierarchical clustering can be outlined as follows:

Step 1: Initially, each item is assigned to its cluster, resulting in N clusters. The distance between each group is the same as the distance between the items.

Step 2: Similar clusters are identified and merged into one cluster to reduce the overall number of clusters.

Step 3: The distance between each new cluster and every other cluster is calculated.

Step 4: Steps 2 and 3 are repeated until all clusters have been merged into a single cluster, resulting in a hierarchical tree structure.

3.2. Noise Removal using Level Set Function. The 99th percentile minimum and maximum [25] technique reduces noise in Magnetic Resonance Imaging (MRI) images for detecting brain tissue and tumor presence. This technique involves identifying intensity values that correspond to the 99th percentile of the intensity distribution in the MRI image. A piecewise linear transformation is applied to segment the image into different regions, and a linear transformation is separately applied to each region. This process helps enhance or modify the image's appearance by separating brain tissues based on contrast, adjustment, and color. The equation 3.1 state the condition,

$$y = \left\{ \begin{array}{c} \left\{ \begin{array}{c} 0 \\ \frac{x-T_{low}}{T_{high}-T_{low}} \\ 1 \end{array} \right. \quad , \quad x < T_{low} \\ T_{low} < x < T_{high} \\ X > T_{high} \end{array} \right\} \tag{3.1}$$

Let I be a cropped MRI image defined on a domain Ω. Tumor formed with zero level set
$C = \{(x \in \Omega \mid (x) = 0)\}$
Points inside the contour is $\emptyset\,(x) < 0$ and outside $\emptyset\,(x) > 0$
Level set Function $\emptyset : \Omega \to R$ [22] energy function is $\epsilon(\emptyset)$ is defined in equation 3.2

$$\epsilon\,(\emptyset) = \mu \mathcal{R}_p + \lambda \mathbf{L}_g\,(\emptyset) + \alpha A_g(\emptyset) \tag{3.2}$$

where $R_p\,(\emptyset) \triangleq \int \Omega \frac{1}{2}(\nabla\emptyset - 1)^2 dx$ The regularization term, which includes $|\nabla\emptyset| \simeq 1$ the length term $\mathscr{L}(\emptyset) = \int \Omega g^{\delta}(\emptyset)|\nabla\emptyset|dx$ and involves the function g integrated along the zero level contour, is represented dx. The area term is then computed as $\mathcal{A}\,(\emptyset) = \int \Omega g^{H}\,(-\emptyset)\,dx$

The method is tested on MRI images to segment tumors in a slice-by-slice manner. The segmentation is refined using the contour representation of the tumor formed with the zero level set.

3.3. Applying Hybrid Fuzzy C means and K means clustering:. The Fuzzy C Means (FCM) is a unsupervised technique that is utilized to distinguish between the background tumor-affected region and healthy brain tissue. Hybrid Fuzzy C-K means (HFCKM) provides several advantages over traditional Fuzzy C Means (FCM) by ensuring smooth and continuous segmentation and avoiding the creation of isolated regions in the image. HFCKM involves pre-processing, initialization, cluster assignment, spatial regularization, cluster update, iteration, and segmentation, which provides more accurate results than traditional FCM. This increased accuracy allows for improved diagnosis and more effective treatment planning for patients.

Algorithm 1: Pseudocode for Fuzzy C-Means and K-Means

Step 1:	The image is pre-processed to remove noise and artifacts.
Step 2:	The number of clusters is determined, and the cluster centers are initialized randomly.
Step 3:	Each pixel is assigned a membership value to each cluster, indicating the degree to which the pixel belongs to the cluster.
Step 4:	The membership values are calculated using the Euclidean distance metric, which controls the degree of membership of each pixel to each cluster.
Step 5:	Spatial regularization is included to ensure that nearby pixels have similar membership values. This can be validated using a Gaussian filter.
Step 6:	The cluster center is updated based on the new membership values.
Step 7:	Steps 3 to 6 are repeated until all the clusters are covered.
Step 8:	Finally, the membership values are thresholded to create the segmented results.
Step 9:	Additionally, the K means the concept is applied to HFCM.
Step 10:	Each pixel in the image is assigned to the nearest cluster center based on the Euclidean distance between data points and the cluster.
Step 11:	The cluster center is updated based on the values of each pixel assigned to each cluster.
Step 12:	Steps 2 and 3 are repeated until all the clusters have been covered.
Step 13:	Finally, the tumor-affected area is represented based on the identified clusters.

To perform HFCMKM, the following steps are presented in Algorithm 1.

The proposed HFCMKM technique is applied in two steps. In the first step, the hybrid FCM algorithm assigns a membership value to each pixel for each cluster based on the degree to which the pixel belongs to the cluster. Spatial regularization is included to ensure that nearest pixels have similar membership values. The Euclidean distance metric is used to calculate membership values, and a Gaussian filter can validate the results.

Next, the K means technique is applied to represent brain tissues as K clusters. Each pixel is assigned to the nearest cluster center based on the Euclidean distance between data points and the cluster. Finally, the HFCMKM technique combines the advantages of both FCM and K means clustering to obtain robust and flexible results.

Overall, the proposed HFCMKM technique (Algorithm 2) provides an efficient and effective way to handle complex data in brain image segmentation. It overcomes the limitations of traditional FCM and K means techniques and can lead to accurate diagnosis and treatment planning for clinical applications.

Finally, the tumor values are clustered using following equations

$$\bigcup_{j=1}^{n} = \sum_{i=1}^{n} W_i \; X_i \tag{3.5}$$

W_i denotes the weightage of X_i

For n data points, the number of U values are estimated in equation 3.6

$$\bigcup_{j=1}^{m} \begin{pmatrix} U_1 \\ U_2 \\ U_n \end{pmatrix} \tag{3.6}$$

The data points U_i are sorted by distance and weighted. Then, they are divided into k equal sets, where k represents the number of clusters. The mean value of each set is taken as the initial centroid. The proposed approach, named HFCMKM, utilizes the distance metric to determine the value of the members in each group. The suggested weighted algorithm selects centroids based on the following formula, which improves the selection of initial centers. HFCMKM divides the image into clusters and employs the centroids to represent the artificially constructed clusters, followed by a re-estimation of the segmented output.

Algorithm 2: HFCMKM

Step 1 Input MRI images [I_n] // cleaned level set images
Step 2: $U^1 = [U_{nj}]$
Step 3: Fix number of clusters C
Step 4: Fix Max Iteration =T
Step 5: Fix delta value ϵ
Step 6: Initialize randomly U^0
Step 7: t=1 to T do

$$U_{nj} = \frac{1}{\sum_{k=1}^{C}\left(\frac{I_n - c_j}{I_n - c_k}\right)} \quad \frac{2}{m-1} \tag{3.3}$$

Calculate the new cluster center C_j

$$C_j = \frac{\sum_{i=1}^{N} U_{nj}^{m} \ . \ I_n}{\sum_{i=1}^{N} U_{nj}^{m}} \tag{3.4}$$

Calculate the new objective function.
if $|J^1 - J^{t-1}| < \in$ then
Break;
else
Continue;
end if
end for

3.4. Hybrid Feature Extraction using GLCM and RDWT.

3.4.1. Gray level Co-occurrence Matrix. There are several characteristics that aid in distinguishing brain tumors, including statistical color features, GLCM texture-based features, and RDWT-based low-level features [12]. GLCM, which represents the frequency of pairs of pixel values in an image, has been particularly useful in characterizing the texture of brain tumors. By using GLCM as a feature extraction method, the texture properties of a brain image can be captured and utilized in a machine learning model for classifying different regions of the image. This process involves several steps, such as image preparation, gray level quantization, co-occurrence matrix calculation, feature extraction, and machine learning model training[12].

$$\text{Contrast} = \sum_{a,b=0}^{N-1} S_{a,b}(a-b)^2 \tag{3.7}$$

$$\text{Homogeneity} = \sum_{a,b=0}^{N-1} \frac{S_{a,b}}{1+(a-b)^2} \tag{3.8}$$

$$\text{Correlation} = \sum_{a,b=0}^{N-1} S_{\ a,b}\left[\frac{(a-\mu_a)(b-\mu_b)}{\sqrt{(\sigma_a^2\ )(\sigma_b^2)}}\right] \tag{3.9}$$

$$\text{Angular Second Moment (ASM)} = \sum_{a,b=0}^{N-1} s_{a,b}^2 \tag{3.10}$$

$$\text{Energy} = \sqrt{ASM}[12] \tag{3.11}$$

3.4.2. Redundant discrete wavelet transform. Next step involves RDWT, It is a popular technique used under the feature extraction of image processing. RDWT is a method of decomposing an image in to wavelet coefficients. It includes Lower level and higher level frequency bands. Lower level represents the features where as higher level represents the finer details. It also used to extract the features for further analysis. This redundancy allows more flexibility and accurate feature extraction results under segmentation process.

$$\text{Mean}(\mu) = \frac{1}{N^2} \sum_{i,j=1}^{N} I(i,j) \tag{3.12}$$

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{\sum_{i,j=1}^{N} [I(i,j) - \mu^2}{N^2}} \tag{3.13}$$

Finally, the features of both GLCM and RDWT are combined and produce the result in the form of hybrid feature matrix[12]. The Proposed HFCMKM is a highly accurate segmentation technique that ensures smooth and continuous segmentation, avoids creating isolated regions in the image, and provides more accurate results than traditional FCM. These advantages lead to improved diagnosis and effective treatment planning for patients

3.5. Deep Learning Based Probabilistic Neural Network (DPNN). The deep probabilistic neural network (DPNN) is a type of neural network that combines the strengths of deep learning and probabilistic modeling, making it a valuable tool in feature extraction and classification. In the context of brain tumor image segmentation, the DPNN can differentiate between tumor and non-tumor tissue based on features such as intensity, texture, and shape. During the feature extraction process, the DPNN learns to assign a probability value to each class based on their features. When presented with a new image, the DPNN uses these probabilities to classify each pixel as either tumor or non-tumor[3] The DPNN performs these classification operations once, learning to train with the input's local characteristics at a higher level. The network also provides information on each layer, such as its size, parameters, and filters. By combining all layers, the DPNN achieves a satisfactory output. Overall, the DPNN is a powerful tool for accurately segmenting brain tumors from medical images.

4. Result Discussion. This system utilizes the BraTS-2020 dataset to give a full evaluation of findings and efficacy compared with previous techniques. Moreover, assessments of numerous objective outcomes as well as visual objective achievements are offered. The MatlabR2021a tool was used in combination with a GPU processor to perform this study. Here on the dataset, the algorithms are taught in realistic scenarios. The experiments are executed using a Windows 10 operating system equipped with a NVIDIA Tesla P100 CPU. The 10-fold cross-validation method is used for learning the suggested model. All of the algorithms were developed using Matlab R2021a with a learning rate of 0.02 and 10-fold cross validation over 1000 iterations.

4.1. Dataset Used for Evaluation. The BraTS-2020 dataset is used to evaluate the proposed network's efficiency. There are 369 images used for training, 125 for validation, and 169 for testing in multi-modal brain MRI studies. Each study contains 80 T1-weighted (T1) images, 80 T1ce-weighted (T1ce) images, 80 T2-weighted (T2) images, and 209 Flair sequence images. The annotation for training studies is made available for online evaluation and the final segmentation competition, but not for validating and test runs.

The proposed HFCMKM method evaluates the brain tumor dataset with following parameter metrics such as accuracy, precision, recall, f1-score and confusion matrix. The proposed method HFCMKM is compared with existing methods such as MCA-CS [2], LDNSD [2] and DB-CNN [2].

Accuracy. It is used to evaluate the classification of correct brain tumor images accurately.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} X100 \tag{4.1}$$

Table 4.1: Metric measures for classification of brain tumor image

Methods Used	Accuracy	Precision	Recall	F1-score
MCA-CS	88	79	89.45	92
LDNSD	84.76	76	87	94.75
DB-CNN	92	88	91	93.09
HFCMKM	**97.23**	**93**	**95.78**	**97.88**

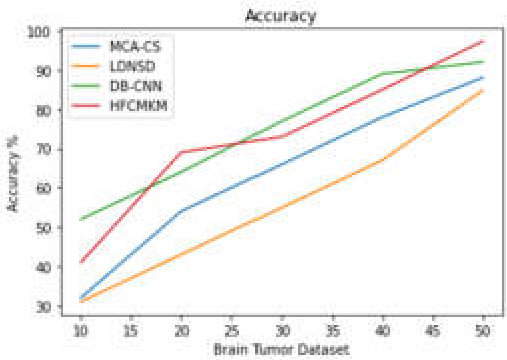


Fig. 4.1: Accuracy

Precision.

$$\text{precision} = \frac{TP}{TP + FP} X100 \tag{4.2}$$

Recall.

$$\text{recall} = \frac{TP}{TP + FN} \tag{4.3}$$

Table 4.1 shows that Classification of four types of brain tumor images like T1-weighted, T1ce-weighted, T2-weighted and Flair sequences using proposed work.

Table 4.1 presents the results of an evaluation of four different machine learning models for a specific task. The task is not specified, but it is assumed to be a classification problem, as the metrics reported are accuracy, precision, recall, and F1-score. These metrics are commonly used for evaluating the performance of classification models.The four models evaluated are MCA-CS, LDNSD, DB-CNN, and HFCMKM. For each model, the table reports the accuracy, precision, recall, and F1-score achieved by the model.

Accuracy is the percentage of correctly classified instances out of all instances. Precision is the percentage of correctly classified positive instances out of all instances classified as positive. Recall is the percentage of correctly classified positive instances out of all actual positive instances. F1-score is the harmonic mean of precision and recall.

Based on the results in the table, the highest accuracy is achieved by the HFCMKM model with a score of 97.23%. The highest precision is achieved by the DB-CNN model with a score of 88%. The highest recall is achieved by the HFCMKM model with a score of 95.78%. Finally, the highest F1-score is achieved by the HFCMKM model with a score of 97.88%.

Figure 4.1 shows the graphical representation of the accuracy which is compared with existing methods. X-axis represents number of images in dataset and Y-axis represents accuracy percentage for tumor prediction.

The HFCMKM model outperforms the other models evaluated in terms of accuracy, recall, and F1-score. However, the DB-CNN model achieves the highest precision score. The choice of the best model will depend on the specific requirements of the task at hand and the trade-off between different performance metrics.

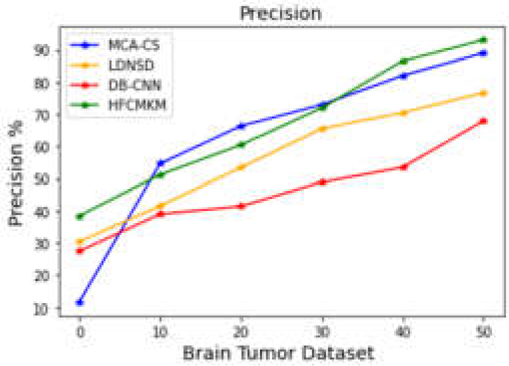


Fig. 4.2: Precision

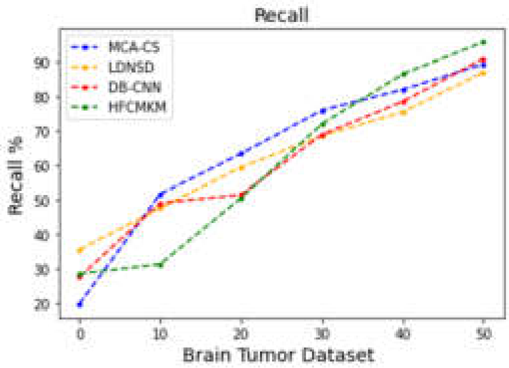


Fig. 4.3: Recall

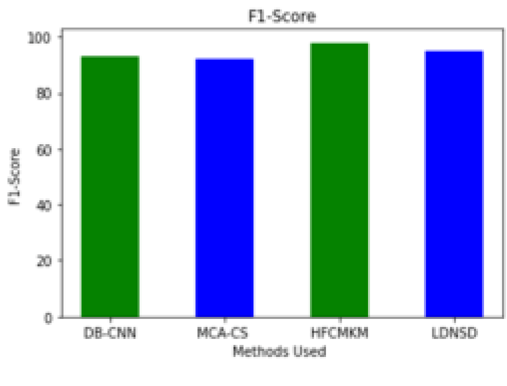


Fig. 4.4: F1-Score

The confusion matrix classified the correct and incorrect classification of brain tumor images. The Classification of four types of brain tumor images like T1-weighted, T1ce-weighted, T2-weighted and Flair sequences evaluated in confusion matrix and shown in figure 4.5.

The use of larger datasets can result in high computational complexity. The proposed approach of combining HFCMKM segmentation and classification performance needs to be improved. There are various metric methods

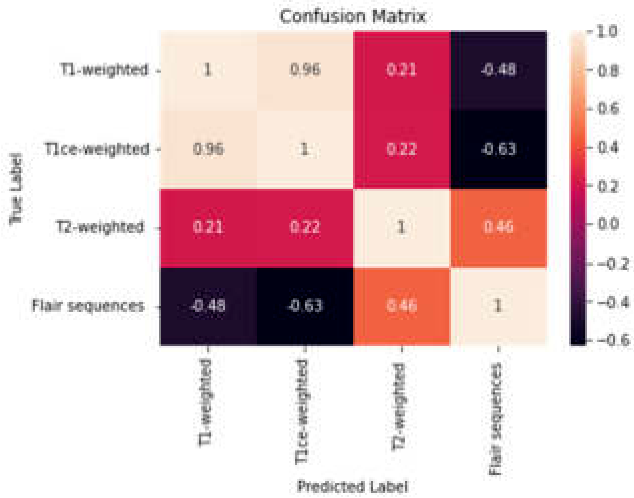


Fig. 4.5: Confusion Matrix

that have been published under DLPNN, but their main limitation is their high complexity. This paper proposes a novel approach that combines clustering methods. Traditional clustering methods have limitations that can affect their accuracy in providing the desired output. By combining the strengths of adaptive clustering methods, high-quality results can be obtained. Therefore, this paper presents a new approach, the HFCMKM method, which involves preprocessing, noise detection using Level Set, applying FCMKM, and finally, combining these strengths to improve the tumor region's accuracy.

5. Conclusion. Finally, this paper we proposed HFCMKM approach to enhance the quality and accuracy of the MRI Brain image to identified the tumor region perfectly based on the above features as we discussed while. It performs in a good way and lead to correct path for clinical diagnosis. This method also helps to localize the tumor location accurately. Initially this work implemented the Hierarchical Clustering to enhance the quality of the image. In addition, Level set function used to remove noise and pre-processing of the source images. Finally combining strengths of both HFCM and K means based segmentation is used to localize the tumor. Furthermore, Hybrid features is extracted from the segmented images using GLCM and RDWT approaches. Finally, DLPNN is used to classify benign and malignant by using this feature. Finally, the proposed method is compared with the various existing methods to prove that the proposed method can be adopted for real time application.

REFERENCES

[1] S. Deepak and P. Ameer, *Automated categorization of brain tumor from mri using cnn features and svm*, Journal of Ambient Intelligence and Humanized Computing, 12 (2021), pp. 8357–8369.

[2] R. S. Devi, B. Perumal, and M. P. Rajasekaran, *A hybrid deep learning based brain tumor classification and segmentation by stationary wavelet packet transform and adaptive kernel fuzzy c means clustering*, Advances in Engineering Software, 170 (2022), p. 103146.

[3] Q. Hu, S. Hu, and F. Zhang, *Multi-modality medical image fusion based on separable dictionary learning and gabor filtering*, Signal Processing: Image Communication, 83 (2020), p. 115758.

[4] M. K. Islam, M. S. Ali, M. S. Miah, M. M. Rahman, M. S. Alam, and M. A. Hossain, *Brain tumor detection in mr image using superpixels, principal component analysis and template based k-means clustering algorithm*, Machine Learning with Applications, 5 (2021), p. 100044.

[5] A. R. Khan, S. Khan, M. Harouni, R. Abbasi, S. Iqbal, and Z. Mehmood, *Brain tumor segmentation using k-means clustering and deep learning with synthetic data augmentation for classification*, Microscopy Research and Technique, 84 (2021), pp. 1389–1399.

[6] M. A. Khan, I. Ashraf, M. Alhaisoni, R. Damaševičius, R. Scherer, A. Rehman, and S. A. C. Bukhari, *Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists*,

Diagnostics, 10 (2020), p. 565.
[7] A. Khosravanian, M. Rahmanimanesh, P. Keshavarzi, and S. Mozaffari, *Fast level set method for glioma brain tumor segmentation based on superpixel fuzzy clustering and lattice boltzmann method*, Computer Methods and Programs in Biomedicine, 198 (2021), p. 105809.
[8] S. Maheswaran, S. Sathesh, G. Saran, and B. Vivek, *Automated coconut tree climber*, in 2017 International Conference on Intelligent Computing and Control (I2C2), IEEE, 2017, pp. 1–6.
[9] K. Maheswari, A. Balamurugan, P. Malathi, and S. Ramkumar, *Hybrid clustering algorithm for an efficient brain tumor segmentation*, Materials Today: Proceedings, 37 (2021), pp. 3002–3006.
[10] M. A. Naser and M. J. Deen, *Brain tumor segmentation and grading of lower-grade glioma using deep learning in mri images*, Computers in biology and medicine, 121 (2020), p. 103758.
[11] P. S. Raja et al., *Brain tumor classification using a hybrid deep autoencoder with bayesian fuzzy clustering-based segmentation approach*, Biocybernetics and Biomedical Engineering, 40 (2020), pp. 440–453.
[12] M. Ramprasad, M. Z. U. Rahman, and M. D. Bayleyegn, *A deep probabilistic sensing and learning model for brain tumor classification with fusion-net and hfcmik segmentation*, IEEE Open Journal of Engineering in Medicine and Biology, 3 (2022), pp. 178–188.
[13] Z. U. Rehman, M. S. Zia, G. R. Bojja, M. Yaqub, F. Jinchao, and K. Arshid, *Texture based localization of a brain tumor from mr-images by using a machine learning approach*, Medical hypotheses, 141 (2020), p. 109705.
[14] S. Rinesh, K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, Y. A. Waji, et al., *Investigations on brain tumor classification using hybrid machine learning algorithms*, Journal of Healthcare Engineering, 2022 (2022).
[15] L. Rundo, L. Beer, S. Ursprung, P. Martin-Gonzalez, F. Markowetz, J. D. Brenton, M. Crispin-Ortuzar, E. Sala, and R. Woitek, *Tissue-specific and interpretable sub-segmentation of whole tumour burden on ct images by unsupervised fuzzy clustering*, Computers in biology and medicine, 120 (2020), p. 103751.
[16] V. Sabitha, J. Nayak, and P. R. Reddy, *Withdrawn: Mri brain tumor detection and classification using kpca and ksvm*, 2021.
[17] V. Sanjay and P. Swarnalatha, *A survey on various machine learning techniques for an efficient brain tumor detection from mri images*, IJEER, 10 (2022), pp. 177–182.
[18] K. S. Sankaran, M. Thangapandian, and N. Vasudevan, *Brain tumor grade identification using deep elman neural network with adaptive fuzzy clustering-based segmentation approach*, Multimedia Tools and Applications, 80 (2021), pp. 25139–25169.
[19] C. J. J. Sheela and G. Suganthi, *Accurate mri brain tumor segmentation based on rotating triangular section with fuzzy c-means optimization*, Sādhanā, 46 (2021), p. 226.
[20] R. Sindhiya Devi, B. Perumal, and M. Pallikonda Rajasekaran, *A robust hybrid fusion segmentation approach for automated tumor diagnosis and classification in brain mr images*, Journal of Intelligent & Fuzzy Systems, 42 (2022), pp. 6063–6078.
[21] R. Vankdothu and M. A. Hameed, *Brain tumor segmentation of mr images using svm and fuzzy classifier in machine learning*, Measurement: Sensors, 24 (2022), p. 100440.
[22] A. Verma, M. Ansari, P. Tripathi, R. Mehrotra, and S. A. Shadab, *Brain tumor detection through mri using image thresholding, k-means, and watershed segmentation*, in Computational Intelligence in Healthcare Applications, Elsevier, 2022, pp. 267–283.
[23] B. Vivek, S. Maheswaran, P. Keerthana, S. Sathesh, S. Bringeraj, R. A. Sri, and S. A. Sulthana, *Low-cost raspberry pi oscilloscope*, in 2018 international conference on intelligent computing and communication for smart World (I2C2SW), IEEE, 2018, pp. 386–390.
[24] L. Xiao, H. Qi, C. Fan, and Z. Ai, *Energy noise detection fcm for breast tumor image segmentation*, IEEE Access, 8 (2020), pp. 62895–62904.
[25] Y. Zhang, B. Jiang, J. Wu, D. Ji, Y. Liu, Y. Chen, E. X. Wu, and X. Tang, *Deep learning initialized and gradient enhanced level-set based segmentation for liver tumor from ct images*, IEEE Access, 8 (2020), pp. 76056–76068.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Mar 10, 2023
Accepted: May 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 1065–1076, DOI 10.12694/scpe.v24i4.2220

CONFIDENTIAL TRAINING AND INFERENCE USING SECURE MULTI-PARTY COMPUTATION ON VERTICALLY PARTITIONED DATASET

KAPIL TIWARI,[*] NIRMALYA SARKAR [†] AND JOSSY P GEORGE[‡]

Abstract. Digitalization across all spheres of life has given rise to issues like data ownership and privacy. Privacy-Preserving Machine Learning (PPML), an active area of research, aims to preserve privacy for machine learning (ML) stakeholders like data owners, ML model owners, and inference users. The Paper, CoTraIn-VPD, proposes private ML inference and training of models for vertically partitioned datasets with Secure Multi-Party Computation (SPMC) and Differential Privacy (DP) techniques. The proposed approach addresses complications linked with the privacy of various ML stakeholders dealing with vertically portioned datasets. This technique is implemented in Python using open-source libraries such as SyMPC (SMPC functions), PyDP (DP aggregations), and CrypTen (secure and private training). The paper uses information privacy measures, including mutual information and KL-Divergence, across different privacy budgets to empirically demonstrate privacy preservation with high ML accuracy and minimal performance cost.

Key words: Privacy-Preserving Machine Learning (PPML), Vertically Partitioned Datasets Secure Multi-Party Computation (SMPC), Confidential Inference, Differential Privacy (DP)

1. Introduction. Today's evolved world is continuously reinventing itself. As globalization and digitalization with Machine Learning and Artificial Intelligence have brought countries closer, personal space and privacy boundaries have been blurred. Digitalization has led to the rat race of countries trying to computerize their citizens' data. This digitalization has unlocked several productivity, accuracy, and efficiency avenues while raising the dreaded privacy question.

While Oxford defines privacy as "the state of being alone and not watched or disturbed by other people", privacy has evolved over the decades; concerning machine learning, privacy is not only of the data owners but also of model owners and clients[23]. A data owner is an individual or an organization willing to share personal data to facilitate the creation of Machine Learning (ML) applications. A model owner is a collective term used to describe the inventors of the ML model. Clients are the end-user and consumers of the Model. True privacy is one where the privacy of all these contributors is maintained and safeguarded[23].

Several methods and protocols have been developed to achieve privacy for all the contributors, and collectively these mechanisms are called Privacy-Preserving Machine Learning (PPML)[20, 18, 16, 14]. The complexity of PPML only increases when various forms of data are accessed in multiple formations. Under conventional thoughts, data is collected from various sources and owners and then combined horizontally as incremental records. Here different sources give us more data points to test, but this is not the only way. Most of the available work in the research fraternity is around Regression problems and Neural networks[7, 3, 19]. Classification, especially Gradient Boosting[2, 12, 18], has recently gained immense popularity. All the methodologies mentioned earlier are being deployed in recommendation systems, anomaly detection, predictive analytics and many more. They all simultaneously face a considerable challenge: tackling Vertically Partitioned Datasets.

To begin with, Vertical Partitioned Data can be distinguished from Horizontal Partitioned Data based on observation and features. Observations are distributed amongst various data owners/providers in Horizontal Partitioned Data for the same features. Contrary to this, in vertically partitioned data, various entities own different features/attributes of information for the same set of entities[10].

Often, individual data held by the institution cannot bring about drastic changes, but combined with other sources, it can give deeper insights. In most countries, there exists a digital database of the citizen. In the

[*]Department of Computer Science, CHRIST University, India (kapiltiwarires1@outlook.com)
[†]Department of Computer Science, CHRIST University, India
[‡]Department of Computer Science, CHRIST University, India

case of India, the Adhaar id acts as a unique identifier for such a database. Policy planners can get more significant insights from it when analyzed alongside the health and insurance and banking sectors' databases. These insights can help create impactful policies and social welfare; one can develop policy recommendation systems if done right. In this example, the Adhaar database, healthcare database and finance database are various data sources catering to the same set of users but having different data. This is a classic example of vertically partitioned data. Even though insights from vertically partitioned data can be of great value, it also poses significant threats to the privacy and performance of ML models.

Policy making is not the only domain wherein vertically partitioned data is utilized; several other fields like fintech and security are interested. Their primary concern is preserving all privacy: the clients that will eventually use the services, the model owner or the creator of ML applications, and the data owner.

Attempts to maintain privacy at various levels have resulted in multiple methodologies and practices. In the aggregation stage, privacy is attempted to be safeguarded via anonymization, Homomorphic Encryption, and Differential Privacy[1]. In the training and inference stages, homomorphic Encryption, Differential Privacy are again popular alongside Secure Multiparty computation techniques. Federated Learning[23] is another commonly used method that has gained popularity, especially regarding vertical partitioning datasets.

Anonymization removes information likes Name, Adhaar number, and address. This method is not functional with Vertically partitioned datasets as PII becomes key in aligning the datasets from various sources. Homomorphic Encryption (HE)[24] is the process of having encrypted data. Even though HE is effective, it comes with high computational costs and time complexities. Federated Learning is training a centralized ML algorithm on decentralized data. Here, the model is shared across various clients and is trained locally. Federated Learning has several challenges, including that expensive communication and system heterogeneity. As the model is being shared with multiple data owners, continuous communication is necessary, but transmission via a network is slower than local computation. Moreover, there exists system heterogeneity, which poses a threat. Data owners vary in hardware, connectivity and power. This can lead to unwanted connection breaks and privacy threats[25].

Based on the statistical method, Differential Privacy (DP) adds perturbation to increase privacy. Secure-Multi-Party Computation (SMPC)[21], uses collaborative computing technology with multiple parties to solve privacy concerns.

Incomplete feature information retained by a single participant and a challenging training procedure are significant issues with vertically partitioned data and needs quick resolution[9]

This paper presents the solution for preserving privacy for collaborative machine learning stakeholders using secure multi-party computation techniques, especially for the vertically partitioned dataset. The method, named CoTraIn-VPD and implemented using the Python language and a few open-source libraries, has showcased the effectiveness in preserving the privacy of collaborative machine learning stakeholders, including data owners, model builders, and inference clients. The paper describes the specific validation mechanisms to prove the privacy gain with CoTraIn-VPD using information metrics such as mutual information and KL-Divergence.

The paper is organized as follows. The context of the study is discussed in the next section, emphasizing previous research and any gaps. While section 3 describes the technique using motivation, architecture, execution flow, implementation details, and lab setup. Section 4 covers the findings of the experiments and discusses the insights and results in fact, followed by a conclusion.

2. Background. Vertically partitioned dataset consists of multiple data owners having mutually exclusive columns, features, or variables for a given population. In a real-life scenario, data is split across multiple data providers such as local and government agencies; Unique Identification systems like Aadhar, Income Tax records, life insurance firms, and transport records carry different information about an individual, which can be collaborated to produce a qualitative machine learning model to generate meaningful insights. There are multiple secure and private ways this collaboration over the vertically partitioned dataset can be carried out, like certain data aggregation techniques like anonymization; however, these techniques are not fool-proof and can still leak private information. Differential Privacy techniques can increase the privacy measured by a privacy budget but adversely affect training and inference performance apart from impacting model accuracy[23]. Some of the proposals presented in the past to solve the problem needed more empirical findings and implementations. Hell at el.[11] Came up with the first implementation using linear regression over Secure Multi-Party Computation but used the Homomorphic Encryption technique that was too expensive to be implemented practically due to heavy computations. There was another implementation by Bogdanov et al. [4], But it was limited to 10 features. Charlotte et al.'s [5] algorithm to train a logistic regression model on an encrypted dataset using a homomorphic encryption technique proved inefficient in model accuracy and training time.

There were a few HE-based techniques for secure training and inference, like Chen et al., Li and Sun[13], Carpov et al.[6], [15], but all of them either lacked model accuracy or performance or were too expensive to implement. To securely train the partitions, there were few federated learning-based techniques like Hardy et al. [17] They applied additive HE to train the model but could not improve the loss of model accuracy due to approximation techniques. Mandal et al. [17] built a regression model using an additive secret-sharing process over high dimensional data but was limited to the horizontally distributed dataset. Liu et al. [8] established a platform to support different systems in developing ML models collaboratively over vertically partitioned datasets. Cock et al. [8] suggested a securely trained LR model for dispersed parties but used a trusted third-party initializer to assign random weights across two computing servers.

PPML consists of private training and inference and should be dealt with separately using secure multi-party computation techniques to preserve compute cost and keep the accuracy loss in control. Today, implementation has yet to empirically prove the effectiveness of a fast multi-party computation technique, especially for the vertically partitioned dataset setting.

To summarize, preserving privacy during model training and inference for machine learning stakeholders, like owners of data and models and inference clients, is an active area of research where existing work lacks implementation and empirical findings.

3. Motivation and Process.

3.1. Motivation. Vertically partitioned data spread across multiple data owners need an efficient and accurate privacy-preserving technique for model training and inference. The solution should scale with the number of partitions and model owners and cater to multiple inference users quickly, securely, and privately. Studies have shown that ensemble learning on vertically partitioned datasets enhances the accuracy of inference results. Hence, a technique for confidential inference and training is a need, especially when the dataset is vertically partitioned. In the past, vertically partitioned datasets brought complexity and performance penalty to privacy-preserving machine learning solutions; that's where the proposed technique would bring accuracy and private gain.

3.2. Technique. A typical vertically partitioned dataset would have multiple data owners having different schemas for details about a single entity. The proposed technique uses the Secure Multi-Party Computation (SMPC) technique to train the intermediate models that privately address data owners' privacy concerns. Similarly, the intermediate model and secret are shared in SMPC clusters with multiple secure nodes such that no single node gets to know the complete model, thereby preserving the privacy of model owners. The same SMPC cluster carries out the private inference when client values are secret and shared over the same secure nodes, so a secure node gets a fraction of inference input. The inference output is generated at each secure node, and later it gets aggregated by a result aggregator, which applies differential privacy-based secure aggregation on the final inference output. The degree of privacy is controlled by 'epsilon' or privacy budget. The final inference output is later shared over the network in an encrypted format.

3.3. Architecture. As shown in Fig 3.1, the vertically partitioned data has three splits named Feature Split A (FSA), Feature Split B (FSB), and Target List. The split has different schemas where FSA and FSB have a list of feature values or x values, while the TL has the target value or y value as one of the columns. As depicted in the figure, FSA and Target List are securely trained using SMPC, which generates an intermediate model named Ma; similarly, Mb is generated by SMPC-based secure training of FSB and Target List. The application of SMPC has ensured the privacy of FSA, FSB, and Target List data owners. The model's Ma and Mb are secretly shared over another SMPC-based cluster of total N Secure Nodes, namely SN_a, SN_b, SN_c...SN_n such that M_a is divided into N multiple splits M_{aa}, M_{ab}, M_{ac}...M_{an} and shared across secure nodes. The secret share ensures the privacy of model owners, as no secure node has the complete model.

The inference input X is similarly divided across N secure nodes such that SN_a, SN_b, SN_c...SN_n gets X_a, X_b, X_c...X_n values, and it makes the inference value private too. The secure nodes compute the intermediate

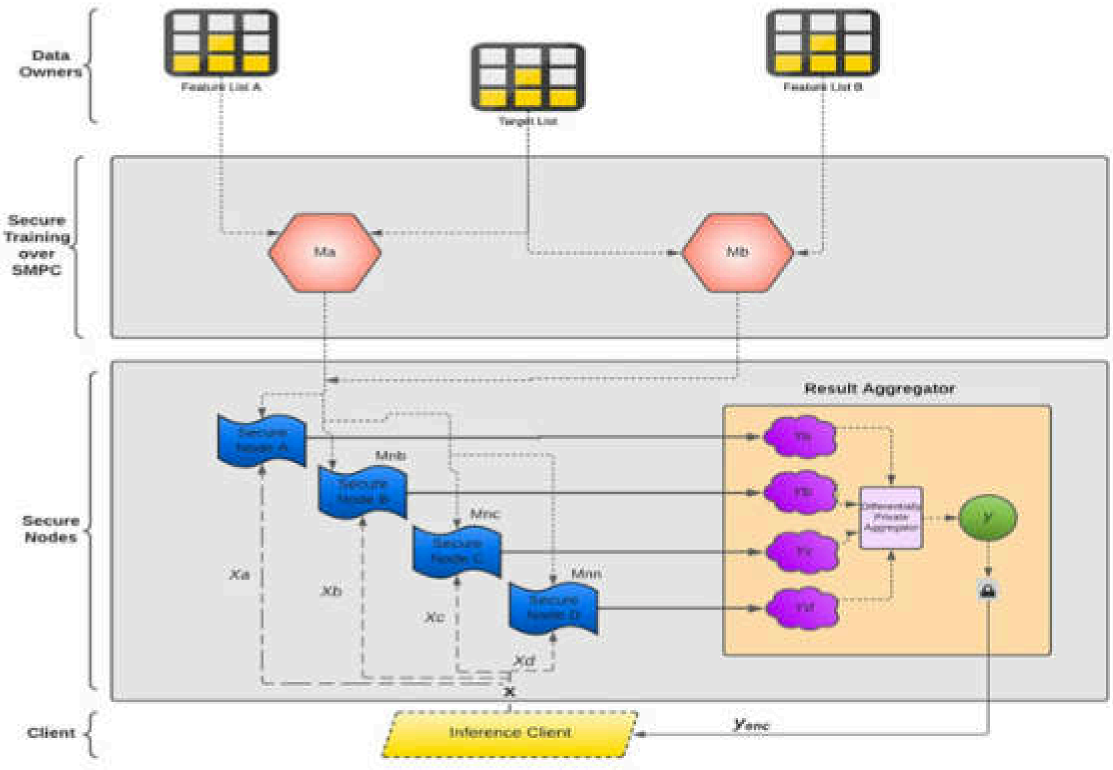


Fig. 3.1: CoTraIn-VPD Architecture

inference outputs like Y_a, Y_b, Y_c...Y_n, which is aggregated by a trusted result aggregator who applied differential privacy aggregation to arrive at the final inference output Y', which is later shared over the network in an encrypted format $Y'_{encrypt}$ to preserver network security. Nowhere the data, model, or inference input has been shared in its original form and value, additionally the collaborative computation occurs on different secure nodes make the technique privacy-preserving.

The methodology of CoTraIn-VPD can be broken down into the following steps:

1. Vertical partitioning of the data: The dataset is partitioned into two or more parts, with each part held by a different party. The partitioning is done such that each party has access to only a subset of the features (columns) of the data, while the full set of records (rows) is distributed across the parties.
2. Local model training: Each party trains a local model on its own data using standard machine learning techniques. The local model is trained on the subset of features that the party has access to, and the output is a model that predicts the target variable based on that subset of features.
3. Secure aggregation: The local models are combined in a secure and privacy-preserving way to obtain a global model that can predict the target variable based on all the features. This is achieved using techniques such as secure multi-party computation (MPC) or homomorphic encryption.
4. Fine-tuning: The global model is fine-tuned using a small amount of jointly-held data to improve its accuracy. This is done in a privacy-preserving way using techniques such as differential privacy or federated learning.

3.4. Mathematical Explanation. As per the proposed training technique, two different splits get trained into a secure multi-party computation setting simultaneously, and the same can be parallelized. Hence, the training time t_{train} is constant and does not depends on the number of vertical splits vs If K is a constant and

t_{train} is training time, we can say:

$$t_{\text{train}} \propto K$$

Similarly, the model is securely distributed to the number of secure nodes and the secure computation occurs at each secure node which later gets aggregated to form the final inference result. Although inference time would increase with secure nodes count because of the computation and aggregation complexity. However, the inference time does has no correlation with number of models participating in the SMPC cluster. Hence, if the time taken to inference is $t_{inference}$ and the secure nodes count is N_{sec_nodes}.

$$t_{\text{inference}} \propto N_{\text{sec}\_nodes}$$

The aggregation adds the statistical noise based on the privacy budget epsilon, which, too, does not impact the time taken to inference. Hence, t does not depend on epsilon.

The CoTraIn-VPD approach has several ethical implications, particularly with regards to data privacy and ownership. Below are some of the potential ethical concerns and considerations:

1. Data privacy: The use of vertically partitioned datasets introduces privacy concerns, as each party may have access to sensitive information that they are not authorized to see. The CoTraIn-VPD approach attempts to address this by limiting each party's access to only a subset of the features, but there is still a risk of sensitive information being leaked if the secure aggregation and privacy-preserving techniques are not implemented correctly.
2. Data ownership: The partitioning of the data means that each party has ownership over their subset of the data. This can lead to issues around data sharing and access, as parties may not want to share their data with others. It is important to establish clear ownership and usage rights for each party, and to ensure that consent is obtained from all parties before any data sharing takes place.
3. Bias and fairness: The use of local models can introduce bias into the global model, as each party may have their own biases and assumptions that are reflected in their local model. This can lead to unfair treatment of certain groups or individuals. It is important to carefully consider the features used by each party and to ensure that the global model is fair and unbiased.
4. Transparency and accountability: The use of secure aggregation and privacy-preserving techniques can make it difficult to understand how the global model is making its predictions. This can make it difficult to hold the parties involved accountable for any errors or biases in the model. It is important to establish clear guidelines for transparency and accountability, and to ensure that the parties involved are able to explain and justify their decisions.
5. Informed consent: In order to participate in vertically partitioned data sharing, parties must give informed consent. This means that they must be fully aware of the potential risks and benefits of sharing their data, and must understand how their data will be used and protected. It is important to ensure that all parties involved have given informed consent, and that any changes to the data sharing agreement are communicated clearly and transparently

3.5. Experimental Setup. As part of the simulation, a virtual machine on Azure cloud with Ubuntu OS was used. The machine's configuration was D2sV3, 2vCPU and RAM as 8GiB. Open-source libraries like PyTorch, PySyft, PyDP, CrypTen, and SyMPC (OpenMined) were used, and Jupiter Notebook was the development environment. The language used was Python 3.9 and dataset was Boston Housing Dataset, with parameters including no. of model's owners: 1to 4, no. of Secure Nodes: 2 to10, and a Result assembler. Pseudocode is stated below,

4. Result and Analysis. Table 4.1 shows the results of running the CoTraIn-VPD approach without differential privacy on two vertically partitioned datasets (VP1 and VP2). The results are presented for different numbers of secure nodes and different numbers of splits in the vertical partitioning. For VP1, with two splits, the inference time ranges from 3 seconds for 3 secure nodes to 11 seconds for 10 secure nodes. The mean squared error (MSE) loss for VP1 is 18.83. For VP2, with three splits, the inference time ranges from 1.42 seconds for 1 secure node to 15.85 seconds for 10 secure nodes. The MSE loss for VP2 is 14.96.

```
Function Secure_Train_Models (n_SecureNodes,n_Models, smpc_Protocol=None)
{
    # split data vertically into multiple feature and one target set based on the columns
    Data_feature_split_A = split columns wise A features
    Data_feature_split_B = split columns wise B features
    Data_feature_split_C = split columns wise C features
    Data_target_split = split columns wise target values

    # securly train each feature set with the target set
    model_A = SecureTrain (Data_feature_split_A, Data_target_split)
    model_B = SecureTrain (Data_feature_split_B, Data_target_split)
    model_C = SecureTrain (Data_feature_split_C, Data_target_split)

    # privately share the model to Differential_Private_And_Secure_Inference function
    n_Models.append(model_A)
    n_Models.append(model_B)
    n_Models.append(model_C)
}
```

```
Function Differential_Private_And_Secure_Inference(n_SecureNodes,n_Models,inference_Data,smpc_Protocol=None)
{
    # Setup the SMPC session for the computation
    if smpc_Protocol == None
        Initialize SMPC session by creating n_SecureNodes vitual machines
        and using SMPC protocol = SPDZ
    Else
        Initialize SMPC session by creating n_SecureNodes vitual machines
        and using SMPC protocol = smpc_Protocol
    Endif
    # Secret share each model to SMPC n_SecureNodes
    For each model in n_Models
        secret share model to SMPC n_SecureNodes resulting into secure model
        append secure model into n_SecureModels
    Endfor
    # Secret share inference data to SMPC n_SecureNodes
    secret share inference data x to SMPC n_SecureNodes resulting secure_Inference_Data
    # Evalute inference at each secure node and generate results
    For each secureModel in n_SecureModels
        calculate inference result with secureModel on secure_Inference_Data
        append the result secure_Results
    Endfor
    # differentially private aggregation of result into final result
    apply differential private aggregation to secure_Result to arrive at final_Result
    Encrypt the final_result to enc_Final_Result
    Return enc_Final_Result
end function
```

Fig. 3.2: Pseudo Code for CoTraIn-VP

Overall, the table shows that increasing the number of secure nodes or splits in the vertical partitioning can increase the inference time, but does not necessarily lead to a reduction in MSE loss. It is important to balance the trade-off between inference time and accuracy when selecting the optimal configuration for the CoTraIn-VPD approach. It should be noted that the results presented in this table 4.1 are without the use of differential privacy. In real-world scenarios, the use of differential privacy may be necessary to protect the privacy of the vertically partitioned datasets. The results may differ when differential privacy is applied.

Table 4.2 shows the results of running the CoTraIn-VPD approach with differential privacy on two vertically partitioned datasets (VP1 and VP2). The results are presented for different numbers of secure nodes and different numbers of splits in the vertical partitioning. For both VP1 and VP2, the inference time is consistent across all configurations, at 0.97 seconds. This is because differential privacy introduces a noise mechanism that adds random noise to the computations, leading to more consistent inference times. However, the MSE loss

Table 4.1: Results without Differential Privacy

Vertical Partitio ning	Secure Node	Inference Time	MSE Loss	Vertical Partitio ning	Inference Time	MSE Loss
VP1 (2 splits)	3	0.968101978	18.82760239	VP 2 (3 splits)	1.422831297	14.96163082
	4	1.772582531	18.82761192		2.535790205	14.96164703
	5	2.774291992	18.82762718		4.1898036	14.96165276
	6	3.980890751	18.82762527		5.812644005	14.96162415
	7	5.856841803	18.82763672		7.877051115	14.96162033
	8	7.015391827	18.82762909		10.2261157	14.96164131
	9	9.025359154	18.82760811		12.91781378	14.96164799
	10	11.13081789	18.8276062		15.84927869	14.9616251

Table 4.2: Results with Differential Privacy

Vertical Partitio ning	Secure Node	Inference Time	MSE Loss	Vertical Partitio ning	Inference Time	MSE Loss
VP1 (2 splits)	3	0.968101978	18.82760239	VP2 (3 splits)	0.968101978	18.82760239
	4	1.772582531	18.82761192		1.772582531	18.82761192
	5	2.774291992	18.82762718		2.774291992	18.82762718
	6	3.980890751	18.82762527		3.980890751	18.82762527
	7	5.856841803	18.82763672		5.856841803	18.82763672
	8	7.015391827	18.82762909		7.015391827	18.82762909
	9	9.025359154	18.82760811		9.025359154	18.82760811
	10	11.13081789	18.8276062		11.13081789	18.8276062

for both VP1 and VP2 remains the same as without differential privacy, at 18.83 and 14.96, respectively. This suggests that the use of differential privacy has not significantly impacted the accuracy of the CoTraIn-VPD approach.

Overall, the results in Table 4.2 demonstrate the effectiveness of the CoTraIn-VPD approach with differential privacy in achieving accurate predictions while preserving the privacy of the vertically partitioned datasets. It is important to note that the use of differential privacy may introduce additional computational costs and considerations, such as selecting the appropriate privacy parameters and noise mechanisms

4.1. Insights.

4.1.1. Partition vs Performance. We got stereotype time taken across various values of secure nodes using different values privacy budget (epsilon 0.2, 0.6, and 0.8). The experiment revealed the time taken for 2 splits was marginally better than 3 splits. However, the trend showed the time taken increases linearly across various privacy budgets with an increase in secure nodes. This explains that the number of splits does not create a huge performance loss across various privacy budgets. The technique does not impact the performance while keeping the inference computation private for all the stakeholders.

4.1.2. Partition vs Privacy. Including differential privacy at the result aggregation increases the privacy for inference clients, model owners, and data owners without impacting accuracy. The experiment shows that the MSE loss remains within the benchmark and does not depend on the number of splits. Including differential privacy brings randomness to the inference output but does not deviate significantly when the privacy budget is increased. The technique showcased accuracy price is a bare minimum and does not depend on the number of splits.

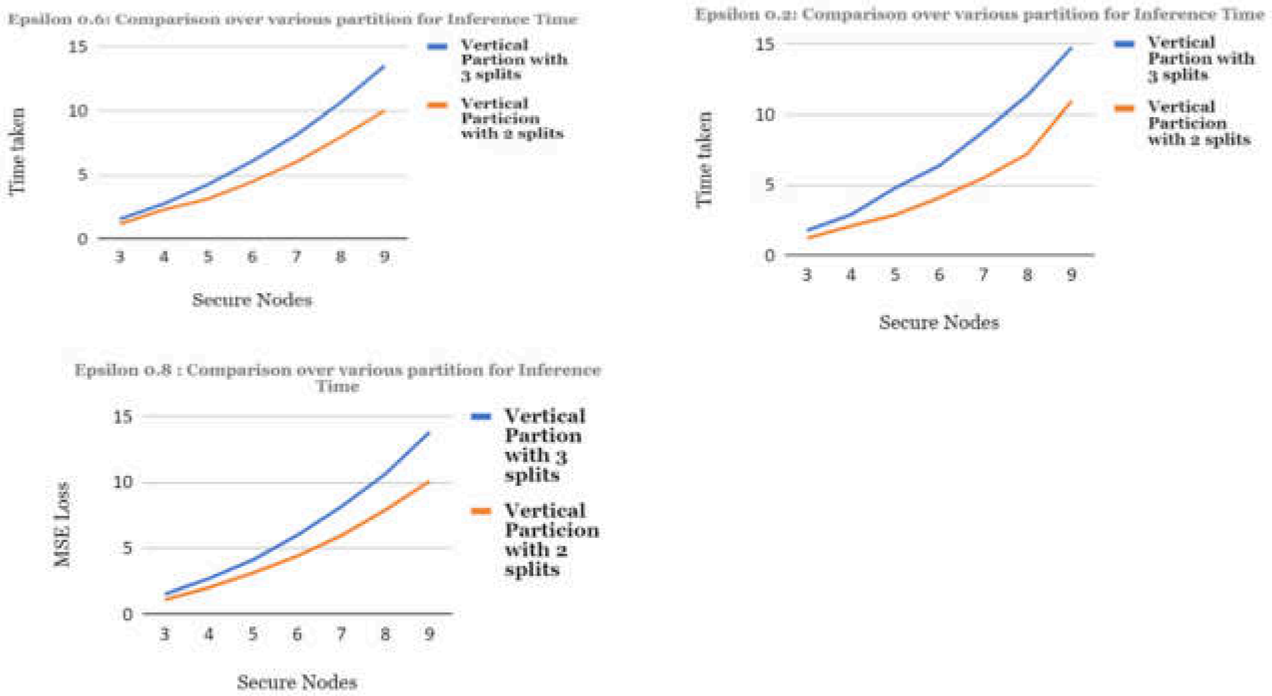


Fig. 4.1: Partition v/s Performance : (a) Epsilon=0.2 (b) Epsilon=0.6 (c) Epsilon=0.9

4.1.3. Privacy vs Accuracy across Partitions. The graph showcased the decrease in MSE loss when epsilon is increased, and this remains the observation across different splits. The technique has proved that the number of splits does not impact the accuracy of inference output. Still, differential privacy increases privacy with a negligible loss of accuracy that the privacy budget can also control. The randomness observed with the privacy budget was in line with the non-private variation of model inference.

4.1.4. Privacy vs Performance across Partitions. The graph represents the mean inference time taken across various privacy budgets is marginally higher than 2 splits. Although, the time taken remains constant across various privacy budgets, which means the technique can give a consistent performance across different values of privacy budgets.

4.1.5. Secure Nodes vs Performance. The observations align with CoInMpro's [22] finding that growing secure nodes increases processing across various nodes, resulting in a linear time increase. We tested 10 secure nodes and found a linear increase in the time taken to inference. However, this did not correlate with the initial number of partitions.

4.1.6. Secure Nodes vs Accuracy . Differential privacy injects randomness into the inference result; however, the study concluded that increasing the secure nodes does not correlate with the accuracy of the output. Across various splits, similar behaviour was observed, and accuracy remained within an acceptable deviation infused due to differential privacy.

4.2. Measure of Privacy. To measure the privacy we need to quantify amount of information reveal by the model after applying the CoTraIn-VPD technique about the training data. With increase in

information leakage the privacy assured by the technique goes down. We have measure the information leakage through Mutual Information concept of information theory, which measure the amount of information two random variable have. In theory, we can prove gain in privacy by the PPML technique, if we find feature vs target value Mutual Information decreases with CoTraIn-VPD when compared with a non-private model.

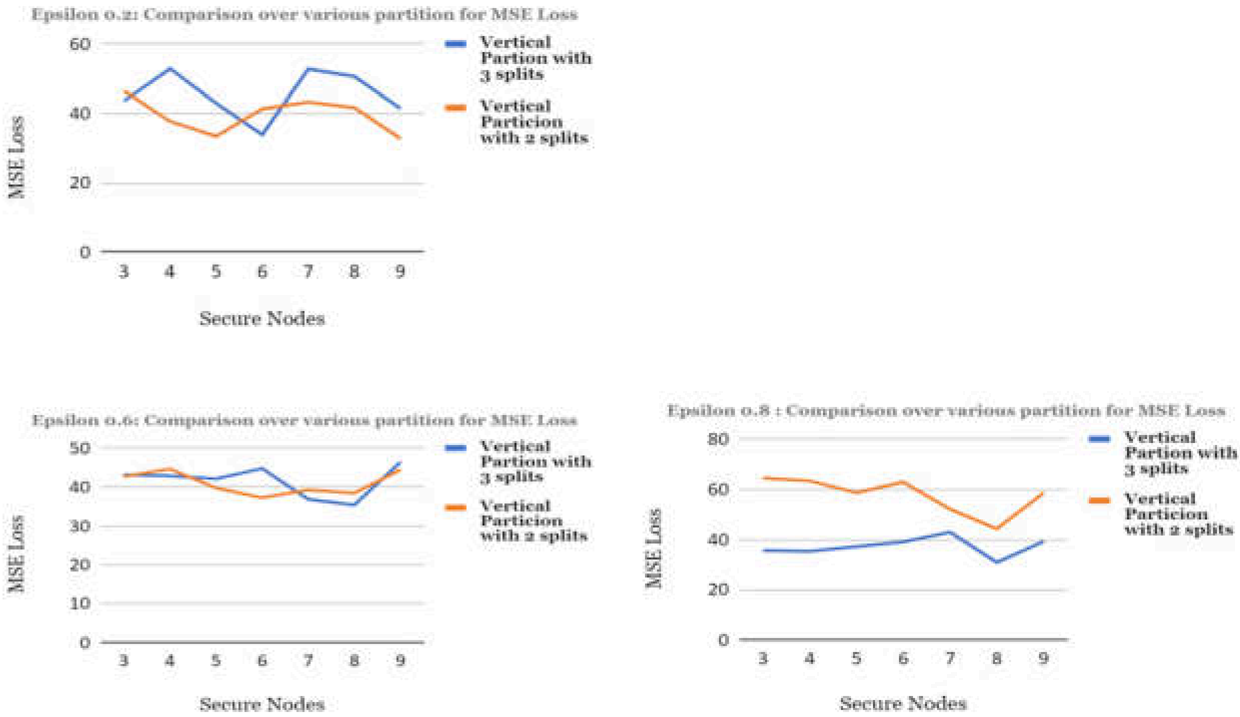


Fig. 4.2: Partition v/s Privacy: (a) Epsilon=0.2 (b) Epsilon=0.6 (c) Epsilon=0.9

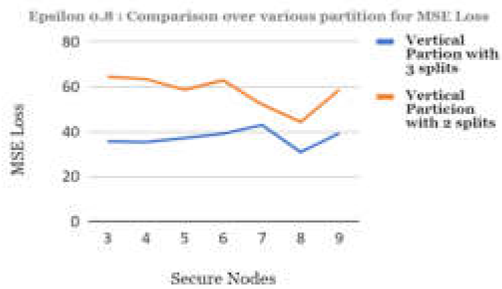


Fig. 4.3: Epsilon v/s Accuracy across Split

Figure 4.7 and 4.8 shows the MI decreased with CoTraIn-VPD across different privacy budget for different splits as compare to the non-private model denoted by privacy budget 0. It is also observed that privacy gains when we go more number of splits. KL-Divergence is another technique to find information leakage by measuring the difference between probability distribution of features and target variable.

In Figure 4.9, the experiments shown the value of KL-Divergence remains below 20 across different privacy budget over different splits.

5. Conclusion. Digitalization across sectors has opened up opportunities for easier collaboration where each sector carries a distinct set of information about individuals, customers, or firms. There is huge scope for improving productivity, efficiency, and synergy across sectors using collaborative machine learning, but is affected because of privacy concerns. There are minimal implementation-proven solutions available today that claim to solve collaborative machine learning across a vertically split dataset without paying huge penalties

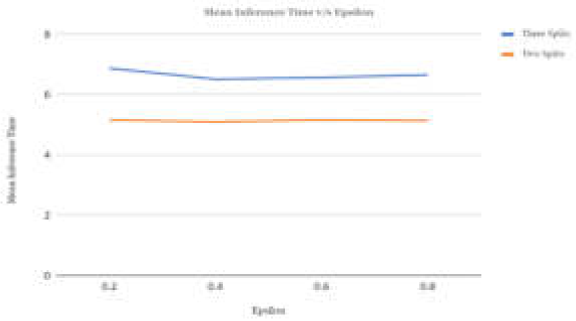


Fig. 4.4: Epsilon v/s Performance across Splits

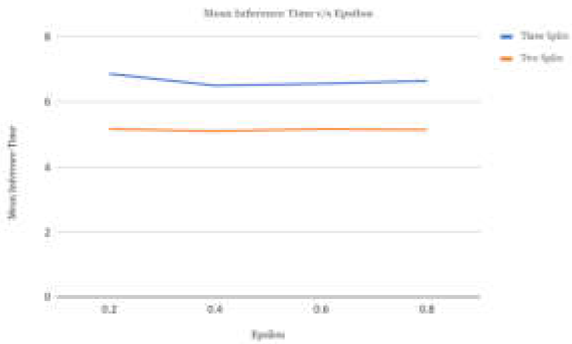


Fig. 4.5: Privacy v/s Performance across 3 splits

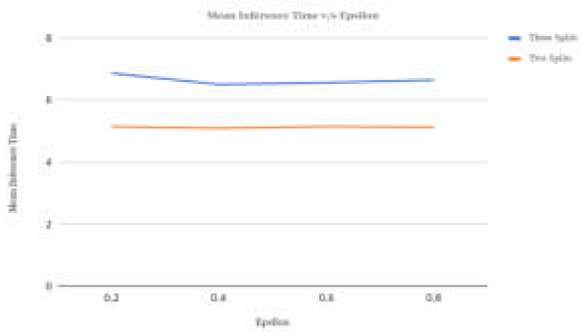


Fig. 4.6: Privacy v/s Accuracy across 3 splits

regarding accuracy and performance or measuring the privacy gain. The paper presented a technique that empirically showed privacy-preserving machine learning for a vertically partitioned dataset using secure multi-party computation techniques. The technique named CoTraIn-VPD, Confidential Training and Inference using secure multi-party computation for the vertically partitioned dataset, trains the vertical split dataset using a secure multi-party computation framework named Crypten. The code written in Python uses open-source libraries such as SyMPC for the SMPC framework, and pyDP for differential privacy features. We ran exhaustive experiments across multiple splits over Azure machine learning VMs running Ubuntu OS. The experiments showed CoTraIn-VPD technique has effectively preserved the privacy of the vertical split data owners, model

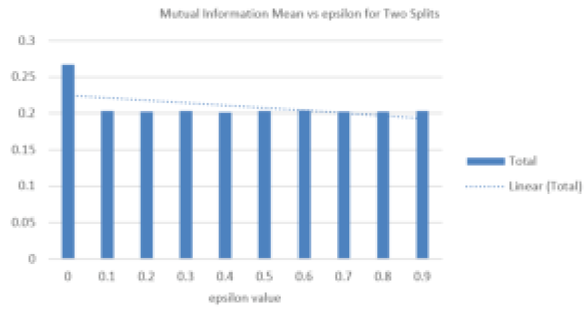


Fig. 4.7: Mutual Information across privacy budgets for two splits

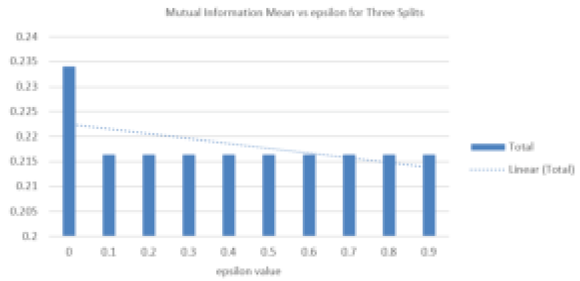


Fig. 4.8: Mutual Information across privacy budgets for three splits

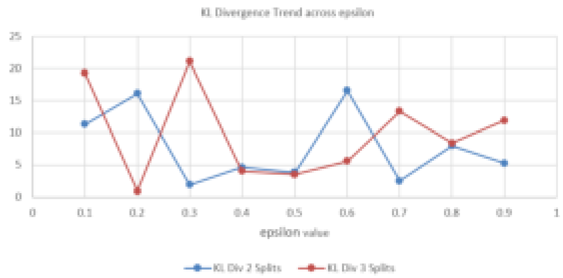


Fig. 4.9: KL-Divergence across privacy budgets and different vertical splits

owners, and inference clients with no impact on accuracy and a marginal linear impact on performance with a growth in secure nodes. The experiment proved the privacy gain using information theory metrics like Mutual Information and KL-Divergence, where information leakage decreased by applying the proposed technique. The experiments focused on two and three splits majorly but can be extended to a higher number of splits and other training and inference algorithms like logistic regression. Using a larger dataset can further extend the research to strengthen the claims of technique effectiveness.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, *Deep learning with differential privacy*, in Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.

[2] N. Agrawal, A. Shahin Shamsabadi, M. J. Kusner, and A. Gascón, *Quotient: two-party secure neural network training and prediction*, in Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 1231–1247.

[3] B. K. Beaulieu-Jones, W. Yuan, S. G. Finlayson, and Z. S. Wu, *Privacy-preserving distributed deep learning for clinical data*, arXiv preprint arXiv:1812.01484, (2018).

[4] D. Bogdanov, L. Kamm, S. Laur, and V. Sokk, *Rmind: a tool for cryptographically secure statistical analysis*, IEEE Transactions on Dependable and Secure Computing, 15 (2016), pp. 481–495.

[5] C. Bonte and F. Vercauteren, *Privacy-preserving logistic regression training*, BMC medical genomics, 11 (2018), pp. 13–21.

[6] S. Carpov, N. Gama, M. Georgieva, and J. R. Troncoso-Pastoriza, *Privacy-preserving semi-parallel logistic regression training with fully homomorphic encryption*, Cryptology ePrint Archive, (2019).

[7] M. Chase, R. Gilad-Bachrach, K. Laine, K. Lauter, and P. Rindal, *Private collaborative neural network learning*, Cryptology ePrint Archive, (2017).

[8] M. De Cock, R. Dowsley, A. C. Nascimento, D. Railsback, J. Shen, and A. Todoki, *High performance logistic regression for privacy-preserving genome analysis*, BMC Medical Genomics, 14 (2021), pp. 1–18.

[9] Y. Deng, X. Jiang, and Q. Long, *Privacy-preserving methods for vertically partitioned incomplete data*, in AMIA Annual Symposium Proceedings, vol. 2020, American Medical Informatics Association, 2020, p. 348.

[10] A. Gascón, P. Schoppmann, B. Balle, M. Raykova, J. Doerner, S. Zahur, and D. Evans, *Secure linear regression on vertically partitioned datasets.*, IACR Cryptol. ePrint Arch., 2016 (2016), p. 892.

[11] R. Hall, S. E. Fienberg, and Y. Nardi, *Secure multiple linear regression based on homomorphic encryption*, Journal of Official Statistics, 27 (2011), pp. 669–691.

[12] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, {*GAZELLE*}*: A low latency framework for secure neural network inference*, in 27th USENIX Security Symposium (USENIX Security 18), 2018, pp. 1651–1669.

[13] Z. Li and M. Sun, *Privacy-preserving classification of personal data with fully homomorphic encryption: an application to high-quality ionospheric data prediction*, in International Conference on Machine Learning for Cyber Security, Springer, 2020, pp. 437–446.

[14] J. Liu, M. Juuti, Y. Lu, and N. Asokan, *Oblivious neural network predictions via minionn transformations*, in Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, 2017, pp. 619–631.

[15] Y. Liu, T. Fan, T. Chen, Q. Xu, and Q. Yang, *Fate: An industrial grade platform for collaborative learning with data protection*, The Journal of Machine Learning Research, 22 (2021), pp. 10320–10325.

[16] E. Makri, D. Rotaru, N. P. Smart, and F. Vercauteren, *Epic: efficient private image classification (or: Learning from the masters)*, in Topics in Cryptology–CT-RSA 2019: The Cryptographers' Track at the RSA Conference 2019, San Francisco, CA, USA, March 4–8, 2019, Proceedings, Springer, 2019, pp. 473–492.

[17] K. Mandal and G. Gong, *Privfl: Practical privacy-preserving federated regressions on high-dimensional data over mobile networks*, in Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop, 2019, pp. 57–68.

[18] P. Mohassel and Y. Zhang, *Secureml: A system for scalable privacy-preserving machine learning*, in 2017 IEEE symposium on security and privacy (SP), IEEE, 2017, pp. 19–38.

[19] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, *Scalable private learning with pate*, arXiv preprint arXiv:1802.08908, (2018).

[20] M. S. Riazi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar, *Chameleon: A hybrid secure computation framework for machine learning applications*, in Proceedings of the 2018 on Asia conference on computer and communications security, 2018, pp. 707–721.

[21] S. Shukla and G. Sadashivappa, *Secure multi-party computation protocol using asymmetric encryption*, in 2014 International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2014, pp. 780–785.

[22] K. Tiwari, K. Bisht, and J. P. George, *Coinmpro: Confidential inference and model protection using secure multi-party computation*, in Data Science and Security: Proceedings of IDSCS 2022, Springer, 2022, pp. 1–14.

[23] K. Tiwari, S. Shukla, and J. P. George, *A systematic review of challenges and techniques of privacy-preserving machine learning*, Data Science and Security: Proceedings of IDSCS 2021, (2021), pp. 19–41.

[24] R. Yonetani, V. Naresh Boddeti, K. M. Kitani, and Y. Sato, *Privacy-preserving visual learning using doubly permuted homomorphic encryption*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2040–2050.

[25] J. Zhang, C. Li, J. Ye, and G. Qu, *Privacy threats and protection in machine learning*, in Proceedings of the 2020 on Great Lakes Symposium on VLSI, 2020, pp. 531–536.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Apr 6, 2023
Accepted: Jun 24, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1077–1086, DOI 10.12694/scpe.v24i4.2229

DESIGN OF A DRAG-REDUCING SYSTEM FOR SMART BAR CONVEYING CHAIN PLATE

JIANGYI PAN*, JIANJUN YANG†, AND WENJUAN LU‡

Abstract. This paper presents a solution to the issue of high frictional resistance between the chain plate and guide rail in cigarette conveying systems, which can result in chain plate failure. The proposed solution is an intelligent drag reduction system that incorporates lubrication and automatic dust removal. The system includes a conveyor chain plate, supporting roller assembly, air knife dust removal assembly, chain plate lubrication assembly, and controller. The controller manages the dust removal assembly's air source output and the lubrication of the chain plate assembly through the opening and closing of an electromagnetic valve, which controls the oil output to decrease resistance. Experimental testing of the system shows that it effectively reduces the chain plate's frictional resistance in the cigarette conveying line, resulting in improved system efficiency

Key words: cigarette conveyor chain plate; drag reduction system; controller

1. Introduction. Currently, the intelligent cigarette conveyor chain resistance-reducing device used in the cigarette conveying system mainly detects the data fed back by the running resistance of the conveying chain plate of each channel. It compares the feedback data with the set threshold according to the feedback data. Suppose the resistance value is greater than the threshold. In that case, the system will compress the air in the corresponding channel, automatically remove dust and add lubricating oil, thereby reducing the resistance of the chain plate. Currently, the commonly used lubrication methods in the bar cigarette conveying system include brushing or artificially adding lubricating oil to the driving sprocket of each motor power head. For dust removal, this method will usually transfer dust to other channels, causing secondary pollution, which is inefficient and increases labor, which is not conducive to dust removal.

Lim[6] present a detailed optimal control design based on the general finite element approach for a structure's integrated design and control system. Wang[11] investigate the effect of the stretching-bending coupling of the piezoelectric sensor/actuator pairs on the system stability of intelligent composite plates. The objective of muFly[8] is to build a fully autonomous micro helicopter comparable to a small bird in size and mass. The trajectory tracking problem of a closed-chain five-bar robot is studied [2]. In the end, Maslák[7] presents a design for converting current labeling using bar codes into labeling using RFID tags. Tsuge[10] uses path synthesis techniques to design four-bar linkage modules to constrain the movement of a 3R chain. The fundamental research question is: how could an intelligent city production system change supply chain design? In answering this question Kumar[5] developed an integrative framework for understanding the interplay between smart city technological initiatives (big data analytics, the industrial Internet of things) and distributed manufacturing on supply chain design. Motivated by the desire for high-throughput public databases[9, 4] design incentive-compatible protocols that run "off-chain", but rely on an existing cryptocurrency to implement a reward and/or punishment mechanism. Abdel-Basset[1] present DEMATEL and AHP in neutrosophic environments to deal effectively with incomplete, uncertain, and incomplete information. Chung[1, 3] consider an innovation in dynamic supply-chain design and operations:connected smart factories that share interchangeable processes through a cloud-based system for personalized production.

Currently, the plastic chain plate online lubrication system configured for the cigarette conveying system of the cigarette factory realizes online automatic detection of the chain plate running resistance and comprehensive analysis. The average running state of the chain plate is the chain plate running state. The detection device

*China Tobacco Guangdong Industrial Co. LTD. Meizhou Cigarette Factory, China (jiangyipan21@outlook.com)
†Inner Mongolia Kunming Cigarette Co. LTD., China
‡Dongguan University of Technology, Dongguan City, Guangdong Province, China

detects that the chain plate pulse data is relatively uniform. The abnormal operation of the chain plate is mainly manifested by bumping, creeping and breaking. When the pulse data of the chain plate detected by the chain plate operation state detection device is relatively large, it indicates that the chain plate has to bump or to creep. When the pulse continues to output, it means that the chain plate is broken and stopped. Standard dedusting methods are: manually blowing the dust scattered on the chain plate surface by using a compressed air gun on the surface of the cigarette conveying line and transporting it for a long distance. Generally, the drive head of the conveying line will be arranged in multiple sections. Only manual blowing can be carried out on the segmented chain plate to reduce the running resistance of the chain plate. The dust purged has yet to be recycled in a centralized way, causing secondary pollution. In addition, it is necessary to add lubricating oil manually, The usage can not be accurately controlled, and the inner side of the chain plate is not evenly smeared. If too much smearing is done, it is easy to penetrate the surface of the chain plate and pollute the delivered cigarette products; When there is dust on the chain plate, the lubricating oil will firmly adhere the dust to the chain plate, thus increasing the running resistance of the chain plate. Low working efficiency and high labor intensity.

The contribution of this research lies in the design and implementation of an intelligent drag reduction system for cigarette conveying, which addresses the issue of high frictional resistance between the chain plate and guide rail. By incorporating lubrication and automatic dust removal, the proposed system reduces the frictional resistance of the chain plate, thereby improving the efficiency of the conveying system.

The system design includes several components, such as the conveyor chain plate, supporting roller assembly, air knife dust removal assembly, chain plate lubrication assembly, and controller, which work together to achieve the intended reduction in resistance. The controller's ability to manage the dust removal assembly's air source output and the lubrication of the chain plate assembly through an electromagnetic valve allows for precise control of the oil output and ensures that the system operates optimally.

The experimental testing of the system further validates its effectiveness in reducing the chain plate's frictional resistance, which is crucial for the smooth and efficient operation of cigarette conveying systems. Therefore, this research's contribution is significant in terms of improving the efficiency and reliability of cigarette conveying systems and may have broader applications in other industries that rely on similar conveying systems.

The plastic chain plate online lubrication system is a set of integrated systems specially designed for the plastic chain plate of cigarette conveying lines and application scenarios.

1.1. System model. The plastic chain plate online lubrication system is a complex system that consists of several different components working together to achieve efficient and reliable operation. The system includes several sets of online lubrication and dust removal devices, automatic lubricating oil supply devices, centralized dust collection devices, single independent drive heads, and the main control system.

The online lubrication and dust removal devices are responsible for applying lubricant to the chain plates and removing any dust or debris that may have accumulated on them. The automatic lubricating oil supply devices are used to control the flow of lubricant to the chain plates, ensuring that the plates are adequately lubricated without being over-lubricated, which can result in additional issues such as build-up and accumulation of debris.

The centralized dust collection devices are designed to collect and dispose of any dust or debris that may be generated during the operation of the system. The single independent drive heads are responsible for driving the chain plates and ensuring they move along the conveyor as required.

The control substations are used to manage and control the individual components of the system, ensuring that they operate as intended. Finally, the main control system is responsible for overseeing the entire system's operation and coordinating the various components to ensure that they work together efficiently.

Figure 1.1 provides a graphical representation of the system's components and how they are connected, allowing for a better understanding of the system's overall design and operation. Overall, the plastic chain plate online lubrication system is a complex but crucial system that enables the efficient and reliable operation of conveyor systems that rely on plastic chain plates.

The line lubrication and dedusting device mainly consist of the lubricating oil pot, the oil pot high-level detection sensor, the oil pot low-level detection sensor, the precision electric pump, the single-channel compressed

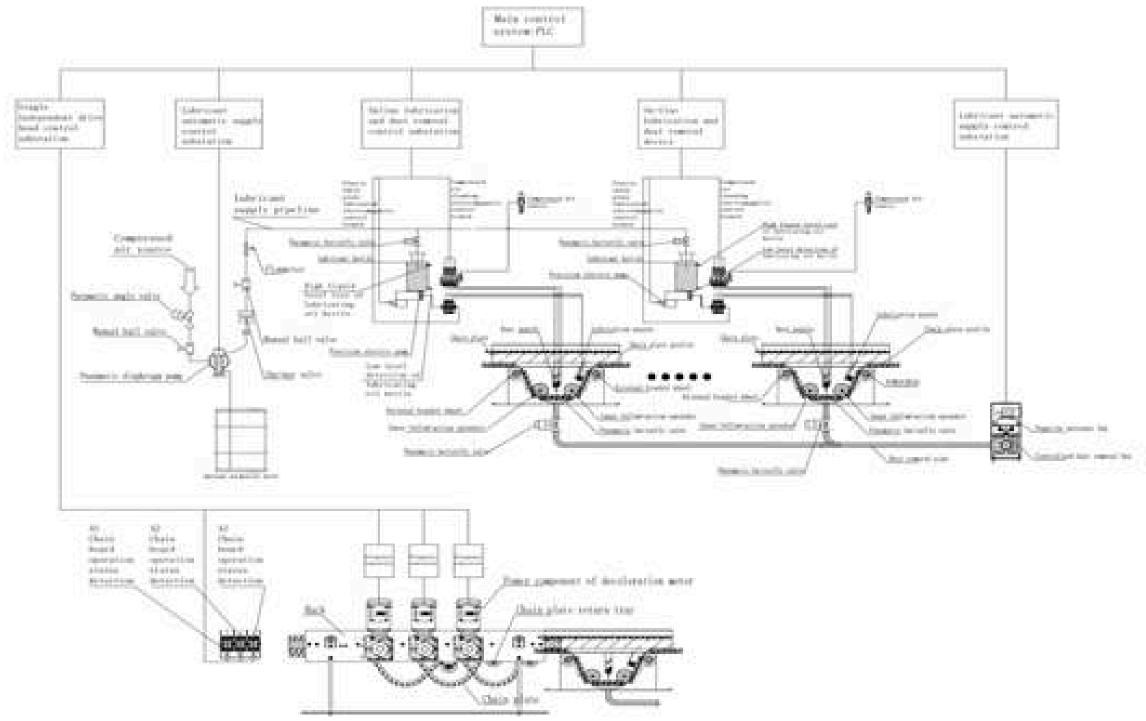


Fig. 1.1: The overall design of the system

air dedusting control solenoid valve group, the lubricating oil circuit control solenoid valve group, the dust blowing nozzle, the outer supporting roller 01, the outer supporting roller 02, the inner supporting roller 01, the inner supporting roller 02, the centralized dust hopper, the lubricating nozzle, and the online lubrication and dedusting control substation.

The working principle of the online lubrication and the dedusting device is to change the chain plate of this section from the original straight section to a U-shaped track by reasonably arranging the outer supporting roller 01, outer supporting roller 02, inner driven sprocket 01, and inner driven sprocket 02. The purpose is to avoid the installation position of the dust-blowing nozzle and the lubricating nozzle and to facilitate daily maintenance and overhaul. The internally driven wheels 01 and 02 adopt the sprocket structure. The advantage is that the sprocket engages with the inner side of the chain plate so that the chain plate will not deviate from left to right during operation, which will lead to failure during the U-shaped track return. The lubricating oil pot is equipped with an oil pot high liquid level detection sensor and an oil pot low liquid level detection sensor. When the oil pot is at a low liquid level, the system automatically starts the automatic lubricating oil supply device to send lubricating oil to the high liquid level alarm, and the supply is suspended. The single-channel compressed air dedusting control solenoid valve group starts the solenoid to connect the compressed air source to purge and dedust the inner side of the chain plate when the system needs single or multiple-channel chain plate dedusting. At the same time, the centralized dust suction hopper directly below is used to collect the dust purged to avoid secondary contamination and clean the dust expressed by the chain plate. The lubricating oil circuit control solenoid valve group starts the solenoid to switch on the precision electric pump when the system needs to add lubrication in a single way. Then the precise electric pump will deliver the oil to the lubricating nozzle at a constant speed in a fixed amount to lubricate the chain plate.

In this paper, an intelligent rod cigarette conveying chain plate resistance reduction system is designed to reduce the frictional force between the system channel and the chain plate to provide a simple structure, easy

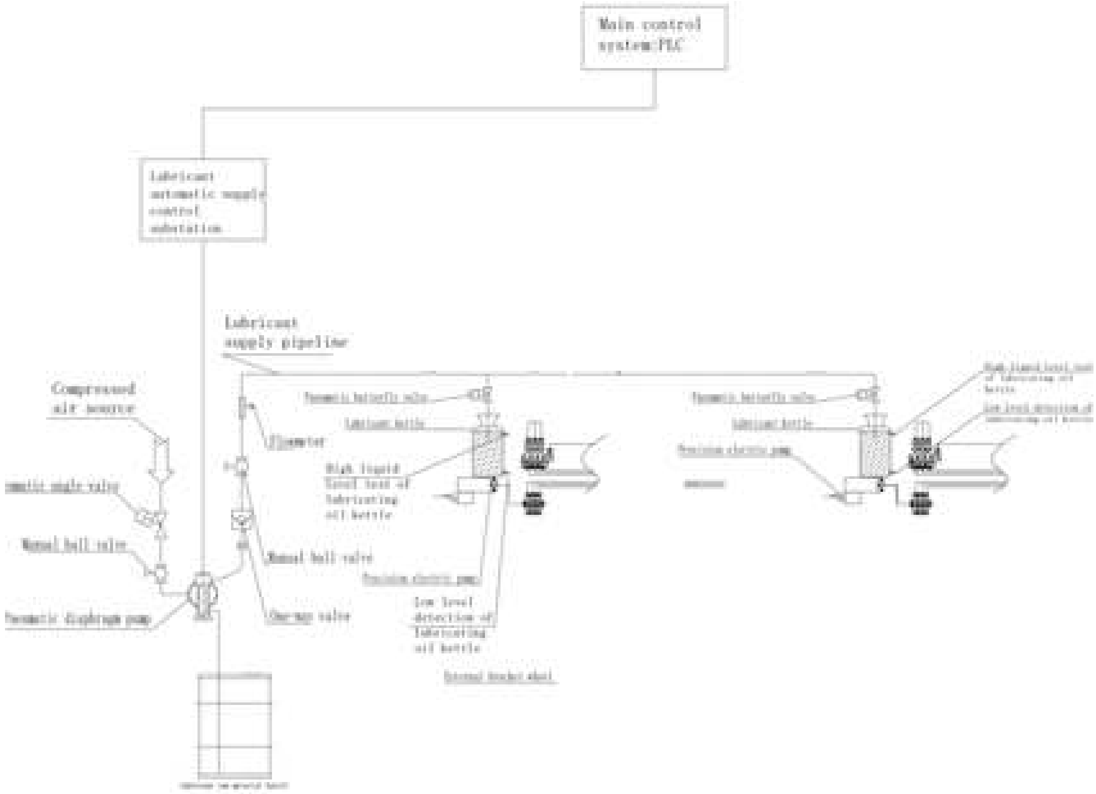


Fig. 2.1: The automatic lubricating oil supply device

to use, and low-cost generation method for the cigarette conveying generation system. Improve the tobacco bar conveying system's production efficiency and chain plate life.

2. The design of the system.

2.1. Process design drawing. The automatic lubricating oil supply device is mainly composed of lubricating oil raw material barrel, pneumatic diaphragm pump, flowmeter, manual ball valve, one-way valve, pneumatic angle valve, lubricating oil supply pipeline, pneumatic ball valve, lubricating oil pot, automatic lubricating oil supply control substation, etc.

For the control process of the resistance reduction system of the smart cigarette conveyor chain plate, it removes the dust one by one according to the number of channels of the conveyor line, and the time for dust removal can also be set according to the length of the conveyor line. The process flow is shown in Figure 2.1. In addition, before lubricating the conveyor chain plate, the chain plate needs to be dusted. After the dust inside and outside the chain plate is blown clean, a reasonable time can be set according to the length of the chain plate. Since the total amount of lubricating oil is constant, different time needs to be set for each conveying channel to lubricate the chain plates. Under normal circumstances, the drag reduction system of the smart bar conveying chain plate runs for one day for dust removal, while for the chain plate, it runs for a lubricating operation after two days. In addition, the chain plate resistance reduction system is provided with a man-machine interface, and the operator can switch the working mode of the system, that is, the manual mode or the automatic mode, through the button.

The working principle of the automatic lubricating oil supply device is that when the main control system feeds back the low liquid level signal of the lubricating oil pot of this station from an online lubricating and dedusting control substation, the pneumatic ball valve of the branch of this station will be opened automatically, and the pneumatic diaphragm pump will automatically supply the lubricating oil through the pipeline to the oil pot until the high liquid level signal is detected. The flowmeter set at the outlet of the pneumatic diaphragm pump is the limit protection value of the equipment when adding lubrication each time because the volume between the high liquid level and the low liquid level of the lubricating oil pot is determined. When the refueling exceeds the set value and a high liquid level alarm is received, the system judges the high liquid level meter fault alarm to avoid pollution and waste caused by the overflow of lubricating oil; The flowmeter also has the function of accurate measurement and statistics, facilitating the lean management of the workshop. The manual valve is set to facilitate daily inspection, the pneumatic angle valve controls the start and stop air source of the pneumatic diaphragm pump, and the one-way valve is set to prevent leakage caused by oil circuit backflow.

The figure depicts various components of the chain plate resistance reduction system. The dust removal and lubricating oil supply assembly is represented by the numeral 1, the chain plate lubrication assembly by 5, the oil pot by 12, the electric plunger pump by 13, the dust removal solenoid valve by 14, the oil supply solenoid valve by 15, and the inner chain plate fan knife dust removal assembly by 41, and the outer chain plate air knife dust removal assembly by 42.

The specific process involved in the operation of the system is as follows: The system targets the power of a single motor on each conveyor line and is equipped with a dust removal air knife and nozzle on both the inner and outer sides of the power head. The system then adjusts the working time of the resistance reduction system device by collecting the motor current signal of each channel. If the collected current signal falls within the normal range, the conveyor chain plate is dusted first, and the dedusting time can be set freely. After three cycles of normal operation, the system sprays lubricating oil on the chain plate.

If the motor current changes for a specific conveying channel exceeds the normal range, the dust removal operation will be carried out for two cycles initially. If the current change does not return to normal, the system will start spraying lubricating oil until the current returns to the normal range. If the system's current does not decrease at this point, it indicates that an obstacle is blocking the system, and the system will issue an alarm.

In summary, the figure shows the components of the chain plate resistance reduction system and the specific process involved in its operation, which includes dust removal, lubrication, and monitoring of the motor current signals to ensure efficient and reliable conveyor operation

2.2. Design of supply components. The schematic diagram of the new dust removal and lubricating oil supply components is shown in Figure 2.1. In figure, 11 is the installation base plate, 121 is the high oil level detection sensor, and 122 is the low oil level detection sensor. The supply assembly part of this paper is provided with a high oil level detection sensor at the upper end of the oil pot and a low oil level detection sensor at the lower end. The oil level sensor uses an ultrasonic sensor, and the low oil level detection sensor mainly detects when the lubricating oil in the oil pot is consumed to the maximum liquid level, and issues an alarm to prompt to add lubricating oil.

2.3. Design of conveyor chain plate device. The diagram of the conveyor chain plate device is shown in Figure 2.3. In the figure, 2 is the chain plate of the conveyor line, 31 is the outer supporting wheel assembly, 32 is the inner supporting wheel group, 41 is the air knife dust removal assembly of the inner chain plate, and 42 is the air knife dust removal assembly of the outer chain plate. The lubricating component 5 in the chain plate device is composed of a grease nozzle mounting frame and a grease nozzle nozzle, and the specific components are a grease nozzle nozzle, an oil inlet nozzle, and a grease nozzle bracket, the number of which is 6 respectively. In addition, the air knife dust removal assembly of the chain plate includes the outer chain plate air knife dust removal assembly and the inner chain plate air knife assembly, and inside the chain plate device, the lubricating component is installed inside it, and its transmission line channel is based on the oil nozzle. Set the sprinkler head. Therefore, the length of each conveying channel is different because the conveying line is affected by its position.

2.4. Design of the dust removal component of the outer chain plate. The outer dust removal air knife assembly of the chain plate device is shown in Figure 2.4. In the figure, 421 is the outer air duct mounting frame, 422 is the outer air knife, and 423 is the outer air inlet connector. The inner chain plate dust removal air knife assembly mainly includes the inner air duct mounting frame, the inner air knife and the inner air inlet joint. Among them, the dust removal solenoid valve controls the air source output of the inner dust removal air

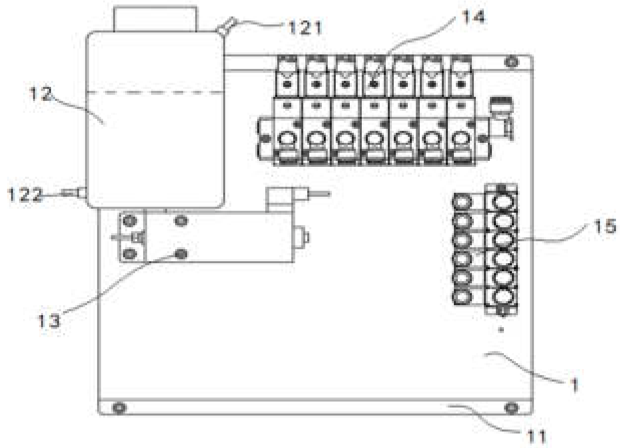


Fig. 2.2: Partial schematic diagram of dust removal and lubricating oil supply components

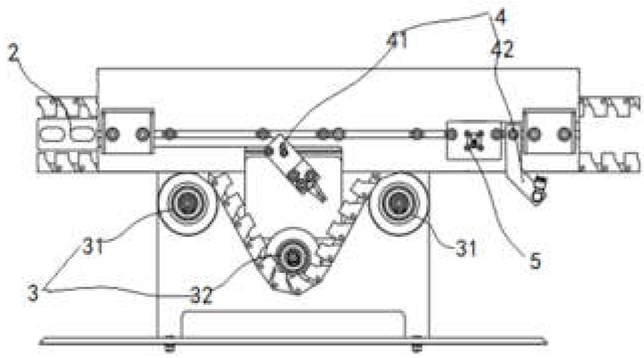


Fig. 2.3: Partial schematic diagram of conveyor chain plate device

knife and the chain plate, and the outer dust removal air knife is controlled by a general dust removal solenoid valve. Because the inner side of the chain plate contains a lot of dust, the resistance of the chain plate to run increases. For this reason, the air knife dust removal component on the inner side is equipped with an air knife according to the channel of the conveying line. In the same way, the inner air knife is also set with a corresponding solenoid valve according to the dust removal component to remove the outer dust and set an air intake air source on the outer chain plate to remove dust from all the conveying channels.

2.5. Design of the dust removal component of the inner chain plate. The schematic diagram of the dust removal assembly of the inner chain plate is shown in Figure 2.5. The standard width of the designed chain plate is 75mm, and the width of the air knife is 62mm. The material is made of ABS plastic material to obtain the corresponding shape by injection molding, and 16 return grooves are set between the ribs. The notch is output, thereby forming a compressed air flow, which covers the inner side of the entire chain plate. When the air knife assembly of the inner chain plate removes dust, the inner side contains a lot of dust, which increases the running resistance of the chain plate. Therefore, setting air knives at the channel openings of each conveying line can speed up the outflow of dust.

2.6. Design of the controller. The main control unit adopts PLC, which has a powerful DP function. Since there are many slave stations, 414-2DP, which requires a powerful DP function, is selected as the control core. At the same time, it is equipped with two digital input modules 421, each with 32 There are 32 switch

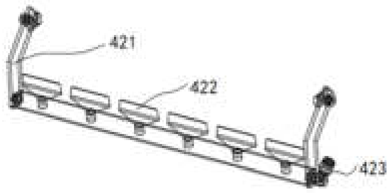


Fig. 2.4: Schematic diagram of the dust removal assembly of the outer chain plate

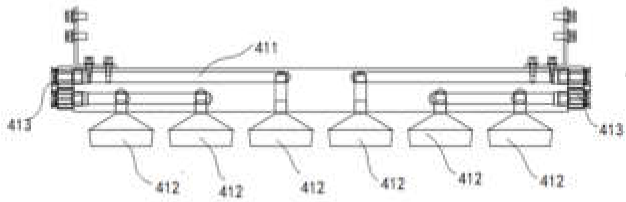


Fig. 2.5: Schematic diagram of the dust removal assembly of the inner chain plate

quantity output points in each block of 422 and 32 switch quantity input points. This CPU has various functions such as large-capacity timers and counters, fast processing speed, self-diagnosis and self-protection, and scalability, which are sufficient to meet the technological and functional requirements of the system. Because there are many slave stations, an IM467 with integrated DP processing function is selected as the master station. This module is selected as another master station, because it has more powerful DP function, which makes the system have stronger processing and coordination ability for DP network.

2.7. Design of online lubrication system of plastic chain plate. At present, the online lubrication system of the plastic chain plate adjusted to the tobacco conveying system of the tobacco industry has online automatic detection and comprehensive analysis of the working resistance of the chain plate. The abnormal operation of the chain plate is mainly manifested by stumbling, creeping, breaking, and other phenomena. When the chain plate running condition detection device detects that the pulse data has a relatively large peak fluctuation, the chain plate is tripping. or crawling. When the pulse is continuous, this indicates that the chain plate has broken and stopped. Standard dust removal methods are to manually blow the dust scattered on the surface of the chain plate with a compressed air gun on the surface of the tobacco transmission line and transmit it to a long distance. Generally, the driving head of the transmission line is placed in many parts. Only the segmented chain plate can be manually cleaned to reduce the working resistance of the chain plate. Is available and the cleaned dust is not recycled centrally, causing secondary pollution.

In addition, it has to be done manually. When adding lubricating oil, the method of manual acquisition cannot precisely control the amount of use, and it does not rub the inside of the chain plate evenly; when it is rubbed too much, it can easily penetrate the surface of the chain plate and contaminate the tobacco products being transported, and when there is dust on the chain plate, the lubricating oil will stick to the dust well. , sticks to the chain plate and increases the working resistance of the chain plate. Work productivity is low, and the labor intensity is high. Online Plastic Chain Plate Lubrication System is an integrated system for plastic chain plates and application scenarios in cigar conveyor lines. The plastic chain plate online lubrication system mainly consists of several sets of online lubrication and dust removal devices and their control substations, automatic lubrication oil supply devices, and their control substations, centralized dust collection devices and their control substations, single-channel independent driving head and their control substations, main control system, and other components.

2.8. Design of line lubrication and dust removal device. The working principle of the centralized dust collection device is that multiple sets of online lubrication and dust removal devices set in the plastic chain plate online lubrication system are connected in series through the dust removal pipeline, and each online lubrication and dust removal device is set with a centralized dust hopper and a branch pneumatic butterfly valve. When the system starts the dust removal function of the online lubrication and dust removal device, the negative pressure fan is also started at the same time, which will purge the dust through the centralized dust suction hopper, collect the dust purged to avoid secondary pollution and also clean the dust expressed by the chain plate.

In-line lubrication and dust removal equipment are mainly composed of a lubrication oil tank, oil tank high fluid level detection sensor, oil tank low fluid level detection sensor, precision electric pump, single-channel compressed air dust removal control solenoid valve group, and lubrication oil circuit. Control solenoid valve group, blower. Dust nozzle, outer support shaft 01, outer support shaft 02, inner support shaft 01, inner support shaft 02, centralized dust bucket, lubrication oil nozzle, online lubrication, and dust control substation. The working principle of the online lubrication and dust removal device is to change the chain plate of this part from the original straight part to a U-shaped path by placing the outer support roller 01, outer support roller 02, and internal drive gear in a reasonable way. 01, the internal drive gear 02. , its purpose is to avoid the installation position of dust-blowing nozzle and lubrication nozzle, to facilitate daily maintenance and repair. The advantage of using the gear structure of internal drive wheels 01 and 02 is that the chain plate meshes with the inside of the chain so that the chain plate does not tilt from left to right during use, which can cause damage during the return of the U-shaped trajectory. The lubricating oil tank has a high oil tank fluid level detection sensor and a low oil tank fluid level detection sensor. When the oil tank fluid level is low, the system automatically activates the lubrication oil supply device. Will raise the oil to a high fluid level, give an alarm and temporarily stop the supply. A group of single-channel compressed air dust control solenoid valves is used in the system to clean single-channel or multi-channel chain plates from dust, initiate an electromagnetic connection to the compressed air source, clean the inside of the chain plate, and remove dust., and at the same time a centralized dust hopper directly below. A vacuum cleaner collects the blown dust and cleans the dust the chain plate represents to prevent secondary pollution. The lubricating oil circuit control solenoid valve group is used to start the electromagnetic connection of the precision electric pump when it is necessary to add lubricant to one channel in the system, and the precision electric pump is transported and charged evenly and quantitatively to the lubricating oil nozzle. chain plate.

2.9. Design of automatic lubricating oil supply device. Automatic lubricating oil supply equipment usually consists of a lubricating oil raw material barrel, gas diaphragm pump, flow meter, manual ball valve, check valve, gas angle valve, lubricating oil pipeline, gas ball valve, lubricating oil tank, and automatic lubricating oil device. . components such as supply control substations are shown in Figure 2.6. The working principle of the automatic lubricating oil supply device is that when the main control system returns the low-level signal of the station lubricating oil tank, the pneumatic ball valve of the station branch will automatically open. The gas diaphragm pump will automatically supply the lubricating oil to the lubricator through the pipeline until the high fluid level signal is detected. Adjusting the flow meter at the outlet of the gas diaphragm pump determines the volume between the high fluid level and the low fluid level of the lubricant, so it is the safety limit value of the equipment each time the oil is added. level alarm, the system filters high-level gauge faults and alarms, avoiding contamination and waste from lubricant overflows; the flow meter performs accurate measurement and statistical functions, which is convenient for workshop management. The manual valve is configured to facilitate daily inspection, the gas angle valve is used to start and stop the air source of the gas diaphragm pump, and the check valve prevents the oil circuit from flowing back and leaking.

2.10. Design of single-channel independent drive head. The working principle of the single independent driving head is to arrange the power components of the reduction motor one by one according to the conveying channel, which is used to drive the chain plate to move forward, and the return idler of the chain plate is used to protect the chain plate from entering the profile smoothly. When the driving head drives the chain plate, the normal running state of the chain plate is the running state of the chain plate. The detection device detects that the pulse data of the chain plate is relatively uniform. The abnormal operation of the chain plate is mainly manifested by bumping, creeping and breaking. When the pulse data of the chain plate

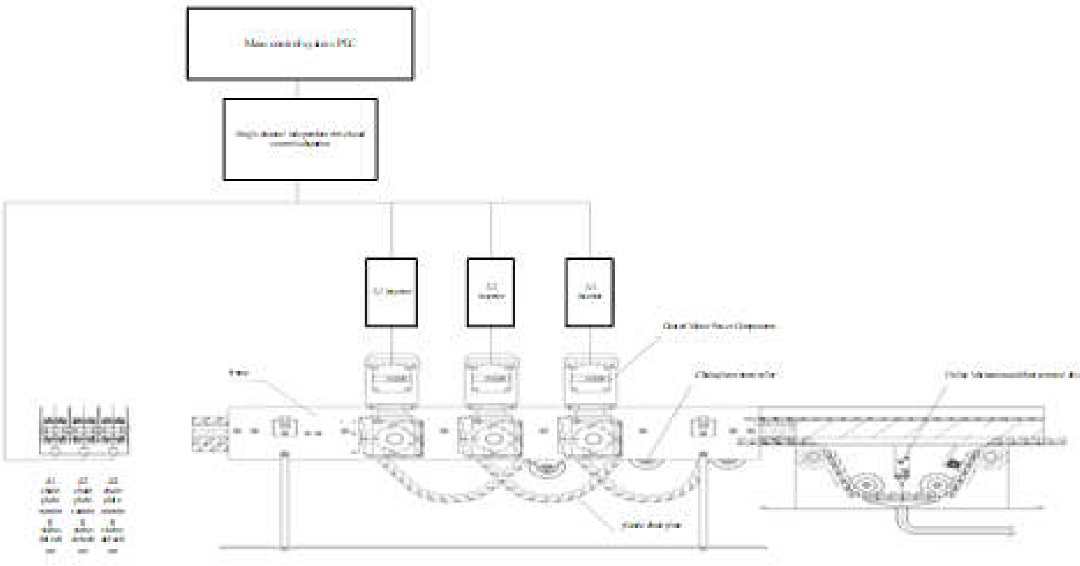


Fig. 2.6: Single-channel independent drive head

detected by the chain plate operation state detection device is relatively large, it indicates that the chain plate has to bump or to creep. When the pulse continues to output, the chain plate is broken and stopped. The single-channel independent drive head control substation suspends the reducer through the frequency converter and gives an alarm.

The single-channel independent drive head comprises gear motor power components, frequency converters, chain plates, racks, return rollers, chain belt running detection components, and single-channel independent drive head control substations. 9. The working principle of the single-channel independent drive head is to arrange the gear motor feed components one by one along the transmission channel, which is used to drive the chain plate forward, and the chain plate is used to protect the return roller. Prevent the chain plate from entering the profile smoothly. When the drive head operates the chain plate, the normal state of the chain plate, the operating state detection device detects that the pulse data of the chain plate is relatively uniform. The abnormal operation of the chain plate is mainly manifested by stumbling, creeping, breaking and other phenomena. When the chain plate running condition detection device detects that the pulse data of the chain plate has a relatively large peak fluctuation, it means that the chain plate is tripping. or crawling. When the pulse is continuous, it means that the chain plate has stopped, and the single-channel independent drive head control substation suspends the reducer through the frequency converter and gives an alarm.

3. Result Summary. The research aimed to solve the problem of high frictional resistance between the chain plate and guide rail of the cigarette conveying system. An intelligent drag reduction system with lubrication and automatic dust removal was designed, which consisted of several components, including a conveyor chain plate, supporting roller assembly, air knife dust removal assembly, chain plate lubrication assembly, and controller.

Experimental tests were conducted on the system, and the results showed that it effectively reduced the frictional resistance of the chain plate of the cigarette conveying line. The system achieved this by reducing the amount of dust on the chain plate through the use of air knife dust removal assemblies and lubricating the chain plate through the use of lubrication assemblies. The controller controlled the electromagnetic valve to achieve this.

The research also discussed the working principle of the single independent driving head, which was used to drive the chain plate forward and protect it from entering the profile smoothly.

The paper also presented a system model that provided a detailed description of the various components of the system and how they work together to reduce frictional resistance.

The research successfully developed an intelligent drag reduction system with lubrication and automatic dust removal that effectively reduced the frictional resistance of the chain plate in the cigarette conveying line. The results of the experimental tests provided evidence of the system's effectiveness, and the detailed system model provided a clear understanding of how the various components worked together to achieve this.

4. Conclusions. In this paper, based on the actual needs of workshop production and process, the intelligent rod cigarette transportation resistance reduction system is developed so that the conveyor chain plate reduces the resistance between components in sorting cigarettes and enhances the working life of the components. And improve the generation efficiency. The system has the characteristics of a simple structure, stable operation, safety, and reliability. , When there is a fault in the process of conveying the cigarette, an alarm message can be issued in time, and the controller is used to automatically control the air knife dust removal component and the storage capacity of the lubricating oil to reduce the dust and remove the lubricating oil in time. The system improves the effective operation rate of the overall production equipment and enhances the quality monitoring system of the production operation environment.

REFERENCES

[1] M. Abdel-Basset, G. Manogaran, and M. Mohamed, *Internet of things (iot) and its impact on supply chain: A framework for building smart, secure and efficient systems*, Future generation computer systems, 86 (2018), pp. 614–628.
[2] L. Cheng, Z.-G. Hou, M. Tan, and W.-J. Zhang, *Tracking control of a closed-chain five-bar robot with two degrees of freedom by integration of an approximation-based approach and mechanical design*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42 (2012), pp. 1470–1479.
[3] B. D. Chung, S. I. Kim, and J. S. Lee, *Dynamic supply chain design and operations plan for connected smart factories with additive manufacturing*, Applied sciences, 8 (2018), p. 583.
[4] A. Kothapalli, A. Miller, and N. Borisov, *Smartcast: An incentive compatible consensus protocol using smart contracts*, in Financial Cryptography and Data Security: FC 2017 International Workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21, Springer, 2017, pp. 536–552.
[5] M. Kumar, G. Graham, P. Hennelly, and J. Srai, *How will smart city production systems transform supply chain design: a product-level investigation*, International Journal of Production Research, 54 (2016), pp. 7181–7192.
[6] Y.-H. Lim, S. V. Gopinathan, V. V. Varadan, and V. K. Varadan, *Finite element simulation of smart structures using an optimal output feedback controller for vibration and noise control*, Smart materials and structures, 8 (1999), p. 324.
[7] O. Maslák and I. Andrisková, *Labeling options for postal shipments by passive rfid technology*, in CBU International Conference Proceedings, vol. 3, 2015, pp. 475–481.
[8] D. Schafroth, S. Bouabdallah, C. Bermes, and R. Siegwart, *From the test benches to the first prototype of the mufly micro helicopter*, Journal of Intelligent and Robotic Systems, 54 (2009), pp. 245–260.
[9] Y. Sun, Q. Guo, S. Zhao, K. Chandran, and G. Fathima, *Context-aware augmented reality using human–computer interaction models*, Journal of Control and Decision, (2022), pp. 1–14.
[10] B. Y. Tsuge and J. M. McCarthy, *Synthesis of a 10-bar linkage to guide the gait cycle of the human leg*, in International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, vol. 57137, American Society of Mechanical Engineers, 2015, p. V05BT08A083.
[11] S. Wang, S. Quek, and K. Ang, *Vibration control of smart piezoelectric composite plates*, Smart materials and Structures, 10 (2001), p. 637.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Apr 11, 2023
Accepted: May 21, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 1087–1096, DOI 10.12694/scpe.v24i4.2383

MULTI-THREADED DATA COMMUNICATION IN JAVA FOR ADVANCED COMPUTING ENVIRONMENTS

JINYING WANG* AND JING YANG†

Abstract. The performance of operating systems like computers requires the proper functioning of the computer language interpreter. This interpreter follows various types of programming languages that make the performing of computer programming easy and effective. The application of the programming language of Java helps in processing multiple tasks at once. This research analyzes the novelty of Java-based data communication models in advanced computing services. This performance makes the saving of the resources used for the development of the programming language. All these development includes the performing of the multiple threads communication data processing. These multiple threads help distribute the single processed input in the multi-channel language processing, thus helping the work competition in time. It also reduces the cost of maintaining the programming languages reduced. Thus, the implementing cost of resources required for programming performance is reduced. Therefore, this implementation impacts the programmer to become more indented to use the language transformation process of Java. Moreover, it creates a more effective representation of the audio or visual content represented by a multi-tasking operating system. In this process of development in the language transformation of the operating system, the ability of the operating system for data processing improves. The systematic process of this language transformation helps in systematically transforming multiple programs at once.

Key words: Multi-thread, data communication, Operating system, programming language, Data processing

1. Introduction. The transformation of input language in the computer requires a conversion of ordinary language to binary computer language. It requires the making of the transformation of data from ordinary language to computer language to examine the information. The languages used for the transformation of information are called computer languages. These languages are JAVA, C++, C, python and so on. This language helps run the inputted command into the computer to process those commands in the operating system. These languages are the group of symbols and rules to translate the algorithmic language to execute those programs through the coding language in the computer. This programming language helps to create a communication bridge between the computer operating system and humans. Multi-threading languages help in the making of multiple programs processed through a single systematic way of processing multiple programming codes sequentially. This helps in making arranging all the input programming codes through a multi-channel transformation of language to binary language. The application of the multi-threading languages of JAVA allows for processing the programming languages through the multi-channel operating language processor.

A significant contribution of this research lies in its exploration of a novel Java-based data communication model within the realm of advanced computing services. This model's unique characteristics hold the potential to revolutionize the way resources are conserved during programming language development. Including multi-threaded communication, and data processing adds to the complexity of the development process, yet it offers substantial benefits. Multiple threads for data communication enable the efficient distribution of singularly processed inputs across a multi-channel language processing environment. This distribution mechanism contributes to the timely completion of tasks, promoting efficient work competition and aiding in meeting deadlines. Furthermore, the research posits that this approach can effectively lower the cost associated with maintaining programming languages as the workload becomes more streamlined and resource-efficient.

2. Objectives. This study properly identifies and describes some of the basic objectives. This includes the basic concept of the technology used to improve computer data programming languages. In the data communication between the operating systems and the human operations is processed by the application of

*Qinhuangdao Vocational and Technical College, Qinghuangdao, China, 066100
†Qinhuangdao Vocational and Technical College, Qinghuangdao, China, 066100 (jingyangres1@outlook.com)

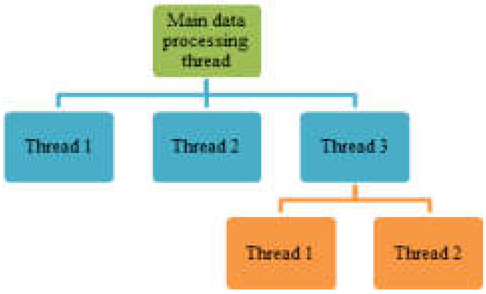


Fig. 3.1: Multithreading data processing language

coding languages in the operating system. This includes the processing of different types of multi-channel programming languages like the application of JAVA for improving communication in computer operations. This helps in the development of the operating environment of the computer user in the making of data processing more easily and quickly. Some of the objectives of the technologies of multi-channel JAVA operating languages in the communication between computers and humans are as follows:

1. To elaborate on the concept of Multithreading data processing language
2. To state the impact of Java multithreading in the computing environment
3. To identify the side effects of computing languages for making communication between computers and human
4. To describe some of the remedies for the issues of data processing languages in computer
5. To examine the impact of multithreading communicating language technology in making the ease of data processing through the operating system
6. To identify the factors of Java multithreading language for setting a systematic data processing language

3. Methodology. In information-based data collection, the programming languages used for the data processing are collected from the detailed observation process. This includes gathering information from the newspaper, articles, and journals based on the coding language information. It represents all the information about the programmers' multi-channel or multi-threading languages. This procedure is used for running many inputted data in a systematic and fluent data processing process. Thus examining all the results of operating through multi-channel data processing by Java sets the easiest way of setting communicating programming language.

Multithreading Data Processing Language. As shown in figure 3.1 the programming language helps to create a communication bridge between the computer operating system and humans. Multi-threading languages help process multiple programs through a single systematic way of processing multiple programming codes sequentially [9]. This helps in arranging all the input programming codes through a multi-channel transformation of language to binary language. The distribution of complex data through multiple programs is processed through a single systematic way of processing multiple programming codes. This helps in the execution of those programs through the multiple processing threads in the computer [15].

This also helps in the making of the representation of the allocated task to be performed by numerous treading in the programming process. This helps in reflecting the final output effectively as early as possible by developing the ability of the operating system to perform the programming task systematically by observing all the circumstances of the language programming and the needs of the programmers.

Therefore it showed that the processing of single coding used in the data processing helps in the making systematic processing of systematic data In an operating system as shown in the above table 3.1. This multiple threading helps in making the capability of the OS interact with multiple users at a single time. This requires numerous inputs from the coding programs running in the operating computer system [5]. It also maintenance of the numerous users interacting in programming at a single time without interpreting any one of them. Therefore

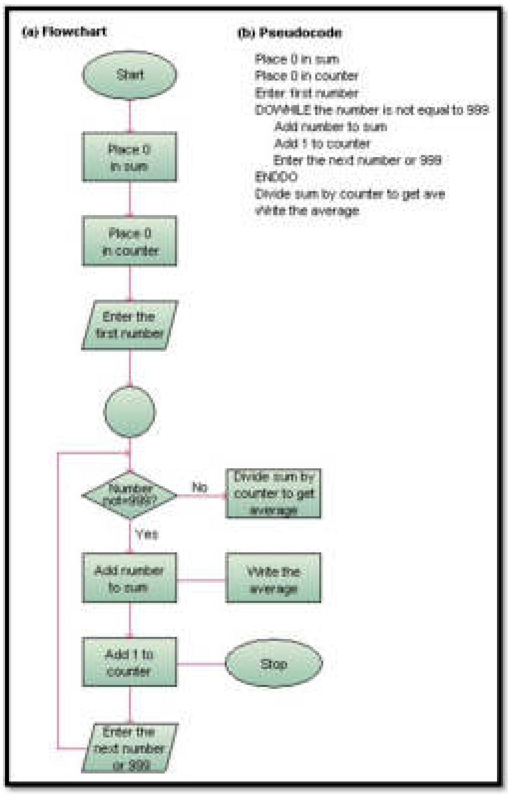


Fig. 3.2: Multithreading computer programming algorithm

Table 3.1: Function and formation of multiple threading

Multithreading	
Concept	Creates a multiple spreading of single data processing through multiple coding channels to increase the calculation ability of the operating system
Function	Distributes all the single-channel programming codes systematically
Formation	Generates threads by observing all the consequences of Input languages

the application of the programming multiple threads of the programming languages makes the development of the programming languages and the programming mediums through the utilization of minimum resources and time.

4. Effect of Multi-threading Communication Language Technology. The application of multiple threading makes the reduction of the possessions required for the calculation and transformation of coding languages as pointed out in Figure 3.2. This also increases the increase of efficiency of the coding transformation. This processes all the threads of the input and processes them in a single processed data operation. It handles the entire user's input at a single time without interrupting any of the users [6]. The performance of the multiple threads in the operating of the programming languages makes the distribution of difficult data through multiple programs processed through a single systematic way of processing multiple programming codes. This helps in the execution of those programs through the multiple processing threads in the computer.

Therefore, it makes the processing of multiple inputs processed by a coding language at a single operation. As shown in table 4.1 this basically increases the operating system's speed and efficiency to provide the best

Table 4.1: Benefits of multiple threading

Benefits of Multiple Threading	Development of the language processing performance Reduction of possessions used in the programming Smooth access to multiple programming applications Simplifying the structures of the programming operation

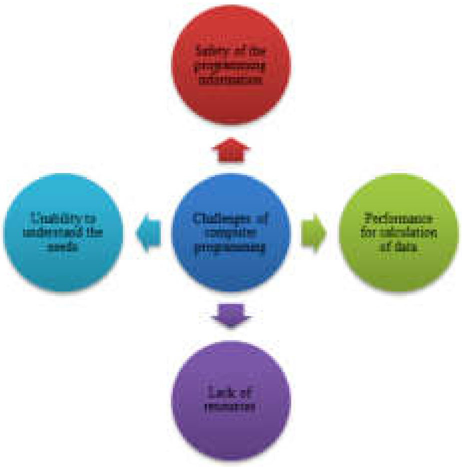


Fig. 5.1: Challenges in computer programming language for data processing

performance by the inputted coding. Its main function is to distribute the single threading task into multiple channeling for better execution of the task quickly and efficiently [10, 14]. It also includes the quick transformation of processes obstructed by the blocked threads to the running threads and allows the users to interact with the programming easily. The development programming structure for the execution of the communication between computer programming and human development helps in the making of smooth programming operations to be performed by the programmers [6]. The most important benefit is that the performance of the single complex language transformation helps in performing the execution through various threads. Thus, the task gets performed quickly by making distribution of them through multiple channels of language transformation.

5. Challenges in Computer Programming Language for Data Processing. The application of programming languages helps in the computation of the input data in the operating systems easily. It also faces many issues in the functioning of the data processing through arrays. This also becomes one of the hardest obstructions to the programmers performing the performing task as shown in Figure 5.1. It requires the starting resource collection for the running of the data smoothly allowing the programmers need to examine the needs of the program [3]. This includes a lack of knowledge about algorithm programming. This issue affects the completion of the programs within the allocated time and as per the requirements

Also increases the security issues generated by the programmers as it gets created by the software programming language. The performance of the programmers through the programming makes the development programming structure for the execution of the communication between computer programming and human development. Also helps in the making of smooth programming operations to be performed by the programmers.

6. Mitigation of Data Processing Issues through Java Multi-threading. The applications of Java multiple threads help to perform numerous programming operations in a single processing of data through programming language as the threads are easily accessible as pointed out in Figure 6.1. This includes the simplifying of programming execution at the time of inputting the information in the programming [2, 12]. This also develops the difficulties faced by the language interpreter thus making the reduction of the production and

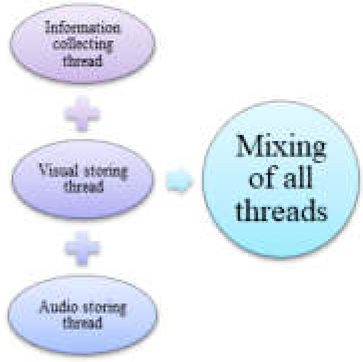


Fig. 6.1: Application of Java multi-threading

Table 6.1: Advantages of using multiple threading in Java

Advantages of using Multiple threading in Java	Performing numerous tasks at once Reduces the maintenance cost of programming Increases efficiency of programming Develops the performance of the operating system Completes allocated programming tasks on time

maintenance cost. This all makes the improvisation of the performance of the operating system.

The quick transformation of processes obstructed by the blocked threads to the running threads allows the users to interact with the programming easily as represented in Table 6.1. The development programming structure for the execution of the communication between computer programming and human development helps in the making of smooth programming operations to be performed by the programmers. As the structure develops the maintenance cost and the complexity decrease [7]. Also, increase the efficiency of the programmers during the execution and making of the language processor. Thus the successful multiple-thread communication created between computer programming and the human gets the efficiency of providing the best outcome by the programming in limited time and in limited utilization of resources.

7. Impact of Java multi-threading in the computing environment. The application of multiple threads makes the development of numerous software applications. This helps in the performing of the programming task to be implemented in the numerous sources as shown in the above figure 7.1. This source of data distribution is known as data threads. It helps to display various types of content such as visuals of animated cartoons, audio recordings, or displaying of a video [15].

This performance of the programming through the different threads makes the performance of different programming tasks required for different purposes at a single time. Figure 7.2 shows multithreading operations. Thus, the programmers save the time of making the development in a programming language. The programmer also provides the necessary information to the consumers and this makes them get satisfied with the final outcome.

As shown in table 7.1 it represents the benefits of implementing the Java multiple threading processes in the programming. This process of the programming language makes the communication model for the development of the data calculation in the operating system. The operating system of the computer CPU gets the improvement in the processing of the input languages for processing the content required by the consumer [1].

This figure 7.3 mainly gives benefits to the programmers for making the development in the execution process of the programming language this multiple threading of advanced Java computation increases their efficiency. This also makes the programmers able to supply all different types of content in a single operation.

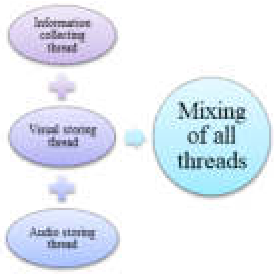


Fig. 7.1: Impact of Java multi-threading in the computing environment

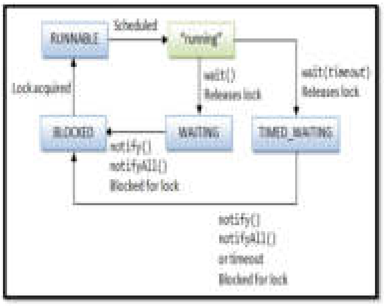


Fig. 7.2: Java multi-threading in the computing environment algorithm

Table 7.1: Impact of Java Multi-threading in the computing environment

Impact of Java Multi-threading in the computing environment	
Operating system software	Programming of Word document Programming animated content Programming the video content Increasing the efficiency of the programmers

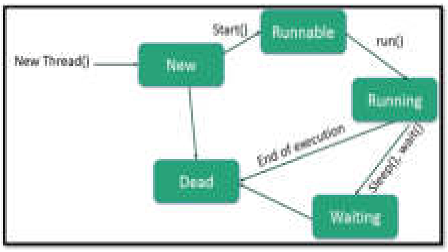


Fig. 7.3: Multiple threading in Java

Therefore it gives the opportunity to perform multitasking efficiently in a short period of time. The difficulties of the programming task also get solved by the application of Java programming computing language.

8. Factors of Java Multithreading Language. The application of Java programming language helps in increasing the capability of a program. The operating system allows multiple users to use the programmed software application at a single time despite representing multiple copies of the programs at the time of running

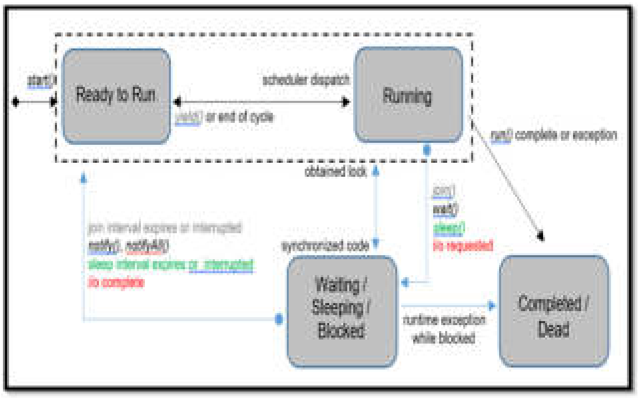


Fig. 8.1: Threads states in Java

the operating software. This includes the functioning of the threads in the multitasking process of the programming language [8]. The factors for the development of Java give benefits to the programmers for making the development in the execution process of the programming language this multiple threading of advanced Java computation increases their efficiency.

This performance of the programming through the different threads makes the performance of different programming tasks required for different purposes at a single time. Thus, the programmers save the time of making the development in a programming language [1]. Various threads and their state of action are presented in figure 8.1. This represents all the information about the multi-channel or multi-threading languages used by the programmers of Java. This procedure is used for running many inputted data in a systematic and fluent data processing process. Thus, examining all the results of operating through multi-channel data processing by Java sets the easiest way of communicating programming language

9. Problem Statement. The programming procedures in the making of development of software-based programmers make the improvement in the functioning of tasks by programming languages. The obstruction in the performing of the task through the computer-based operating system stops the success of the programming language. This study introduced the performance of Java multiple threading programming through the different threads making the performance of different programming tasks required for different purposes at a single time [11]. Thus the programmers save the time of making the development in a programming language. The programmer also provides the necessary information to the consumers and this makes them get satisfied with the final outcome. This helps in solving all the issues. It gives the opportunity of performing multitasking efficiently in a short period of time [12]. The difficulties of the programming task are also solved by the application of the Java programming computing language. This introduction of the multiple thread application in the programming process of Java makes the development of the procedures required for constructing the data communication by the advanced techniques of Java.

10. Result. This research paper delivers an overall concept of the multithreading data processing language and evaluates the side effects of the computing environments for providing proper communication between humans and computers. Parallelization techniques have been determined with the help of the hierarchical hybrid parallel that is helping to implement quad-core professor [4]. There are four different automatic parallelization techniques included in this research study those are included Cetus, Par4all, and S2PMOACC, S2P. Those tools play a crucial role in delivering efficiency of every tool [4].

The above figure 10.1, 10.2, and 10.3 represent the market size and serial matrix multiplication has been analyzed with the help of the various sizes. MPI process has been done to evaluate the hybrid parallel code among hybrid and serial versions regarding execution time. All the graphs denote the improvement of the executing time for the overall matrix [4]. In the 1000x1000 matrix size, the database and single GPU are running together and delivering proper solutions. S2PMOACC takes 8 to 9 seconds for execution and Cetus is

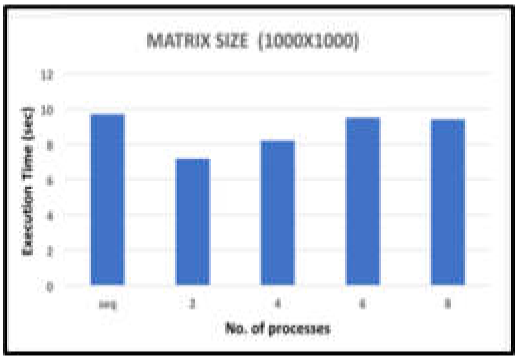


Fig. 10.1: Execution time for 1000x1000 matrixes

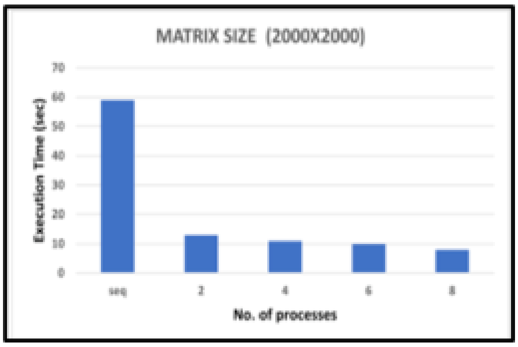


Fig. 10.2: Time matrix for 2000x2000

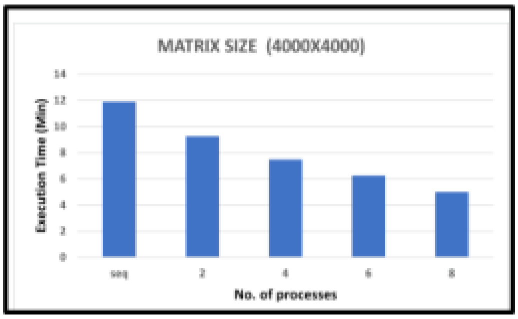


Fig. 10.3: Overall time matrix for 4000x4000

completed in 13 seconds and in the end, Par4All is doing among 22sec [4]. The above figures resulted in that proposed executive model with the help of a smaller number of resources. The research paper has developed GPU cores and determined various kinds of transactional tools.

The above figure shows the measurement of the performance matrix and it measures values in an accurate and direct way. It has been observed that those tools are proper for data analysis purposes and based on the features and capabilities included tools [4].

The above figure is representing the total amount of traffic determined for compromise and hacking SSH services. The use of the IoT tools for the Industry 4.0 is delivering security challenges that are exchanging and

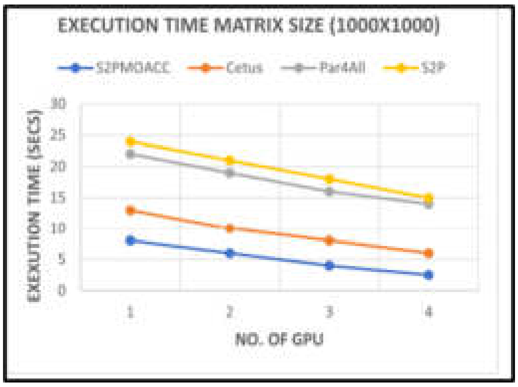


Fig. 10.4: Performance of DMM

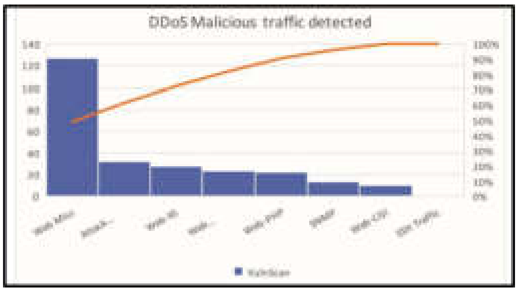


Fig. 10.5: Results of DDoS attack for detecting web services

generating results [13]. The threats of the IoT come from the lack of investment, training, security schemes, and staff capacity. IPS and IDS tools are helping to test and deliver proper support to the IoT system and helping to detect DDoS attacks.

11. Result discussion. From this research, it can be concluded that there are various multi-threaded programming in Java that are based on delivering effective communication, reducing the usage of resources, and delivering applications for responsiveness [13]. Multithreading and multiprocessing techniques have been used in this research paper to get a better output for achieving multitasking. From the research, it has been observed that multithreading saves time and performs various operations and does not affect other threads. It is mainly a CPU-based feature that allows delivering processes and delivering efficiency for computing resources [13]. For running simultaneously of the programs, multi-threaded models have played a crucial role. There are mainly three types of multithreading models include one to one relationships, many-to-one relationships and many-to-many relationships. Among those, many-to-many models have played a significant role in this research paper and it has helped to develop shortcomings for the many-to-one and one-to-one models [13]. There are four major advantages of advanced Java those are included process simplification, imaging and network-centric aspects. Java is developing GUI applications and delivering modern GUI applications that are based on 3D graphical applications. Along with this, a research paper has delivered knowledge regarding the factors for the Java multithreading languages [13].

12. Conclusion. This all concludes the making of multi-threaded communication of data languages through the transformation of the language in the programming environment. These multi-threading languages help process multiple programs through a single systematic way of processing multiple programming codes sequentially. It also helps arrange all the input programming codes through a multi-channel transformation of language to binary language. This helps in performing the programming of multiple tasks to be performed by the programmers with high efficiency in the allocated time. It also makes the easy accessibility of programming languages to be used by the programmer to create a communication bridge between computer technologies or operating systems and humans.

The research primarily focuses on the application of the Java programming language, potentially limiting its generalizability to other programming languages. Different languages may have unique features that impact their suitability for multi-threaded communication and data processing. Conducting real-world case studies and implementation scenarios across various domains and industries could validate the effectiveness and practicality of the proposed Java-based data communication model in actual operational environments.

REFERENCES

[1] A. Abuabdo and Z. A. Al-Sharif, *Virtualization vs. containerization: Towards a multithreaded performance evaluation approach*, in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), IEEE, 2019, pp. 1–6.

[2] 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.

[3] K. A. Darabkh and O. M. Amro, *New routing protocol for half-duplex cognitive radio ad-hoc networks over iot environment*, in 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE, 2020, pp. 1–5.

[4] M. F. Farooqui, M. Muqeem, A. Sultan, J. Nazeer, and H. A. Abdeljaber, *A fuzzy logic based solution for network traffic problems in migrating parallel crawlers*, International Journal of Advanced Computer Science and Applications, 14 (2023).

[5] E. Gamess and S. Hernandez, *Performance evaluation of different raspberry pi models for a broad spectrum of interests*, International Journal of Advanced Computer Science and Applications, 13 (2022).

[6] B. Hu and L. Xingguo, *Real-time simulation and optimization of elastic aircraft vehicle based on multi-gpu workstation*, IEEE Access, 7 (2019), pp. 155659–155670.

[7] T. N. Khasawneh, M. H. AL-Sahlee, and A. A. Safia, *Sql, newsql, and nosql databases: A comparative survey*, in 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, 2020, pp. 013–021.

[8] G. Labrèche and C. G. Alvarez, *Saasyml: Software as a service for machine learning on-board the ops-sat spacecraft*, in 2023 IEEE Aerospace Conference, IEEE, 2023, pp. 1–9.

[9] L. S. Nair, *An analytical study of performance towards task-level parallelism on many-core systems using java api*, in 2021 6th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2021, pp. 1255–1259.

[10] Z. Ournani, M. C. Belgaid, R. Rouvoy, P. Rust, and J. Penhoat, *Evaluating the impact of java virtual machines on energy consumption*, in Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 2021, pp. 1–11.

[11] E. Pobee, X. Mei, and W. K. Chan, *Efficient transaction-based deterministic replay for multi-threaded programs*, in 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), IEEE, 2019, pp. 760–771.

[12] P. Pufek, H. Grgić, and B. Mihaljević, *Analysis of garbage collection algorithms and memory management in java*, in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2019, pp. 1677–1682.

[13] J. J. Ruíz-Lagunas, A.-H. Anastacio, M. R. Reyes-Gutiérrez, H. Ferreira-Medina, T.-M. Cristhian, and P.-V. Omar, *How to improve the iot security implementing ids/ips tool using raspberry pi 3b+*, International Journal of Advanced Computer Science and Applications, 10 (2019).

[14] 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.

[15] J. Xu, B. Randell, A. Romanovsky, R. J. Stroud, and A. F. Zorzo, *Supporting and controlling complex concurrency in fault-tolerant distributed systems*, arXiv preprint arXiv:2111.06339, (2021).

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jun 30, 2023
Accepted: Sep 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 1097–1105, DOI 10.12694/scpe.v24i4.2384

BLENDED TEACHING SYSTEM FOR COLLEGE ENGLISH WITH MULTI-SOURCE DATA FUSION IN EDGE COMPUTING

YA FU*

Abstract. The modern education system involves the online education system with the traditional method of studies. This increases the education system's versatility and helps increase the quality of studies. The use of modern digital technology also helps the students to gather and accumulate information about their study topics. The use of edge computation helps the education system to access all the information in comparatively less time. This increases the performance of the students and teachers in analyzing the different things in the study. Multi-source data fusion is important for comparing synchronized data from different sources. This fusion system helps to allocate the data through the database and helps the students to use data from a distinctive database. This system has saved the time of students to search different contents. They are getting ready-made content from the database as per their requirements. This type of technology is essential to increase students' performance. These systems collectively develop the education system more easily and will enhance students' interest. The students could learn English using different related content and practice sheets. Digital education helps students to get appropriate sources of study materials for developing quality.

Key words: multisource, edge computing, blended teaching, data fusion

1. Introduction. Blended teaching system is a complex teaching system in which the traditional and modern systems of teaching are included. The importance of the blended teaching system is quite high to increase the students' confidence as this helps them to understand the matter from different online sources as well as from the classroom. Data gathering from the other sources is demarcated as multi-source data fusion. In modern times, the educational system is not only dependent on the traditional classroom but it includes the online education system as well as guides the students to gather the information by using modern technology. The competition among the students has arisen from the earlier times, and the mode of study has changed from the traditional to the online. For increasing the sources of study materials and the quality, the implication of a blended teaching system is a must.

The above figure represents all the interrelated things of the BTS. This helps to understand the features of this modern education system [2]. This also helps to understand the things that are interrelated with this education system. This includes the real and virtual classes for developing a clear sense of students. Modern education systems like BTS also develop communication skills for representing their views. The processes of delivering content are through offline and online methods[5]. The offline teaching method is according to the traditional education system, and online methods are followed in the online education system. It also increases the thinking capability of all the students. BTS education system helps to increase the complicated thinking process. This helps the students to increase their efficiency in their studies as well as in the decision-making process of life. BTS helps the students by increasing their talents and enhancing their scopes for achieving higher success in life [24]. BTS is the essential strategy of education, and it increases the efficiency of students and makes them all-rounder.

2. Objectives.
1. To analyze the meaning of the blended teaching system (BTS) and its characteristics
2. To know the impact of the blended teaching system on the modern students
3. To analyze the importance of multi-source data fusion in the field of education
4. To check the actual impact of a blended teaching system on the learning of the English language
5. To analyze the importance of edge computing in the education system

*Basic Teaching Department, Henan Technical College of Construction, Zhengzhou, 450064, China (yafucollege@outlook.com)

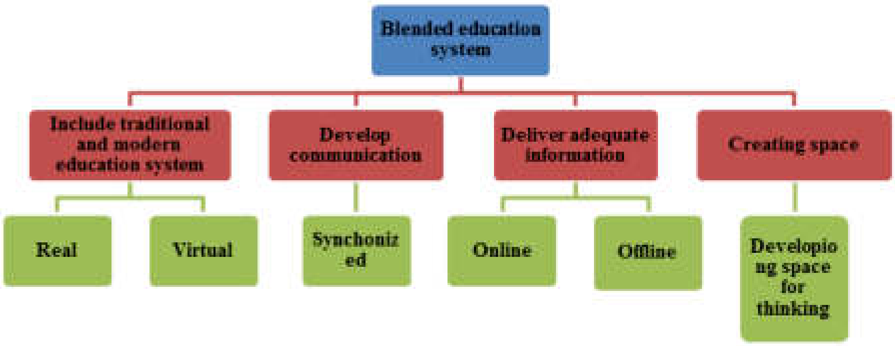


Fig. 1.1: Blended education system

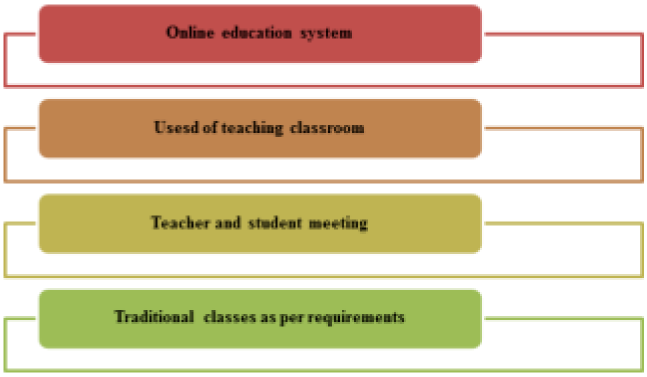


Fig. 4.1: Characteristics of BTS

6. To know the relationship between the blended teaching system and edge computing

3. Methodology. A systematic review of the different sources helps to understand the correlated matters of the blended teaching system. The qualitative analysis of the different sources through the process of observation and analysis helps to determine the most influential factors that increase the importance of BTS in the developmental process of students [16]. The qualitative study expands the knowledge of the research topic and helps analyze the effective factors behind the research topic. All the available sources have systematically aligned according to the required values and analyzed the research results through understanding all the study materials details. Qualitative analysis does not include any kind of calculation and helps the researcher to find the related influential factors through the process of observation only [10]. This method of analysis has allowed to complete the task with adequate quality.

4. Characteristics of blended teaching system (BTS). BTS is mainly based on the modern system of education as well as on the traditional teaching process. This helps the students to analyze all the study materials. The purpose of this BTS is to increase the understanding capability of students [18, 4]. This includes the processes to obtain or collect the information on the study material from the classroom and from different online sources. This enhances the quality of the education system in other educational organizations. BTS has developed by following the ideology of both teaching methods and those are traditional and modern ways of teaching. The development of this teaching system is helping the students to increase their analysis capabilities and enhance their knowledge of different subjects [22, 15]. The requirement of this education system is essential for developing the actual knowledge of the students as well as increasing the educational performance of students. It has several characteristics which are represented in the following figure.

Table 5.1: Positive impacts of BTS on students

SL NO	Effect of BTS
1	Increase the number of engaged students
2	Enabling students to understand the real-time feedback
3	More clear knowledge of students
4	Time management
5	Collaborative teaching to increase their performance

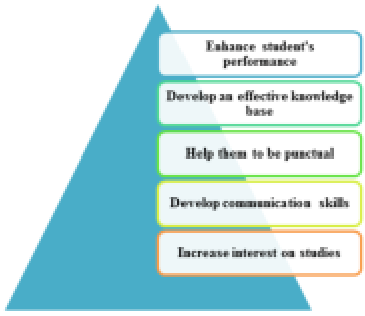


Fig. 5.1: Consequence after the implication of BTS

Figure 1.1 has represented the features of BTS and it helps to know the actual characteristics of this education system. This is a more complex and modern version of education where the online and offline education system has been agglomerated to enhance the teaching method. Digital technology helps to collect qualitative study material from different online sources [9]. It increases the scope of the students to compare their traditional teaching classes with the teachers of different online platforms such as YouTube. BTS is a modernized system of education with several unique features compared to other educational systems such as student-teacher meetings, the conduction of conferences, online guidance, a digital submission system, Online education and classes, and so on [19]. The main objective of this education is to increase student's knowledge by using different required processes.

5. Effect of BTS on the process of performance development of modern students. The ideology of the BTS education system is to engage students and teachers through traditional methods and modern methods for enhancing the performance of education. This system of education also enhanced the relationship between the teacher and students as this includes the system of conferences and suggests the physical presence of teachers and students in the same place.

Table 5.1 has described the advantages of the BTS and its impact on the character development of the students. This helps the students to develop better communication, time management or punctuality, and enhancement of knowledge in different subjects ton increase overall performances [14]. This is the most ideal process of teaching as this provides ways to develop learning habits as well as helps the students to develop suitable characteristics for getting more opportunities in the future. A collaborative system of education is essential to evaluate the actual development of students. Online education is not as capable as this may decrease the interest and motivation of the students [13]. The implication of this type of teaching process should be preferable in all education centers to increase the engagement of students with their studies. It directly boosted the quality of the traditional education system by collaborating the modern online education system. BTS is more prominent for its effect on the student's performance and brings a drastic change among the students.

Figure 4.1 has represented the actual scopes of the BTS that increased the performances of students such as increasing the knowledge of students, making them punctual, developing communication skills, increasing

Table 6.1: Benefits of using the methods of MSDF

SL NO	Important factors related to MSDF
1	Producing results from different databases
2	Increase the quality of information
3	Specific interface
4	Increase the quality of data

Fig. 6.1: Effect of MSDF implication in the education system

interest in studies and so on[14]. The increasing interest of the students helps them to engage in educational activities for more time. The traditional education system is comparatively less formative as in this education system the teacher can guide the students in the classroom only. BTS has several positive sides such as increasing the satisfaction of students to understand all the complicated study topics through the use of internet facilities and digital systems [24]. The modern system of teaching becomes more effective after the collaboration of all the teaching processes in the same place.

6. Impact of multi-source data fusion (MSDF) in the education system. MSDF is the process of management of complex situations through the use of massive and core evidence. MSDF process helps to manage complicated situations through enhanced clarity by enabling the usage of different authorized data [11]. This system helps the students and teachers to collect sufficient sources to understand a topic from different aspects. The implication of this process increases the sources of information on the same topic and helps the student to know the basic things about the study topic. Data fusion systems are interrelated with the management of data from different databases to increase the accuracy of the research and stated statement [21]. This means this method increases the accuracy of learning and decreases the possibility of occurring error in the learning process. This increases the confidence of the student to gather the information by accessing different databases.

Table 6.1 has described the details of the MSDF as this helps to collect the right information from different sources and increases the accuracy of information. This system is basically interrelated with the search system for getting important and required information from databases [1]. This makes the search process easier and more accurate in comparison to other processes.

The above-mentioned Figure 5.1 has represented the effect of MSDF on the education system. The MSDF increases the quality sources of the study and affects the development of students' knowledge [23]. This is a system to gather the right information by comparing all the available sources of databases.

7. Relation of BTS and learning of the English language. English is the most usable language in the whole world and it needs to be learned by the students. The learning of language is totally dependent on the teacher and the availability of suitable study materials [3]. Quality study materials are really important for learning good English. The learning of a language also depends on the practices of the student and it develops through the repetitive practice of the language. BTS helps the students to increase their speaking capability in English and helps them to evaluate their own quality [16]. This helps to determine the most effective practice methods for increasing the strengths of the students in the communication and usage of the proper abbreviation of English.

Table 7.1: Impact of BTS on the Learning of Language

Effect of BTS on the studying process	**Consequences of effects**
Increases the sources of study materials	Increase knowledge
Improving the understanding capability of students	Develop clear knowledge
Provide practice set	Helps to remember all things
Make the students capable to analyze their own growth	Evaluation of growth

Fig. 8.1: Advantages of edge computing

Table 7.1 has represented the effect of BTS on the learning process of language. The learning of language is basically dependent on the sources of study materials, understanding process, and repetitive practices. BTS provides all these things to the students to get higher satisfaction from the study as well as helps them to analyze their depth in English. In this education system, teachers could suggest following better skill development styles for promoting the knowledge of the English language. This teaching process helps the students by providing them with better learning materials and increasing their speaking capability through conducting special conferences for analyzing the developing status of all students [18]. BTS helps to learn all the subjects by engaging the students with practice, enabling them to align the study with the realm of life situations, developing the practical sense among the students, and making them capable of critical thinking.

8. Effect of edge computing in the modern Studies and System of education. Edge computing is the process in which the latency of the system has decreased. This increases the capability of any system by decreasing the factors that affect the performance of the system [17]. This helps the students and teachers to connect their systems and makes the virtual conversation much easier. This affects the performance of teachers and students in the long term.

Figure 6.1 has represented the advantages of the commuting edge in the different sectors such as it decreases the latency means it increases the speed of the work. This increases the overall performance of the operated system and related activities [7]. This also impacts the education system by enabling students and teachers to connect with the fast system.

Figure 6.1 has represented the impact of edge computing on the education center. Edge computing increases the system of education by increasing the speed of work and quality.

9. The relation between edge computing and BTS. BTS is dependent on the quality of edge computing and all the systems of education are dependent on edge computing. Edge computing increases the communication spewed and enhances the quality [8]. BTS is an interrelated process of the digitalized system and a traditional system of education.

Figure 8.1 has represented the relationship between edge computing and BTS. This enhances the accessibility of information and technology from different sources at comparatively low costs. Edge computing helps educational organizations to apply the ideology of BTS in a fruitful way [25]. This enhances the experiences of

Table 8.1: Effect of Edge Computing on Studies

Relation between edge computing and studies
Making the fast connection of all members in virtual classes
Enhance the speed of search engines and other machinery
Decreases the possibility of data leakage
Decreases the cost of education

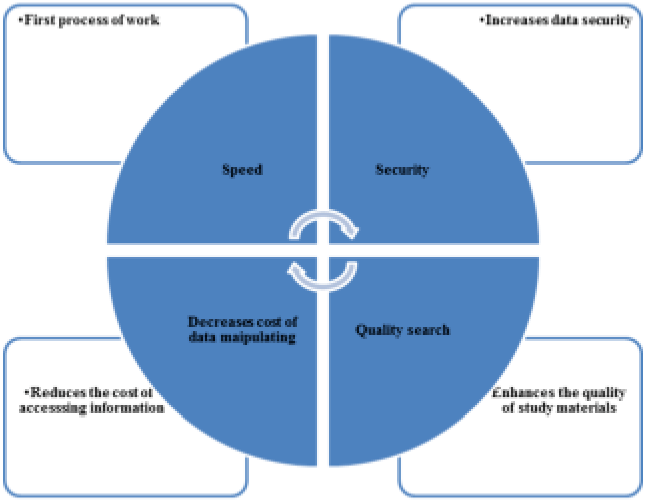


Fig. 9.1: Relation between the computing edge and blended teaching system

students, as this system is fast and durable in comparison to traditional digital technologies. It provides better security to the system of the school and secures the system from any kind of external effect [7]. BTS system becomes more performative after the implication of the edge computing process.

10. Result. In the result section, multisource data fusion and edge computing bring improvement in the teaching system throughout the world. A blended teaching system brings revolutionary changes in a country's education system. Here are the competitive characteristics of BTS, an educational institute that brings improvement in classroom teaching quality.

As per the above graphical image, edge computing improved the classroom teaching quality through the help of searchability. Here, an edge computing-based teaching system provides effective communication between teachers and students [10]. Through this advanced technology, students can solve different learning activities. On the other hand, this technology brings reformation to the education system where students get competitive advantages. Through this edge computing-based education system, students can manage their time which creates positive impacts on students' performances. Through the help of this multisource data fusion, students can get huge amounts of data that can improve the quality of the teaching system. Here students get real-time feedback from the teachers that helps to fill the gaps and improve their performances.

As per the above image, edge-based computing teaching methods in classrooms bring context optimization. Through the help of context optimization, teachers can provide relevant study materials to the students as a result; this multisource data fusion and edge-based computing system create positive impacts on the student's performances. Here, increasing the number of students who adopt edge computing-based teaching systems, those students get competitive advantages [12]. On the other hand, BTS provides different learning practices that include a collaborative learning process that brings strong relationships between students and teachers that help to analyze different educational contexts easily. Sometimes, this fusion learning system brings punctuality among students that create positive impacts on the student's performances. For improvement of learning per-

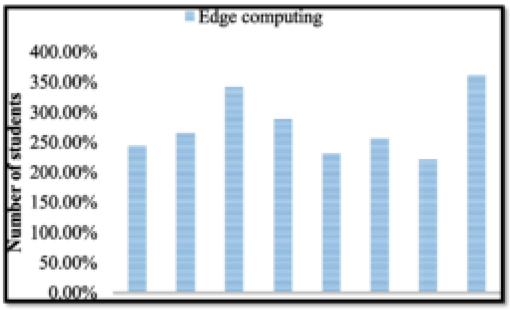


Fig. 10.1: The searchability of the classifier's process improvement of an edge-computing classroom teaching quality evaluation

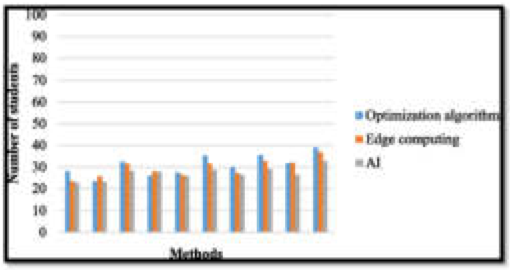


Fig. 10.2: The analysis of the optimization in the context of edge computing-based classroom teaching methods

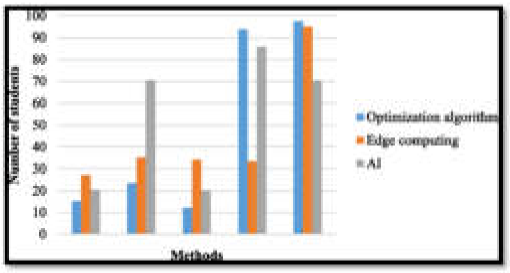


Fig. 10.3: Improvement of resource creation using Edge computing in the modern educational system

formances, students from different educational fields adopt BTS which includes edge-based computing learning systems and multisource data fusion. The availability of study materials increases the interest level among students which creates positive impacts on the context analysis aspects as a result, students get positive developments in the future. This technology reduces the cost of context analysis helps an educational institute to bring more improvement in the teaching process.

The above image provides information about the creation of the resource through edge computing-based teaching systems as well as multi-sources data fusion. As per the image, the increasing number of students who adopted edge computing helps them to make effective resources for their studies. More than 90 percent of students adopt this teaching system to get extra benefits as a result; students bring improvement in their performances. Here, a blended teaching system helps provide competitive advantages for the student and teachers where the availability of a vast amount of data helps to make effective resources for students and teachers [20]. By getting effective resources, students can improve their time management and help them to reduce the time consumed for source creation. By reducing the time consumption to make effective resources, students can spend their extra time on extracurricular activities that increase the interest level in English study among students. On the other hand, through the edge computing teaching system, students can produce results from the different types of databases that bring quality to the learning performances among students. By making appropriate study materials, students who adopt BTS get effective advantages as compared to the traditional learning system.

11. Problem statement. The main problem is to increase the reach of education to all. The cost of developing the modernized education system is high and this factor affects the extension of this system of education. This is a problem because there needs huge investment for developing an effective BTS [6]. This is the main issue of the education system as they need huge investments to expand the modernized education system in all places. The government of the country needs to implement a strategy and plan for enabling the students to use the modern education system. This increases the knowledge, performance, and interest of students and helps them to develop more clear concepts on different topics.

12. Conclusion. The blended education system is so advanced and it increases the quality of education by providing better learning opportunities. The implication of modern digitalized technology helps this education system to be successful in nature. The students access the information by using the internet facilities and they discussed the matter with the teacher. Edge computing increases include mechanical support and digitalized functions to increase the efficiency of students. On the other hand, multisource data fusion helps to analyze the different sources of data and enhances the quality of students' presentations. These processes are necessary to enhance the knowledge and ideal characteristics of human beings among students.

REFERENCES

[1] B. Adriano, J. Xia, G. Baier, N. Yokoya, and S. Koshimura, *Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 sulawesi earthquake and tsunami in palu, indonesia*, Remote Sensing, 11 (2019), p. 886.

[2] W. S. Albiladi and K. K. Alshareef, *Blended learning in english teaching and learning: A review of the current literature*, Journal of Language Teaching and Research, 10 (2019), pp. 232–238.

[3] E. Anastasiou, A. Castrignanò, K. Arvanitis, and S. Fountas, *A multi-source data fusion approach to assess spatial-temporal variability and delineate homogeneous zones: A use case in a table grape vineyard in greece*, Science of the Total Environment, 684 (2019), pp. 155–163.

[4] 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.

[5] B. Bruggeman, J. Tondeur, K. Struyven, B. Pynoo, A. Garone, and S. Vanslambrouck, *Experts speaking: Crucial teacher attributes for implementing blended learning in higher education*, The Internet and Higher Education, 48 (2021), p. 100772.

[6] J. Cárcel-Carrasco and C. Gómez-Gómez, *Qualitative analysis of the perception of company managers in knowledge management in the maintenance activity in the era of industry 4.0*, Processes, 9 (2021), p. 121.

[7] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, *Edge intelligence: The confluence of edge computing and artificial intelligence*, IEEE Internet of Things Journal, 7 (2020), pp. 7457–7469.

[8] L. Kong, J. Tan, J. Huang, G. Chen, S. Wang, X. Jin, P. Zeng, M. Khan, and S. K. Das, *Edge-computing-driven internet of things: A survey*, ACM Computing Surveys, 55 (2022), pp. 1–41.

[9] L. A. Krasnova and V. Y. Shurygin, *Blended learning of physics in the context of the professional development of teachers*, International Journal of Technology Enhanced Learning, 12 (2020), pp. 38–52.

[10] F. Li and C. Wang, *Artificial intelligence and edge computing for teaching quality evaluation based on 5g-enabled wireless communication technology*, Journal of Cloud Computing, 12 (2023), pp. 1–17.

[11] R. W. Liu, Y. Guo, J. Nie, Q. Hu, Z. Xiong, H. Yu, and M. Guizani, *Intelligent edge-enabled efficient multi-source data fusion for autonomous surface vehicles in maritime internet of things*, IEEE Transactions on Green Communications and Networking, 6 (2022), pp. 1574–1587.

[12] R. Mahmud, S. Pallewatta, M. Goudarzi, and R. Buyya, *Ifogsim2: An extended ifogsim simulator for mobility, clustering, and microservice management in edge and fog computing environments*, Journal of Systems and Software, 190 (2022), p. 111351.

[13] P. Peng, *Construction of the blended teaching model of ideological and political course in colleges and universities*, in 2021 6th International Conference on Modern Management and Education Technology (MMET 2021), Atlantis Press, 2021, pp. 17–20.

[14] B. Philipsen, J. Tondeur, N. Pareja Roblin, S. Vanslambrouck, and C. Zhu, *Improving teacher professional development for online and blended learning: A systematic meta-aggregative review*, Educational Technology Research and Development, 67 (2019), pp. 1145–1174.

[15] K. Polhun, T. Kramarenko, M. Maloivan, and A. Tomilina, *Shift from blended learning to distance one during the lockdown period using moodle: test control of students' academic achievement and analysis of its results*, in Journal of physics: Conference series, vol. 1840, IOP Publishing, 2021, p. 012053.

[16] R. Rafiola, P. Setyosari, C. Radjah, and M. Ramli, *The effect of learning motivation, self-efficacy, and blended learning on students' achievement in the industrial revolution 4.0*, International Journal of Emerging Technologies in Learning (iJET), 15 (2020), pp. 71–82.

[17] P. P. Ray, D. Dash, and D. De, *Edge computing for internet of things: A survey, e-healthcare case study and future direction*, Journal of Network and Computer Applications, 140 (2019), pp. 1–22.

[18] F. M. Sari and A. Y. Wahyudin, *Undergraduate students' perceptions toward blended learning through instagram in english for business class.*, International Journal of Language Education, 3 (2019), pp. 64–73.

[19] D. Shi, *A study on the influence of blended teaching of an integrated english course based on flipped classroom on alleviating students'learning anxiety*, Psychiatria Danubina, 34 (2022), pp. 467–467.

[20] J. Singh, K. Steele, and L. Singh, *Combining the best of online and face-to-face learning: Hybrid and blended learning approach for covid-19, post vaccine, & post-pandemic world*, Journal of Educational Technology Systems, 50 (2021), pp. 140–171.

[21] J. Song, C. Zhao, S. Zhong, T. A. S. Nielsen, and A. V. Prishchepov, *Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques*, Computers, Environment and Urban Systems, 77 (2019), p. 101364.

[22] 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.

[23] Z. Wang and F. Xiao, *An improved multi-source data fusion method based on the belief entropy and divergence measure*, Entropy, 21 (2019), p. 611.

[24] R. Zhang, *Exploring blended learning experiences through the community of inquiry framework*, (2020).

[25] X. Zhang, Z. Cao, and W. Dong, *Overview of edge computing in the agricultural internet of things: key technologies, applications, challenges*, Ieee Access, 8 (2020), pp. 141748–141761.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 1, 2023
Accepted: Sep 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1107–1116, DOI 10.12694/scpe.v24i4.2385

AI-ASSISTED CHINESE LANGUAGE TEACHING SYSTEM IN DISTRIBUTED SENSOR NETWORKS

HUAFENG SHAN*

Abstract. The Chinese learning system gained enormous importance to support the population of China with recent technologies. The AI and ML-assisted classes are useful for learners. However, it can cut down the employment of teachers. Various planning and act was passed to reform the Chinese education system with artificial intelligence and machine language. On the other hand, this education process is interesting, customized, and unsupervised based. It was observed that the human touch study process is more satisfying and clears doubts easily. The key object of data gathering sensor-based network is accumulating all the data from the nodes and sending those sensed data to the base station. These nodes are basically working while focusing on one thought: all the nodes present in the sensor-based system are well aware of all the other nodes currently working in the same network and transmitting data in the base station. Although AI comes with a few challenges like lack of resources, training, and various security issues, if these are overcome with proper strategies, AI can be the equipment to reign in the 21^{st} century. Additionally, LLMs and NLP modeling are described in the results that helped to understand the use of different models. Moreover, the Benefits of LLMs and NLP for Chinese language training are discussed in the study.

Key words: AI, Machine learning, Chinese language, Learners, Impact, Distributed sensor, LLMs, NLP

1. Introduction. The educational systems are becoming more dependent on the automated learning process and it has changed the form of the Chinese learning process. The learners are becoming more dependent on the customized learning system, and it has a positive impact on the learning system [1]. It has been observed that the upcoming generations will be more dependent on the intelligent network teaching system model that is based on AI. Artificial intelligence is a process of perceiving, synthesizing, and inferring information that is demonstrated by non-human things. The uses of artificial intelligence and machine learning can decrease the process of traditional learning and will enable the smart learning process [5]. However, it has a side effect as it can reduce the human touch in the educational system, and many teachers can lose their jobs. Therefore, it can be said that behind every smart use, there is a side effect always available. In below are the well sides of artificial intelligence and machine learning assistance in Chinese learning, and they can create a positive impact on the student's mind.

The research is motivated by a complex interplay of factors such as technological advancement, educational reform, employment concerns, human-centric learning, and the potential benefits of AI and language technologies. By exploring these facets, the research seeks to provide insights into how technology can be harnessed to enhance education while addressing associated challenges. The Chinese language teaching with Intelligence creates wide interest among learners. The potential impact of this research is multifaceted. From an academic standpoint, the study can contribute to the existing body of knowledge by shedding light on the nuanced connections between urban green spaces and mental health outcomes. On a practical level, the findings could lead to evidence-based recommendations for designing and implementing green spaces that optimize mental health benefits. Such insights can foster healthier communities, reduce healthcare costs associated with mental health issues, and improve the overall quality of life for urban residents. Top of FormBottom of Form

2. Objectives.
1. To identify the importance of AI and machine language in the Chinese language learning process
2. To know more about the process of reform that makes by AI and ML in the computer-assisted instruction system in Chinese education
3. To analyze the impact of AI and machine language in the education system of Chinese learning

*Zhengzhou Railway Vocational Technical College, Zhengzhou, 450000, China (huafengshan1@outlook.com)

Fig. 2.1: AI and ML varieties in the learning system

Table 4.1: Uses of machine in education system

Uses	Definition
Structured or unstructured	Machine learning can be either structured or unstructured way and that can be supervised by someone or not.
Robot navigation	There is someone needed who can use machine
Smart classes	While people are thinking of learning a new language, that must be interesting for them

4. To investigate the use of distributed sensor networks to gather data about learners
5. To address the challenges that have been faced by Chinese language learners due to AI and Machine learning
6. To explore the methods that have been used by the learners using AI for the customized learning process

3. Methodology. The research process was done based on the secondary qualitative method and the information was collected about AI and ML. The whole research is related to machine learning and AI uses related to Chinese learning in distributed sensor networks [1]. Therefore, it is essential to gather information only related to the AI. Machine learning is a process that requires centralized sensors to handle instructional strategies. The data was collected with the help of Google Scholar, articles, blogs, and journals. In this way, it could find the information related to the topic, which was an appropriate data collection process for this study [2]. Without proper data collection, no research can have a successful outcome.

4. Importance of Machine Learning and AI-assisted Classse in Chinese Learning. Figure 4.1 indicates that there are different types of importance available for machine learning and AI-assisted classes in Chinese learning. The students can learn about this language by staying at any place in the world, and digital medium is only the way to learn things [3, 6]. Students can easily enjoy these smart classes as there are different types of videos and gaming systems available in between the learning, and they can customize their learning hours. Artificial intelligence offers an effective education through online and automated instructors' routine tasks are available there. In this learning process, the reforming section might be available, and real-time decision-making processes, AI assistance, and robot navigation all are available [16, 7]. The supervised learning processes can easily give proper classification or clarification related to the topic. Having knowledge of machine learning can make a person successful in their career. This AI and machine learning education is effective for those students who are studying online and have enough knowledge about machines [5]. The uses of machine learning and AI can enable one to make a successful career as the uses of technology are increasing and uses almost in every organization.

5. Impact of AI and Machine Learning Language in the Education System of Chinese Learning. There is always an impact available behind every successful work and that has to be faced by everyone. On the other hand, there are many teachers available who are connected with this field and their only earning source is this. The increasing number of uses of AI and ML can easily cut down the teachers and human touch from the education sectors. In the future, students will only get an education from machines, and many people

Fig. 4.1: AI and machine learning uses in the education field

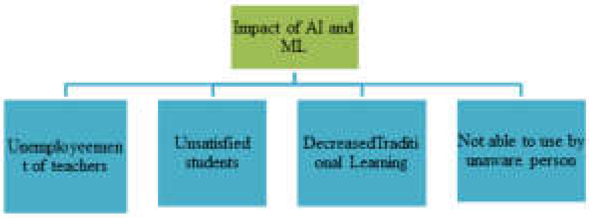


Fig. 5.1: Side effects of digital learning

will be facing unemployment.

Figure 5.1 says that online systems are mainly unsupervised learning and it was found that [9]. AI and machine learning a robotic systems that cannot able to work individually without anyone's help. Learners' and educators' connection matter a lot and the way it is possible online, is not possible in the online education system much as unemployment will happen due to machine use, people will be depressed more, and uncertain circumstances can happen at that time [10]. Therefore, it can be suggested that online education with the help of AI and ML can be also done based on supervised learning. Machine uses can be harmful to life also and as this can only be run by humans, therefore, having basic knowledge about it is necessary.

6. The Process of Reforming the Computer-assisted Chinese Education by AI and ML. With the help of artificial intelligence and machine learning, the education system can create a positive impact on the student's life. The Chinese education system is one of the fields that can create a positive influence on the student's life and they can increase the use the online education [10]. In the year 2019, the Chinese State Council developed two important plans to drive continued reform in and advancement of China's education sector. They have been trying to reform the learning system since China's opening up in 1978.

Figure 6.1 indicates two major plans of the education system in China. These plans are the "China's Education Modernisation 2035" Plan and the "Implementation Plan for Accelerating Education Modernisation (2018-2022)". The main ambition to reform the education system is to create sustainability in Chinese learning till 2035. By developing artificial intelligence, the education system can easily identify talented people among them. It has been observed that the main impact of machine learning and artificial intelligence was developed by mainland China [8]. AI and ML can able to replace human interaction in the learning field and that can give the students enough interesting work. Based on these education processes and reforming systems, AI and

Table 5.1: Impact and definition of machine uses increase

Impact	Definition
Unemployment of teachers	Machine can replace the place of teachers
Unsatisfied students	students can only clear their doubts and be satisfied while there are teachers available
Traditional learning	The increasing uses of artificial intelligence and machine learning in the education system can be decreased the process of traditional learning [8].
Unaware about machine use	It is not possible for one who has no idea about using systems.

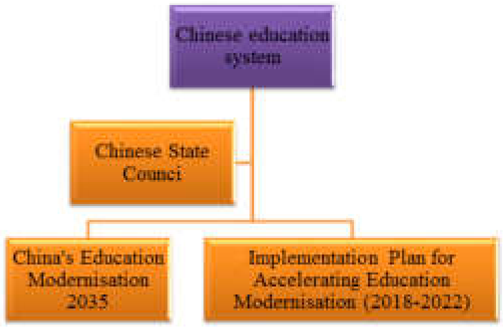


Fig. 6.1: Plan passed by Chinese state council

ML can be sustained many times. Reforming the education system with Artificial intelligence and machine are one type of significant as well as it can be harmful for human. It will reduce the work opportunities for many people and that can create huge unemployment [4].

7. Investigating the Utilization of the Networked with Distributed Sensor in Order to Gather Data about the Learners. The main focus of gathering the data in the network with distributed sensors is to transmit the data that is being sensed from each of the current nodes to another base station. Data gathering is basically a process of utilizing the nodes to their maximum capacity before they stop working or die, and the selected network becomes out of operation or inoperable at all [19]. A full complete round indicates that a specific base station is able to collect all the data that are being sent from the sensor nodes stationed outside. This is also used to reduce delay and minimize energy usage in the network that comes with the sensor [18]. There are different types of data gathering, as it is mentioned below:

Data Transmission: All the nodes with sensors transmit the called data directly to one place, the base station. However, there are some disadvantages of this whole process that cannot be ignored, such as:

1. It consumes a lot of energy and that makes it a lot more expensive than others do. The more energy it will consume, the more costly it will become.
2. The delay in the media is large also, as transmitting; the information that has been accumulated from the nodes to the base station that is located far away from the nodes would take a much longer time.
3. Whenever it is transmitting the data directly to the far-away base station, the performance becomes poorer.

PEGASIS: It stands for "Power-Efficient gathering For Sensor Information System", it is an enhancement for adapting in low energy in the hierarchy in clustering. It is founded on the belief that all the nodes that are connected in the network are aware of every other node that already exists in the network [18]. The main goal of PEGASIS is to decrease the distance of the noise transmission. Apart from that, it also manages to lower

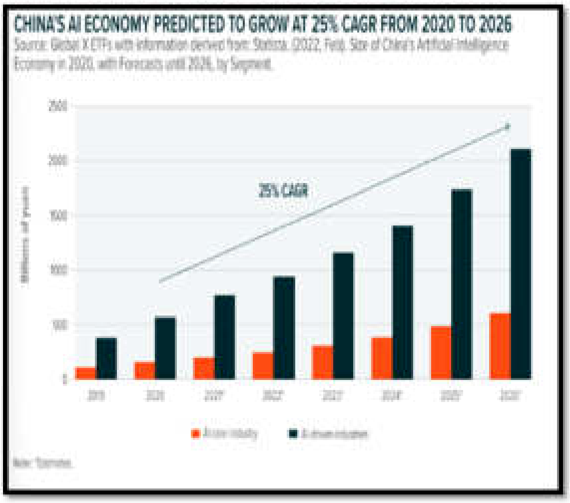


Fig. 7.1: Prediction of AI economy of China

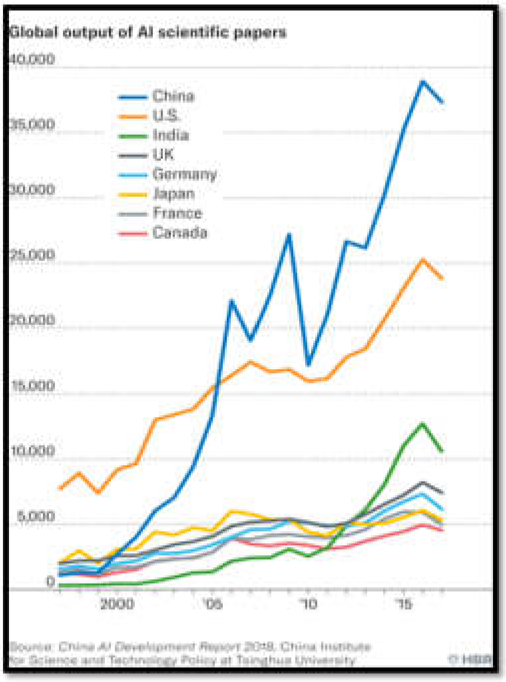


Fig. 7.2: AI development of China, 2018

the overhead of the overall broadcasting. It decreases the total number of sensed messages that are required to be sent to the base station [20]. It also makes sure that the energy is distributed equally throughout all the nodes that are connected in the network.

8. Evaluating the Key Issues Faced by Learners of the Chinese Language due to Machine Learning and AI. AI and machine learning are growing rapidly in China as the research on this topic is growing and ending up in real-life implications. The adoption of AI has spiked in the last few years in various

Table 7.1: Implication of data gathering sensor network

The implication of proper data gathering in the network with a sensor	Checking the current environment Mechanization of the industries Smart homes in the urban areas Checking on the health Reacting promptly to a disaster or crisis [22]

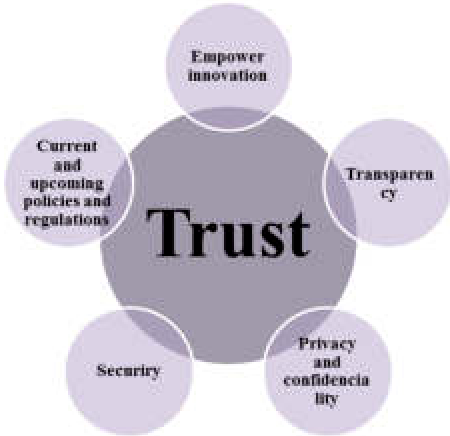


Fig. 8.1: Preserving privacy in the machine learning

industries, even beyond the tech industries [13]. These industries include retail, finance, telecommunication, and governance. The government in China has released a statement already that is very supportive in terms of the policies and guidelines for fostering more investments in the fields that are related to AI and machine learning.

As AI has some great sides that will help learners to learn new things quickly, there are some challenges too, that need to be agreed upon while incorporating AI in the classroom all over the world including China. The initial challenge that is faced in the case of using AI in the classroom is a technical expert is always needed in the room [11]. There are teachers who are not experts in these matters or are still not comfortable with using AI broadly and may face difficulties in applying this technology in their own practices of teaching. These teachers may need extra help, support, and training for them to adapt to the new technology [11]. Many universities and schools do not have the resources to buy these costlier machines and technologies and maintain them continuously. These institutes may need extra funding to incorporate them into their teaching practices in the classrooms [12]. Left apart from the concerns related to security, privacy, and the job demand market is also there. The challenges and limitations are discussed as follows,

1. The integration of technology could inadvertently widen the digital divide, as access to advanced technology might not be uniform across all regions and socioeconomic backgrounds. This could lead to unequal learning opportunities for different segments of the population.
2. While AI can enhance personalized learning, ensuring the accuracy and reliability of AI-driven content and assessments could be a challenge. AI systems might lack the nuanced understanding of human instructors, potentially impacting the quality of education.
3. Gathering and analyzing student data through sensor networks raises concerns about student privacy and data security. Safeguarding sensitive information and preventing unauthorized access become crucial considerations.
4. Integrating AI and technology into the education system requires substantial teacher training. Not all educators might be comfortable with or knowledgeable about these technologies, leading to potential

Table 9.1: Process of understanding, measuring, and mitigating the privacy breaching in machine leaning

Understand	**Measure**	**Mitigate**
Conduct threat modeling and attack research Identify confidentiality properties and guarantees Understand regulatory requirements	Capture vulnerabilities quantitatively Develop and imply framework to monitor possible risks and mitigation success	Develop and imply techniques to decrease the risks of privacy Meet legal and compliance rules and regulations

resistance or suboptimal utilization.

5. While technology can enhance learning efficiency, an overreliance on AI-assisted learning might lead to a reduction in direct human interactions between teachers and students. The social and emotional aspects of learning could be compromised.
6. Language models and NLP might struggle with capturing the intricacies of the Chinese language and its cultural context. Ensuring accurate language processing and cultural relevance in AI-driven content could be challenging.
7. The use of AI in education raises ethical questions, especially when it comes to data privacy, bias in algorithms, and potential manipulation of learning experiences. Ensuring ethical AI deployment is essential.
8. Integrating AI seamlessly into an existing educational framework is a complex task. Developing user-friendly interfaces and platforms that educators, students, and parents can navigate easily is a significant challenge.

9. Identifying the Strategies to Mitigate the Issues faced by Chinese Language Learners for the Tailor-made Learning Process. There are a few strategies that can be recommended to incorporate the AI system in the classroom of Chinese learners. The first thing is needs to be done here is to find a provider that is reliable and partner with them. Only then AI can be integrated into the classroom successfully [15]. This partner can be a company from a technological background, any NGO, or any local university. Rather than implying AI in the whole system or curriculum in the first go, it is suggested that the teachers start using it in smaller sections, and gradually expand it [14]. This will also allow the teachers to gain more experience on how to handle these technologies and build confidence in them. Eventually, it will end up building more confidence and refining the whole practice of teaching.

It is the responsibility of the teachers to encourage the students to think critically and make ethical decisions [17]. The students should follow the guidelines for implying AI ethically in their studies and in their real life while making critical thinking about it. They should also consider its usage in a broad spectrum and the possible consequences. This will also help the students to become more aware of the usage of the technologies and will be well-equipped digital citizens who can navigate through any challenges [17]. They will be enabled to grab the opportunities the digital era has to offer.

10. Pproblem Statement. This research addresses the issue that Chinese learners are facing due to the usage of AI and machine learning. This is an issue because AI and machine learning are growing at a light speed and putting their hand in every place possible. This is an issue in contemporary times because AI comes with a few default issues, such as privacy breaches, unethical usage, security and privacy, and so on. This research sheds light on the key issues that Chinese learners are facing due to the usage of AI in customized learning and how to mitigate those issues and use AI in the favor of the students.

11. Results. ***Inclusion of LLMs model***

Large language models (LLM) are one of the most Implemented models for designing AI-based designs. Thus, Along with natural language processing (NLP) LLMs are best suited for AI-Assisted Chinese Language Teaching Systems. In addition, the inclusion of LLMs allows a wide range of themes. It was found that most of the linguistic-based AI interphones allied the data until 2021 [8]. For example, chat GPT-3, uses one of the AI-based models that handle data until 2021 [8]. However, with the expansion of different themes data range

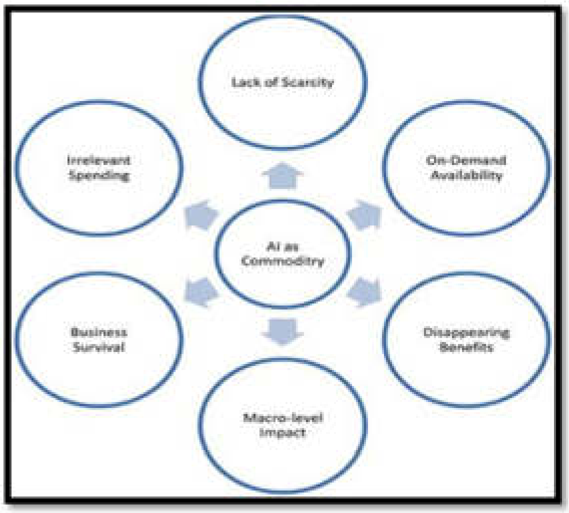


Fig. 9.1: AI as commodity

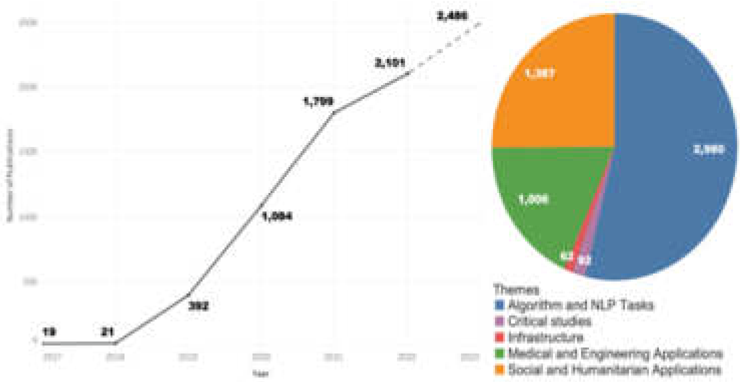


Fig. 11.1: Research themes and trends for LLMs

is increasing for such models.

Figure 11.1 is related to different themes and trends of LLMs, additionally, the data is presented in co-relation with NLP tasks. Moreover, the above graph indicates that the LLMs field can be distinguished in several different themes that are related to other subfields. Therefore, a large field of data, algorithms, and mass can be included in the natural Chinese language processing system. Furthermore, with the help of such models, a wide range of information is fed into the AI modeling. From the above graph, it can be seen chances of inclusion of algorithms in the NLP task and in the field of LLM is 54%. In addition, humanitarian applications and social application topics have a 25% representation [8].

Such studies indicate a wide range of applications for LLMs, for instance, Identifying controversial speeches and humanitarian research is one of the major themes. In addition, identifying controversial speeches and identifying translated language and sentimental analysis are some of the primary features [15]. In addition, an LLM pre-training and fine-tuning of an AI-based linguistic model helps to identify different tasks and provide responses accordingly [8]. In order to such advanced models and differentiate linguistics for natural language processing, LLMs are one of the most effective models. Moreover, moreover, with such a design of the natural language proposing system a network of different linguistic abilities can be achieved. For instance, comparing a

Fig. 11.2: LLMs research and its popularity over different regions

theme based on different strata of data provides the most preferable responses for a language processing system. In addition, there are different factors that allow us to compare data by comparing all the possibilities factors. Therefore, with such abilities an improved system of language learning is achieved.

Figure 11.2 of the statistical analysis provides an overview of LLMs research and its popularity over different regions. Additionally, it highlights the networking between the different systems for the year 2022 [8]. Thus it can be seen that with a widespread network, improved databases can be achieved. Thus, designing the Chinese language and Implementing an LLM module is preferable. In addition, the main interaction of such a model will be related to the students. Thus, an appropriate distribution of the following model can be seen in the image above figure [8]. Figure 11.2 further shows a cluster of different nations, China and the USA being the leading nations for using LLMs and natural language processing for linguistics. Moreover, it was found that with the increment in the databases, a better training model can be formulated for the AI-based linguistic training system. Additionally, a wide distribution of data is required in order to provide different strata of information[8]. Hence, the Implementation of such a model in the ai-assisted Chinese language teaching system is important. Moreover, implementing LLMs along with NPL provides an improved data set to work with.

12. Conclusion. In conclusion, AI and machine learning in the classrooms of Chinese learners may offer an opportunity that is both unique and excellent for the students as well as the teachers. AI has the possibilities and potential to offer learners engaging and tailor-made learning experiences. This will also help them to develop skills that are necessary for the 21st century such as problem-solving and critical thinking. However, this also comes with a range of issues and if it can be overcome, this will be a revolutionary change.

REFERENCES

[1] A. Alam, *Should robots replace teachers? mobilisation of ai and learning analytics in education*, in 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), IEEE, 2021, pp. 1–12.

[2] I. L. Alberts, L. Mercolli, T. Pyka, G. Prenosil, K. Shi, A. Rominger, and A. Afshar-Oromieh, *Large language models (llm) and chatgpt: what will the impact on nuclear medicine be?*, European journal of nuclear medicine and molecular imaging, 50 (2023), pp. 1549–1552.

[3] 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.

[4] A. Bandi, *A review towards ai empowered 6g communication requirements, applications, and technologies in mobile edge computing*, in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2022, pp. 12–17.

[5] M. Chen, X. Wei, Y. Gao, L. Huang, M. Chen, and B. Kang, *Deep-broad learning system for traffic flow prediction toward 5g cellular wireless network*, in 2020 International Wireless Communications and Mobile Computing (IWCMC), IEEE, 2020, pp. 940–945.

[6] W. Chen, R. He, G. Wang, J. Zhang, F. Wang, K. Xiong, B. Ai, and Z. Zhong, *Ai assisted phy in future wireless systems: Recent developments and challenges*, China Communications, 18 (2021), pp. 285–297.

[7] N. Dey, P. N. Mahalle, P. M. Shafi, V. V. Kimabahune, and A. E. Hassanien, *Internet of Things, smart computing and technology: a roadmap ahead*, Springer, 2020.

[8] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, *A bibliometric review of large language models research from 2017 to 2023*, arXiv preprint arXiv:2304.02020, (2023).

[9] Y. Gao, L. T. Yang, J. Yang, H. Wang, and Y. Zhao, *Attention u-net based on bi-convlstm and its optimization for smart healthcare*, IEEE Transactions on Computational Social Systems, (2023).

[10] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, *What is semantic communication? a view on conveying meaning in the era of machine intelligence*, Journal of Communications and Information Networks, 6 (2021), pp. 336–371.

[11] M. Li, J. Gao, C. Zhou, X. S. Shen, and W. Zhuang, *Slicing-based artificial intelligence service provisioning on the network edge: Balancing ai service performance and resource consumption of data management*, IEEE Vehicular Technology Magazine, 16 (2021), pp. 16–26.

[12] K. Muhammad, H. Ullah, M. S. Obaidat, A. Ullah, A. Munir, M. Sajjad, and V. H. C. De Albuquerque, *Ai-driven salient soccer events recognition framework for next generation iot-enabled environments*, IEEE Internet of Things Journal, (2021).

[13] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, *Holistic network virtualization and pervasive network intelligence for 6g*, IEEE Communications Surveys & Tutorials, 24 (2021), pp. 1–30.

[14] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, *Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19*, IEEE reviews in biomedical engineering, 14 (2020), pp. 4–15.

[15] F. Ufuk, *The role and limitations of large language models such as chatgpt in clinical settings and medical journalism*, Radiology, 307 (2023), p. e230276.

[16] 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.

[17] J. Wan, X. Li, H.-N. Dai, A. Kusiak, M. Martinez-Garcia, and D. Li, *Artificial-intelligence-driven customized manufacturing factory: key technologies, applications, and challenges*, Proceedings of the IEEE, 109 (2020), pp. 377–398.

[18] R. Wang, Y. Zhang, G. Fortino, Q. Guan, J. Liu, and J. Song, *Software escalation prediction based on deep learning in the cognitive internet of vehicles*, IEEE Transactions on Intelligent Transportation Systems, 23 (2022), pp. 25408–25418.

[19] W. Wang, T. Sato, V. Piuri, M. Aloqaily, and K. Yu, *Guest editorial special issue on advanced cognitive computing for data-driven computational social systems*, IEEE Transactions on Computational Social Systems, 9 (2022), pp. 1589–1593.

[20] Z. Wang, D. Jiang, and Z. Lv, *Ai-assisted trustworthy architecture for industrial iot based on dynamic heterogeneous redundancy*, IEEE Transactions on Industrial Informatics, 19 (2022), pp. 2019–2027.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 1, 2023
Accepted: Sep 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1117–1126, DOI 10.12694/scpe.v24i4.2386

SENTIMENT ANALYSIS OF ONLINE PRODUCT REVIEW USING DEEP LEARNING IN DISTRIBUTED SENSOR NETWORKS

JUN YAO*

Abstract. Recently, sentiment analysis has been a major business practice for an organization that helps analyze online reviews of different products. An organization can understand the customer's perception of their organizational products by analyzing reviews. On the other hand, this analysis process also helps to understand the customer's emotions about the organizational products. In this process, deep learning and distributed sensor networks play crucial roles in analyzing emotions. Five different steps of this analysis will provide accurate information about the customers' emotions on products. This analysis will help increase the product's value by understanding the customer's perception of where negative feedback improves the products. Through the help of this analysis, an organization will get several benefits that enrich its organizational image. On the other hand, this analysis process will face polarity issues, tone issues, comparative sentence analysis issues, and understanding idioms and emojis issues. The implementation of live API and proper sentiment analysis tools help to increase the effectiveness of this analysis.

Key words: Distributed Sensor Networks, Machine Learning, Decision Tree Algorithm, Natural Language Processing (NLP), Computational Linguistics, Support Vector Mechanics

1. Introduction. Sentiment analysis of online product reviews refers to the automated process of understanding the opinion as well as sentiments of customers about the products [17]. Through this process, an organization can automatically analyze the online reviews from consumers about its products and services and make them sorted in terms of positive as well as negative and neutral. By using deep learning, organizations implement different algorithms such as Linear Regression, Naive Bayes, and Support Vector Machines which analyze online reviews about the products [21].

Illustrate the effectiveness of sentiment analysis, consider the case of an e-commerce platform seeking to enhance customer satisfaction. The company collected thousands of product reviews and utilized sentiment analysis techniques to extract meaningful insights. The company employed a combination of deep learning algorithms and natural language processing (NLP) tools to analyze customer reviews. The reviews were categorized as positive, negative, or neutral sentiments. The sentiment analysis process was supported by distributed sensor networks (DSN) for data collection.

By implementing sentiment analysis, the e-commerce platform gained actionable insights into customer perceptions. They identified specific pain points in their products and services, leading to targeted improvements. Negative sentiment patterns related to delivery delays were identified, prompting the company to address logistics issues. As a result, customer satisfaction scores increased by 20% over six months, leading to improved customer loyalty and increased sales.

This case study underscores the importance of sentiment analysis in understanding customer sentiments. It also highlights the value of real-time data collection through DSN, which provides timely feedback for operational improvements. By utilizing deep learning and NLP, the company identified problems and formulated effective solutions based on customer feedback. Including such a case study would provide readers with a tangible example of how sentiment analysis can drive tangible improvements in real-world scenarios.

The contribution of the research,

1. Our research introduces the integration of distributed sensor networks (DSN) to sentiment analysis, enabling real-time data collection and analysis. This innovative approach ensures that organizations promptly gain actionable insights into customer sentiments and market trends, facilitating agile decision-making.

*School of Mathematics and Computer Science, Tongling University, Tongling, China, 244061, email: junyaolive@outlook.com

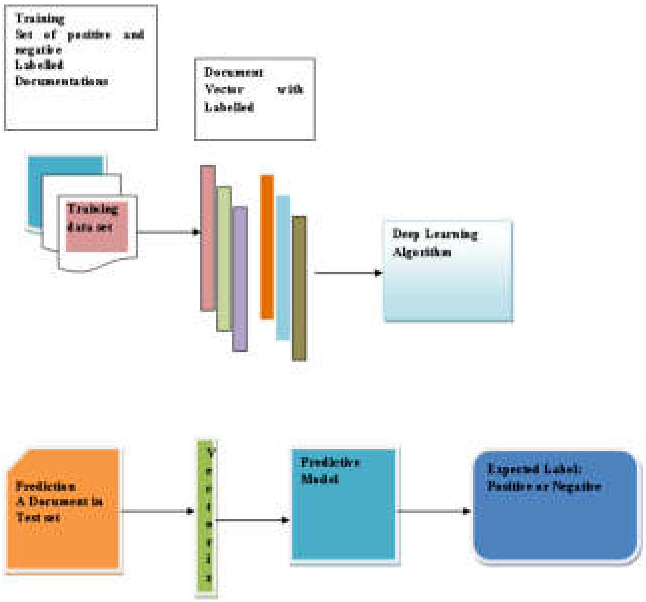


Fig. 1.1: Process of Sentiment analysis through Deep learning algorithm by using DSN

Table 1.1: Type of sentiment analysis for online products reviews

Types	**Description**
Knowledge-Based	This type of sentiment approach provides information about emanating emotion
Statistical	This approach mainly uses different types of deep learning and machine learning algorithms to ensure accuracy in the sentiment detection aspects [1].
Hybrid	This type of approach is mainly based on statistical as well as knowledge-based techniques used for sentiment detection.

2. We address the challenges posed by idiomatic expressions and emojis in sentiment analysis, enhancing the accuracy of our analysis. By employing advanced NLP techniques, we offer a comprehensive solution to capturing nuanced sentiments, which is crucial for understanding modern online communications.
3. By providing a comprehensive framework that combines deep learning algorithms with DSN, our research offers a new perspective on sentiment analysis that caters to the needs of businesses operating in rapidly changing market landscapes.

As per the above image, all the online reviews are used as data sets, documented as per their level. Here, distributed sensor networks (DSN) play a vital role in sentiment analysis through deep learning [14]. Using different ML algorithms, an organization can understand customers' expectations of its products, which helps in the product improvement segment. Here, through the predictive model, all kinds of expressions of customers on products are categorized in terms of positive and negative perspectives [5]. Organizations use different programming language expertise, including Python, machine learning, and R language. On the other hand, organizations extract information through the DSN and analyze that data through natural language processing (NLP), computational linguistics, text analytics, and classifying the polarity of the opinion processes. DSN is used as an autonomous sensor to analyze different online reviews from customers about the products and pass data and information to advanced technology through this networking system.

As per Table 1.1, in the sentiment analysis aspects, there are a total of three types of approaches used that include Knowledge based technique, statistical technique, and hybrid technique.

The heart of sentiment analysis lies in the application of deep learning algorithms. Decision tree algorithms, Support Vector Machines (SVM), and neural networks are commonly employed to classify textual data into different sentiment categories. Decision trees, for instance, partition data based on specific features and follow a hierarchical structure to arrive at classification decisions. SVM, on the other hand, constructs hyperplanes to separate data into distinct categories. Neural networks, with their intricate layers of interconnected nodes, are known for capturing complex patterns in text data.

NLP plays a crucial role in preprocessing and understanding textual data. Techniques like tokenization, stemming, and lemmatization aid in breaking down sentences into individual words and reducing them to their root forms. Stop-word removal helps eliminate common words with little semantic value. Part-of-speech tagging allows for grammatical analysis, which is vital in accurately capturing context and meaning.

Challenges and Limitations:

1. Sentiment analysis heavily relies on the quality and diversity of training data. Biased or unrepresentative datasets can lead to skewed results.
2. Textual nuances like sarcasm and irony can be challenging for sentiment analysis algorithms to detect accurately, as they often require an understanding of contextual cues.
3. Sentences can have different meanings based on the surrounding context. Algorithms might struggle to grasp the complete context, leading to misclassification.
4. While emojis and idiomatic expressions provide rich emotional context, they can be difficult to interpret algorithmically. NLP tools might not be well-equipped to handle these elements effectively.
5. Sentiment analysis might struggle with linguistic variations, slang, and regional differences, particularly in diverse online content.
6. Different industries and domains might have specialized terminology that is not covered in general sentiment analysis models.

To mitigate some of these limitations, hybrid approaches that combine statistical and knowledge-based techniques have been explored. These approaches attempt to leverage the strengths of both methodologies to achieve more accurate sentiment analysis results. By providing an in-depth explanation of the methodologies, you equip your readers with a clear understanding of the technical underpinnings and challenges of sentiment analysis. This not only adds credibility to your work but also offers valuable insights for researchers and practitioners seeking to implement sentiment analysis effectively.

2. Objectives. The objectives of the research study are:

1. To understand the concept of sentiment Analysis of Online Product Reviews
2. To understand the five main steps of sentiment Analysis of Online Product Reviews through deep learning in DSN
3. To identify the benefits of sentiment analysis of online product reviews by using deep learning in DSN
4. To evaluate the role of deep learning and DSN in online product reviews through the sentiment analysis process
5. To Understand different challenges raised during sentiment analysis of online product reviews in DSN
6. To recommend effective strategies to increase the use of deep learning in online product reviews in DSN

3. Methodology. In this study, the secondary research method was used to understand the benefits of sentiment analysis of online product reviews through DSN and deep learning. Here a secondary data collection process was used where all the collected data is taken from authentic existing sources such as websites, books, journals, and newspapers [13, 3]. This methodology section used qualitative analysis processes to understand the different types of sentiment analysis for online product reviews. Interpretivism is the research philosophy used for this study to understand the subjective aspects of this study. Here different deep learning algorithms are used for the data analysis segment. Through the help of the analysis process, unbiased results for this study have been seen.

Our research offers a novel approach by integrating deep learning algorithms with distributed sensor networks (DSN) for sentiment analysis of online product reviews. While sentiment analysis has been extensively explored, the incorporation of DSN adds a real-time data collection dimension that enhances the accuracy and timeliness of our analysis. This unique combination not only captures customer sentiments but also provides insights into evolving trends, making our approach particularly suited for dynamic and fast-paced market environments.

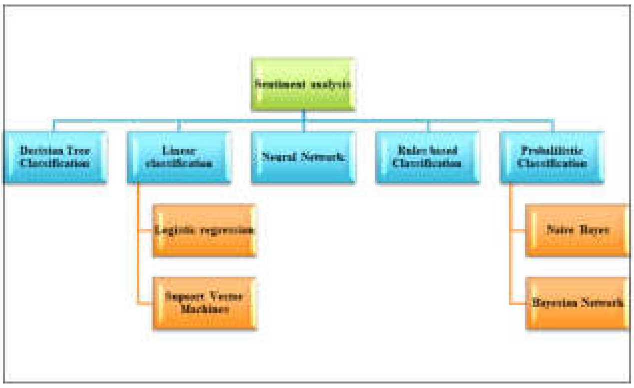


Fig. 4.1: Using different Deep learning classifications in the sentiment analysis process

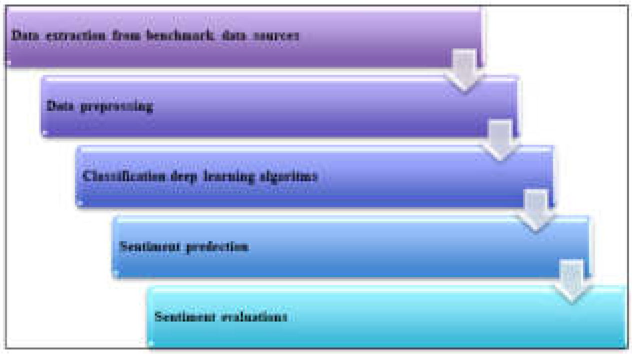


Fig. 5.1: Process of sentiment analysis through deep learning by using DSN

4. Concept of Sentiment Analysis of Online Product reviews in DBS. Sentiment analysis is an opinion-mining process that helps to understand the expression of human beings by analyzing text. By using this process, an organization can understand its product quality by analyzing the customer's reviews. The DSN helps an organization extract text from online reviews about products and all the extracted data is used as a data set [23, 24]. Through the help of the DSN networking system, all the data transfers to the deep learning aspects where different types of algorithms analyze those texts in terms of human beings' expression.

Figure 4.1 provides information about different types of deep learning processes used for sentiment analysis to analyze online product reviews. Here four different deep learning processes are used that are related to the decision tree classification, neural network, linear classification, and rules-based classification. On the other hand, logistic regression and SVM are used for the linear classification deep learning process to analyze human beings' expressions through textual language [12].

5. The Five Main Stapes of Sentiment Analysis of Online Product Reviews Through Deep Learning in DSN. In sentiment analysis, five steps help an organization evaluate sentiment from online product reviews.

Figure 5.1 provides information about the steps to complete sentiment analysis. In the first step, data is extracted from different benchmark data sources by using live API. This live API allows extracting data from the online platform of the organization. After extracting data, DSN is used to prepare data and all the extracted keywords are transferred to the deep learning process through open networking. After the transferring data, deep learning classification has been seen and different algorithms are used for sentiment prediction [18]. As per the expression of human beings, deep learning categorized the reviews in terms of positive, negative and

Table 6.1: Benefits of sentiment analysis through deep learning and DSN on online product reviews

Serial No.	Benefits
1	Help to understand the customer's perception of products
2	To get information on current market status from customers' perceptions.
3	It improved the crisis management system for an organization [22]
4	Help to understand the brand strength
5	To make a sales and marketing plan

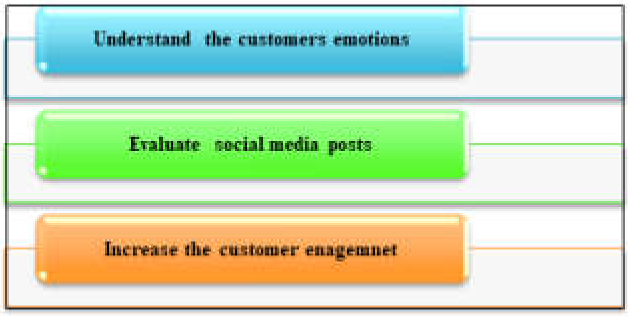


Fig. 6.1: benefits get by an organization through sentiment analysis on online product reviews

neutral. In terms of deep learning classification

6. Benefits of Sentiment Analysis of Online Products Reviews by Using Deep Learning in DSN. There are several benefits seen due to the sentiment analysis of different product reviews through the use of deep learning and DSN. Through the help of sentiment analysis, an organization understands correspondence concerns about the products from the customers. By understanding the perception of human beings about products, organizations can understand the value of products [8].

The above table provides information about the benefits of this analysis process. Here, DSN helps to extract data from the specific platform and all the extracted data analysis in terms of positive and negative polarity. Through the help of this analysis process, an organization can maintain its brand reputation by making improvements to its products. Here, online reviews are analyzed through the decision tree algorithm that analyzes the contextual expression of human beings [25]. By analyzing reviews, an organization can understand the consumer's perspective about their products which brings enhancement to the product's quality.

Figure 6.1 provides information about the benefits an organization gets by analyzing online reviews of its products. Sentiment analysis brings enhancement to customer engagement by fulfilling the customer extraction. This analysis process provides practices to understand human beings' emotions for products [18].

7. The Role of Deep Learning and DSN in Online Product Reviews through the Sentiment Analysis Process. In sentiment analysis, deep learning and DSN play a crucial role. Here these two advanced technologies provide their contribution to the data fusing and collaboration segment. On the other hand, the deep learning process provides effective practices to understand human beings' expressions through contextual segments.

Table 7.1 provides information about the role of the deep learning process and DSN in the sentiment analysis process. Here, this analysis helps in the textual analysis and classification segment. Through the classification, organizations can categorize online reviews in terms of neutral as well as positive and negative feedback. By understanding the customer's negative reviews, an organization can make improvements to products that help to increase customer engagement [16]. On the other hand, DSN helps to extract human beings' expressed data and transfer it to the deep learning process where different types of linear algorithms are used to provide accurate information about the products to the organization. The deep learning processes understand the value

Table 7.1: Role of DSN and deep learning in the sentiment analysis for online product reviews

Serial No.	Importance of deep learning and DSN
1	Textual analytics and classification
2	Understand the Value of expression
3	Help to understand the market trend by understanding the customers' perspective on products [11]

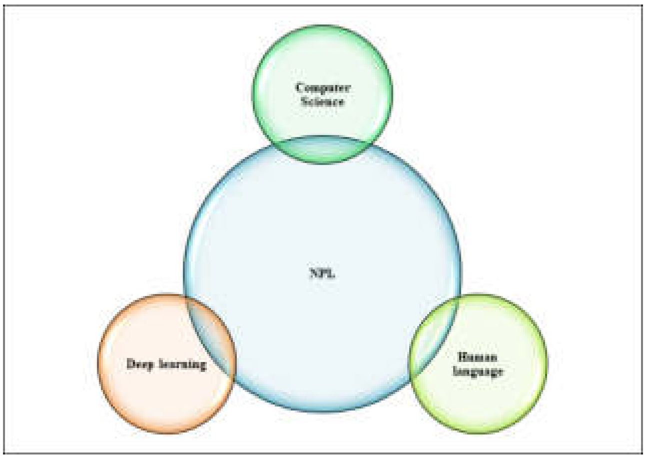


Fig. 7.1: Natural language processing (NLP) in sentiment analysis

of human beings that bring revolutionary changes to forecast the market trend.

Figure 7.1 offers information about the NPL sentiment analysis that provides organizational practices to analyze human beings' opinions about products. In NPL, DSN provides a set of data to the Deep learning process which helps in polarity classification. In terms of current market analysis, sentiment analysis plays a vital role. Here, DSN transferred market-related data to machine learning and all the data was analyzed through linear algorithms that provide effective market forecasting to the organization [9]. On the other hand, through deep learning, an organization can analyze large amounts of valuable data in a fraction of a second through an automatic process. Here this analytic process

8. Challenges Raised during Sentiment Analysis of Online Product Reviews in DSN. In sentiment analysis, there are different types of issues.

Table 8.1 provides information about the challenges raised in sentiment analysis through deep learning and DSN. In the review, if negations words such as cannot, never, and were not then deep learning classification did not respond as per the words as a result, the sentiment analysis did not happen in the proper ways. On the other hand, sometimes due to the employee's bias, barriers are raised to the sentiment analysis of online product reviews for an organization [19].

9. Effective Strategies to Increase the Use of Deep Learning in Online Product Reviews in DSN. In terms of mitigation of issues, an organization should take some effective initiatives that will increase the effectiveness of sentiment analysis for online reviews on products.

The above image provides information about the strategies taken by an organization to increase the effectiveness of sentiment analysis. Here, the organization should implement effective sentiment analysis software to understand the difficult written word. On the other hand, the implementation of topic-based sentiment analysis provides effective round-based analysis that understands complex reviews of products [10]. Through this initiative, an organization can mitigate polarity issues that are raised during sentiment analysis. To solve emoji issues, an organization should use an improved API. Through effective API, deep learning can understand different types of special characters as well as emojis. By providing training in different algorithm models,

Table 8.1: Challenges faced raised in sentiment analysis for products reviews

Factor	Issue
Tone	The deep learning process faced issues during sentiment analysis due to interpreting verbally. Here, due to the large volume of data, sometimes, the deep learning process did not recognize subjective as well as objective responses in the online product reviews.
Polarity	Some words such as hate and love provide negative and positive polarity scores. If there are any confusing words seen in the data set then sentiment analysis did not happen effectively.
Emojis	In the online reviews, if there are any emojis seen then it creates barriers to the analysis process. Here NPL provides a language-specific analysis process as a result; sentiment analysis did not happen positively [20].
Idioms	Deep learning and artificial intelligence did not understand figures of speech as a result, when idioms were used in online product reviews, deep learning programs did not understand the expressions of human beings.

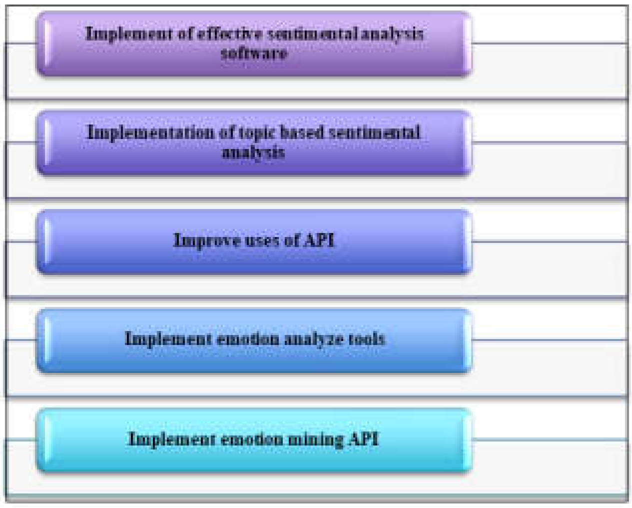


Fig. 9.1: Recommended strategies to increase the effectiveness of sentiment analysis

an organization can increase the functionality of sentiment analysis that provides analysis of human beings' expressions about the organizational products. By analyzing product value, an organization can understand the importance of products in the market that will help to get competitive advantages in the current market.

To mitigate the negation words issue, the organization should provide training to understand the double negation words as a result, understanding the human being expressed through textual context has been seen [26]. On the other hand, text analytics helps an organization reduce biases. Here, this analytical tool can understand the real emotions of employees through the help of a feedback system that eliminates human beings' errors.

10. Result. Distribution sensor network increases the sales of different organizations. Deep learning and analysis process helps recognize the customers' satisfaction level. Sentiment analysis is essential to know the honest feedback of the users. Sentiment analysis is important for all organizations as this represents the satisfaction level of the customers. The results and reviews of the customer have been used as the resources of the analysis. These reviews help to detect the quality of the different products and services that satisfy the customers. According to the view of [11], customer satisfaction could be analyzed using the machine and proper algorithm using online reviews. Online reviews suggest product quality and new buyers should check all the reviews before buying any type of commodity. Based on the views of [2], understanding of all types of

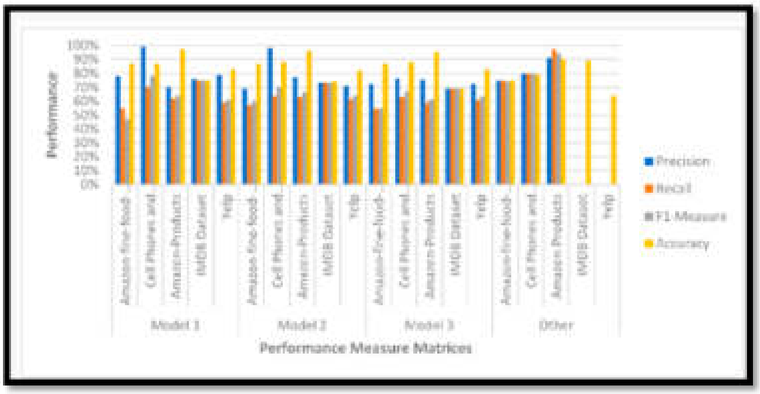


Fig. 10.1: Performance measure matrices

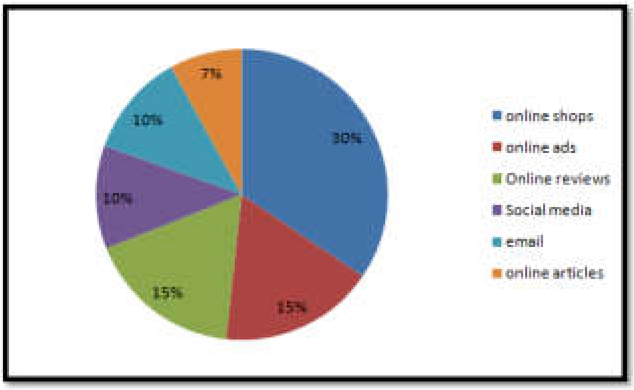


Fig. 10.2: Importance of different sources of reviews

reviews helps the organization to bring new modifications for increasing the quality of the products. This helps the organizations to understand the problems that face the customers and through understanding the reviews received from the online sources.

The above figure 10.1 has represented the precision, recall, F1 measure and accuracy. All these things are related to customer feedback on the different products of these companies. Precision values of the organizations represent the preciseness of the work and they are used to develop the product through maintaining the quality. Based on the views of [7], sentimental analysis totally depends on the nature of products and such as the quality of the products. The preciseness of Amazon fine food is 78% and the cell phone companies are 100%. This demarcates the satisfaction level of the employees. The main criteria of sentiment analysis are to examine the different openings of the customer from different sources. Customer satisfaction is demarcated as the most essential component for an organization. This represents the results of the product that is produced by the organization. The organizations are used to analyze this to change the things on the product based on the negative reviews on the product.

The figure represents the importance of different sources after analysis. Online shops are used by 30% of customers. The other sources are online advertisements which are used by the existing customers 15%, online reviews 15%, 10% from social media, email 10%, and 7 % by online articles. This denotes that the sources of the information used by the customers are not equally valued. The importance of different sources of information about the products are online shops, online advertisements, reviews, email, online articles, and email. People used to verify the quality of products by understanding the reviews. According to the views of

[14], customer satisfaction and emotions are related to the quality and longevity of the products. The sales and the brand value of the organization depend on the sales, support, and quality of the products. Based on the views of the organizations used to analyze their review to create the changes and solve the issues faced by the customers. This process of experiment helps the organization to determine the happiness and satisfaction of the customers. The deep learning on this content helps analyze the actual impact of the organizational development and the market share. According to the views of [15], understanding the customer's feedback is important for an organization to increase its market share. This process helps different organizations to analyze the impact of their products on the customer. Good customer reviews help motivate customers to buy high-rated products. Based on the views of [7], analysis of customer satisfaction is complicated as this includes negative as well as positive reviews. This study has found that the satisfaction level could be demarcated from the different sources and used by the organization for understanding their performance.

11. Problem Statement. Organizations should enforce effective API as well as proper sentiment analysis software to get effective analysis processes for online reviews from customers and stakeholders [6]. On the other hand, the real-time process of data is another issue associated with this study that mitigates through the help of the implementation of NPL. Analysis of online product reviews is one of the major and crucial tasks for a particular business that helps to understand the value of products as well as the current trend of data. In recent days, the increasing volumes of online reviews on products created barriers to the smooth sentiment analysis process by using deep learning and DSN [4]. Here, the due to improper infrastructure creates barriers to sentiment analysis. The main issue is to analyze a large volume of data with comparative sentences. Computational linguistics also helps to solve issues of DSN during transferring large volume data that enriches the effectiveness of sentiment analysis for online product reviews, through sentimental analysis, an organization can understand the customer expectation on its products that can improve the sales activities for the organization.

12. Conclusion. Sentiment analysis emerges as a transformative analytical process that revolutionizes the evaluation of product reviews for organizations. This method unfolds through five pivotal steps, guiding the meticulous analysis of textual content present in reviews. The essence lies in comprehending the perspectives of individuals, enabling organizations to glean profound insights into customer sentiments regarding their offerings. Unraveling these sentiments involves the judicious application of diverse algorithms, notably decision tree algorithms and Support Vector Machines (SVM). These sophisticated computational tools facilitate the dissection of customer feedback, delineating opinions into the dimensions of positivity, negativity, and neutrality. Such insightful categorization empowers organizations with a comprehensive understanding of their customers' nuanced feedback, transcending the conventional binary categorization.

The ramifications of sentiment analysis are far-reaching, underpinning several compelling benefits for organizations. Notably, it offers a panoramic view of customers' viewpoints, paving the way for product enhancement and strategic improvements. Moreover, this analysis serves as a lighthouse, guiding crisis management endeavors through the early detection of concerns and grievances. The strategic foundation is further fortified through the construction of effective marketing blueprints, intricately aligned with the tenor of customers' sentiments.

Acknowledgement. 1. Anhui Province 2011 Provincial Natural Science Research Project: Research on high-dimensional time series pattern data mining method No. KJ2011Z380.

2. The Scientific Research Programs at Universities in Anhui Province (Grant No.2022AH051749)

REFERENCES

[1] E. S. Alamoudi and N. S. Alghamdi, *Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings*, Journal of Decision Systems, 30 (2021), pp. 259–281.

[2] M. Andronie, G. Lăzăroiu, M. Iatagan, C. Uță, R. Ștefănescu, and M. Cocoșatu, *Artificial intelligence-based decision-making algorithms, internet of things sensing networks, and deep learning-assisted smart process management in cyber-physical production systems*, Electronics, 10 (2021), p. 2497.

[3] 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.

[4] M. Berrimi, M. Oussalah, A. Moussaoui, and M. Saidi, *Attention mechanism architecture for arabic sentiment analysis*, ACM Transactions on Asian and Low-Resource Language Information Processing, 22 (2023), pp. 1–26.

[5] S. Chatterjee, D. Goyal, A. Prakash, and J. Sharma, *Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application*, Journal of Business Research, 131 (2021), pp. 815–825.
[6] D. Cirqueira, F. Almeida, G. Cakir, A. Jacob, F. Lobato, M. Bezbradica, and M. Helfert, *Explainable sentiment analysis application for social media crisis management in retail*, (2020).
[7] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, *Sentiment analysis of persian movie reviews using deep learning*, Entropy, 23 (2021), p. 596.
[8] R. de Kervenoael, R. Hasan, A. Schwob, and E. Goh, *Leveraging human-robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots*, Tourism Management, 78 (2020), p. 104042.
[9] S. S. Deshmukh, *Progress in machine learning techniques for stock market movement forecast*, in Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022), vol. 105, Springer Nature, 2023, p. 69.
[10] A. M. Jacks, *A Study into How Elementary School Principals Across Virginia Reduce or Eliminate Exclusionary Discipline for Students with Disabilities*, PhD thesis, Virginia Tech, 2019.
[11] P. K. Jain, R. Pamula, and G. Srivastava, *A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews*, Computer science review, 41 (2021), p. 100413.
[12] M. L. Joshi and N. Kanoongo, *Depression detection using emotional artificial intelligence and machine learning: A closer review*, Materials Today: Proceedings, 58 (2022), pp. 217–226.
[13] A. O. U. Kalu, L. C. Unachukwu, and O. Ibiam, *Accessing secondary data: A literature review*, (2019).
[14] K. Kottursamy et al., *A review on finding efficient approach to detect customer emotion analysis using deep learning analysis*, Journal of Trends in Computer Science and Smart Technology, 3 (2021), pp. 95–113.
[15] S. Kumar, M. Gahalawat, P. P. Roy, D. P. Dogra, and B.-G. Kim, *Exploring impact of age and gender on sentiment analysis using machine learning*, Electronics, 9 (2020), p. 374.
[16] H. Li, K. L. Xie, and Z. Zhang, *The effects of consumer experience and disconfirmation on the timing of online review: Field evidence from the restaurant business*, International Journal of Hospitality Management, 84 (2020), p. 102344.
[17] J. Luo, S. Huang, and R. Wang, *A fine-grained sentiment analysis of online guest reviews of economy hotels in china*, Journal of Hospitality Marketing & Management, 30 (2021), pp. 71–95.
[18] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, *Survey on ai-based multimodal methods for emotion detection.*, High-performance modelling and simulation for big data applications, 11400 (2019), pp. 307–324.
[19] M. S. Neill and S. A. Bowen, *Ethical listening to employees during a pandemic: new approaches, barriers and lessons*, Journal of Communication Management, 25 (2021), pp. 276–297.
[20] N. J. Prottasha, A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, and M. Baz, *Transfer learning for sentiment analysis using bert based supervised fine-tuning*, Sensors, 22 (2022), p. 4157.
[21] A. M. Rahat, A. Kahir, and A. K. M. Masum, *Comparison of naive bayes and svm algorithm based on sentiment analysis using review dataset*, in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), IEEE, 2019, pp. 266–270.
[22] G. A. Ruz, P. A. Henríquez, and A. Mascareño, *Sentiment analysis of twitter data during critical events through bayesian networks classifiers*, Future Generation Computer Systems, 106 (2020), pp. 92–104.
[23] M. Safaei, S. Asadi, M. Driss, W. Boulila, A. Alsaeedi, H. Chizari, R. Abdullah, and M. Safaei, *A systematic literature review on outlier detection in wireless sensor networks*, Symmetry, 12 (2020), p. 328.
[24] 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.
[25] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, *A survey on concept-level sentiment analysis techniques of textual data*, in 2021 IEEE World AI IoT Congress (AIIoT), IEEE, 2021, pp. 0285–0291.
[26] B. Zhang, J. Zhu, and H. Su, *Toward the third generation artificial intelligence*, Science China Information Sciences, 66 (2023), p. 121101.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 1, 2023
Accepted: Sep 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1127–1136, DOI 10.12694/scpe.v24i4.2433

CLINICAL INDICATOR ANALYSIS FOR PREDICTING PATHOGENIC PNEUMONIA INFECTION IN NEWBORNS WITH DISTRIBUTED SENSOR NETWORKS DATA ANALYTICS

YA ZHANG* AND WENBIN DONG†

Abstract. Neonatal infections are prevalent as newborn children are prone to a range of infections because of their absence of immunity. Being susceptible to various diseases, the immune system of infants is not adequately developed to fight against protozoa bacteria, viruses and parasites. Pathogenic pneumonia is one of the most common infections identified within the neonatal group. It is a lung infection occurring in the neonates, which can start after a few hours of delivery or even after a week. The infection can even occur due to the normal flora found in the genital tract of the mother, and the respiratory distress caused by pathogenic pneumonia can even lead to the infant's death. The study examined the clinical indicator for assessing the occurrence of pathogenic pneumonia in infants. With the help of sensor networking in data analytics, the prediction of such a disorder has been assessed in-depth in the article.

Key words: Sensor Networks, Neonates, Pneumonia, Data Analytics, Prediction of Disease

1. Introduction. Digital disruption has been a major advancement in the medical field, which has led to the growth of disease identification and increased the efficiency of the workers in the sector. With the help of digital transformation within the healthcare departments, improvement of the experience of the patients has been enabled. Such an aspect not only increases the overall demand of such a sector, but also increases the trust and loyalty of the consumers [2]. Making greater discoveries and innovation of drugs through repeated trials and examining the results with the advanced hardware and software has been made feasible due to the inculcation of digital means into the healthcare sector.

The above mentioned figure illustrates the different kinds of tools and digital technology applied in the healthcare system to increase the overall efficiency of the industry [10]. Automation, big data, artificial intelligence, the Internet of Things, virtual reality, and telemedicine have been integrated into the sector to increase workflow and provide better communication between doctors and patients. Data analytics have played a vital role in predicting diseases at the early stages, which increases the overall potential of the medical procedure to aid the patients.

The vulnerability of newborns to a range of infections due to their underdeveloped immune systems is a pressing concern in the field of neonatal healthcare. The lack of immunity against various pathogens makes neonates susceptible to infections that can have severe consequences on their health and survival. Among these infections, pathogenic pneumonia in neonates stands out as a critical issue. Pathogenic pneumonia can arise shortly after birth or within the first week, leading to significant respiratory distress and potential fatality. This pressing challenge necessitates a deeper understanding of the clinical indicators and predictive measures that can aid in the early detection and management of pathogenic pneumonia.

The gravity of the situation calls for research efforts that not only delve into the clinical indicators of this infection but also explore innovative approaches to predict its occurrence. The integration of sensor networking and data analytics presents a promising avenue to enhance our understanding of the complex interplay of factors leading to pathogenic pneumonia in neonates. By leveraging technological advancements, researchers and healthcare professionals can potentially develop early detection systems that aid in timely intervention, thereby reducing the morbidity and mortality associated with this infection. The potential impact of such

*Southwest Medical University ,Luzhou,646000,China, email: yazhangres21@outlook.com
†Southwest Medical University ,Luzhou,646000,China

Fig. 1.1: Different digital mediums used in healthcare

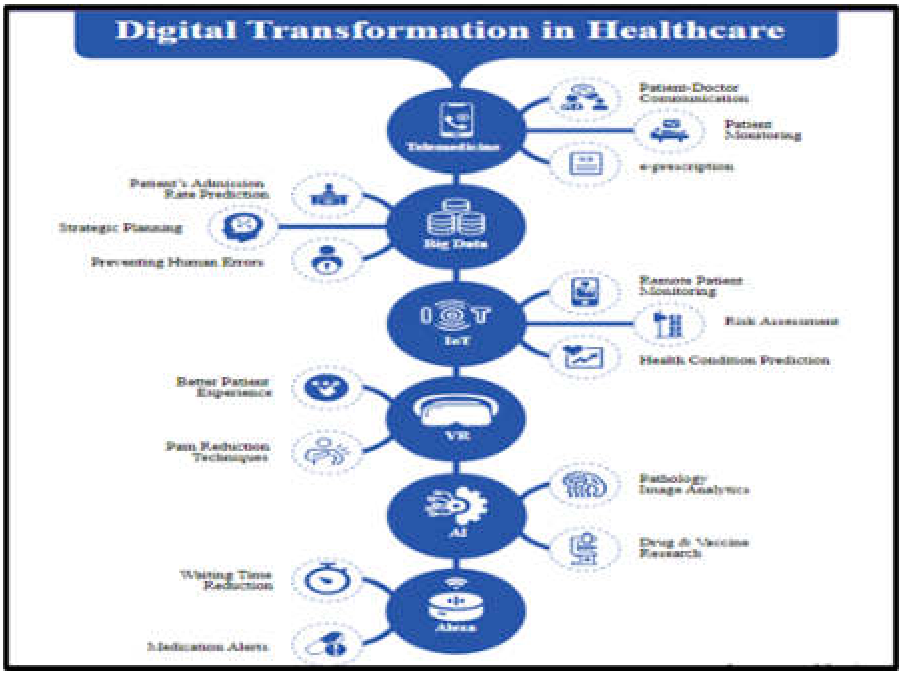


Fig. 2.1: Impact of digital transformation in healthcare

research is significant, as it addresses a critical gap in neonatal care and contributes to the well-being of the most vulnerable members of our population.

2. Objectives. The objectives which have been developed for the study are as follows:
1. To examine the occurrence of pathogenic pneumonia infection in neonates
2. To inspect the sensory networks developed through data analytics
3. To analyze pathogenic pneumonia prediction in newborns with sensor networking
4. To scrutinize the benefits which are achieved through early disease prediction with sensor networking

This work contributes to the field in several significant ways, showcasing both novelty and potential impact.
1. The research recognizes the transformative potential of digital mechanisms in healthcare systems, such as artificial intelligence, machine learning, and robotics. However, its unique contribution lies in expanding these digital advancements to the specific domain of neonatal healthcare. By applying sen-

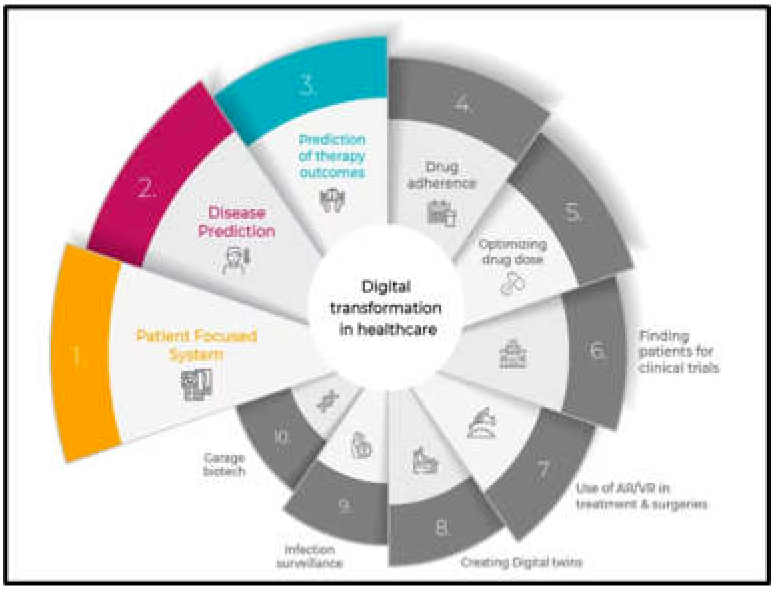


Fig. 3.1: Benefits of digital methods in medical system

sor networking and data analytics to predict pathogenic pneumonia, the study advances the use of cutting-edge technology to address a critical healthcare challenge mainly affecting the most vulnerable population.

2. A noteworthy contribution of this work is its innovative approach to disease prediction in neonates. By utilizing past patient data, the study introduces a predictive model that aids in understanding disease progression and patterns. The novelty here is the application of such predictive modeling to neonatal infections, particularly pathogenic pneumonia. This can lead to early interventions, improving treatment outcomes, and reducing newborn mortality rates.
3. The research further extends its contribution by proposing the integration of sensor networks into the healthcare framework. This approach involves collecting real-time patient data through sensors, transmitting it to cloud-based systems, and leveraging data analytics for comprehensive interpretation. Applying this sensor-based approach tailored explicitly to neonatal pneumonia is novel and offers a novel perspective on disease management.
4. The study goes beyond conventional diagnostic methods by incorporating deep learning techniques. Integrating deep learning tools like DenseNet-121 for multimodal data analysis allows for a sophisticated and nuanced assessment of patient conditions. This represents a significant advancement in diagnostic capabilities, particularly in identifying the patterns associated with pathogenic pneumonia.
5. The research addresses a gap in neonatal healthcare by providing a tailored solution for predicting pathogenic pneumonia. Neonates have distinct healthcare needs, and this work recognizes that by designing a prediction model specific to their vulnerabilities. The proposed methodology could serve as a blueprint for other neonatal healthcare predictive models, further advancing the field of neonatology.

3. Methodology. A range of digital mechanisms have been integrated into the healthcare system, increasing the industry's overall efficiency. For instance artificial intelligence and machine learning help doctors to be provided with large amounts of analyzed data, which can recognize the patterns for improving the overall outcome of the healthcare system [4]. On the other hand, robotics and three-dimensional printing have also enabled the creation of prosthetics and implants, which can assist patients and doctors alike [5]. Optimisation of the workflow within the industry along with reduced expenses due to the inculcation of a larger number of healthcare workers, are produced as a result of digital transformation.

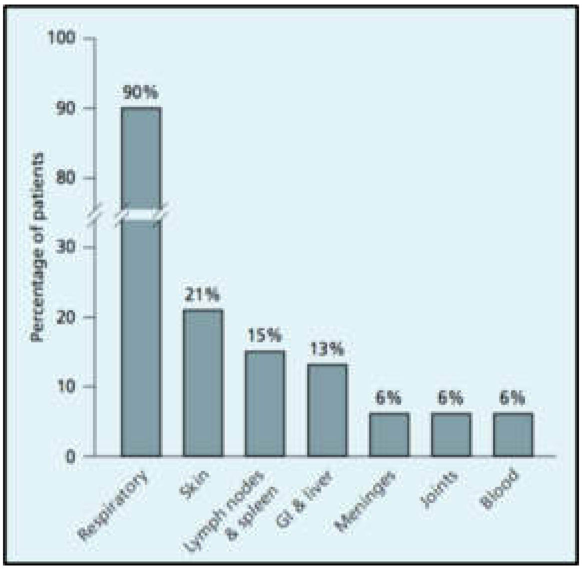


Fig. 4.1: Range of infections occurring in neonates

One of the vital advantages achieved with the inclusion of digital transformation is disease prediction through the assessment of past data from the patients [11]. By understanding the patterns and examination for the progression of the disease, data analytics has enabled researchers and doctors to gain a better understanding of the occurrence of the disorder within the patients. In such a manner, the alterations within the medical journey for the patients in terms of medicine and tests can be easily performed by the doctors, with the help of the interpreted data [3, 1].

4. Occurence of Pathogenic Pneumoniae Infection in Neonates. *Streptococcus pneumoniae* is the causative agent of pathogenic pneumonia in neonates. It has been regarded as one of the top pathogens that induce a high rate of mortality and morbidity in infants [30, 6]. The occurrence percentage of the disease has been found between 1% and 11%. However, the potential of mortality of the disease has been significantly high [7]. Two different variations of the diseases are noted to occur within the neonates, namely Early-onset pneumonia and Late-onset pneumonia. The latter group of disease is also termed ventilator-associated pneumonia because the induction of the disease is due to the elongated period of endotracheal intubation.

From the above figure, it can be assumed that respiratory diseases prove to be the highest-ranking disease sector that can impact the neonates. With about 90% of patient percentage, it is more than four times more likely to occur, than the second-ranking disease sector of skin infections [12]. The respiratory status of the patient drastically decreases with the increase in the longevity of the disorder, where respiratory secretions are observed.

The basic diagnosis of neonatal pneumonia occurs in the chest X-ray of the new needs along with the blood cultures and gram staining of tracheal aspirate. Tachypnea and chest recession also occurs in the patient, which gives a clear indication of the disease progression [8]. However, a range of issues have been identified in diagnosing neonatal pneumonia because of the lack of strict positive results. For instance, it has been recorded that the inspection by gram stain of tracheal aspirate is not an extremely prevalent test because of its decreased efficiency. It has been recorded that only about 2 to 5 percent of neonatal pneumonia cases can be identified with the help of such a test [13]. With the gram staining procedure, a lumbar puncture needs to be integrated with the medical diagnosis to increase the overall potential of the identification.

Such cumbersome procedures impact the health of the neonate because of the decreased immunity and body strength. For such reasons, disease prediction of the cases for such an age is extremely beneficial for decreasing the rate of mortality within the neonate [9].

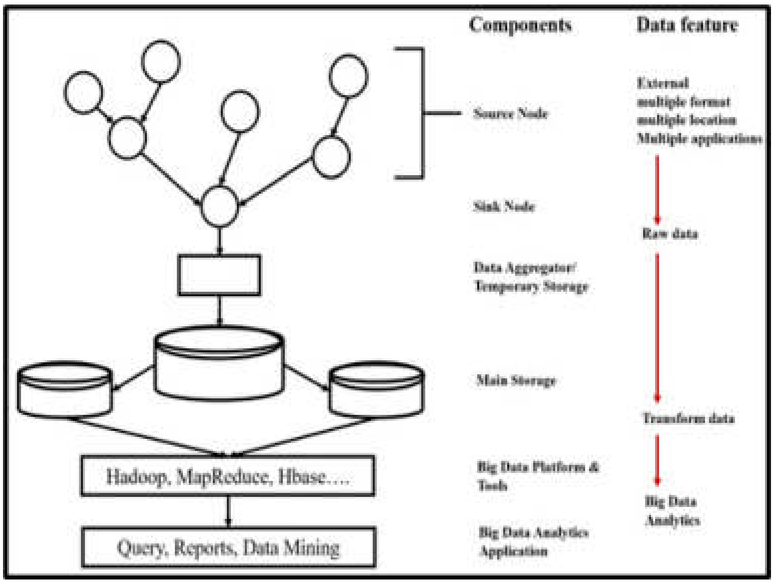


Fig. 5.1: Data transfer framework through the help of sensor data analytics

5. Sensor Networks Development through Data Analytics. Sensor data analytics is a sector of data analytics that aids in the interpretation of the past data for the generation of trends and patterns. Sensors are used for the collection of the data and information, which is sent to the cloud, for being computed and stored [14]. The analytical tool is used for the computation of the information and helps in the identification of the characteristics of the data. Through the help of such a medium, an in-depth assessment of the data can be achieved, which illustrates the linkage found between the aligning elements of the data [29].

As noted in the above image, a range of data systems are presented in data analytics for the transfer of information from one component to another. The sink node, present in the second layer of the data integration system, helps in the collection of the data from the sensor node [19]. The sensor node can obtain information from a range of sources and integrate it into the temporary storage or data aggregator present in the cloud. The manipulation of the aggregated data can be done with the help of the big data system which utilizes the chief storage for the alterations. The data which is hence transformed, are sent to big data platforms such as Hadoop and MapReduce. They find its application in the generation of queries and reports along with data mining activities.

6. Predicting Pathogenic Pneumoniae Infection in Newborns with Sensor Networks. IoT or the Internet of Things proves to be extremely necessary for the monitoring of information [31]. Through the application of data analytics, the analytical tools have the capacity of formatting and storing the data, as per the trends and patterns. Pinpointing of the different areas which can be improved with the help of data analytics can also be enabled. With the help of such a tool in healthcare, the assessment of the clinical data of the patients can be achieved. Uploading the data into the cloud server and the interpretation of the evidence can be achieved with the help of the data analytic system [24]. The software has the capacity to monitor the data against the occurring conditions, and optimise the evidence for accurate utilisation. In such a manner, the notions of asset utilisation can be achieved, where the segregated data can be interpreted by the doctors for reaching the desired outcome.

The detection strategies used for the diagnosis of pneumonia can be segregated broadly into two categories of machine learning based predictions and deep learning based classifications, as noted in the above figure 7.1 [17]. Edge computing environments have been seen to implement sensor data for increased decision making, in the healthcare system [20]. Through the collection of the signals with the help of the sensor, the information is passed onto the temporary and main storage. Network protocols such as Hypertext Transfer Protocol and

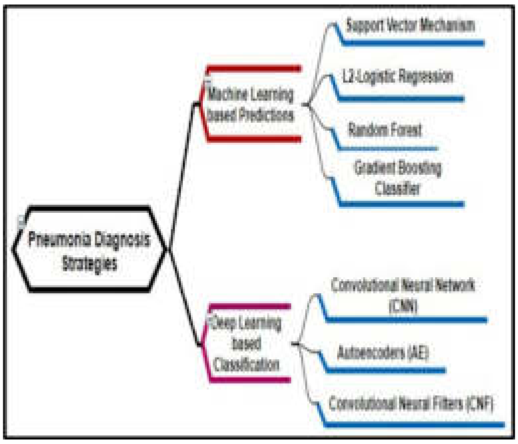


Fig. 6.1: Schematic representation of the data analytics for diagnosis

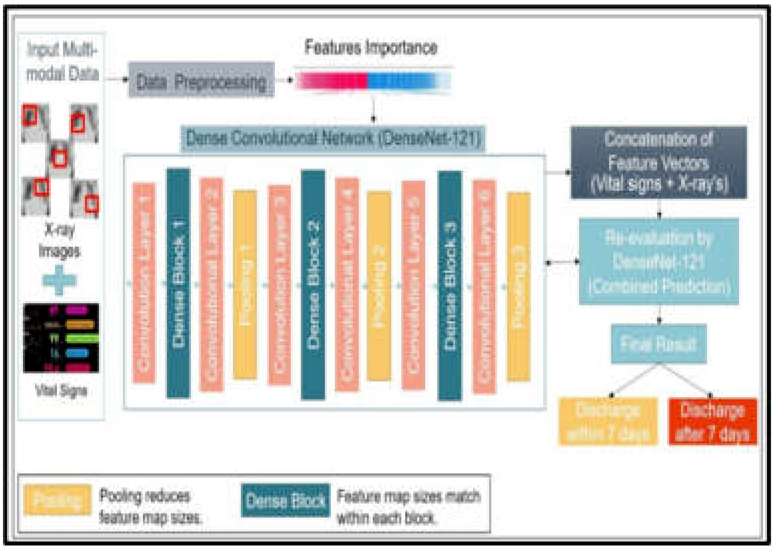


Fig. 6.2: Data transfer cycle followed in data analytics for diagnosis

Constrained Application Protocol are utilised by the data sensors for the transfer of information, which enables in the sharing of the data of the patients. One of the most successful sensors which has been applied in the detection of neonatal pathogenic pneumonia is that of multimodal data analysis sensors [15].

Pathogen detection with the help of multimodal data analysis for neonatal pneumonia is achieved due to the usage of deep learning classification. In such a case, the signs and symptoms of the neonates are integrated into the sensor which classifies the status of health for the new net against the presented information. The improvement or the deterioration of the patient's condition can be made achievable with the help of the presented data where the disease prediction and the stage can be easily identified. This is enabled due to the prevalence of data analytics within the sensor system and helps in obtaining a highly accurate prediction of the outcome of the patient [25].

A range of image extraction can also be availed with the help of data analytics through the utilization of the technique of deep learning [32]. With the inclusion of a large amount of multimodal data in the form of X-rays and the vital signals of the patient, the various interpretations can be enabled with the help of data

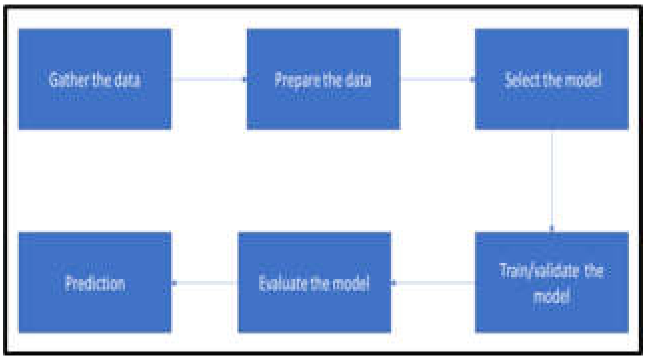


Fig. 7.1: Steps for solving of problems with data analytics

analytics through the process of data pre-processing. The key factors and the identification of the patterns can be integrated into the deer learning tool of DenseNet-121, which allows a lengthy interpretation of the data with the help of the dense layer architecture of the DenseNet-121 tool [21]. Such a digital medium provides the identification of the key patterns amongst the multimodal data and relates it with the vital signals of the neonates for examining the condition of pathogenic pneumonia infection within the patient.

7. Benefits of Early Disease Prediction with Sensor Networking. Sensor networking has been seen to provide an extensive edge in disease prediction [23]. Because of the fact that the whole system is automated, and the sensors pick up the digital information without any human intervention, the presence of human errors is absent. Such an aspect drastically increases the probability of reaching an error-free outcome that can be trusted by the doctors.

The advancement of tools and techniques for data analytics has allowed the doctors to achieve an early prediction, and with the auto-generated outcomes, the changing of the medical procedures for decreasing the mortality of the patients can be obtained [33]. Such a fact proves to be extremely desirable for increasing the overall experience of the patients and for inducing a greater chance of reaching the desired outcome [28].

8. Result. Artificial intelligence also enables the doctors to increase their evaluation frequency through the integration of data within the digital models and inducing data analytics for altering medication and method of assessment. Image processing and feature extraction of the X-rays and the symptoms respectively can be included into data analytics by the doctors for understanding the patterns within the new model infection and the alterations within the neonates, as per the stage of infection [16]. Several variables are identified using categorical grouping within the multimodal system and a range of tests such as univariate and multivariate regression can be performed with the data analytics system. Such an aspect is induced with the help of the pattern recognition and interpolation of the data within the data analytics system can be performed for algebraic operations [26]. The utilisation of computer vision operations can also be a major source of data for the sensors which helps the doctors to understand the exact condition of the stage of infection as per the results obtained by the tool performing data analytics.

As noted in the above figure, a large number of variables that are included within the data analytics are integrated for the correct interpretation of the characteristics of the neonates [22]. For instance, the age group of the patients is included within the data analytics for understanding the exact age bracket for the occurring symptoms. The extent of susceptibility of the new need to pneumonia infection is identified based on the age group of the parents, the presence of impact immunity, or the occurrence of chronic respiratory diseases. The prevalence of other infectious agents such as viruses and pathogens are also examined within the neonate to be included in the speed of data analytics. The clinical phenotype and the presence of causative pathogen for pneumonia, falling under the primary findings of the doctor are included along with the blood test reports [18]. Through the inculcation of such a range of data within the data analytic system, the doctors have the capacity to understand the extent to which the pathogenic pneumonia infection within the neonate can occur.

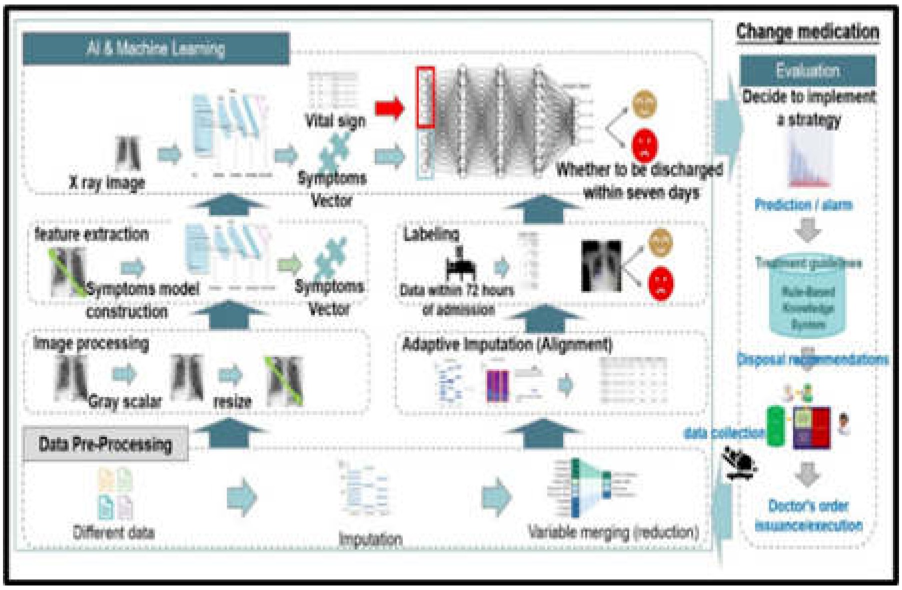


Fig. 8.1: Relationship between AI, Machine learning, and alliteration of medical procedures for the neonates

The usage of such variables which act as clinical indicators proves to be extremely necessary as the primary data being integrated acts as the basis upon which data analytics with the help of the sensors, would occur.

The ethnicity of the patients along with the presence of smokers in the household have been identified to have an impact on the probability of the neonate being born with pneumonia. On the other hand, the presence of impaired immunity and the prevalence of hospital-related respiratory diseases have also been identified to have a direct effect on the generation of neonatal pneumonia. Through the integration of such acute information within the sensor of data analytics, examination of the probability of the neonate being affected by pneumonia can be enabled. Such an aspect is recorded within the data analytics system to provide an in-depth examination of the occurrence of neonate pneumonia [27]. With the rise in the probability of the trends and patterns seen in the data, the changing of medications from an early hour can be enabled by the doctors.

The incorporation of the data such as the presence of influenza in larynx, the involvement of infection in the upper, the presence of causative pathogen for pneumonia in the pleural effusion, along with the blood test results examined in an extensive manner within the digital system for reaching a confirmed result. The integration of such data from the sensors of the data system enables the smooth computational flow of information and evidence to the data analytics system. With the occurrence of an in-depth interpretation of the aligning variables of the study, the doctors have the probability of understanding the presence or absence of neonatal pneumonia.

9. Problem Statement. One of the key limitations that has been identified in the article is the lack of financial information regarding the inclusion of sensor networking in the medical chains. Due to the fact that the induction of digital transformation withdraws a significant amount of finance from the facilities, the exact budget needed for setting up such a facility within the medical unit has not been jotted down in the study.

10. Conclusion. Hence, the study largely focused on the usage of digital transformation within the healthcare sector for achieving improved patient experience and helping in pathology image analytics. Risk assessment and remote patient monitoring by the reduction of human errors by doctors have also been identified to play major roles because of the integration of digital transformation within the healthcare sector. The inclusion of sensor data analytics for the prediction and confirmation of pathogenic pneumonia within newborns has been examined in an extensive manner. The utilization of clouds for the storage of data, the capturing of information with the help of sensors, and the provision of computer information as an output of data analytics have been analyzed throughout the length of the study.

It's important to note that while sensor networking and data analytics offer valuable insights into predicting pathogenic pneumonia in neonates, the accuracy and reliability of such predictions may be influenced by factors such as the availability and quality of data, the diversity of pathogens, and the variability in individual neonatal responses to infections. Further research can explore the integration of advanced diagnostic techniques, such as genetic testing and rapid pathogen identification, to enhance the precision of predicting pathogenic pneumonia in neonates. Additionally, investigating the effectiveness of early intervention strategies based on the predictive models could contribute to reducing the impact of this life-threatening infection on newborns.

REFERENCES

[1] 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.

[2] L. Cadot, H. Bruguière, E. Jumas-Bilak, M.-N. Didelot, A. Masnou, G. de Barry, G. Cambonie, S. Parer, and S. Romano-Bertrand, *Extended spectrum beta-lactamase-producing klebsiella pneumoniae outbreak reveals incubators as pathogen reservoir in neonatal care center*, European Journal of Pediatrics, 178 (2019), pp. 505–513.

[3] D. Dai, H. Chen, X. Dong, J. Chen, M. Mei, Y. Lu, L. Yang, B. Wu, Y. Cao, J. Wang, et al., *Bronchopulmonary dysplasia predicted by developing a machine learning model of genetic and clinical information*, Frontiers in Genetics, 12 (2021), p. 689071.

[4] X. Dong, P. Zhou, and G. Zhu, *Common pathogens and drug resistance of neonatal pneumonia with new multichannel sensor*, Contrast Media & Molecular Imaging, 2022 (2022).

[5] P. Fang, K. Gao, J. Yang, T. Li, W. Gong, Q. Sun, and Y. Wang, *Prevalence of multidrug-resistant pathogens causing neonatal early and late onset sepsis, a retrospective study from the tertiary referral children's hospital*, Infection and Drug Resistance, (2023), pp. 4213–4225.

[6] L. P. Garcia, I. J. C. Schneider, C. de Oliveira, E. Traebert, and J. Traebert, *What is the impact of national public expenditure and its allocation on neonatal and child mortality? a machine learning analysis*, BMC Public Health, 23 (2023), pp. 1–11.

[7] J. Glaser and S. Shaw, *Digital transformation success: what can health care providers learn from other industries?*, NEJM Catalyst Innovations in Care Delivery, 3 (2022).

[8] G. Gopal, C. Suter-Crazzolara, L. Toldo, and W. Eberhardt, *Digital transformation in healthcare–architectures of present and future information technologies*, Clinical Chemistry and Laboratory Medicine (CCLM), 57 (2019), pp. 328–335.

[9] S. Hermes, T. Riasanow, E. K. Clemons, M. Böhm, and H. Krcmar, *The digital transformation of the healthcare industry: exploring the rise of emerging platform ecosystems and their influence on the role of patients*, Business Research, 13 (2020), pp. 1033–1069.

[10] S. A. G. Kaabi and H. K. Musafer, *An experimental mouse model for phage therapy of bacterial pathogens causing bacteremia*, Microbial Pathogenesis, 137 (2019), p. 103770.

[11] K. Klinker, M. Wiesche, and H. Krcmar, *Digital transformation in health care: Augmented reality for hands-free service innovation*, Information Systems Frontiers, 22 (2020), pp. 1419–1431.

[12] A. Kouroubali and D. G. Katehakis, *The new european interoperability framework as a facilitator of digital transformation for citizen empowerment*, Journal of biomedical informatics, 94 (2019), p. 103166.

[13] S. Kraus, F. Schiavone, A. Pluzhnikova, and A. C. Invernizzi, *Digital transformation in healthcare: Analyzing the current state-of-research*, Journal of Business Research, 123 (2021), pp. 557–567.

[14] N. Kumar, G. Akangire, B. Sullivan, K. Fairchild, and V. Sampath, *Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront*, Pediatric research, 87 (2020), pp. 210–220.

[15] L. Le Guennec, M. Coureuil, X. Nassif, and S. Bourdoulous, *Strategies used by bacterial pathogens to cross the blood–brain barrier*, Cellular microbiology, 22 (2020), p. e13132.

[16] Y.-C. Liu, H.-Y. Cheng, T.-H. Chang, T.-W. Ho, T.-C. Liu, T.-Y. Yen, C.-C. Chou, L.-Y. Chang, and F. Lai, *Evaluation of the need for intensive care in children with pneumonia: machine learning approach*, JMIR Medical Informatics, 10 (2022), p. e28934.

[17] I. C. Marques and J. J. Ferreira, *Digital transformation in the area of health: systematic review of 45 years of evolution*, Health and Technology, 10 (2020), pp. 575–586.

[18] M. Massaro, *Digital transformation in the healthcare sector through blockchain technology. insights from academic research and business developments*, Technovation, (2021), p. 102386.

[19] S. MATHUR, *Prediction and spatial interpretation of meteorological factors with neonatal mortality rate*, (2022).

[20] R. M. McAdams, R. Kaur, Y. Sun, H. Bindra, S. J. Cho, and H. Singh, *Predicting clinical outcomes using artificial intelligence and machine learning in neonatal intensive care units: a systematic review*, Journal of Perinatology, 42 (2022), pp. 1561–1575.

[21] A. M. Momany, E. Jasper, K. E. Markon, M. A. Nikolas, and K. K. Ryckman, *Latent class analysis to characterize neonatal risk for neurodevelopmental differences*, Journal of Child Psychology and Psychiatry, 64 (2023), pp. 100–109.

[22] C. O'Sullivan, D. H.-T. Tsai, I. C.-Y. Wu, E. Boselli, C. Hughes, D. Padmanabhan, and Y. Hsia, *Machine learning*

applications on neonatal sepsis treatment: a scoping review, BMC Infectious Diseases, 23 (2023), pp. 1–10.
[23] K. Rassels and P. French, *Bio-remote sensing in predicting infection in neonates with thermal imaging and machine learning*, in 2022 Smart Systems Integration (SSI), IEEE, 2022, pp. 1–4.
[24] W. Ricciardi, P. Pita Barros, A. Bourek, W. Brouwer, T. Kelsey, and L. Lehtonen, *How to govern the digital transformation of health services*, European journal of public health, 29 (2019), pp. 7–12.
[25] N. Rixe, A. Frisch, Z. Wang, J. M. Martin, S. Suresh, T. A. Florin, and S. Ramgopal, *The development of a novel natural language processing tool to identify pediatric chest radiograph reports with pneumonia*, Frontiers in Digital Health, 5 (2023), p. 1104604.
[26] C. J. Saul, D. Y. Urey, and C. D. Taktakoglu, *Early diagnosis of pneumonia with deep learning*, arXiv preprint arXiv:1904.00937, (2019).
[27] R.-K. Sheu, L.-C. Chen, C.-L. Wu, M. S. Pardeshi, K.-C. Pai, C.-C. Huang, C.-Y. Chen, and W.-C. Chen, *Multi-modal data analysis for pneumonia status prediction using deep learning (mda-psp)*, Diagnostics, 12 (2022), p. 1706.
[28] K. M. Shuaib, S. A. Shahid, A. A. Javed, and M. Zaid, *Pneumonia detection through x-ray using deep learning*, IOSR Journal of Computer Engineering (IOSR-JCE), 22 (2020), pp. 08–11.
[29] J. C. Smith, A. Spann, A. B. McCoy, J. A. Johnson, D. H. Arnold, D. J. Williams, and A. O. Weitkamp, *Natural language processing and machine learning to enable clinical decision support for treatment of pediatric pneumonia*, in AMIA Annual Symposium Proceedings, vol. 2020, American Medical Informatics Association, 2020, p. 1130.
[30] 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.
[31] H.-C. Wang, C.-C. Liao, S.-M. Chu, M.-Y. Lai, H.-R. Huang, M.-C. Chiang, R.-H. Fu, J.-F. Hsu, and M.-H. Tsai, *Impacts of multidrug-resistant pathogens and inappropriate initial antibiotic therapy on the outcomes of neonates with ventilator-associated pneumonia*, Antibiotics, 9 (2020), p. 760.
[32] J. Wang, H. Zhang, J. Yan, and T. Zhang, *Literature review on the distribution characteristics and antimicrobial resistance of bacterial pathogens in neonatal sepsis*, The Journal of Maternal-Fetal & Neonatal Medicine, 35 (2022), pp. 861–870.
[33] Y. Wu, S. Mascaro, M. Bhuiyan, P. Fathima, A. O. Mace, M. P. Nicol, P. C. Richmond, L.-A. Kirkham, M. Dymock, D. A. Foley, et al., *Predicting the causative pathogen among children with pneumonia using a causal bayesian network*, PLoS Computational Biology, 19 (2023), p. e1010967.

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
© 2023 SCPE. Volume 24, Issues 4, pp. 1137–1144, DOI 10.12694/scpe.v24i4.2434

MULTIMODAL ANALYSIS OF IMAGE TEXTS IN COLLEGE ENGLISH TEXTBOOKS USING VISUAL GRAMMAR THEORY FOR REAL-TIME EMBEDED SYSTEMS

DEXIA LI*

Abstract. The existence of images in the English textbook can bring several benefits to colleges. Research has been done on multimodality to analyze the impact of the image text within English textbooks. This study uses visual grammar to identify image and text elements within textbooks. The multimodal framework is involved in "Van Leeuwen's Grammar Visual Design, which has been used in this study to identify multimodal elements that can appear in the English textbook. Image or visual elements can provide pedagogical advantages to college students to understand the text.

Key words: Visual Grammar, Multimodal Analysis, Image Texts, real-time embedded system

1. Introduction. The multimodal analysis of the text image is one of the packages of software to analyze, learn, and teach multimodal texts. Multimodal analysis is one of the key software packages in the market that not only provide theoretical understanding but also the practicability of learning and teaching in different circumstances [1]. In the case of linguistics, multimodal analysis has become one of the key approaches to global language learning. The process of multimodal analysis is found using different forms of language, such as texts, images, and other forms. However, the process of image text is found most crucial as it helps to uncover the maximum number of semiotic resources.

The employment of appealing visual features in instructional materials has become increasingly common in today's quickly expanding digital world, notably in college English textbooks. It is commonly acknowledged that combining visual graphics and textual text may considerably improve the overall learning experience and successfully assist in a better grasp of complex subjects. As a result, this paper goes into the interesting worlds of multimodal analysis and visual grammar theory, focusing on their enormous influence on analyzing picture texts found in college English textbooks. Furthermore, this investigation focuses on the amazing applicability of these analytical frameworks to the dynamic world of real-time embedded systems.

The above figure presents the multimodal analysis framework involved in the context, disclosure, and genre, including the textual and visual layout of images. The research also discusses the use of English Grammar Theory in the process of multimodal text analysis. The above image shows the different structures revolving around multimodal text analysis and the role of visual grammar in this section [3, 2].

The presence of images within college English textbooks has become increasingly recognized for its potential to enhance the learning experience. The convergence of visual and textual elements in educational materials has prompted research into multimodality, focusing on how the combination of image and text impacts comprehension and engagement. This study aims to extend the understanding of this phenomenon by utilizing the principles of visual grammar to delve into the intricacies of multimodal analysis of image-text interactions within college English textbooks. Moreover, this exploration is conducted with an awareness of the implications of real-time embedded systems, which can potentially facilitate the integration and interpretation of these multimodal elements.

The steps of the Multimodal analysis framework of the image text have been given in the above table.

The aim of this study is a multimodal analysis of the image texts in English textbooks using the theory of visual grammar. While this study has shed light on these two domains' potential synergies and applications, a more detailed exploration of how real-time embedded systems can effectively facilitate and enhance multimodal

*School of International Education, Xianda College of Economics and Humanities Shanghai International Studies University, Shanghai 200083, China (dexialiresearc1@outlook.com)

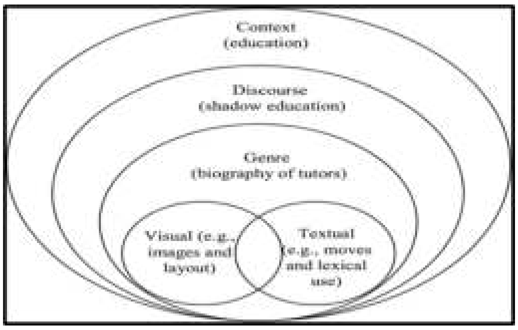


Fig. 1.1: Multimodal analysis framework of Image text

Table 1.1: Steps involved in Multimodal analysis framework of Image text

Steps	Consideration
Identification of Mode	This includes various visual modes like colors, images typography and layout
Semiotic analysis	Analysis of individual visual elements or semiotic resources like framing, composition, syntax as well as symbolism
Intermodal analysis	Exploring relationships and interactions between various modes of image text
Meaning-making	This step is related to the interpretation of meanings that are conveyed by image text and the audience has a crucial role here
Contextual factors	this Consider the cultural, social as well as historical context that can be present around the image text.
Audience analysis	This analysis reflects on how audiences can be different in terms of responding to and interpreting the image text, which depends on the personal, social, and cultural contexts.
Critical reflection	This step involves critical reflection as well as analysis of image text. Here, is an assessment of the potential implications, ethical considerations, and effects associated with the utilization of the visual modes of image text.

text analysis would provide a deeper understanding of their interplay. This could involve delving into the technical intricacies of integrating real-time processing capabilities within text analysis frameworks, ensuring seamless synchronization and efficient data flow. Additionally, investigating practical use cases where real-time embedded systems can mitigate challenges and optimize processes in multimodal text analysis, such as real-time sentiment analysis of social media streams or live transcription and translation of multimedia content, could further underscore the significance of this connection. By addressing these aspects, the study would contribute to a more comprehensive comprehension of the potential at the intersection of real-time embedded systems and multimodal text analysis.

2. Objectives.

1. To evaluate the concept of the multimodal analysis of image texts and the theory of visual grammar
2. To determine the impact of using the theory of visual grammar in the multiple analysis of image texts for English college textbooks
3. To identify the impacts of using visual learning on the “real-time embedded system” .

Motivation: The motivation behind this study lies in the growing recognition of images’ role in facilitating effective learning, particularly in the context of college English education. As students engage with complex textual content, visual elements can serve as valuable aids in comprehension. However, the nuanced ways images and texts interact require a deeper analysis considering linguistic and visual components. Furthermore, the emergence of real-time embedded systems presents opportunities to dynamically enhance the interaction

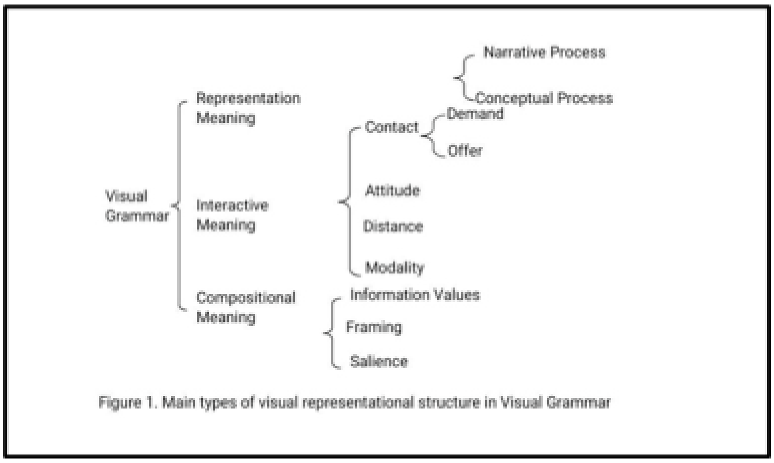


Fig. 4.1: Visual grammar framework

between image and text, potentially leading to more immersive and interactive learning experiences.

Contribution: This study contributes to the existing body of knowledge by applying the theory of visual grammar, specifically "Van Leeuwen's Grammar of Visual Design," to analyze image-text relationships within college English textbooks. By employing this framework, the study aims to systematically identify and categorize multimodal elements, shedding light on how images enhance textual understanding. Moreover, considering real-time embedded systems in this context adds a futuristic dimension, proposing ways in which technological advancements can further augment the multimodal learning experience.

3. Methodology. A descriptive research design has been employed to conduct the secondary qualitative research. This method is an effective way to code, screening as well, and interpret. This research design was in line as data have been collected, categorized as well and analyzed.

Secondary thematic analysis has been done as a data analysis process to conduct the study to determine the multimodal analysis of the image texts within English textbooks and the benefits of using the theory of visual grammar [4, 13]. Therefore, subjective data have been gathered and analyzed by making themes based on the study's objectives.

4. Concept of the Multimodal Analysis of Image Texts and Visual Grammar Theory. Based on the above figure, it can be seen that the main type of visual representation is visual grammar. The concept of multimodal analysis is embedded in the use of semiotic modes. These models are the primary factors of textbooks that work as an essential factor in analyzing image texts [7]. Due to this reason, the relationship between images and texts is one of the key things in the multimodal analysis of text images. The arrangement of their images, the position of images and texts in the visual fields, and the role of the images in the meaning-making process are very important in this case. The above picture discusses the different types of visual grammar used in standard multimodal analysis [6]. The representation of the images, the composition of the images, and the interaction between the images and the texts are the primary visual grammar used in the text analysis process. However, all these forms of images have different subsections and divisions based on the type of grammar used in the analysis process.

Understanding the intricacies of college English textbooks can be done through multimodal analysis, which is the research of various communication modalities within a given context. Along with written content, multimodal analysis also looks at images, sounds, and motions. By utilizing the concepts of visual grammar theory, researchers gain a deeper understanding of the visual components present in various communication channels. By giving researchers a framework for examining crucial elements like composition, color, perspective, symbolism, and visual metaphors, visual grammar theory enables them to understand the hidden visual cues contained in picture texts in college English textbooks.

Table 5.1: Elements of visual grammar

Elements	Description
Objects	The basic element of visual grammar can be concrete or abstract
Structures	Patterns of the visual grammar should be created based on the basic element as well as patterns
Activities	The process can be represented based on the basic element and the patterns
Relations	This includes the relationship among the process, patterns, and elements

5. Uses of Theory Visual Grammar for the Multimodal Analysis in College English Textbooks. The use of visual grammar has been found as one of the main tools for analyzing college English textbooks. The use of this technology in textbook representations is found to consist of conceptual and narrative representations [9]. The representation can be done correctly based on the structure of the display, texts, and the interrelation between texts and images. It has been found that cohesiveness is one of the key factors that can be achieved by the proper implementation of the visual grammar theory. The use of visual characteristics in college English textbooks can be better understood using the framework provided by visual grammar theory [10]. Using this approach, students can examine a variety of examples, charts, graphs, photographs, and other visual elements that are frequently included in textbooks in greater detail. By carefully examining the composition, color schemes, and visual metaphors used in these picture texts, researchers can decipher not just the intended emphasis but also the emotions and underlying concepts portrayed [11]. This thorough investigation is essential in enabling teachers to provide more useful teaching resources, which in turn aids students in deriving pertinent and significant information from the visual texts they encounter.

The assimilation between synonyms, repetition, and visual models creates a perfect cohesiveness of text and images, which helps students as well as teachers to convey the messages in the textbooks. In the case of English textbooks, it has been found that the higher the cohesiveness between the texts and images, the greater the utilization of knowledge [11]. Students can comprehend more knowledge from a textbook, which provides synchronization between the texts and the images present in the English textbooks. Multimodal analysis of picture texts using visual grammar theory has several advantages in the setting of real-time embedded systems. It enhances learners' comprehension and retention, to start. By examining the visual elements in real-time embedded systems featured in college English textbooks, students can learn difficult subjects and remember information better. Multiple entrance points for comprehension are made possible by the combination of written and visual data, suiting different learning styles.

6. Advantages of Analysis of the Image Text by using Visual Grammer Theory in case of the Real-time Embedded System. Platforms of interactive language learning can be the "real-time embedded system". This is the platform where learners can get the opportunity to engage in "real-time conversation". This system is able to analyze patterns of speech, give instant feedback as well as can offer vocabulary and grammar.

The above figure depicts the advantages of the multimodal analysis for the real-time embedded system. This system is capable of adopting the needs of an individual as well as fixing the progress of learners by monitoring the performance of learners. "Real-time embedded system", therefore can be integrated to analyze the image text for providing instant feedback and an interactive experience of learning [14]. College students can interact with the system by uploading images or capturing images of their textbooks or any specific image texts. These systems can employ the vision of the computer algorithms for processing texts and images as well as extract visual features for analysis.

The embedded system serves the pedagogical support to instructors and learners. For instance, tutorials of visual grammar can be system generated, and interactive exercises or supplementary materials can be developed based on the analysis of the image texts. Therefore, learners can access these resources anytime. On the other hand, it enables independent learning and provides various additional opportunities such as reinforcement and practice. This has advantages in the preventive algorithm in scheduling. However, it can be a disadvantage in the learning courses, as this system cannot prioritize any tasks [16]. One task can go through multiple rounds

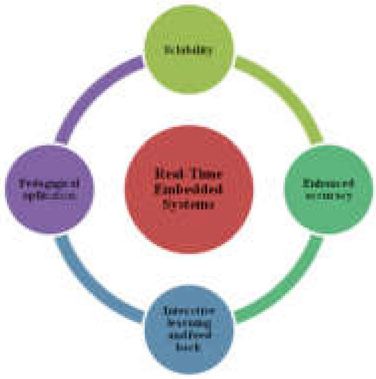


Fig. 6.1: Advantages of the real-time embedded system

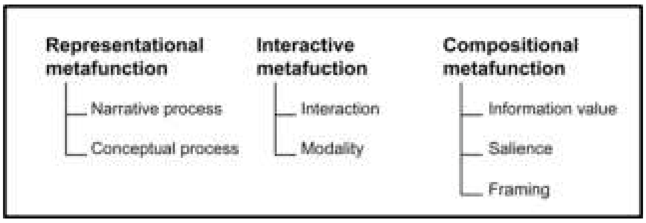


Fig. 7.1: Categories of the visual grammar theory

for completion; therefore, it can be said that this approach is relatively straightforward and easy. Improved Teaching Experience: Adding visually appealing image texts to college English textbooks can increase student engagement with the subject matter and lead to a better learning environment.

7. Theoretical Framework. Van Leeuwen's Grammar Visual Design

This framework focuses on the grammar as well as principles that are underlying visual design and communication. This theory provides a systematic analysis of the different types of visual images as well as their interaction for catering and managing within visual texts. In the case of visual grammar, this involves different categories as well as principles to analyze visual communication [18, 5]. This grammar involves the understanding of the process of the visual elements that can be combined and organized for developing and conveying a specific meaning. This includes several categories like compositional structures, representational structures as well and interactive Meta functions.

The above figure depicts that three categories of the visual grammar theory have different processes representational conceptual and narrative processes. While narrative representation shows every participant that is connected to one another by lines, which are called vectors. On the other hand, representational representation can show a stable concept.

8. Result. The visual grammar theory has been used in the image work to communicate messages just as words do. Like every language, the English language has its grammar as well as an understanding of the uses of images for communicating clearly by visual grammar [17].

The above figure depicts a relationship between teaching diversification in English and teaching effectiveness based on the above graph it can be said that diversification of the various modes of teaching as well as teaching effectiveness is more remarkable which has an impact on the rising intention of the student learning [8].

The real-time embedded system can be operated based on several principles like quick response, predictability, deadline of the task as well as operational failures. This system can work in fixed constraints of time. Apart from this system can be predictable or deterministic and does not allow any deviations. Meeting deadlines for

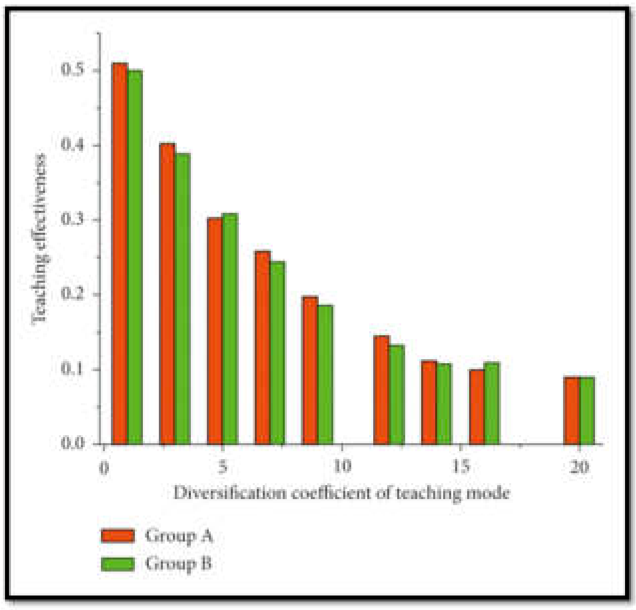


Fig. 8.1: Relationship between the teaching diversification and its effectiveness

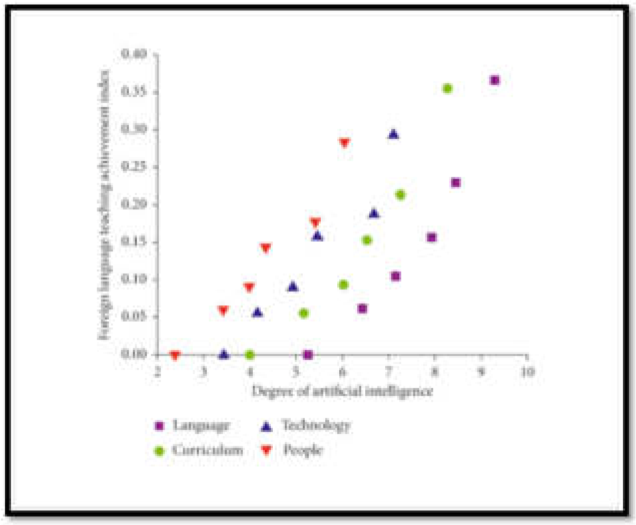


Fig. 8.2: Relationship between foreign language and uses of artificial intelligence

tasks is more crucial for this system than other characteristics. If any tasks fail to meet the deadline, then it can have a negative impact on the users as well as even lead the fatal results. Abstract objects of visual grammar are included in the lines, points, surfaces as well and volumes [18]. Volumes of the abstract objects in English learning textbooks involve three dimensions that are lines, points, and surfaces.

The above figure presents a scatter diagram to establish the relationship between the uses of AI in the case of real-time embedded systems. From the above figure, it can be seen that there are four-dimensional systems such as language, technology, curriculum as well as people or audiences [15]. This manual cooperation can increase the effectiveness of English teaching. In order to use multimodal analysis can produce multimodal

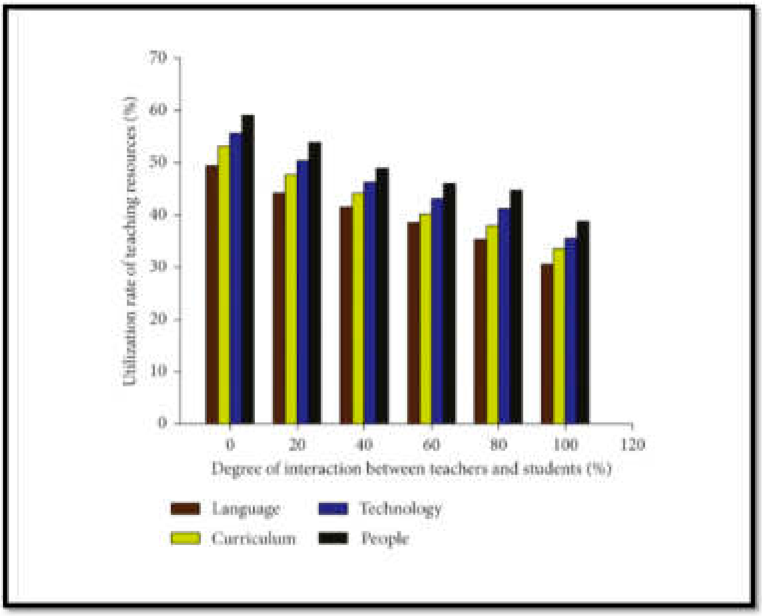


Fig. 9.1: Relationship between the various teaching resources and interaction among teachers and students

recognition within teaching space, time, resources, and audiences.

9. Discussion. A real-time embedded system can be defined as a particular type of embedded system that works with real-time computing through an RTOS embedded system. This system follows a quick response pattern to the external stimulus to work within a proper period. In the case of multimodal text analysis for English textbooks, the use of real-time embedded systems can be found very important [14].

The above figure presents the relationship between teaching resources and the highest teaching resource is its people than technology. The main basis of English learning is pure language however in the real-time embedded system the companion language is one of the crucial components. This has a role in the real-time embedded system as a supplementary, auxiliary, as well as reinforcing role within the transmission of the English language [12, 8]. This transmission can be done by using color, shape, spatial layout, sound, font size, and others.

Real-time embedded systems can be developed through the symbiosis of English textbooks based on the use of visual grammar. If the multimodal analysis is found fruitful, the cohesive communication between the textbook materials can accelerate the process of real-time embeddedness. Students' visual literacy skills are improved when visual grammar theory is used [8, 15]. Through multimodal analysis, students develop the ability to evaluate and grasp the visual elements in picture texts. Learners, who comprehend the foundations of visual composition, color, and symbolism may use and assimilate visual information more selectively. Since complex visuals are frequently seen in real-time embedded systems, this knowledge is extremely helpful.

10. Conclusion. Based on the above discussion it can be said that several empirical studies focus on the representation of digital English textbooks. Multimodal analysis can be divided into two groups, which depend on the process of making conceptualizing meaning. Multimodal elements encourage fostering multimodal relationships to foster a relationship between the readers and text producers. Visual grammar theory can be used examine how to textual and visual content can encode as well as communicate ideas about the real world in terms of representation. Real-time embedded systems can be involved in the soft and hard real-time as well as the non-preemptive and preemptive scheduling. "Real-time embedded system" is effective as it can simplify several processes as well as increase the quality of life.

While this study provides valuable insights, it is important to acknowledge its limitations. The scope of this analysis primarily focuses on the theoretical aspects of visual grammar theory and its application to multimodal analysis. Further empirical studies could delve into the practical implementation of these theories in actual digital English textbooks, assessing their impact on student engagement and comprehension.

Moreover, the exploration of real-time embedded systems opens avenues for future research. As this study merely touches upon the potential of real-time embedded systems in enhancing various processes and improving quality of life, in-depth investigations can delve into the intricacies of their integration within educational settings. Future studies may also consider exploring the challenges and ethical considerations associated with the integration of technology, particularly in educational contexts. Additionally, further research can delve into the cognitive and pedagogical implications of different types of image-text interactions, contributing to a deeper understanding of effective instructional design. This study paves the way for future investigations into the dynamic fusion of visual grammar theory, multimodal analysis, and real-time embedded systems in educational contexts.

REFERENCES

[1] B. Asghar, M. Azher, and S. Fatima, *Gender portrayal on cover pages of pakistani english children's literature: A visual semiotic analysis*, Journal of Gender and Social Issues, (2022).

[2] 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.

[3] N. Cohn, *Structural complexity in visual narratives: Theory, brains, and cross-cultural diversity*, Narrative complexity and media: Experiential and cognitive interfaces, (2019), pp. 174–199.

[4] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, *Image–text sentiment analysis via deep multimodal attentive fusion*, Knowledge-Based Systems, 167 (2019), pp. 26–37.

[5] J. Le Kernec, F. Fioranelli, C. Ding, H. Zhao, L. Sun, H. Hong, J. Lorandel, and O. Romain, *Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms*, IEEE Signal Processing Magazine, 36 (2019), pp. 29–41.

[6] J. E. Liu and A. M. Lin, *(re) conceptualizing "language" in clil: Multimodality, translanguaging and trans-semiotizing in clil*, AILA Review, 34 (2021), pp. 240–261.

[7] Z. Liu, *Introducing a multimodal perspective to emotional variables in second language acquisition education: Systemic functional multimodal discourse analysis*, Frontiers in Psychology, 13 (2022), p. 1016441.

[8] L. Niu, *Analysis of multimodal teaching of college english under the background of artificial intelligence*, Security and Communication Networks, 2022 (2022).

[9] M. Petryk, M. Rivera, S. Bhattacharya, L. Qiu, and S. Kumar, *How network embeddedness affects real-time performance feedback: An empirical investigation*, Information Systems Research, 33 (2022), pp. 1467–1489.

[10] M. Purgina, M. Mozgovoy, and J. Blake, *Wordbricks: Mobile technology and visual grammar formalism for gamification of natural language grammar acquisition*, Journal of educational computing research, 58 (2020), pp. 126–159.

[11] R. P. Sari and T. Tawami, *Visual-text integration in digital public service announcement: A strategy to fight fake news spread*, Aksara, 35 (2023), pp. 125–137.

[12] S. Sun, Y.-C. Chen, L. Li, S. Wang, Y. Fang, and J. Liu, *Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 982–997.

[13] 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.

[14] X. Yang, S. Feng, D. Wang, and Y. Zhang, *Image-text multimodal emotion classification via multi-view attentional network*, IEEE Transactions on Multimedia, 23 (2020), pp. 4014–4026.

[15] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, *Multimodal image synthesis and editing: A survey*, arXiv preprint arXiv:2112.13592, (2021).

[16] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, J. Tian, M. C. H. Chua, and M. Liu, *An image-text consistency driven multimodal sentiment analysis approach for social media*, Information Processing & Management, 56 (2019), p. 102097.

[17] J. Zhu, Y. Zhou, J. Zhang, H. Li, C. Zong, and C. Li, *Multimodal summarization with guidance of multimodal reference*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 9749–9756.

[18] R. Zhussupova, *Teaching english using multimodal texts*, EDITOR: Dzintra ILIŠKO, Professor, Dr, (2023), p. 28.

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
© 2023 SCPE. Volume 24, Issues 4, pp. 1145–1156, DOI 10.12694/scpe.v24i4.2439

IMPACT OF PHOTOVOLTAIC SYSTEMS ON DISTRIBUTION NETWORKS WITH ADVANCES OF CLOUD, GRID AND CLUSTER COMPUTING

ZHUORAN SONG[*], JIANFENG LI[†], TAO JIANG[‡], AND SICHEN LU[§]

Abstract. The impact of photovoltaic (PV) systems on distribution networks in advances of cloud, grid, and cluster computing was annoying in the study. Additional PV systems are explained with their working principles. Furthermore, A clustering-based model was developed to reduce the error in the calculation. The weather conditions were classified into different synoptic types. Prediction-based models were introduced after classifying the weather conditions in different optic types. For the movie, it was noted that conditions play a Vital role in predicting the output produced by a PV system. Moreover, a coherent discussion about the key elements regarding the impact of PV systems on Distribution Networks for Cloud, Grid, and Cluster Computing is done. Thus, through the discussion, an empirical knowledge-based analysis was presented.

Key words: cluster computing, PV system, solar grid, distribution networks. Secondary data

1. Introduction. The energy landscape is transforming with the global concern over depleting conventional energy sources and the pressing need for sustainable alternatives. Photovoltaic systems (PV) have emerged as a promising solution, harnessing solar energy to generate electricity. As societies strive to minimize their carbon footprint and embrace greener technologies, PV systems have gained prominence due to their renewable nature and potential for widespread deployment [1]. This transition to cleaner energy sources aligns with the broader global objective of mitigating climate change and achieving sustainable development.

Amidst this context, there is a growing recognition of the interplay between advancements in information technology and the energy sector. With its exponential growth and increasing computational demands, cloud computing stands out as a prime example of technology's burgeoning appetite for energy. This confluence of factors underscores the significance of exploring how PV systems can effectively power advanced technologies like cloud computing with their renewable energy generation capacity.

Research Scope: The scope of this study encompasses an in-depth investigation into the integration of PV systems with advanced technologies, focusing mainly on their potential to power cloud computing infrastructure. The research delves into the intricate web of factors, challenges, and opportunities associated with this integration. By examining the technological, environmental, and economic dimensions, the study provides a comprehensive understanding of the feasibility and implications of using PV-generated energy to sustain energy-intensive technologies like cloud computing.

Motivation: The motivation behind this research stems from the urgent need to address the energy challenges posed by the escalating adoption of advanced technologies. Cloud computing offers unprecedented scalability and efficiency but demands substantial energy resources. By leveraging renewable energy sources like PV systems to power cloud infrastructure, we can explore a pathway to more sustainable technology deployment. This study aims to contribute to the ongoing discourse on sustainable energy solutions and their role in supporting digital transformation.

Contribution: This research contributes to the existing body of knowledge by bridging the gap between renewable energy systems and advanced technology requirements. By explicitly focusing on PV systems and their potential to power cloud computing, this study offers insights into the feasibility, benefits, and challenges of such integration. The findings aim to inform policymakers, energy experts, and technology stakeholders about

[*]State Grid Liaoning Electric Power Supply Co., Ltd, Shenyang,110006, China (zhuoransongres1@outlook.com)
[†]State Grid Liaoning Electric Power Supply Co., Ltd, Shenyang, 110006, China
[‡]State Grid Liaoning Electric Power Supply Co., Ltd,Shenyang, 110006, China
[§]State Grid Liaoning Electric Power Supply Co., Ltd,Shenyang, 110006, China

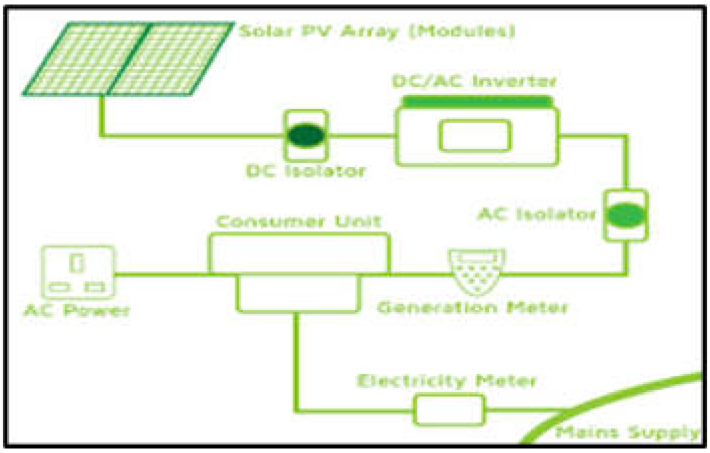


Fig. 2.1: PV system grid

the opportunities for reducing carbon emissions while meeting the energy needs of emerging digital landscapes. Through an empirical analysis of relevant secondary data and a systematic discussion of key factors, this study provides valuable insights into an evolving area at the intersection of energy and technology.

In the subsequent sections, we delve deeper into the factors that underscore the integration of PV systems with advanced technologies, particularly cloud computing. This study seeks to illuminate the path toward a greener and more sustainable technological future by examining technological considerations and a comprehensive analysis of associated factors.

2. Objectives. The objective of the study provided an outline for the overall analysis. In addition, the overall analysis is based on the following objectives:

1. To observe the different components of PV systems
2. To discuss the working principle of the PV system
3. To discuss different models to predict the output of Systems
4. To analyze the best prediction algorithms for PV system output
5. To analyze the problem of the PV system in cluster computing

3. Methodology. The methodology of n research looks into different factors and methods implemented to develop an empirical analysis. Moreover, different research strata and other factors aided the process of creating the study based on the objectives and justified through an appropriate methodology [3]. To comprehend different aspects of PV systems and the impact of the same on cluster computing, the method of secondary qualitative analysis was employed. To collect data for the study, secondary data sources such as past research, articles, news, and other authentic external resources were collected. At the same time method of qualitative analysis was used to analyze the collected data for the study [5]. The secondary quantitative method calls for authentic data and factor-based analysis for the study, therefore. The method of secondary research is beneficial to develop a knowledge-based result that has tangible implications in real life. Therefore, a secondary quantitative method of analysis was used to create the overall analysis in a systematic manner [2].

4. Working process of a photovoltaic (PV) system. For understanding the process of a photovoltaic (PV) system in a comprehensive manner it is essential to understand the components associated with the same. Additionally, it is essential to understand the basic working process of a PV system. Thorough knowledge of such systems provides the premises to understand the impact of the PV system in cluster computing [6].

A photovoltaic (PV) system converts the packets of energy of sunlight into eclectic energy. The basic working principle is that solar light which contains packets of energy or photons falls on the solar panel and through a process called the photovoltaic effect, it is converted into electric current. Each of the panels converts photons into electricity and produces a small amount of electric energy. Thus, small panels are interconnected

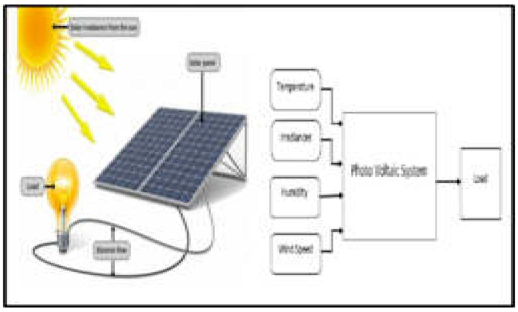


Fig. 3.1: PV system unit

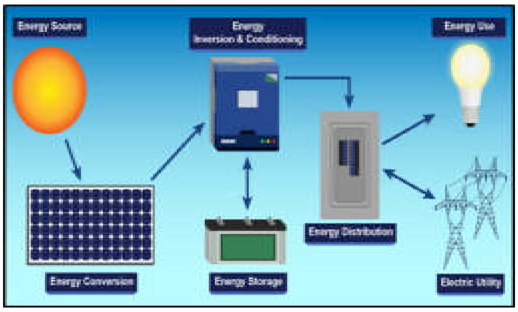


Fig. 4.1: Gride design

in a solar array in order to produce a large amount of energy.

Figure 2.1 of the analysis is related to the grid design of a solar panel. A dollar panel produces DC current as a result. However, most cluster computing technology uses AC current for its functioning [7]. Therefore, an electrical unit grid and inverters are connected with the panels in order to produce concert DC current in AC current.

At the same time, there are different other components that are associated with the main system in order to produce electricity. Following is a brief discussion about the different components:

a) Solar Panel

The solar panel of a PV system consists of solar cells having semiconductor properties in the material. Additionally, the semiconductor material works as a shield that protects the panel from the environment. After the conversion of solar photons into usable electricity through the photovoltaic effect the conducting material collects the electricity. The semiconductor is presented on either of the panel [9, 4]. Additionally, there is a layer of antireflection coating in order to prevent reflection and minimize the loss of energy. The primary component of the solar panel is crystalline silicon which has a 33% capacity of energy producing capacity from sunlight [24, 8]. However, there are other semiconductor materials that have the same functionality however the price for them is higher.

b) Inverters

Inverters of solar units are the components that intake DC electricity and convert that to AC flow. The current produced by a solar panel is DC current, however, in cluster computing AC current is required [11]. Therefore, an inverter unit of the PV system converges DC to AC current. Hence, the inverter is the most important and expensive component after solar panels [10].

Most of the inverters come with a conversion efficiency of 90%. In addition, some of the solar panels can produce more power than that. The safety feature of an inverter is a significant feature that includes fault circuits. Additionally, a ground fault circuit is one of the most important features of a PV system that shuts

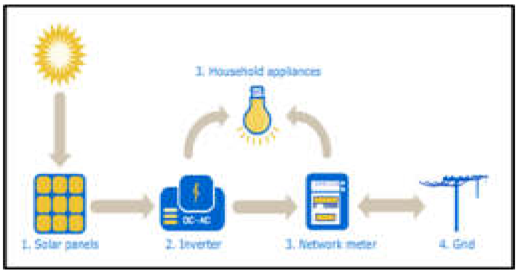


Fig. 4.2: Components of the PV system

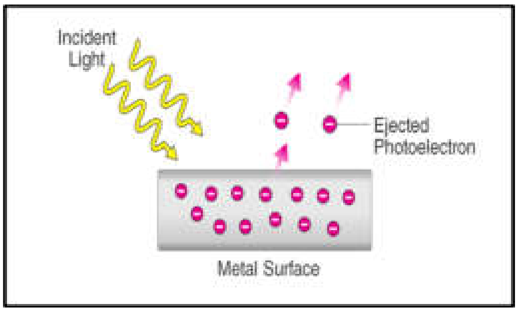


Fig. 4.3: Photovoltaic effect

down the system when there is a fault detected in the grid system.

c) Racking

The rack is the mount that is used to fix the solar array with a base. Usually, the construction of the rack is aluminum or steel. Additionally, one of the significant features of the solar array is of grounding of the overall grid system [12]. The Racking of a solar panel is mechanically grounded with the ground in order to prevent electrocution.

d) Other Important Components

There are other essential components of a PV system that are essential for the systematic functioning of the system. For instance, disconnects, combiners, meters, breakers, and wiring are important for the functioning of the system. The main function of a solar combiner is to connect two or more cables in a single unit [14].

Electrical systems are shielded from surges or overcurrent by circuit breakers or breakers. Breakers can also be manually actuated, serving as an extra disconnect, although they are designed to activate automatically when the current exceeds a certain value [13]. Electric meters track the quantity of energy flowing through a system moreover usage and energy consumption are calculated through the meters. Additionally, there are wirings that must be of the correct size to carry the current and convey electrical energy from and between each component.

5. Photovoltaic effect. The process of Photovoltaic effect is the process of generating electricity from solar energy. When sunlight is reflected off of it. This phenomenon, which results from the solar panel's cells converting sunlight into electrical energy, is what makes solar panels valuable. Types of semiconductors that are used for the PV effect are p-type and n-type. A p-n junction is created between the different types of conductors. The major principle that the Photovoltaic effect works on is the concept of the photoelectric effect [15].

Considering, light is a particle in nature, the photoelectric property can be explained. The photons of light are then converted into energy. Planck's equation connects the energy of a photon to the frequency of the light.

$$E = h\nu = h\frac{c}{\lambda}$$

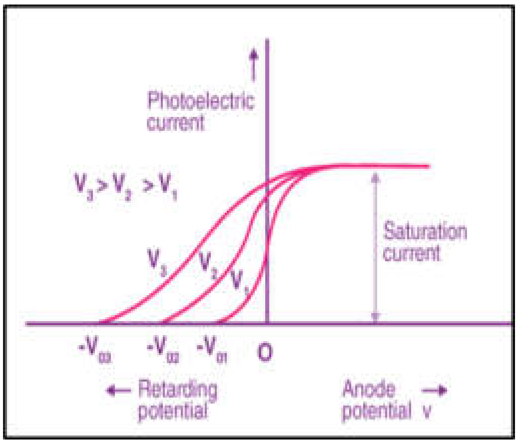


Fig. 5.1: Photovoltaic effect

In the above equation, E stands for the energy of the photon, which is the frequency of the light. In addition, h is Planck's constant, and c is the light speed (in a vacuum). Furthermore, λ is the wavelength of the light.

It can be comprehended that different frequencies of waves from different light sources consist of photons with different energies [12]. For instance, red light has more frequency than blue light. Hence the energy that a photon of blue light has is higher than that of a photon of red light.

The above graph is related to the photovoltaic effect and a threshold of electricity generation is shown in the above graph. The positive potential is steadily increased while the light's intensity and frequency remain constant [10]. When the potential between the metal plate and the collector rises to a certain level and is positive, the photoelectric current increases. In addition, When the potential is raised above the typical value for any rise in the accelerating voltage, the photoelectric current remains unchanged [16]. The current's maximal value is referred to as the saturation current.

The XG-Boost self-organizing map (TS-SOM) represents a sophisticated technique that combines the power of two distinct methodologies: the XG-Boost algorithm and the self-organizing map (SOM). The SOM is a neural network-based approach for clustering and visualizing high-dimensional data, while XG-Boost is an ensemble learning algorithm that excels in predictive modeling.

1. Self-Organizing Map (SOM) Component: At its core, the SOM is an unsupervised learning technique that organizes data points into a grid-like structure based on similarity. Neurons within the grid (also called nodes) compete to represent input data, and the competition results in a topological mapping of the data. SOMs are well-known for their ability to reveal underlying patterns and relationships within complex datasets.
2. XG-Boost Component: XG-Boost, on the other hand, is a boosting algorithm that constructs a strong predictive model by iteratively adding weak learners (usually decision trees) to the ensemble. It addresses issues of bias, variance, and overfitting by combining the predictions of multiple weak models. XG-Boost is widely recognized for its high performance in various machine learning tasks, including classification and regression.

Integration of XG-Boost and SOM: Data Tree Structure In this study, the innovative concept of the "data tree structure of the XG-Boost self-organizing map (TS-SOM)" is introduced. This hybrid structure combines the SOM's grid-like organization of data and the predictive power of XG-Boost. Here's how the integration works:

1. The data tree structure represents a hierarchical arrangement of nodes that captures both the topological relationships derived from SOM and the predictive capabilities of XG-Boost. Each node in the tree corresponds to a cluster identified by SOM, and it contains an XG-Boost model trained on the data points belonging to that cluster.

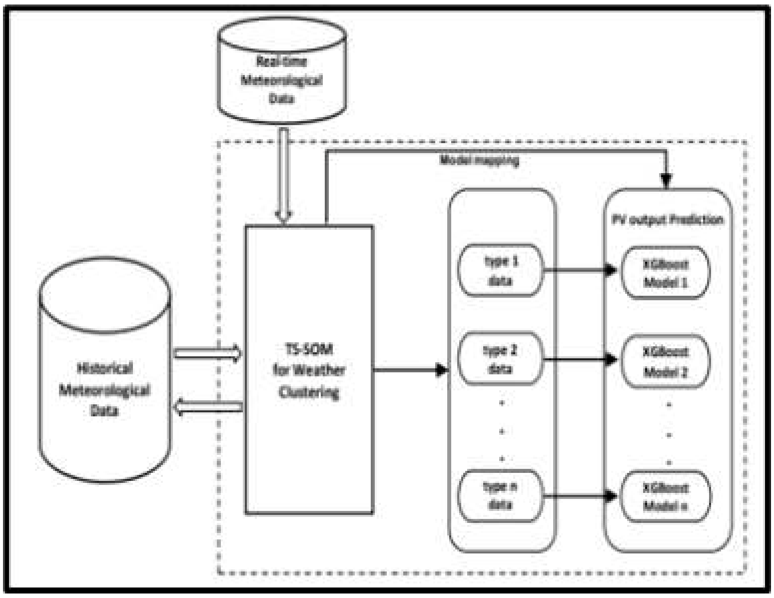


Fig. 6.1: Algorithms used in the study

2. When a new data point is introduced to the TS-SOM, it traverses the data tree structure from the root to a leaf node. At each node, the XG-Boost model provides predictions based on the learned patterns of the data points in that cluster. The prediction is propagated down the tree until the data point reaches a leaf node, and the final prediction is the aggregated result of predictions from multiple nodes.

6. Output calculation of PV system for cluster computing . The data tree structure of the XGBoost self-organizing map (TS-SOM) is used in order to calculate the output of a PV. Based on the machine learning model the output is calculated in different sections [19]. Additionally, with the implication of the aforementioned training models, low computational overhead allows for the precise PV power output forecast, which makes a variety of edge devices appropriate for its usage can be done. The PV output in different circumstances is essential to calculate, Furthermore, an overall output production of the PV system aids the functioning of the cluster computing. Additionally, the grid system and the power output depend on the field of energy [17]. In addition, the performance of output is calculated based on broadly used. Moreover, a Generalised Regression Neural Network (GRNN) was used for prediction accuracy. Additionally, a fair comparison of the system's Long Short-Term Memory (LSTM) was used. Furthermore, SVR is trained with a similar training model to GRNN. Moreover, an identical training approach was used XGBoost approach [18].

The above image is related to the model that was used in the analysis. Moreover, historical data used in the model was output TS-SOM was gathered. It was found that meteorological factors have a huge impact on the system [25]. The PV output curves were seen to emit spikes and fall with abrupt changes in the weather. Thus, such unpredictability of meteorological influences creates hindrances in building a high-performing PV model. Therefore, Given prior understanding as historical data input about the weather changes over time. In addition, and in the meanwhile, is still capable of properly estimating PV power output for the electrical unit [19].

Additionally, the prediction-based model was completed with the capability of handling data loss. It was noted that the PV output system often faces data loss that leads to faulty prediction of the PV output system. moreover, it was noted that there is a hindrance in data transmission due to various reasons. hence handling Data loss is a major component of the PV system. In addition, it was found that there are various components that might be affecting the output results. Such as Global Horizontal Irradiance (GHI), Diffuse Horizontal Irradiance (DHI), and Direct Normal Irradiance (DNI) have a strong impact on the PV power output. All the

Table 6.1: Pearson Correlation

FIELD	PEARSON CORRELATION	DESCRIPTION
ghi	0.9767	Irradiance horizontal global for the centre value
ghi90	0.9744	Irradiance horizontal global at 90%
ghi10	0.9629	Irradiance horizontal global at 10%
dni90	0.8644	90% value direct normal irradiance
dni	0.9275	Centre value direct normal irradiance
dni10	0.9077	10% value direct normal irradiance
air temp	0.3350	Air temperature
cloud opacity	-0.2184	Cloud quantity
azimuth	0.0092	The azimuth angle of the sun. Range: -180∼180
dhi	0.9314	Diffuse irradiance (horizontal)

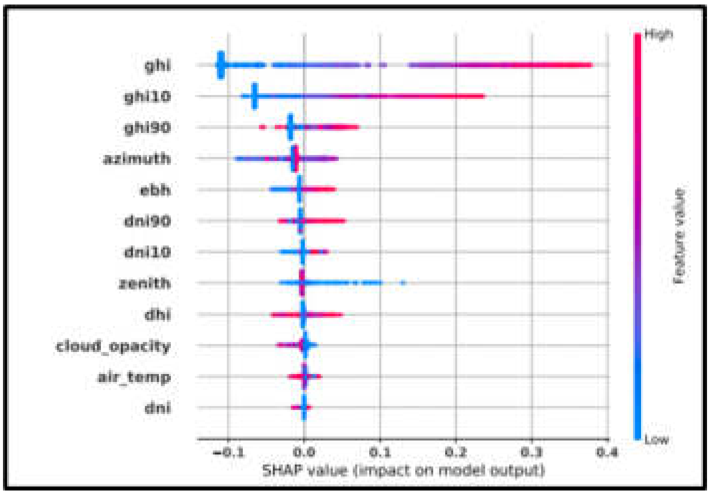


Fig. 6.2: SHAP Value

components are mentioned in Table 6.1 of the study [21].

For the first component, the roots are divided into two subsets of data considering response to the four different types of weather. Compared to the traditional SOM the implemented data tree structure is helped In the spring of the training process and provides faster results [20]. Additionally, parallel processors are employed in order to initiate the process of training which works as a catalyst for this prediction model.

The SHAP Value was calculated based on the four divisions of the data set which is shown in the above figure. Calculating the SHAP values was important in order to produce reliability in the training model Additionally interpretability of the training model is presented through the above figure. The features of the figure are present in a non-ascending order. Moreover, registration is based on the importance of different facts [22].

Additionally, figure 4.1 of the study contains localized Comprehension of the individual weather conditions and the values of contains are presented in the figure and the table. Clustering-based weather conditions were analyzed and compared with other models. In order to produce reliable data XGBoost Neural network was compared with Other regression methods such as super vector regulation (SVR) [25]. In addition, XGBoost was compared with GRNN and LSTN in order to produce comparable results for the prediction model. Furthermore, to produce a fair comparison all of the data sets were trained using a similar training model as XGBoost. The use of LSTM was considered due to its features of Analysing historical data and tackling information in specific

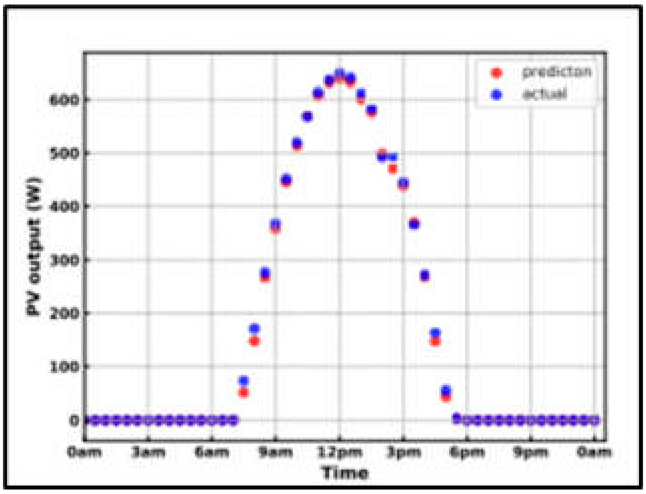


Fig. 6.3: XGBoost

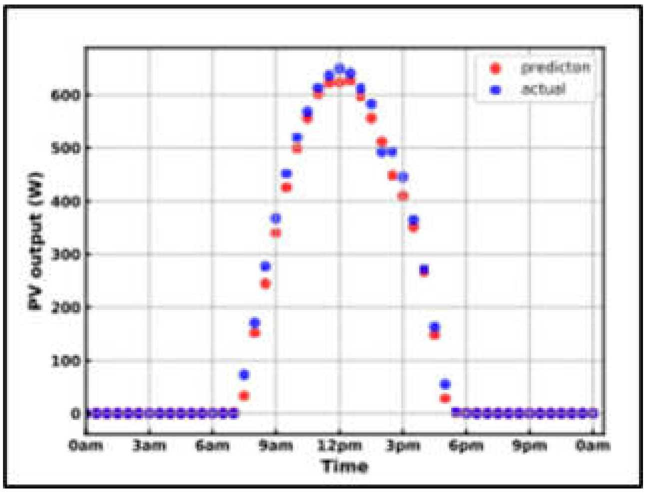


Fig. 6.4: SVR

time series. Additionally, meteorological features of different zones are fed into the model in order to active reliable statistics [22].

In order to produce the power output there are measurement metrics that aided in calculating the broader terms of the model. Measurement moduli such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and determination coefficient were used in the study (R square value)

In the above figure, the output data of the XGBoost is presented. It can be seen that there is similarity and comparatively less gap between the prediction and the actual output. Moreover, the outcome is calculated at 12 pm when the sun is at its peak. Additionally, the power output is maximum in the situation. On the other hand, it can be seen that at 9 pm the prediction is nearly too accurate, and similar results can be seen at the 3 pm mark [22].

The above figure is derived using the SVR and a similar training module. It can be seen that in the 12 pm mark, there is a difference between the actuarial output and the predicted output. Moreover, at the 9 am mark, a similar disparity is observed. At the same time for the 3 pm output, the prediction was close to the

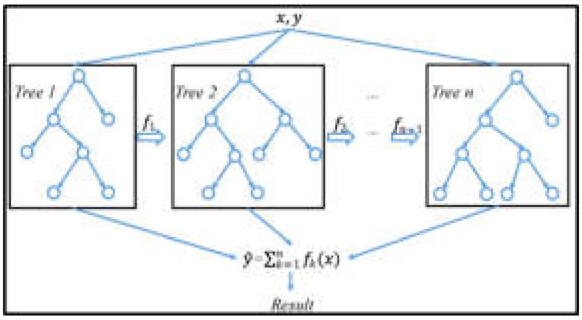


Fig. 6.5: XGboost working model

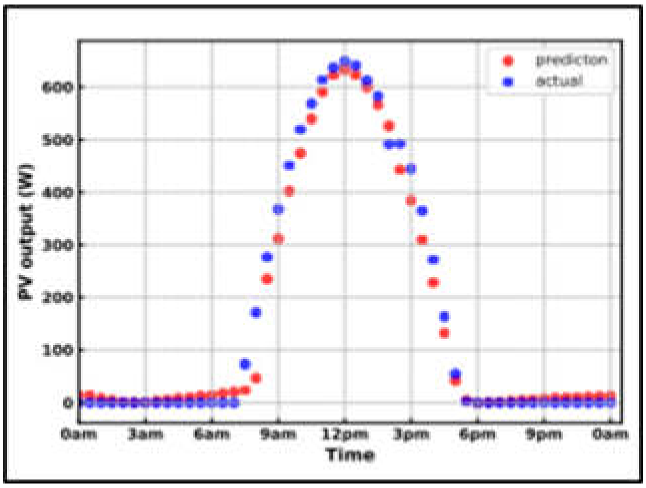


Fig. 6.6: GRNN

actual output. Therefore, it can be understood that. Hence, the disparity between the prediction and actual output is different with comparatively large values.

The above output was achieved with the GRNN model and the training was done through a similar data tree used for training other models. It can be seen that there is a noticeable difference between the actual output and the predicted output of a similar model. Moreover, the differentiation between the prediction and actual output can be noted from the 6 pm mark. Similarly, a major difference was noted at the 9 am mark where the output was nowhere near the actual output. Therefore, it can be said that GRNN showed a massive disparity in the actual and predicted output of the data [23].

In the end, the implementation of LSTM was considered for the prediction, it can be seen that there is a major differential between the prediction model and the actual output. At 3 pm March, there is an overlapping in the prediction and the actual output. However, the difference can be seen in other time marks [20].

From the above table, the related values of the algorithms can be seen. In addition, after a thorough analysis, it can be said that all of the values provided reliable statistics in order to prove liability.

7. Results. From the above analysis, it can be said that the fluctuation of power output needs to be managed in order to use a PV system, for cluster computing. Therefore, implementing an appropriate training model is essential for predicting stability.

It was found that implementing the XGBoost algorithm produces maximum reliability for the model. In addition, it was found that there are different components of a PV system that need to operate in a systematic

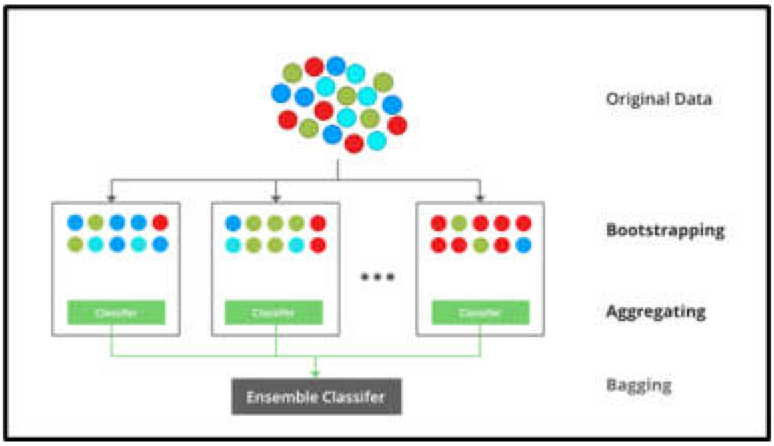


Fig. 6.7: XGboost clustering

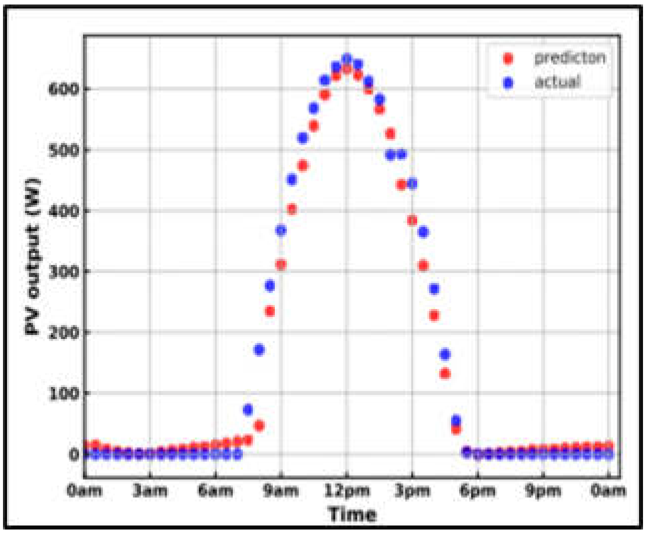


Fig. 6.8: LSTM

order for the sustainable development of the data [9]. Additionally, it was noted that the implication of the PV system depends on the fluctuation. Moreover, the fluctuation in the system can hinder the usability of PV systems in cluster computing. In the above image, fluctuation can be seen. Moreover, the differentiation indicates the use of a reliable model for sustainable and scalable clustering computing with a PV system

Conclusion

Thus, the PV system is analyzed based on different model models and algorithms. Moreover, an actual presentation of the clustering system is presented in the analysis. Additional PV system is explained with their different components. It was found that PV system is significant for cluster computing and influenced by weather conditions. Therefore, XGBoost is recommended for predicting the power generation of the PV system. Future research could delve deeper into the dynamic adaptation of PV systems to varying weather conditions. Developing algorithms and models that can dynamically adjust the PV system's behavior and output based on real-time weather data could significantly enhance its performance and efficiency. While XGBoost has shown promise in predicting power generation, exploring other advanced machine learning techniques, such as deep

Table 6.2: Regression analysis of algorithms

Algorithm	**R2**	**MAE**	**RMSE**	**Time cost(s)**
XGBoost	0.9863	12.21	27.49	0.047
SVR	0.9708	29.12	40.22	0.003
SVR	0.9649	23.22	45.38	0.29
GRNN	0.9824	15.29	31.19	0.090

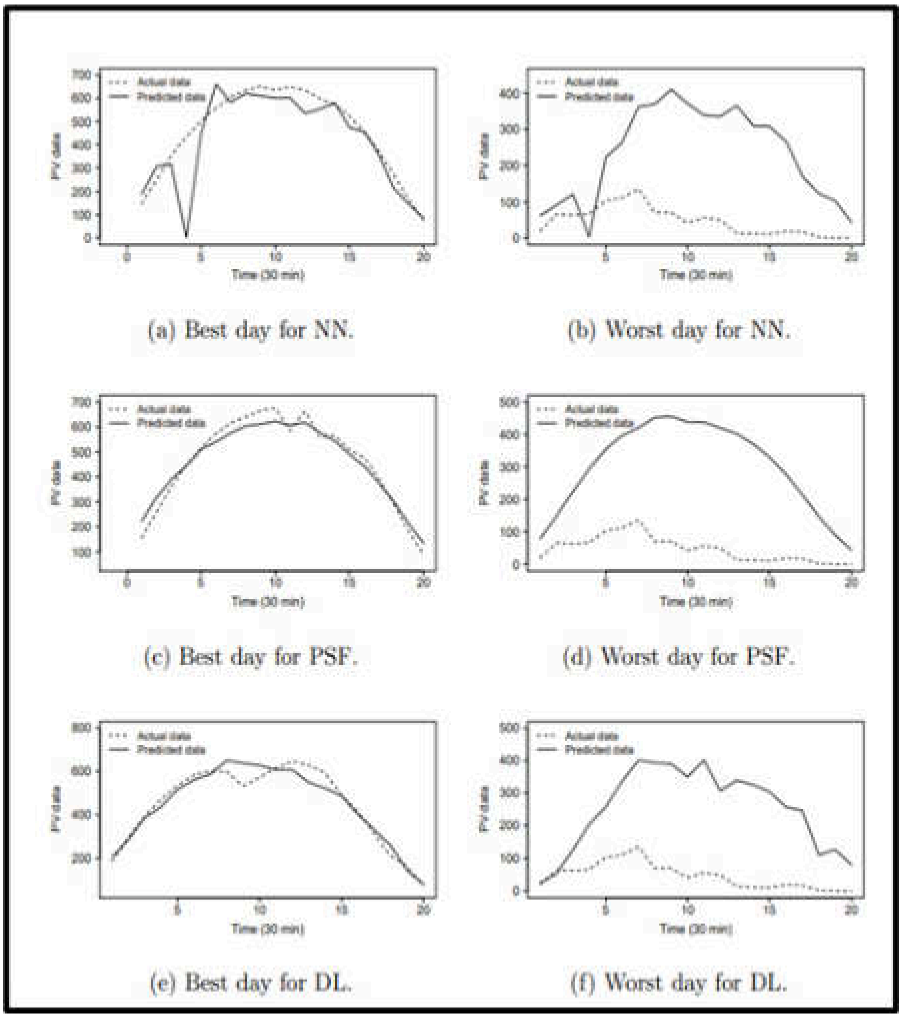


Fig. 7.1: LSTM

learning or hybrid models, could lead to even more accurate predictions. These models could capture intricate relationships between weather patterns and PV output.

REFERENCES

[1] R. Ahmed, V. Sreeram, Y. Mishra, and M. Arif, *A review and evaluation of the state-of-the-art in pv solar power forecasting: Techniques and optimization*, Renewable and Sustainable Energy Reviews, 124 (2020), p. 109792.

[2] M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński, and U. Amjad, *Prediction of pile bearing capacity using xgboost algorithm: modeling and performance evaluation*, Applied Sciences, 12 (2022), p. 2126.

[3] A. Asselman, M. Khaldi, and S. Aammou, *Enhancing the prediction of student performance based on the machine learning xgboost algorithm*, Interactive Learning Environments, (2021), pp. 1–20.

[4] 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.

[5] A. Deng, H. Zhang, W. Wang, J. Zhang, D. Fan, P. Chen, and B. Wang, *Developing computational model to predict protein-protein interaction sites based on the xgboost algorithm*, International journal of molecular sciences, 21 (2020), p. 2274.

[6] M. Emamian, A. Eskandari, M. Aghaei, A. Nedaei, A. M. Sizkouhi, and J. Milimonfared, *Cloud computing and iot based intelligent monitoring system for photovoltaic plants using machine learning techniques*, Energies, 15 (2022), p. 3014.

[7] S. B. Jabeur, S. Mefteh-Wali, and J.-L. Viviani, *Forecasting gold price with the xgboost algorithm and shap interaction values*, Annals of Operations Research, (2021), pp. 1–21.

[8] Y. Jiang, G. Tong, H. Yin, and N. Xiong, *A pedestrian detection method based on genetic algorithm for optimize xgboost training parameters*, IEEE Access, 7 (2019), pp. 118310–118321.

[9] Z. A. Khan and D. Jayaweera, *Smart meter data based load forecasting and demand side management in distribution networks with embedded pv systems*, IEEE Access, 8 (2019), pp. 2631–2644.

[10] S.-C. Lim, J.-H. Huh, S.-H. Hong, C.-Y. Park, and J.-C. Kim, *Solar power forecasting using cnn-lstm hybrid model*, Energies, 15 (2022), p. 8233.

[11] C.-H. Liu, J.-C. Gu, and M.-T. Yang, *A simplified lstm neural networks for one day-ahead solar power forecasting*, Ieee Access, 9 (2021), pp. 17174–17195.

[12] G. Mitrentsis and H. Lens, *An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting*, Applied Energy, 309 (2022), p. 118473.

[13] M. A. Mughal, S. Cui, H. Zhang, M. Carpinello, J. McManus, M. Moreira, and K. Pacheco, *Cloud motion vector system to detect clouds and forecast real-time photovoltaic system performance*, in 2021 IEEE 48th Photovoltaic Specialists Conference (PVSC), IEEE, 2021, pp. 2212–2216.

[14] U. Munawar and Z. Wang, *A framework of using machine learning approaches for short-term solar power forecasting*, Journal of Electrical Engineering & Technology, 15 (2020), pp. 561–569.

[15] A. Paleczek, D. Grochala, and A. Rydosz, *Artificial breath classification using xgboost algorithm for diabetes detection*, Sensors, 21 (2021), p. 4187.

[16] S. Pan, Z. Zheng, Z. Guo, and H. Luo, *An optimized xgboost method for predicting reservoir porosity using petrophysical logs*, Journal of Petroleum Science and Engineering, 208 (2022), p. 109520.

[17] S. Parui, A. K. R. Bajiya, D. Samanta, and N. Chakravorty, *Emotion recognition from eeg signal using xgboost algorithm*, in 2019 IEEE 16th India Council International Conference (INDICON), IEEE, 2019, pp. 1–4.

[18] E. Ravishankar, R. E. Booth, J. A. Hollingsworth, H. Ade, H. Sederoff, J. F. DeCarolis, and B. T. O'Connor, *Organic solar powered greenhouse performance optimization and global economic opportunity*, Energy & Environmental Science, 15 (2022), pp. 1659–1671.

[19] K. Song, F. Yan, T. Ding, L. Gao, and S. Lu, *A steel property optimization model based on the xgboost algorithm and improved pso*, Computational Materials Science, 174 (2020), p. 109472.

[20] P. Song and Y. Liu, *An xgboost algorithm for predicting purchasing behaviour on e-commerce platforms*, Tehnički vjesnik, 27 (2020), pp. 1467–1471.

[21] M. Sun, C. Feng, and J. Zhang, *Probabilistic solar power forecasting based on weather scenario generation*, Applied Energy, 266 (2020), p. 114823.

[22] T. Tao, Y. Liu, Y. Qiao, L. Gao, J. Lu, C. Zhang, and Y. Wang, *Wind turbine blade icing diagnosis using hybrid features and stacked-xgboost algorithm*, Renewable Energy, 180 (2021), pp. 1004–1013.

[23] J. F. Torres, A. Troncoso, I. Koprinska, Z. Wang, and F. Martínez-Álvarez, *Deep learning for big data time series forecasting applied to solar power*, in International Joint Conference SOCO'18-CISIS'18-ICEUTE'18: San Sebastián, Spain, June 6-8, 2018 Proceedings 13, Springer, 2019, pp. 123–133.

[24] 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.

[25] S. Zhang, D. Zhang, J. Qiao, X. Wang, and Z. Zhang, *Preventive control for power system transient security based on xgboost and dcopf with consideration of model interpretability*, CSEE Journal of Power and Energy Systems, 7 (2020), pp. 279–294.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 24, 2023
Accepted: Sep 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 1157–1168, DOI 10.12694/scpe.v24i4.2440

PERSONALIZED EXERCISE PROGRAM DESIGN WITH MACHINE LEARNING IN SENSOR NETWORKS

YAN LU*

Abstract. The use of wearable solutions and sensors has reached the point of modern machine learning (ML) techniques in recent years. It contains the different programs that help to develop personalized opportunities for the betterment of the individuals. The research shed light upon the functions of ML and distributed sensor networks (DSN) in promoting healthcare among users. It can be said that the functions of the ML help to develop good fitness among the users, which results in good health. This dissertation consists of different kinds of ML algorithms and the usage of DSNs for the development of physical exercise programs. It has been found that most individuals use the sensors for their benefit in developing good health. The use of machine learning techniques not only helps to record an individual's health data but also sends emergency information to the nearest medical center for the user's benefit. The use of ML helps to the identification of the location of an individual with the help of GPS measures of real-time distance and users' health activities.

Key words: Clinical trials, EHRs, sensor-based software for health development, challenges of ML and DSN

1. Introduction.

In recent years, intelligent devices such as fitness watches and sensor networks have resulted in the development of fitness among individuals. Machine learning in sensory networks has helped distribute great fitness techniques and information to the world. Personalized exercise programs are helpful to the development of fitness among every age group. Machine learning in distributed sensory networks helps identify accurate medical information about a person. Workout is crucial for keeping one's health and general well-being in good shape. However, creating personalised workout plans that consider each person's demands and goals can be difficult. The field of customized exercise program design (PEPD) has entered a new phase as a result of developments in machine learning (ML) and distributed networks of sensors. PEPD provides customized workout regimens that maximize efficacy and efficiency by utilizing ML algorithms and current information gathered via distributed sensor networks.

The use of personalized exercise programs with the help of machine learning creates an advantage for an individual to adopt different procedures for their own health and fitness.

The motivation behind the research stems from the remarkable convergence of wearable solutions, sensors, and modern machine-learning techniques in recent times. This convergence has unlocked a realm of possibilities for creating personalized interventions that can significantly enhance the well-being of individuals. With the potential to revolutionize healthcare, the study investigates the synergistic functions of machine learning and distributed sensor networks (DSNs) to drive improvements in user health and fitness.

The primary goal of the research is to illuminate the pivotal role played by machine learning and DSNs in advancing healthcare practices, particularly in the context of user-centered applications. The study recognizes the intrinsic link between physical fitness and overall health, and it aims to harness the power of machine learning algorithms to design tailored physical exercise programs. By leveraging various machine learning techniques, the research endeavors to optimize these programs, considering individual variations and preferences.

2. Objectives.

To identify ways PEPD (Personalized Exercise Program Development) helps manage physical activity using Machine Learning

This objective involves investigating how Personalized Exercise Program Development (PEPD) leverages Machine Learning (ML) techniques to manage and enhance individuals' physical activity effectively.

*School of Physical Education, University of Sanya, Sanya 572000, China

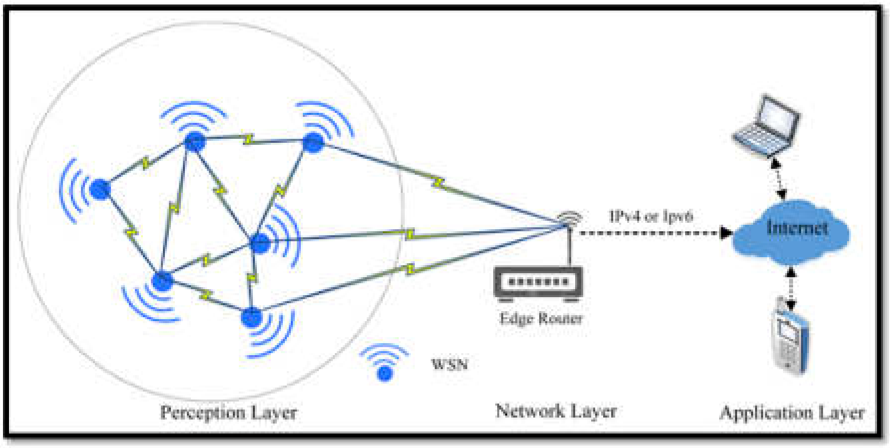


Fig. 1.1: Machine learning for wireless sensors (Source:[4])

The research aims to identify and outline the specific methodologies, algorithms, and strategies employed within PEPD that utilize ML. This might include exploring how machine learning models can analyze user data, such as activity levels, preferences, and health conditions, to create tailored exercise programs. Additionally, the research would delve into how ML helps adapt these programs over time based on the user's progress and feedback. By achieving this objective, the study would contribute to a better understanding of the synergy between machine learning and personalized exercise planning for optimal physical activity management.

To understand the ML functions increasing individual fitness

This objective centers on comprehending the mechanisms through which Machine Learning functions contribute to improving individual fitness levels. The research explores the various ML techniques employed to optimize fitness outcomes. This could encompass studying algorithms that analyze physiological data (such as heart rate, sleep patterns, and caloric expenditure) gathered from wearable sensors. The research might also investigate how ML models identify patterns and correlations within this data to suggest more effective exercise routines, nutritional plans, and recovery strategies. By achieving this objective, the study would provide insights into the role of ML in fostering improved fitness results and potentially advancing the field of fitness science.

To depict the challenges related to DSNs (Distributed Sensor Networks) in healthcare

This objective focuses on exploring the challenges associated with implementing Distributed Sensor Networks (DSNs) in the healthcare context. DSNs involve interconnected sensors placed in various locations to collect and transmit data. DSNs can facilitate real-time health monitoring and data collection in the healthcare domain. The research identifies and analyzes hurdles and obstacles that might arise while deploying DSNs for healthcare applications. Challenges could encompass data security, privacy concerns, interoperability of sensor devices, data accuracy, and reliability. By addressing these challenges, the research aims to provide insights into how DSNs can be effectively utilized in healthcare settings and propose potential solutions to mitigate these challenges.

The research objectives revolve around understanding the role of Machine Learning in personalized exercise planning, uncovering how Machine Learning functions contribute to improved individual fitness, and highlighting challenges related to implementing Distributed Sensor Networks in healthcare. Achieving these objectives would contribute to advancing the field and pave the way for more effective, data-driven, and personalized approaches to promoting health and wellness.

3. Methodology. The methodology of this paper aims to provide a comprehensive understanding of the various functions, benefits, and challenges associated with the utilization of Machine Learning (ML) and Data Science Networks (DSNs) for the development and implementation of personalized programs in the healthcare

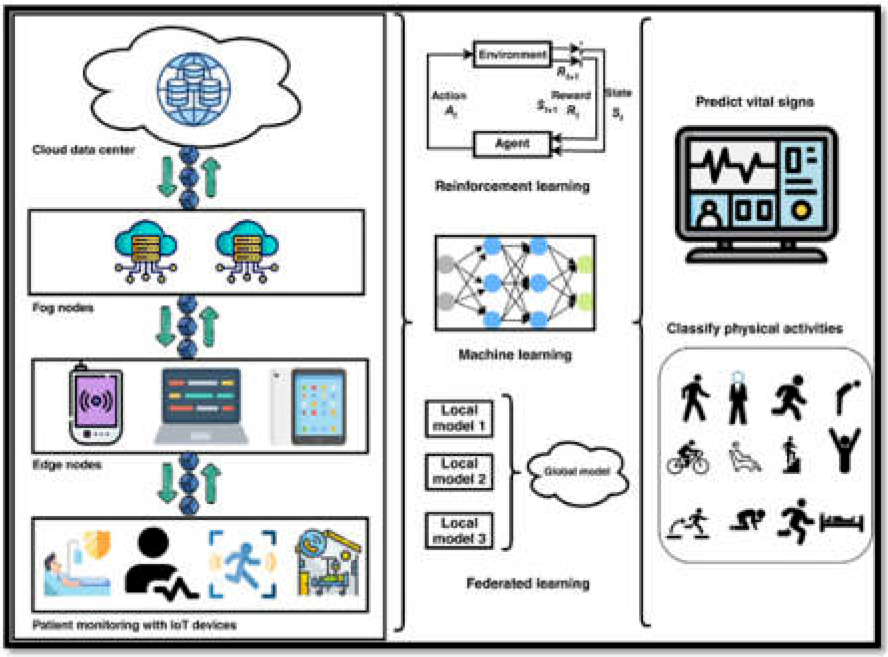


Fig. 4.1: Monitoring ML for fitness (Source: [1])

sector. As depicted in the aforementioned figure, there exist several types of ML functions that prove to be highly advantageous within DSNs, ultimately benefiting individuals as well.

These functions aid in identifying diseases and offer valuable suggestions for their treatment [1]. Furthermore, integrating sensors that monitor bodily functions allows for accurately recording health factors. This research, therefore, delves deeper into the numerous advantages ML and DSNs bring to the table, specifically enhancing an individual's overall well-being and fitness.

4. Summary of Machine Learning and DSNs. The ML and DSN help in the identification and recording of human health information to maintain the adequate flow of fitness work. It can be said that the use of ML helps to determine the body's functions and manages to record the same in the cloud storage.

Data from ML and DSNs is used to analyze the personal information of users, including their fitness goals, medical history, body requirements, and training sessions [4]. The ML and DSN designs the software and meet the needs of the user based on the data evaluated from the medical history. However, the use of ML and DSN has different challenges described in this research. Improved Efficiency, the design of the ML and DSN programs are been made under the requirements of the individual in achieving the primary motive in life. It consists of the basic fitness training every user must perform for a better life. It contains workout routines and generates exercise plans for the user's benefit to maintain good health. This also helps in the achievement of the fitness goals of the individuals.

The above figure shows the use of machine learning and DSN to help in the recording of the data and information related to the health of the individual [8]. These data are been recorded with the help of using biometrics that measure the BPM and blood flow in the body. It also helps in the identification of the actual number of steps taken by the user and calories burnt from the same. Reduction of Injury Risks, Individual limitations, and injury histories are taken into account when designing personal exercise programs with ML and DSNs. This information can be used by the systems to identify and avoid exercises that may create life risks to the user [9]. The goal of this approach is to minimize the chances of injuries and promote safe and sustainable workout practices throughout the year. Behavioral Analysis and Motivation, with the use of machine learning and data science, users' behavior patterns of the individuals, suggest exercise methods, and provide several

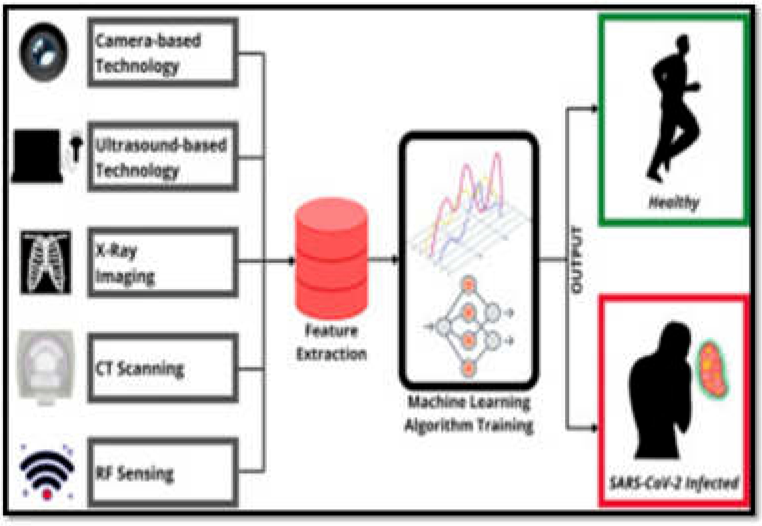


Fig. 4.2: ML Algorithm Training (Source: [5])

motivational factors that are helpful for the development of the fitness of the individual [5]. The use of these systems can keep users engaged and motivated by their exercise programs by providing them with personalized recommendations, reminders, and motivational factors for their own betterment.

5. Machine learning functions that are beneficial to the fitness of individual. *Personalization*, Machine learning techniques are been helpful in the development of personalized programs for users so that they can be able to work out and get an idea of their own health. The use of sensors in machine learning helps to record real-time data and can be customized based on the requirements of the individual to perform the exercise [6]. It helps to record the steps from walking along with the calories that are burnt from the same. It also measures the heart rate of the user to determine their physical health condition. Optimize Training Efficiency, The use of ML for the development of physical health results in the provision of proper training and yoga facilities. It helps to provide relevant information based on the health effects recorded in the sensors [16]. These sensors help the development of the training of the individual based on their abilities and health. These are helpful to the development of physical health which benefits the users.

ML algorithms help in the identification of several aspects that are related to health development based on the feedback and progress of the user. It takes into consideration the different aspects of the health benefits that are helpful to the individual over time. The risk of failing to do exercise and training can be solved with the use of ML as a basic tool. This helps in the identification of health factors and suggests necessary requirements to avoid getting injured and avoiding life risks. The use of ML with the sensors helps to recognize the blood pressure and heart rate of the user at any time [15]. Moreover, it sends real-time data to emergency contacts for quick

Machine learning algorithms individuals stay committed to their fitness routines by providing them with personalized feedback, motivation, and guidance [17]. With the help of algorithms that analyze user behavior, preferences, and historical data, the program can be tailored to match the individual's preference and provide timely encouragement and support to meet their needs.

The use of machine learning helps the development of the health of an individual. It has various benefits that are useful to the mitigation of risk and recording of healthcare information of the individual. The use of ML and DSN helps with the recording and storing of information with individuals for healthcare benefits. A DSN is been a cloud-storing software that keeps a record of the health data of individuals and provides them with personalized exercise programs which help to increase the audience [19]. Further, the use of ML also helps in the identification of possible risks to individuals that can affect their health.

Fig. 5.1: Benefits of ML for physical fitness (Source: [4])

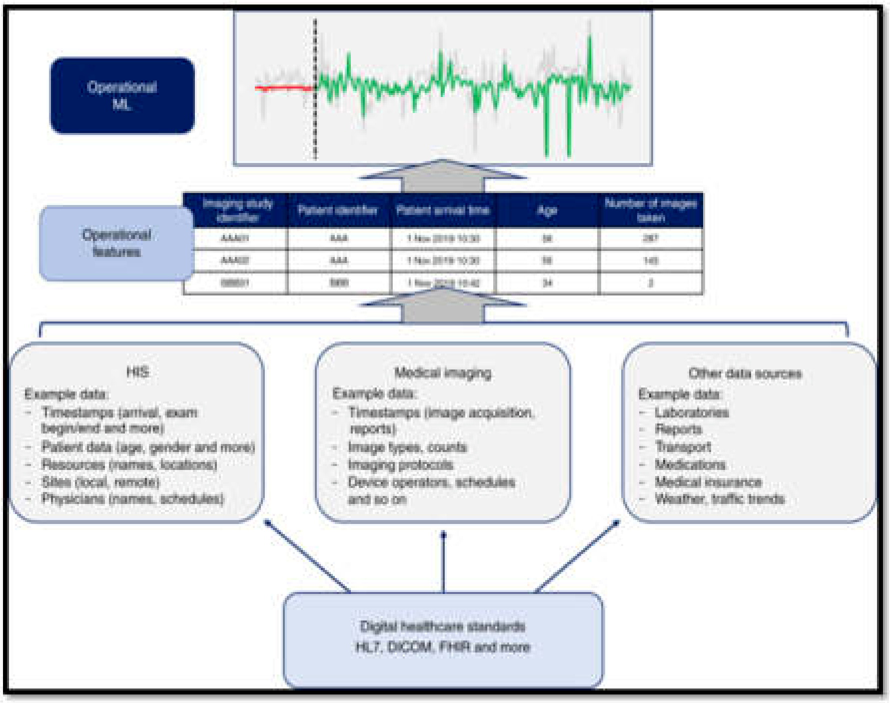


Fig. 5.2: Medical imaging using sensors (Source: [7])

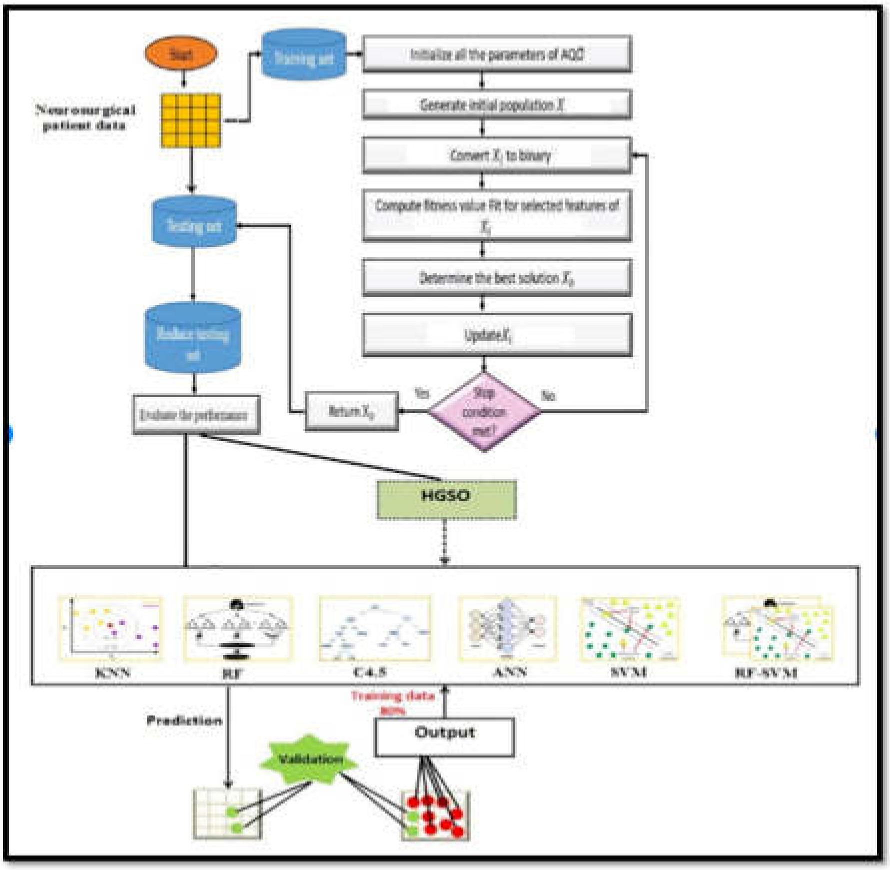


Fig. 5.3: The Machine learning models on AQO (Source: [8])

Machine learning consists of different kinds of features with challenges that are concerned with the physical health development of the user. The features consist of EHRs, BPM calculators, step counting tools, and smart health records that help to record the different health issues in the database of the software used by the user. It also sends reports to emergency contacts in case of any kind of mishappening.

6. Challenges in using ML in healthcare. There are different kinds of challenges that are related to the healthcare of the individual. It refers to the improper recording of the healthcare information of the individuals that may also result in the life risk of the person. The data stored and recorded by the ML software might contain causal information that is ineligible for the benefit of human health [18]. In machine learning, the algorithms might not work properly which can result in the increase of risk and failure of proper fitness to the users. The challenges of ML have been discussed below.

The machine learning algorithms use the different information of the user which is to be maintained by the engineers. The engineers create modern software and sensors that help in the recording of data related to health [12]. The lack of engineers has resulted in the fall of the ML activities among the users which depicts inaccurate health information due to a lack of appropriate sensors. This results in the medical industry creating mishaps among the individuals using the software. It can be further said that most of the smartwatches in the market do not provide appropriate healthcare information for the individual [10]. This is due to the insufficiency of sensors that record the real-time data of the users. The lack of engineers also results in the risk of leakage of health information to a third party resulting in the ineffective flow of work. Distributed networks of sensors are essential for obtaining real-time data during exercises. The data that wearable technology collects include

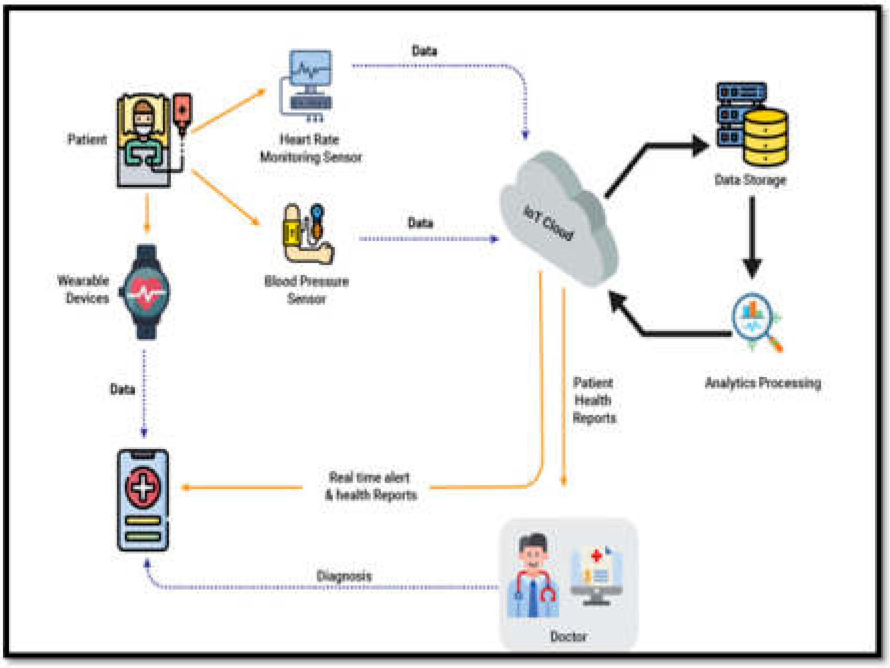


Fig. 6.1: Ways ML can help in managing healthcare (Source: [9])

heart rate, calorie burn, movement patterns, and even muscle activation. Examples include fitness trackers, heart rate tracks, and motion detectors. ML algorithms examine this data to learn more about a person's performance, level of effort, and physiological reactions while exercising.

While performing the task of machine learning, there could be a problem that could lead to the machine learning models not performing well. The bias present in machine learning models can be due to the poor information and data provided to models. For example, if the data that is given to the models contain a lot of information about a particular class and less information about the minority class then it can create a problem among the minority class to use the same for their own benefits in health development [7]. This can create a decrease in the faith in using ML for the identification of healthcare knowledge and adopting the same for the betterment. As a result, the models achieve very high accuracy for upper-class customers while they perform poorly among the lower class.

The lack of information for the development of good software is been one of the major challenges in the use of ML for personalized healthcare. This software uses information from the different hospitals and medical institutions that are updating the different health identification techniques [2]. Without the proper information on the same, the software and the sensors fail to obtain accurate information which results in the downfall of the usage of the machine learning tools. The lack of data and software also results in the failure of the devices to work. This failure can be caused by the attacks of viruses in the system and the software.

The use of unnecessary information creates a risk to the health of the users and the ML also suggests inappropriate options to the users. This cannot be identified by the users which becomes a risk for them in maintaining proper healthcare.

Data are been present all over the place. It is necessary for the ML software to use appropriate data for the identification of the necessary things that are beneficial to human health. Often it has been found that ML uses inaccurate information and data that creates a barrier to the smooth functioning of the human body [3]. It has been further noticed that the use of healthcare software fails to record the data of the individual's health which causes a major drawback in the life of the user. It has been found that the proper execution of data is required to benefit human beings in maintaining good physical health with the use of ML features present in

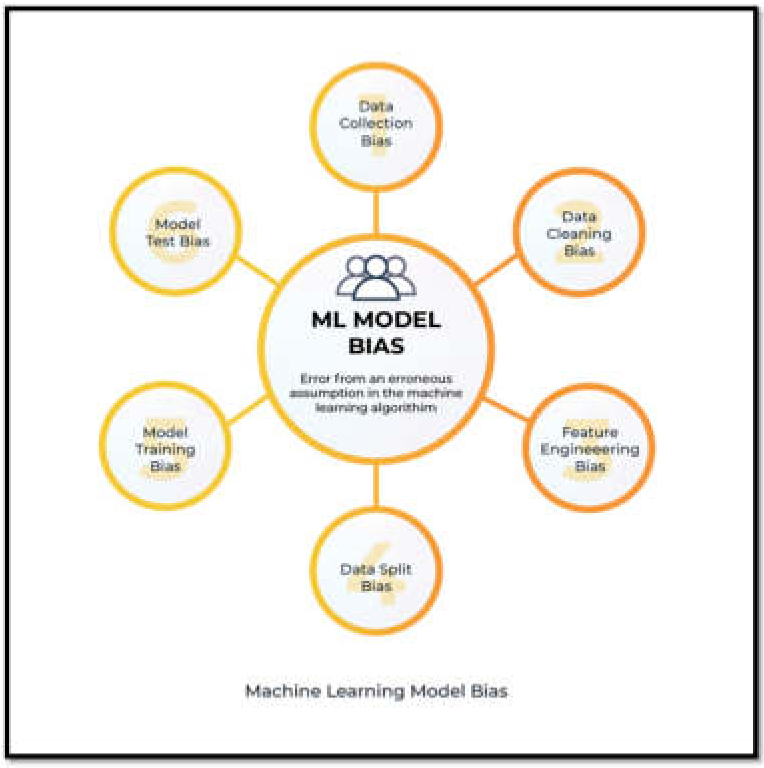


Fig. 6.2: MI model Bias (Source: [10])

different devices.

The use of ML is essential to every human being for maintaining good health. It has been found that the use of many often there has been a lack of maintenance of the data that are stored in the ML software. This results in the failure of accurate usage of the devices. It has been found that the use of ML requires tuning in a repetitive manner. Missing out of the same can result in the failure of working and computing the actual result. It is been one of the major drawbacks of the use of ML software to the development of human health which can be both physical and mental [15]. Therefore it can be said that the use of proper data and techniques along with sufficient tuning will result in the increase of risk among human beings and failure to use ML by the users.

7. Distributed Sensor Networks for Personalized Exercise Program Design. The use of distributed sensor networks helps the use of personalized program designing for the development of health among the users. It has been found that the use of EHRs is helpful for the development of health among individuals. It has been defined as the major factor for recording analyzing and providing effective solutions to the users for the development of good health and maintaining bodily fitness. DSN is presented in smartwatches that are used for the purpose of recording the different levels of fitness and health care functions of an individual. It has been found that the use of fitness watches helps to determine the calories and BPM that indicate the level of fit the user is. It also records the sleep patterns and distance traveled along with the actual location of the user to provide guidance related to fitness wellness. This helps in the monitoring of the user and providence of urgent facilities in case any mishappens occurs. It helps to determine the anomalies and problems that are affecting the health of the individuals. The real-time data of the individual is identified with the help of sensors present in the devices that record the fluctuation of health data and suggest immediate treatments for the same. It also stores the data that can be accessed for a period of 5 years and these data can be restored by accessing the webpage of the device used.

The health of the user is calculated based on the calories burnt and the blood flow level in the body.

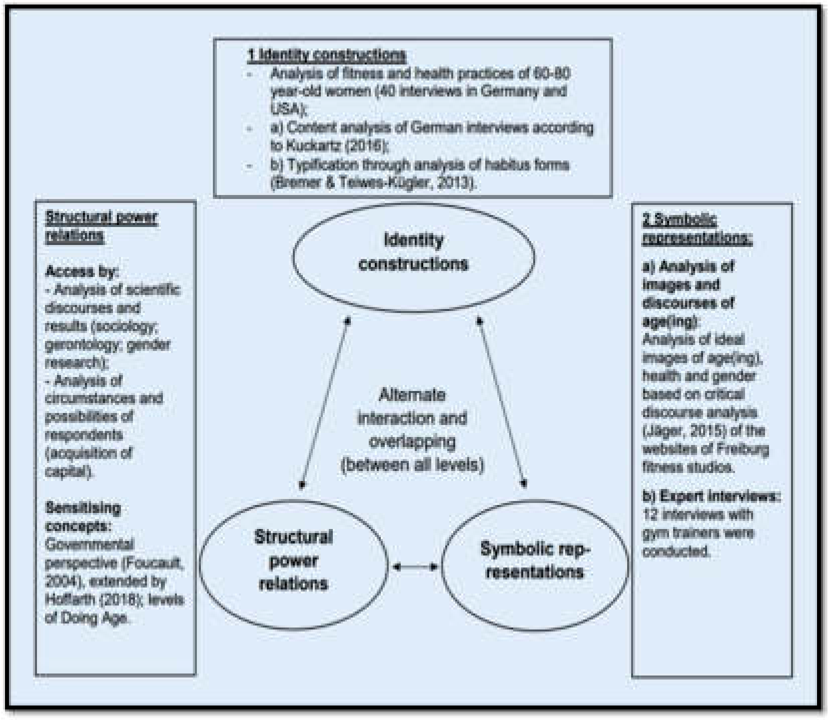


Fig. 6.3: Consequences of Lack of Data and information in the software (Source: [12])

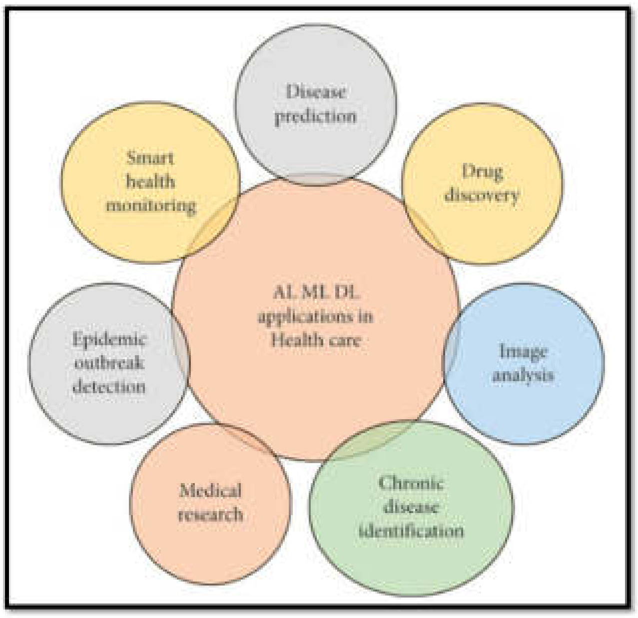


Fig. 7.1: AI ML application features in Healthcare (Source: [15])

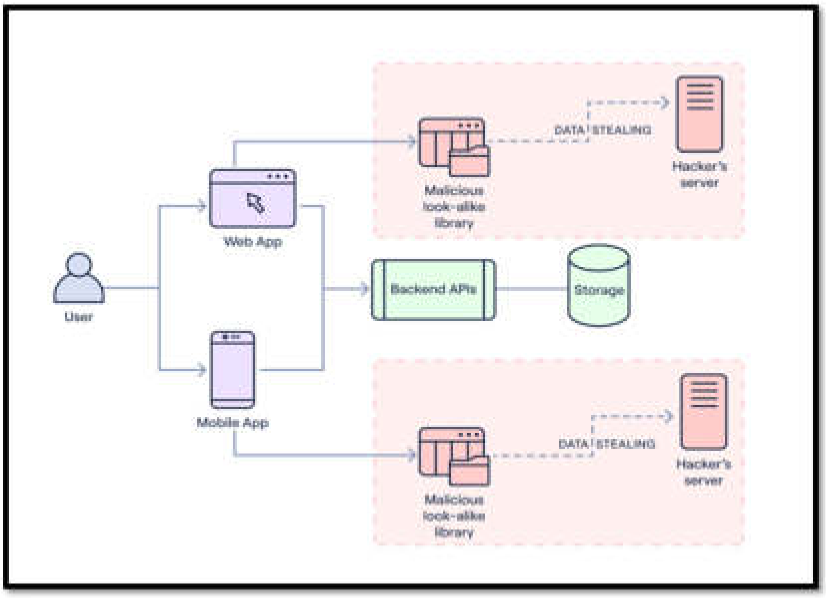


Fig. 8.1: Back end ML features (Source: [11]

Many users of the DSN have found that these are the necessary tools that are beneficial to the development of effective health of individuals. This storing of data also helps in the identification of the psychological and physical health of individuals. The users can set goals and make adjustments for the development of proper well-being in society. These DSN tools contain high-level privacy that is helpful for the effective benefit of the individual [14]. The use of smartwatches and electronic blood pressure measuring tools helps to maintain the privacy of the data evaluated by examining the health of the user.

However, it's important to address privacy and security concerns when collecting real-time data through DSNs.

8. Results. Management of data computed from human beings: Health organizations face massive information-related challenges that can affect patient safety. The different tasks of the health companies create a challenge for the recording of the information of the individuals. Maintaining effective privacy and security, is one of the most important challenges in the healthcare industry is being able to maintain the privacy desired and needed for the patient data and the security for preventing unauthorized access to the database. Privacy is sometimes found to be breached due to the attack of 3rd party in the software of the devices [13]. This creates a misconduct of the several kinds of functions of the individual dealing with the health problems. However, the industry shall keep extreme control and efforts to keep the data private and secure, to keep the trust of the patients and their data safe. A key development in optimizing exercise programs for people is the use of machine learning in distributed sensor networks for personalized exercise program design (PEPD). PEPD develops customized exercise routines that maximize effectiveness, avoid injuries, and encourage enthusiasm by utilizing ML algorithms and real-time data gathered from distributed sensor networks.

Data Retention: Health data must stay accessible for at least five years. It is referred to as one of the major drawbacks of the software. The business is required to keep access to the data of the individual for a long-term period which is found to be missing for the individual. The medical data of the individual is to be kept in the software till the lifetime of the individual. It can be said that the use of these data is helpful for the identification of past health data structures [16]. Therefore, the retention of the same is helpful in the evaluation of the actual performance of the individual and shall get proper treatment from the same.

The research also contains the types and procedures along with risk mitigation strategies that are highly essential to the users for the usage of ML and DSN for their own health benefits. The use of smartwatches is the main component used by individuals to maintain a healthy life by way of recording different health information. The research includes several challenges and procedures are been provided for the mitigation of the risks. This research is been based on the identification of the benefits and challenges of using sensors in smart devices and the use of ML software along with DSNs which helps in the identification of human health and motivates the users for health betterment.

9. Conclusion. The research presented above offers comprehensive insights into the multifaceted landscape of utilizing Machine Learning (ML) and Distributed Sensor Networks (DSN) in the realm of personalized healthcare. By examining the benefits, challenges, and usage protocols associated with ML and DSN, the study uncovers a nuanced understanding of their roles in shaping individual well-being. It is evident that the advantages of integrating ML and DSN for developing Personalized Exercise Program Development (PEPD) are significant. These technologies hold immense promise in redefining how individuals approach fitness and health goals. Despite the clear benefits, it is essential to address the limitations that emerged during the research.

Confidentiality concerns, algorithm precision, and user adoption were identified as noteworthy challenges in implementing PEPD. Confidentiality issues revolve around the sensitive health data being collected and shared, necessitating robust security measures. Algorithm precision, while promising, requires continuous refinement to ensure accurate and relevant exercise recommendations. User adoption presents a hurdle, as not all individuals may be comfortable with wearable technology or the idea of entrusting their health management to algorithms.

Looking ahead, future research should focus on mitigating these limitations. Improved data privacy protocols should be developed to safeguard users' personal information. Algorithm enhancement through advanced machine learning techniques could lead to more precise and adaptable exercise plans. Understanding user psychology and preferences is vital for increasing adoption rates. In the evolving landscape of technology and ML advancements, the potential for PEPD to revolutionize fitness routines remains promising. With concerted efforts to address limitations and refine methodologies, personalized healthcare programs powered by ML and DSN can pave the way for a healthier and more empowered future.

REFERENCES

[1] A. S. Adamson and A. Smith, *Machine learning and health care disparities in dermatology*, JAMA dermatology, 154 (2018), pp. 1247–1248.

[2] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, *Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine*, Database, 2020 (2020), p. baaa010.

[3] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, *Federated learning for healthcare: Systematic review and architecture proposal*, ACM Transactions on Intelligent Systems and Technology (TIST), 13 (2022), pp. 1–23.

[4] M. J. Baucas, P. Spachos, and K. N. Plataniotis, *Federated learning and blockchain-enabled fog-iot platform for wearables in predictive healthcare*, IEEE Transactions on Computational Social Systems, (2023).

[5] S. H. Beck and D. S. Nielsen, *Healthcare professionals' perspectives on the outgoing geriatric team: A qualitative explorative study*, Journal of Ageing and Longevity, 2 (2022), pp. 316–325.

[6] E. Begoli, T. Bhattacharya, and D. Kusnezov, *The need for uncertainty quantification in machine-assisted medical decision making*, Nature Machine Intelligence, 1 (2019), pp. 20–23.

[7] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, *Ethical machine learning in healthcare*, Annual review of biomedical data science, 4 (2021), pp. 123–144.

[8] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, *Big data in healthcare: management, analysis and future prospects*, Journal of big data, 6 (2019), pp. 1–25.

[9] S. Gerke, T. Minssen, and G. Cohen, *Ethical and legal challenges of artificial intelligence-driven healthcare*, in Artificial intelligence in healthcare, Elsevier, 2020, pp. 295–336.

[10] H. Ghayvat, S. Pandya, and A. Patel, *Deep learning model for acoustics signal based preventive healthcare monitoring and activity of daily living*, in 2nd International Conference on Data, Engineering and Applications (IDEA), IEEE, 2020, pp. 1–7.

[11] M. Hartmann, U. S. Hashmi, and A. Imran, *Edge computing in smart health care systems: Review, challenges, and research directions*, Transactions on Emerging Telecommunications Technologies, 33 (2022), p. e3710.

[12] T. McGhin, K.-K. R. Choo, C. Z. Liu, and D. He, *Blockchain in healthcare applications: Research challenges and opportunities*, Journal of network and computer applications, 135 (2019), pp. 62–75.

[13] A. N. Navaz, M. A. Serhani, H. T. El Kassabi, N. Al-Qirim, and H. Ismail, *Trends, technologies, and key challenges in smart and connected healthcare*, Ieee Access, 9 (2021), pp. 74044–74067.

[14] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood, A. Haider, M. Hosseinzadeh, and R. Ali Naqvi, *Machine learning (ml) in medicine: Review, applications, and challenges*, Mathematics, 9 (2021), p. 2970.

[15] S. M. Shortreed, A. J. Cook, R. Y. Coley, J. F. Bobb, and J. C. Nelson, *Challenges and opportunities for using big health care data to advance medical science and public health*, American journal of epidemiology, 188 (2019), pp. 851–861.
[16] J. Waring, C. Lindvall, and R. Umeton, *Automated machine learning: Review of the state-of-the-art and opportunities for healthcare*, Artificial intelligence in medicine, 104 (2020), p. 101822.
[17] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, et al., *Do no harm: a roadmap for responsible machine learning for health care*, Nature medicine, 25 (2019), pp. 1337–1340.
[18] Z. Wu, S. J. Kan, R. D. Lewis, B. J. Wittmann, and F. H. Arnold, *Machine learning-assisted directed protein evolution with combinatorial libraries*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 8852–8858.
[19] H. Yoganarasimhan, *Search personalization using machine learning*, Management Science, 66 (2020), pp. 1045–1070.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Jul 24, 2023
Accepted: Aug 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1169–1176, DOI 10.12694/scpe.v24i4.2535

EXPLORING DATA ACQUISITION AND REAL-TIME ANALYSIS ALGORITHMS IN SMART MANUFACTURING WITH A FOCUS ON AUTOMATION AND INSPECTION TECHNOLOGIES

ZHENSHAN CHEN*, HANJUN ZHENG†, AND JINYAN LIN‡

Abstract. The conducted research project is based on the process followed in the smart manufacturing method. It has been observed that there are different types of methods that have been used in the smart manufacturing process and among these methods most of the methods are directly associated with different modern technology and artificial intelligence as well. Along with the inclusion of automation systems in the smart manufacturing process, the incorporation of inspection technologies is also observed in the case of performing smart manufacturing systems. In association with these, the usage of an acquisition system and real-time analysis process is also observed as well. Independent steps are present inside the phases of production process of the smart manufacturing methods. The incorporation of automation processes significantly enhances the performance of individual phases within the smart manufacturing process, thereby raising the overall production standards. Automation minimizes the likelihood of human errors, data misplacement, and mishandling. Moreover, it proves highly advantageous in the data acquisition process, ensuring error-free data transfer through automated systems.

Key words: Data acquisition software system, Transducer sensor, Superlinear algorithm, Smart manufacturing process

1. Introduction. The term smart manufacturing process refers to the production process where Artificial Intelligence has been used. The smart manufacturing process is capable of performing the production process by following the automation system with the help of AI-based technology. Therefore, a smart manufacturing system is an independent method of producing an item and does not depend on manpower and other human-based factors to produce any item. Due to the presence of such beneficial items, different business sectors all over the world are adopting this technology of smart manufacturing in association with AI. Along with the inclusion of automation systems in the smart manufacturing process, the incorporation of inspection technologies is also observed in the case of performing smart manufacturing systems. Apart from this the conducted project also highlights the application of both the data acquisition process and the real-time analysis algorithms in the smart manufacturing process. Data acquisition involves outperforming an important role in terms of transferring the data in the smart manufacturing process. Moreover, all kinds of algorithm-based statistical and numerical data are further analyzed by real-time data analysis systems. Therefore these two technologies played an important role in the improvement and better performance of the smart manufacturing system.

Smart manufacturing represents a paradigm shift in the way industries operate, leveraging cutting-edge technologies to maximize efficiency and product quality. Automation, driven by artificial intelligence, robotics, and the Internet of Things (IoT), plays a pivotal role in orchestrating an integrated, data-driven production environment. The application of automation technologies not only enhances the performance of individual manufacturing phases but also raises the overall standards of production. With the reduction of human errors, improved data accuracy, and greater precision, automation stands as a cornerstone in the quest for manufacturing excellence.

Simultaneously, inspection technologies have undergone significant advancements, contributing to the real-time assessment of product quality and process performance. These technologies facilitate the constant monitoring of production processes, enabling prompt corrective actions and quality control.

*The Higher Educational Key Laboratory for Flexible Manufacturing Equipment Integration of Fujian Province, Xiamen Institute of Technology, Xiamen, 361021, China (zhenshanres1@outlook.com)
†Xiamen Anscen Technology Co., Ltd., Xiamen, 361021, China
‡The Higher Educational Key Laboratory for Flexible Manufacturing Equipment Integration of Fujian Province, Xiamen Institute of Technology, Xiamen, 361021, China

2. Objectives. The primary objectives of the project are illustrated in the following section.

1. To describe the overview of the data acquisition system
2. To identify the role of data acquisition system in the smart manufacturing process
3. To understand techniques used in the real-time analyzing process of algorithms
4. To find out the application of automation process in different aspects of the smart manufacturing system
5. To recognize different types of methods that are used under the inspection technologies of the smart manufacturing process

The incorporation of automation systems and inspection technologies within the smart manufacturing framework emerges as a key focal point, showcasing how these technologies optimize various phases of the production process. Moreover, this research sheds light on the pivotal role of data acquisition systems and real-time analysis methods in this context, emphasizing their impact on enhancing manufacturing efficiency. By elucidating the individual steps within the production phases, this study underlines the transformative potential of automation, which significantly elevates production standards while reducing the likelihood of human errors and data mishandling. Additionally, the error-free data transfer facilitated by automated systems represents a noteworthy advancement in the smart manufacturing landscape. This research makes a valuable contribution to the understanding and implementation of smart manufacturing processes, setting the stage for more efficient, accurate, and technologically-driven production methodologies.

3. Methodology. Different information based on the smart manufacturing process is discussed in this project. Among different data mostly secondary types of data have been used in this project. With the help of secondary information from other writers about the manufacturing process and the other technology and technical gadgets used in the smart manufacturing process illustrated in this project and by this secondary type information, the final result of the project can also be developed [1]. Hence secondary data is collected for developing the final outcome of the project and this step is a clear indication of performing the qualitative methods in order to collect information for the completion of the project.

Data acquisition systems begin with sensors and transducers that capture physical parameters such as temperature, pressure, voltage, current, or other environmental conditions. These sensors convert physical measurements into electrical signals. In many cases, the raw electrical signals from sensors need to be conditioned to ensure accuracy and compatibility with the data acquisition system. Signal conditioning may include amplification, filtering, analog-to-digital conversion, and other adjustments to the signal.

3.1. Overview of the Data Acquision System. The data acquisition system contains software and different devices as well. This software and devices are involved in measuring the data while transferring the data from one system to another. The application of the sensor system is also observed in the data acquisition system [2]. With the help of a sensor system, the data acquisition system develops the ability to sense the data and its transfer process as well. Moreover, it has been observed that the data acquisition software is involved in performing different functions as well. Acquiring the data, and storing the data in different devices all are performed by a data acquisition software system. Moreover visualizing the data on the device screens are also performed by the software of the data acquisition system [3]. In addition to that the software system is also capable of processing data and modifying the data making it ready for further utilization of the data by the users. In association with understanding the electrical wave, the data acquisition system is capable of transferring the data from one end of the device to the other end of the device as well [23].

The concerned diagram of the above section highlights the different sections of the Data Acquisition system [22]. In order to transfer data from one computer to another the data has to pass by four different phases. Each phase of the data transferring process is involved in processing the data and making the data ready for transferring to another device. The four different phases are the "physical system, transducer sensor, signal conditioning, and the analog-digital converter" [5]. In all four phases of the data acquisition system different signals are developed and by accepting these signals the device of another end of the transferring process can be able to accept the data.

3.2. Utilization of Data Acquisition System in the Smart Manufacturing Process. The application of the data acquisition system is involved in the inclusion of the automation system in the data transferring

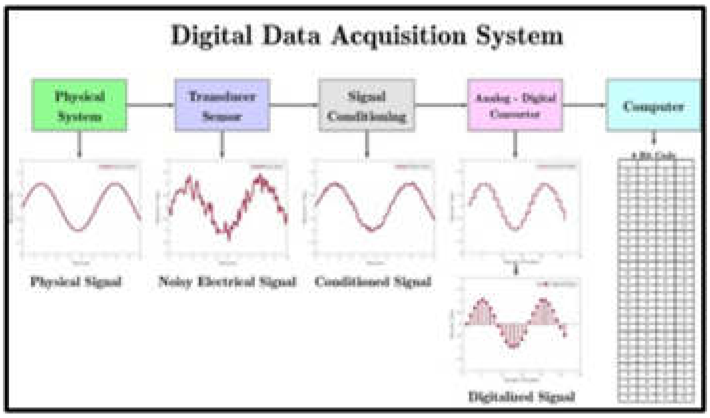


Fig. 3.1: Digital Data Acquisition System

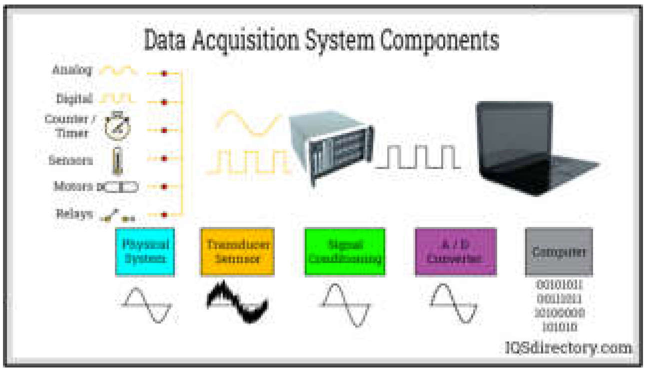


Fig. 3.2: Utilization of data acquisition system

and in the data managing system. The utilization of the data acquisition system is involved in lowering manual errors. Moreover, it has been observed that when the data processing, data managing, and data transferring system is performed by the different software applications [4]. This lowers the participation of any manual recess and furthermore, this also lowers the occurrence of any kind of human error as well. Furthermore, in the case of using this system in the smart manufacturing process, it also benefited by lowering the incidents of any kind of misplacing and misconducting the data. By reducing the misplacing of the data the rate of the accuracy of the transferred data is increased to a high level [6].

Moreover, it has been observed that the usage of digital data acquisition systems is also involved in reducing the overall charges of the data acquisition process that is performed under smart manufacturing procedures [21]. When the methodology and the techniques of the data acquisition process shifted from a manual process to a digital process. In this shifting procedure, the charges of performing different types of physical data managing and data acquisition steps are removed which helps lower the overall estimated costs of the data acquisition process [7]. In addition to that, the inclusion of a digital data acquisition system is also beneficial in terms of performing the retrieval process of the data.

3.3. Techniques followed in Real-time Analysis Algorithms. There are three different techniques that have been observed which are involved in performing the real-time analysis process of algorithms. These three techniques are based on the three different designs of the algorithm structures. Therefore three different algorithm designs are "divide-and-conquer, dynamic programming, and greedy heuristics". Apart from three different algorithm designs, different examples of the real-time analysis of algorithms are "Linear algorithm

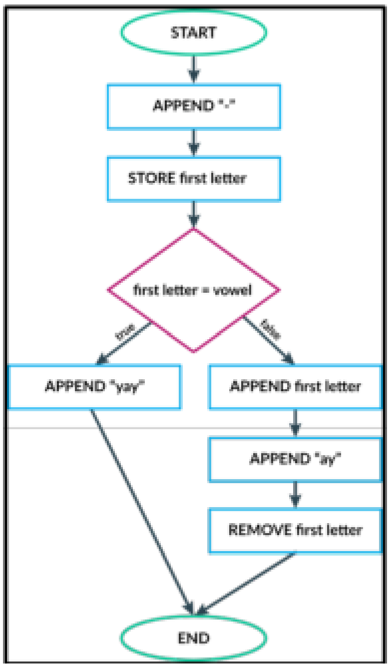


Fig. 3.3: Flow chart of an algorithm analysis

– O(n) and Superlinear algorithm – O(nlogn), Polynomial algorithm – O(n∧c), and Exponential algorithm – O(c∧n)" [8].

This above-developed flow chart fig. 3.3 contains different steps which are involved in developing the visualization of the whole steps that are involved in completing the analysis of different algorithms. In this flowchart, the application of natural language has been observed and as a result of this, these steps involved process also can be performed naturally as well [9]. With the help of using natural languages, this flowchart can be able to connect maximum numbers of users by its easy programming capacity. Apart from that it has been observed that the application of natural languages is also associated with different limitations such as the application of natural languages in the algorithm analysis process causing the presence of difficulties in the whole structure of the process [10].

4. Role of Automation in Smart Manufacturing. The inclusion of an automation system in the smart manufacturing process is involved in creating a fully autonomous production process. Moreover, the automation system in the smart manufacturing process is also capable to develop an interconnection between the steps that are performed in the supply chain process under the smart manufacturing system [11]. The automation process allows the application of different types of robotics, AI, and different other software applications in order to make the smart manufacturing process a fully autonomous [production process.

The inclusion of an automation system in the smart manufacturing process is directly involved in offering different types of benefits. These benefits are further involved in enhancing the efficiency of the smart manufacturing process. Cost-effectiveness is one of the most important benefits that the smart manufacturing process can have after the inclusion of the automation system [12]. Moreover, the inclusion of an automation system is helpful for improving the worker's safety and lowering any kinds of hazardous risks in the manufacturing process as well.

4.1. Types of Inspection Technology. There is a wide range of technologies available which are helpful for performing the inspection process. These inspection technologies are involved in checking the quality of the final products and also give the declaration of the high-quality product. In the case of the smart manufacturing process, the application of Automated inspection technology (AIS) has been observed in the

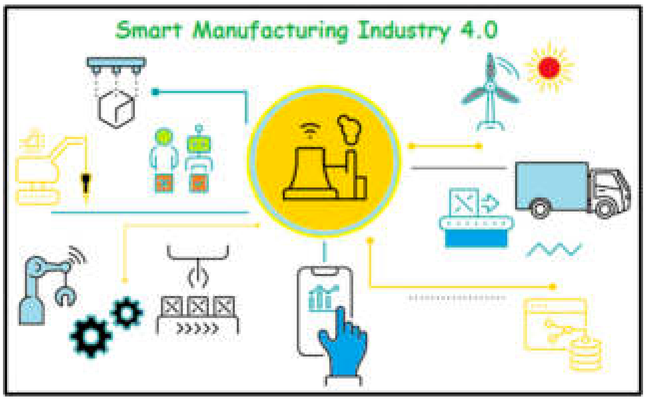


Fig. 4.1: Role of Automation in Smart Manufacturing

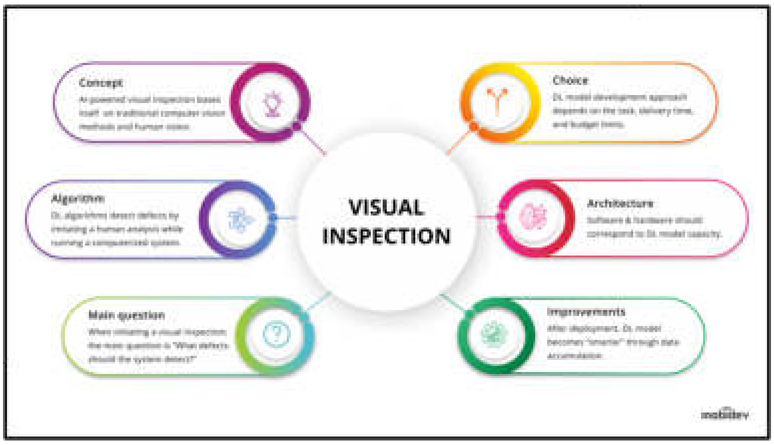


Fig. 4.2: Visual inspection

largest amount [13]. The application of this AIS is involved in improving the quality of the product to a high level. Apart from this method, it has been observed that there are different methods that also contributed to the develop,emy of inspection technologies. Hence the different other technologies that can also be used as examples of inspection technologies are "nondestructive testing, radiography, magnetic particle inspection, automated optical inspection, ultrasound, acoustic emission, visual inspection, and thermography" [14].

Apart from this technology in the smart manufacturing process mainly four different methods are observed which are involved in the inspection process of the quality control process. The four different methods that are performed under the manufacturing process are "pre-production inspection (PPI), first article inspection (FAI), during production inspection (DPI), pre-shipment inspection (PSI), container loading inspection (CLI)" [15]. Proceeding with all these technologies and steps is helpful for enhancing the overall effectiveness of the smart manufacturing process.

5. Results. The result of the project is based on the data and information discussed in the previous section. It has been observed that the inclusion of different automated software applications is considered one of the most important steps in order to maximize the output of the smart manufacturing process [16]. It has been observed that with the help of the automated system, the users of the smart manufacturing process can be able to add an autonomous strategy of performing different steps of the production recess. This helps to increase the rate of production even in the least amount of time. Due to its transferring into an autonomous

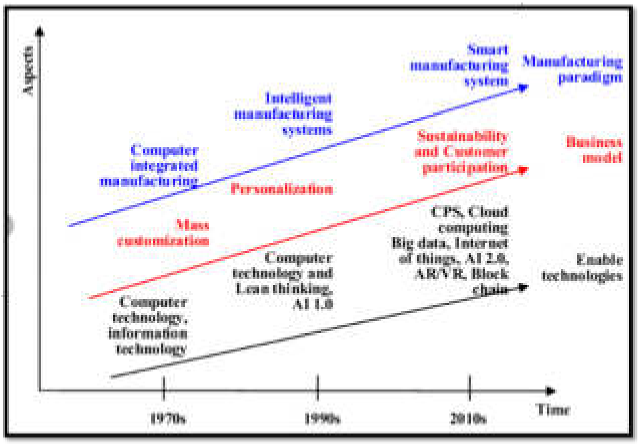


Fig. 5.1: Application of SMART manufacturing process

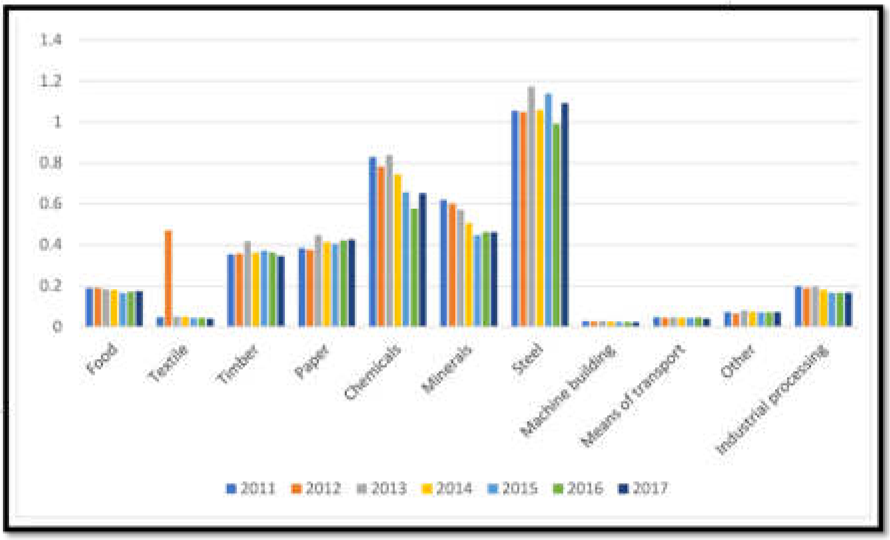


Fig. 5.2: SMART manufacturing process in different business industries

process, the time required to produce a product becomes reduced than the usual one [17].

Moreover, it has been observed that the inclusion of the digital data acquisition system is also involved in enhancing the quality of the data management and the data transferring process. By enhancing the data transferring process the utilization of the data can be also enhanced in the smart manufacturing system [18]. When the utilization of the data becomes better than the final output that is developed by the utilization of the data it also gets better.

The betterment of the data acquisition process is also considered beneficial for analyzing the algorithm data as well. It has been observed that due to the improvement of the algorithm, the data management process of the smart manufacturing system is getting better than the existing one [19]. Apart from that inspection technology is further involved in enhancing the overall quality of the manufacturing process that is totally based on smart technology and autonomous artificial intelligence as well. Therefore it can be stated that the application of all four different technologies finally impacted the positive development of the smart manufacturing system [20].

6. Conclusion. Thus it can be concluded that the concerned project highlights the smart manufacturing process and different technologies that are available and can be included in the process of smart manufacturing. It has to be observed that the inclusion of an automation system is considered one of the most effective software applications that can increase the accuracy rate of the smart manufacturing system and are also involved in improving the steps that are performed under the autonomous production process as well. Apart from this, the data acquisition system is also involved in improving the data-transferring process of the smart manufacturing process. Furthermore, the implementation of a data acquisition system plays a pivotal role in optimizing data transfer processes within the smart manufacturing landscape. The harmonious amalgamation of these technologies goes beyond just optimizing production; it elevates the overall quality, efficiency, and precision of the smart manufacturing process. As a result, this research illuminates the transformative potential of these technologies, charting a path towards more advanced, efficient, and high-quality smart manufacturing practices.

REFERENCES

[1] H. S. Bedi, K. K. Sharma, and R. Gupta, *A review paper on performance analysis of ieee 802.11 e*, in Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019), Springer, 2020, pp. 47–56.

[2] L. L. Bello and W. Steiner, *A perspective on ieee time-sensitive networking for industrial communication and automation systems*, Proceedings of the IEEE, 107 (2019), pp. 1094–1120.

[3] H.-M. Chung, S. Maharjan, Y. Zhang, F. Eliassen, and K. Strunz, *Placement and routing optimization for automated inspection with unmanned aerial vehicles: A study in offshore wind farm*, IEEE Transactions on Industrial Informatics, 17 (2020), pp. 3032–3043.

[4] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and A. Parmentola, *Smart manufacturing systems and applied industrial technologies for a sustainable industry: A systematic literature review*, Applied Sciences, 10 (2020), p. 2897.

[5] J. Falk, D. Hellmanns, B. Carabelli, N. Nayak, F. Dürr, S. Kehrer, and K. Rothermel, *Nesting: Simulating ieee time-sensitive networking (tsn) in omnet++*, in 2019 International Conference on Networked Systems (NetSys), IEEE, 2019, pp. 1–8.

[6] P. Feng, Q. Li, P. Zhang, and Z. Chen, *Private data acquisition method based on system-level data migration and volatile memory forensics for android applications*, IEEE Access, 7 (2019), pp. 16695–16703.

[7] H. Guan, X. Sun, Y. Su, T. Hu, H. Wang, H. Wang, C. Peng, and Q. Guo, *Uav-lidar aids automatic intelligent powerline inspection*, International Journal of Electrical Power & Energy Systems, 130 (2021), p. 106987.

[8] K. Haricha, A. Khiat, Y. Issaoui, A. Bahnasse, and H. Ouajji, *Recent technological progress to empower smart manufacturing: Review and potential guidelines*, IEEE Access, (2023).

[9] I. Kevin, K. Wang, X. Zhou, W. Liang, Z. Yan, and J. She, *Federated transfer learning based cross-domain prediction for smart manufacturing*, IEEE Transactions on Industrial Informatics, 18 (2021), pp. 4088–4096.

[10] X. Kong, K. Wang, S. Wang, X. Wang, X. Jiang, Y. Guo, G. Shen, X. Chen, and Q. Ni, *Real-time mask identification for covid-19: An edge-computing-based deep learning framework*, IEEE Internet of Things Journal, 8 (2021), pp. 15929–15938.

[11] J. Le Kernec, F. Fioranelli, C. Ding, H. Zhao, L. Sun, H. Hong, J. Lorandel, and O. Romain, *Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms*, IEEE Signal Processing Magazine, 36 (2019), pp. 29–41.

[12] Y. Li, K. Li, C. Zhang, J. Montoya, and G.-H. Chen, *Learning to reconstruct computed tomography images directly from sinogram data under a variety of data acquisition conditions*, IEEE transactions on medical imaging, 38 (2019), pp. 2469–2481.

[13] X. Liang, A. Kukko, I. Balenović, N. Saarinen, S. Junttila, V. Kankare, M. Holopainen, M. Mokroš, P. Surovỳ, H. Kaartinen, et al., *Close-range remote sensing of forests: The state of the art, challenges, and opportunities for systems and data acquisitions*, IEEE Geoscience and Remote Sensing Magazine, 10 (2022), pp. 32–71.

[14] S. Liu, S. Wang, X. Liu, C.-T. Lin, and Z. Lv, *Fuzzy detection aided real-time and robust visual tracking under complex environments*, IEEE Transactions on Fuzzy Systems, 29 (2020), pp. 90–102.

[15] E. A. Oyekanlu, A. C. Smith, W. P. Thomas, G. Mulroy, D. Hitesh, M. Ramsey, D. J. Kuhn, J. D. Mcghinnis, S. C. Buonavita, N. A. Looper, et al., *A review of recent advances in automated guided vehicle technologies: Integration challenges and research areas for 5g-based smart manufacturing applications*, IEEE access, 8 (2020), pp. 202312–202353.

[16] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, *Joint power and blocklength optimization for urllc in a factory automation scenario*, IEEE Transactions on Wireless Communications, 19 (2019), pp. 1786–1801.

[17] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, *Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19*, IEEE reviews in biomedical engineering, 14 (2020), pp. 4–15.

[18] S. Shim, J. Kim, G.-C. Cho, and S.-W. Lee, *Multiscale and adversarial learning-based semi-supervised semantic segmentation approach for crack detection in concrete structures*, IEEE Access, 8 (2020), pp. 170939–170950.

[19] S. K. Vishwakarma, P. Upadhyaya, B. Kumari, and A. K. Mishra, *Smart energy efficient home automation system using iot*, in 2019 4th international conference on internet of things: Smart innovation and usages (IoT-SIU), Ieee, 2019, pp. 1–4.

[20] Z. Wang, Y. Bian, S. E. Shladover, G. Wu, S. E. Li, and M. J. Barth, *A survey on cooperative longitudinal motion control of multiple connected and automated vehicles*, IEEE Intelligent Transportation Systems Magazine, 12 (2019), pp. 4–24.
[21] S. Yin, J. J. Rodriguez-Andina, and Y. Jiang, *Real-time monitoring and control of industrial cyberphysical systems: With integrated plant-wide monitoring and control framework*, IEEE Industrial Electronics Magazine, 13 (2019), pp. 38–47.
[22] L. Yu, E. Yang, P. Ren, C. Luo, G. Dobie, D. Gu, and X. Yan, *Inspection robots in oil and gas industry: a review of current solutions and future trends*, in 2019 25th International Conference on Automation and Computing (ICAC), IEEE, 2019, pp. 1–6.
[23] X. Zhou, A. Chen, M. Usman, Q. Chen, F. Xiong, J. Wu, and H. Chen, *Underground pipeline mapping from multi-positional data: Data acquisition platform and pipeline mapping model*, IEEE Transactions on Geoscience and Remote Sensing, (2023).

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 24, 2023
Accepted: Oct 30, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1177–1184, DOI 10.12694/scpe.v24i4.2546

ANOMALY DIAGNOSIS METHOD AND CONDITION ASSESSMENT OF POWER METERING DEVICE BASED ON SSD ALGORITHM

CHAO JIANG[*], JINGHUA WANG[†], YUE WANG[‡], AND WANRU ZHAO[§]

Abstract. The advancement of anomaly diagnosis methods plays a crucial role in classifying and analyzing data, particularly in distinguishing between normal and abnormal patterns. This study explores the utilization of Support Vector Machine (SVM) techniques to facilitate the selection of pertinent data, thus enhancing the accuracy of anomaly detection. Furthermore, this research delves into the condition assessment process, which offers valuable insights into the real-time state of entities passing through power metering devices. Drawing upon a wealth of secondary data sources, this study employs the SSD algorithm to gain a comprehensive understanding of power metering devices and their interrelated aspects. The SSD algorithm, with its diverse anchor points, is revealed as a powerful tool for quantifying the energy flow passing through the power metering device. This approach not only aids in precise energy measurement but also provides essential insights into the functioning of power metering systems. By combining anomaly diagnosis, SVM techniques, and the SSD algorithm, this study contributes to a deeper comprehension of power metering devices' performance and their capacity to accurately measure energy. These insights have significant implications for improving the overall reliability and efficiency of power metering in various applications.

Key words: anomaly diagnosis, secondary data, SSD algorithm, power metering devices

1. Introduction. The anomaly diagnosis method was applied for the detection of the quality of the data. This process differentiates the data into normal and abnormal forms and helps to level the data according to the characteristics. The implementation of the anomaly diagnosis method is important for the development of power metering devices. A power metering device is important for measuring the pulses of energy that pass through it. The development of power metering based on the SSD algorithm helps to manipulate all the things or systems of passing the electric energy in a systematic way. This study has analyzed the implication of the anomaly diagnosis method and condition assessment on the development of SSD algorithm-based power metering devices.

The network control used in the power generations id the crucial backbone of the networks. It is physically isolated from the networks that are outsides. In case of any issue, abnormality or the problems within the network than it can directly impact or influence competiveness in relation with the power plants. It makes the power plants very crucial in detecting or tracking of and the prevention of the challenges and the issues proactively not by just responding to them after they have established. The continuous development and the growth as well as the construction of the power grid and its scale are expanding on the large scale. The amounts of the map that are infrared are assembled by the thermal images. It has become very crucial and vital problem by the usage of the procedures that are intelligent algorithm for the approach massive infrared image data at times.

Anomaly detection, a critical aspect of data analysis and pattern recognition, has gained prominence in various fields and applications, including cybersecurity, industrial automation, and quality assurance. The primary objective of anomaly detection is to identify data points or patterns that deviate significantly from the norm or exhibit unusual behaviour. If left undetected, such anomalies can have far-reaching consequences, from security breaches and equipment failures to financial losses. In recent years, the development of advanced algorithms and machine learning techniques has substantially improved the accuracy and efficiency of anomaly

*School of Economics and Management, Shanghai University of Electric Power, Shanghai 200090, China (chaojiangres@outlook.com)
†School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University Shanghai, 200000, China
‡State Grid Shanghai Marketing Service Center, Shanghai 200030, China
§State Grid Shanghai Marketing Service Center, Shanghai 200030, China

detection systems. These advancements have enabled organisations to proactively identify deviations from expected behaviour and take timely corrective actions.

2. Objectives.

1. To understand the effect of the anomaly diagnosis method on the development of power metering devices which is based on the SSD algorithm
2. To know the impact of condition assessment on the developmental processes of power metering devices
3. To evaluate the issues arrived at regarding the development of SSD algorithm-based power metering devices
4. To suggest effective ways to implicate the SSD-based algorithm in the field of machine learning

As we embark on this exploration, the objective is to unravel the capabilities and implications of these advanced anomaly detection methods, particularly emphasising their potential to improve the accuracy and reliability of power metering systems. This research strives to contribute to the broader field of anomaly detection and its diverse applications, offering insights into how innovative approaches can revolutionise data analysis and decision-making processes.

3. Methodology. This study has been prepared based on secondary sources of data. Secondary sources like journals and articles have been used as sources. This increases the knowledge of the research topic and helps to analyse the details. The implication of the anomaly diagnosis method is important for developing an effective SSD algorithm as the quality of the data depends on the support vector machines (SVM), which differentiate the data according to the parent characteristics [14]. Thus, the secondary sources and method of the data analysis help to know the effect of the anomaly diagnosis method and condition assessment on the development of the power metering device.

3.1. Effect of anomaly diagnosis method on the development of the SSD algorithm-based power metering device. Power metering devices represent a crucial component in contemporary power distribution systems, serving the vital function of accurately and consistently channelling electrical currents and other ions. In the early stages of power metering device innovation, these systems were primarily analogue and mechanical [24]. However, with the rapid technological advancements, modern power metering devices have evolved to incorporate sophisticated machinery and employ the SSD (Solid-State Device) algorithm, greatly enhancing the precision and reliability of energy measurement as it traverses through these metering devices.

As an integral part of this evolution, the anomaly diagnosis process assumes a pivotal role in ensuring the input of accurate data into these devices. This process is instrumental in elevating the accuracy of power metering devices to a high degree, contributing to the overall enhancement of energy measurement and monitoring [21]. This increases the utilisation of the power metering devices as this helps to know the accurate amount of electrical division and supply units. The anomaly diagnosis method used the efficient support vector machine to differentiate the data sources based on their characteristics [16]. The quality of the SSD algorithm is based on the input quality of the data. The quality of the inputted data increased with the implication of the anomaly diagnosis method.

Figure 3.1 represents the steps of the anomaly diagnostic method. The main task of the anomaly diagnostic method is related to the identification of the problems in the data [12]. This process helps in the development of better-quality of data [20]. This is mainly based on the support vector machines which utilized for regression and classification of the data.

The anomaly diagnosis method helps in the development of the algorithm which helps in the development of the power metering device. Power metering device helps to measure the frequency of data passing from the device [19]. This measures the amount of energy that passes through it accurately as well as this helps in the distribution of the energy according to need. Anomaly diagnosis method is mainly divided the normal and abnormal inputted data and helps in the process of SSD algorithm development [11]. The SSD algorithm-based power metering device represents the accurate values than the traditional analogue meters. This helps to keep the records and assists to manipulate the consumption of the total quantity of output [26]. Thus, the implication of the SSD algorithm is important for the development of the power metering device.

The complex background of the sustain is a very crucial factors that impacts the diagnosis and identification of the infrared thermal images in the management of the equipment that are sustainable. The more refined

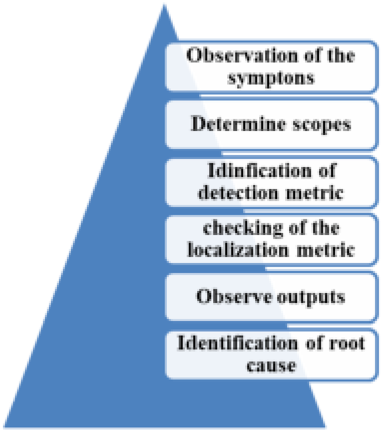


Fig. 3.1: Steps of the anomaly diagnostic method

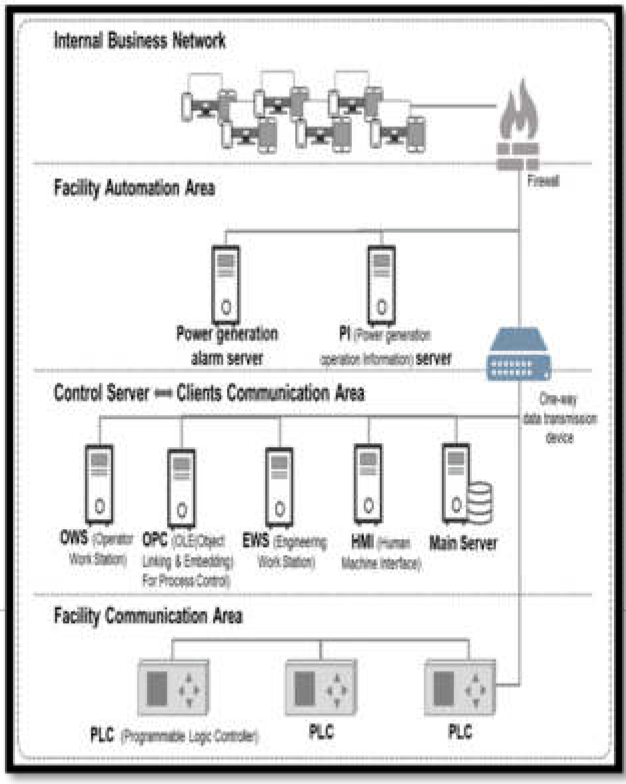


Fig. 3.2: The Network of the SSSD algorithm

elements that are local are characteristics or the features related to the power equipment that are usually obtained from the multi-scale convolution kernel and the substance related to the detection of the algorithms. There is the continuous update and the development in the networks that are neutral as well as convolutional. The deep learning or knowledge approaches and techniques in order to detect or track power supply by the usage data mining method for the achievement to the fault diagnosis

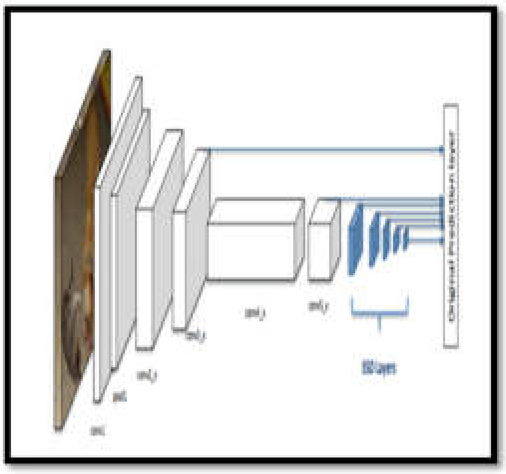


Fig. 3.3: SSD detector

3.2. Meaning of condition assessment process and its implication in the development of power metering device. Condition assessment is the process to measure the condition of the equipment and the processes involved in the development of the algorithm and algorithm based machinery. The condition assessment is mainly divided into two stages such as preliminary and detailed analysis. These distinctive stages of the analysis help in measuring the faults and level of accuracy maintained in the development of the model [6]. This system measures the safety of the processes and helps to understand all the things in detail. The importance of efficient power metering devices is high as this represents the value of the energy passes from the meter [15]. The facilities of the anomaly diagnosis analysis and the quality of the algorithm are determined through the analysis of the condition assessment [28]. Condition assessment is the process to develop a highly qualitative algorithm for the development of SSD based power metering devices. Power metering devices are generally used for the measurement of electrical power. The implementation of the algorithm increases the quality of these machines as this state the highly accurate result [29]. In the modern times, people used to implicate algorithm based machine learning as this helps to calculate the pulses that pass from the machine accurately. Power meters help to analyse the average power of an illumination of a light [15]. This measures the actual power of the laser beam. A power meter analyses the weaknesses and strength of any event. There are different types of pedal-based, hub-based, and crank-based. The implication of these power meters is different to each other.

3.3. The challenges occurred during the development of the power metering device based on the SSD algorithm. The most challenges that occur from the power metering devices are the excessive time taken for integrating the advanced metering infrastructure, legacy in billing and so on. SSD algorithm uses convolutional filters through which it understand the object's scale change very well [27].

Power meter device has been facing many challenges during the distribution of electricity and many houses also use illegal connections through the distribution of pole grids. At present many systems have been used to reduce power theft and through this system power imbalance can be detected.

Power meter devices have grappled with a myriad of operational challenges, encompassing both technical and non-technical issues. Technical losses, stemming from cable faults and overhead line issues, have posed significant operational hindrances [5]. Additionally, non-technical operational losses, such as electricity theft, meter tampering, cable theft, and illegal electrical connections, have further exacerbated these challenges. The culmination of these issues translates into substantial financial losses for governmental bodies and power utilities. Notably, electricity theft stands out as a primary contributor to the financial strain faced by electricity departments, simultaneously exerting a detrimental impact on a country's economic growth.

The resulting electricity shortages frequently lead to power outages, a phenomenon known as load shedding.

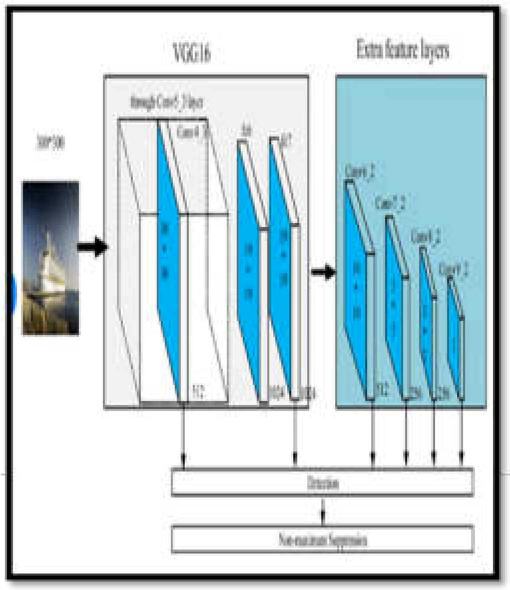


Fig. 3.4: SSD algorithm structure

To address this issue, the adoption of smart power meters, equipped with advanced features, presents a promising solution [1]. Incorporating the SSD algorithm within power meter devices prioritizes image-based techniques. This approach offers the advantage of reducing errors associated with human factors [7]. Image-based techniques rely on scene text detection, enabling automatic inspections that significantly decrease the need for manual intervention, ultimately saving on human resources and enhancing operational efficiency.

SSD algorithm represents an object detector to detect the serial numbers on meter images which are located in residential area. This process also gives a bad impact while capturing images of meters in a complex environment as well as capturing images in an error-prone pattern [3]. As a result, it reduces the quality of the meter images and creates difficulties to understand the images properly.

4. Results. This study has found the effectiveness of the anomaly diagnosis method and condition assessment processes for the development of power metering devices which are based on the SSD algorithm. Anomaly diagnosis methods are applied for the management of the data as that helps in the development of the algorithm [10]. SSD algorithm is an algorithm which detects the object through the process of deep learning. This helps to measure the detection speed as well as detection accuracy [23]. SSD based algorithm is important as this helps to predict the object passing through this. SSD includes different filters to detect the problems and helps in the management of the overall quality of the result [25]. The algorithm-based power metering devices represent the result accurately and it is more conventional than the traditional system included in the power metering devices.

The anomaly diagnosis method classifies all the things based on the character [2]. SSD algorithm is comparatively more robust and manages to determine the objects whose scales change repeatedly. This is different from R-CNN as that classified the objects based on the different pixels of the objects rather than based on the algorithm as like SSD.

The condition assessment of the power metering devices is updated more frequently as all the updated information is available from the monitoring devices. Condition assessment measures the condition of facilities in terms of health and asset [22]. SSD algorithm evaluates all the interrelated things of the power metering devices through the digitalized system. This observes the signals from the receivers and measures the strength and frequency of the beam that passes through it [18].

The power metering devices are mainly empowered to measure the energy that passes through them and

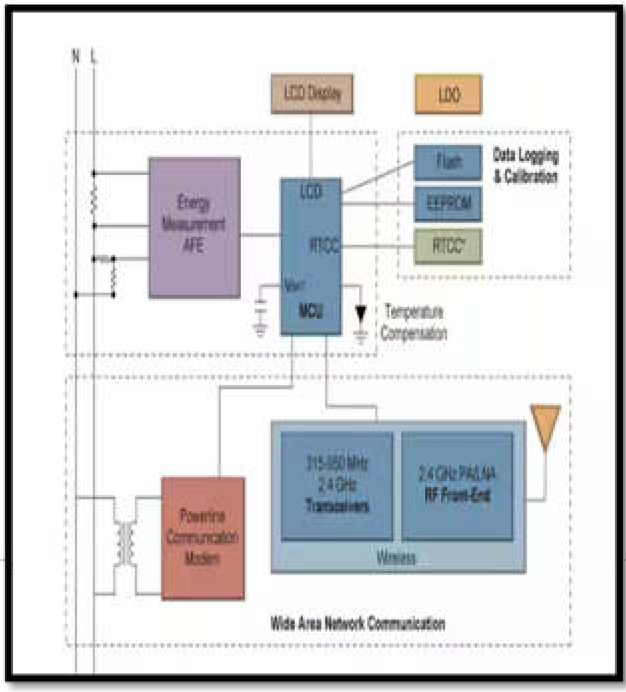


Fig. 4.1: Power metering devices

to equally distribute the energy to all the devices. The power meter is a simple instrument that is considered very useful to measure electrical power and no measured data analysis is needed here [4]. Technicians as well as engineers were given importance on power meter to performing power measurements. It is also considered a cost effective solution for power measurement as well as measure voltage and currents [13]. Power meter is also used for measuring grid power and it measures parameters in direct current as well as alternative current.

The figure 4.2 represents the demand of power meter devices increasing from the 2020 to 2027. The market size of the power meter device was 10.4 billion US dollars to 2020 and itt increased to 11.8 billion US dollars in 2022. SSD is an object detector device which uses grid cell to detect objects in a particular region of an image [9]. Detecting objects represent the class and location prediction of an object through a particular region.

Power metering devices are utilized in the different sectors to increase the safety as well as to focus on energy consumption [17]. The transformation of the traditional power metering to the modern digitalized power metering could be possible through the implication of algorithm. Power metering devices measured the fluctuation of the distribution and measure the health of the energy passing through it [8]. The implication and requirement of the power metering device is considerably high for measuring the issues related to the energy flow.

5. Conclusion. This study has yielded valuable insights into the efficacy of anomaly diagnosis methods and condition assessment processes in the development of power metering devices, employing the innovative SSD algorithm. Anomaly diagnosis methods have proven instrumental in data management, facilitating algorithm development [10]. The SSD algorithm, rooted in deep learning, exhibits remarkable capabilities in detecting objects with impressive speed and accuracy. This algorithm is pivotal in predicting the objects passing through, incorporating a range of filters to identify issues and optimize result quality. Algorithm-based power metering devices have displayed remarkable precision, surpassing conventional systems. Anomaly diagnosis, which classifies objects based on their characteristics, benefits from the robustness of the SSD algorithm, effectively identifying objects with changing scales—a departure from the pixel-based classification used in systems like R-CNN. Power meters serve as invaluable tools for calculating electricity consumption and providing real-time

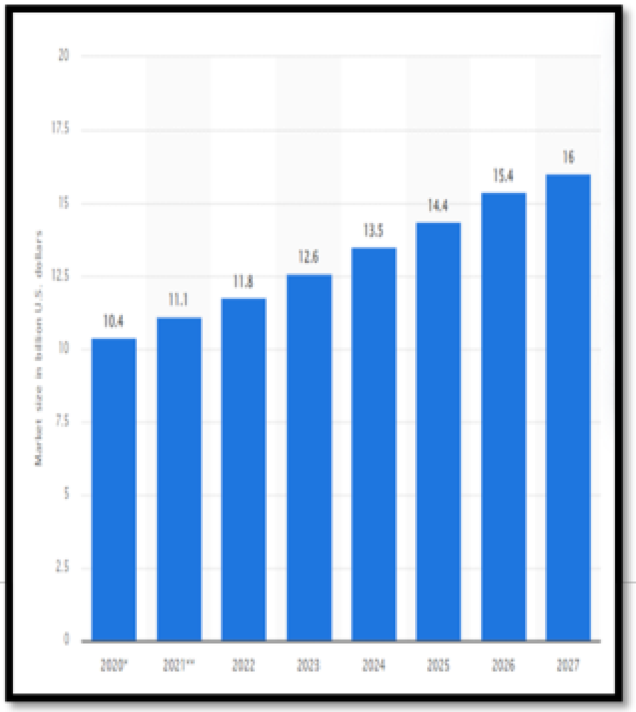


Fig. 4.2: Increasing the demand of power metering devices

data. They are instrumental in monitoring household energy usage and quantifying electricity consumption in kilowatt-hours. The anomaly diagnosis method, incorporating the support vector machine algorithm, not only enhances data quality but also safeguards systems from data breaches and financial losses. This method is notably focused on fraud detection, identifying potential risks and opportunities while positively impacting data measurement in power meter devices.

REFERENCES

[1] F. Abate, M. Carratù, C. Liguori, and V. Paciello, *A low cost smart power meter for iot*, Measurement, 136 (2019), pp. 59–66.
[2] L. Basora, X. Olive, and T. Dubot, *Recent advances in anomaly detection methods applied to aviation*, Aerospace, 6 (2019), p. 117.
[3] J. Cao, C. Song, S. Song, S. Peng, D. Wang, Y. Shao, and F. Xiao, *Front vehicle detection algorithm for smart car based on improved ssd model*, Sensors, 20 (2020), p. 4646.
[4] Y.-Y. Chen, Y.-H. Lin, C.-C. Kung, M.-H. Chung, and I.-H. Yen, *Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes*, Sensors, 19 (2019), p. 2047.
[5] F. del Ama Gonzalo, J. Ferrandiz, D. Fonseca, and J. A. Hernandez, *Non-intrusive electric power monitoring system in multi-purpose educational buildings*, Int. J. Power Electron. Drive Syst, 9 (2019), pp. 51–62.
[6] H. S. Dhiman, D. Deb, S. Muyeen, and I. Kamwa, *Wind turbine gearbox anomaly detection based on adaptive threshold and twin support vector machines*, IEEE Transactions on Energy Conversion, 36 (2021), pp. 3462–3469.
[7] F. Ding, Z. Zhuang, Y. Liu, D. Jiang, X. Yan, and Z. Wang, *Detecting defects on solid wood panels based on an improved ssd algorithm*, Sensors, 20 (2020), p. 5315.
[8] K. Guo and P. Reynaert, *29.2 a 0.59 thz beam-steerable coherent radiator array with 1mw radiated power and 24.1 dbm eirp in 40nm cmos*, in 2020 IEEE International Solid-State Circuits Conference-(ISSCC), IEEE, 2020, pp. 442–444.
[9] X. Hu, H. Li, X. Li, and C. Wang, *Mobilenet-ssd microscope using adaptive error correction algorithm: real-time detection of license plates on mobile devices*, IET Intelligent Transport Systems, 14 (2020), pp. 110–118.
[10] Z. Huang, Z. Yin, Y. Ma, C. Fan, and A. Chai, *Mobile phone component object detection algorithm based on improved ssd*, Procedia Computer Science, 183 (2021), pp. 107–114.

[11] W. Jiang, Y. Hong, B. Zhou, X. He, and C. Cheng, *A gan-based anomaly detection approach for imbalanced industrial time series*, IEEE Access, 7 (2019), pp. 143608–143619.
[12] L. Jin and G. Liu, *An approach on image processing of deep learning based on improved ssd*, Symmetry, 13 (2021), p. 495.
[13] F. Khan, M. A. B. Siddiqui, A. U. Rehman, J. Khan, M. T. S. A. Asad, and A. Asad, *Iot based power monitoring system for smart grid applications*, in 2020 international conference on engineering and emerging technologies (ICEET), IEEE, 2020, pp. 1–5.
[14] D. Li, Z. Zhang, P. Liu, Z. Wang, and L. Zhang, *Battery fault diagnosis for electric vehicles based on voltage abnormality by combining the long short-term memory neural network and the equivalent circuit model*, IEEE Transactions on Power Electronics, 36 (2020), pp. 1303–1315.
[15] Y. Liu, R. Liu, S. Wang, D. Yan, B. Peng, and T. Zhang, *Video face detection based on improved ssd model and target tracking algorithm*, Journal of Web Engineering, (2022), pp. 545–568.
[16] J. Qiu, Q. Du, K. Yin, S.-L. Zhang, and C. Qian, *A causality mining and knowledge graph based method of root cause diagnosis for performance anomaly in cloud applications*, Applied Sciences, 10 (2020), p. 2166.
[17] M. S. Saeed, M. W. Mustafa, U. U. Sheikh, T. A. Jumani, and N. H. Mirjat, *Ensemble bagged tree based classification for reducing non-technical losses in multan electric power company of pakistan*, Electronics, 8 (2019), p. 860.
[18] H. Shan and W. Zhu, *A small traffic sign detection algorithm based on modified ssd*, in IOP Conference Series: Materials Science and Engineering, vol. 646, IOP Publishing, 2019, p. 012006.
[19] H. Sun, H. Xu, B. Liu, D. He, J. He, H. Zhang, and N. Geng, *Mean-ssd: A novel real-time detector for apple leaf diseases using improved light-weight convolutional neural networks*, Computers and Electronics in Agriculture, 189 (2021), p. 106379.
[20] Z. Wang, C. Song, L. Zhang, Y. Zhao, P. Liu, and D. G. Dorrell, *A data-driven method for battery charging capacity abnormality diagnosis in electric vehicle applications*, IEEE Transactions on Transportation Electrification, 8 (2021), pp. 990–999.
[21] K. Yan, *Chiller fault detection and diagnosis with anomaly detective generative adversarial network*, Building and Environment, 201 (2021), p. 107982.
[22] Y. Yang, H. Wang, D. Jiang, and Z. Hu, *Surface detection of solid wood defects based on ssd improved with resnet*, Forests, 12 (2021), p. 1419.
[23] J. Yao, Z. Wang, C. Liu, G. Huang, Q. Yuan, K. Xu, and W. Zhang, *Detection method of crushing mouth loose material blockage based on ssd algorithm*, Sustainability, 14 (2022), p. 14386.
[24] G. Yu, H. Fan, H. Zhou, T. Wu, and H. Zhu, *Vehicle target detection method based on improved ssd model*, Journal of Artificial Intelligence, 2 (2020), p. 125.
[25] T. Yuan, L. Lv, F. Zhang, J. Fu, J. Gao, J. Zhang, W. Li, C. Zhang, and W. Zhang, *Robust cherry tomatoes detection algorithm in greenhouse scene based on ssd*, Agriculture, 10 (2020), p. 160.
[26] S. Zhai, D. Shang, S. Wang, and S. Dong, *Df-ssd: An improved ssd object detection algorithm based on densenet and feature fusion*, IEEE access, 8 (2020), pp. 24344–24357.
[27] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, and N. V. Chawla, *A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data*, in Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 1409–1416.
[28] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, et al., *Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection*, IEEE transactions on medical imaging, 40 (2020), pp. 879–890.
[29] H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, B. Xu, J. Bai, J. Tong, and Q. Zhang, *Multivariate time-series anomaly detection via graph attention network*, in 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 841–850.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 26, 2023
Accepted: Oct 29, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1185–1190, DOI 10.12694/scpe.v24i4.2547

CIGARETTE PRODUCTION LINE FLOW AUTOMATIC CONTROL SYSTEM BASED ON PLC TECHNOLOGY AND EMBEDDED SOFTWARE

DING ZHI[*], LIHUA XIE[†], YONG XIA[‡], AND JIE SHI[§]

Abstract. The automatic controlling system of the production will increase the preciseness of the work and it helps the employees to meet the organizational targets. The programming logic controller (PLC) and software-based technologies controlled everything interrelated with the production processes. This reduces the risk of production and evaluates the quality of the tobacco used to manufacture cigarettes. This study is based on secondary sources as which helps to get adequate information about the study topic. The implementation cost of the PLC is quite high, and proper training of the employees is needed for the manipulation of the PLC system of production. This is a computerised technology that supports the production process of the organisation and reduces the cost of production in the long term.

The term control technology is usually based on the control which is manual and on which the human operator has a general or primary role in following the procedures. These approaches are used for the decision-making and the execution of the programs and the operations, which are done manually.

Key words: PLC, Automatic controlling system. Software-based technology control technology

1. Introduction. In the era of technological development the machineries become an automated system and it helps the tobacco industries to produce more cigarettes. This helps in the cutting process of tobacco and increases the quality and quantity of the final output from the production system. The automatic machines slice the tobacco in the same pieces and insert the filters at the ends portion of the cigarette. The automatic control system helps in the maintenance of all the steps in a systematic way. This study has focused on the implication of the automatic control system in cigarette production and the effect of PLC technology in it.

The inventions in relation with the manufacture of the cigarettes in to system of control are able to control the manufacture of the cigarette-rod and the machines of the filter tipping. The rod making in the cigarette and the manufacture or the production of the filter trippers are very well-know. In recent years, the production or the manufacturing of the cigarette has been increased by the establishment of the double length rods of cigarette.

Automation and control technologies have revolutionized various industries by enhancing precision and productivity while reducing human involvement in manual processes. In the context of manufacturing, such technologies play a pivotal role in ensuring product quality, meeting production targets, and mitigating operational risks. This research delves into the application of automatic control systems, particularly focusing on the utilization of Programming Logic Controllers (PLCs) and software-based technologies to manage and optimize the intricate processes involved in cigarette production.

The tobacco industry, which encompasses the production of cigarettes, relies on advanced control systems to maintain the accuracy and efficiency of its manufacturing processes. Through the automation of key tasks, these systems help in achieving organizational goals, such as improving product quality and increasing overall output. Furthermore, they contribute to risk reduction, ensuring that the manufacturing process meets stringent quality standards. This study builds upon existing knowledge and insights, drawing from secondary sources to gain a comprehensive understanding of the significance of control technologies in the tobacco manufacturing sector. These secondary sources provide valuable information about the intricacies of PLC-based systems and software-driven automation, shedding light on their role in optimizing the production process.

[*]China Tobacco Sichuan Industrial Co., LTD, Chengdu 610020, China
[†]China Tobacco Sichuan Industrial Co., LTD, Chengdu 610020, China (lihuaxieres@outlook.com)
[‡]China Tobacco Sichuan Industrial Co., LTD, Chengdu 610020, China
[§]China Tobacco Sichuan Industrial Co., LTD, Chengdu 610020, China

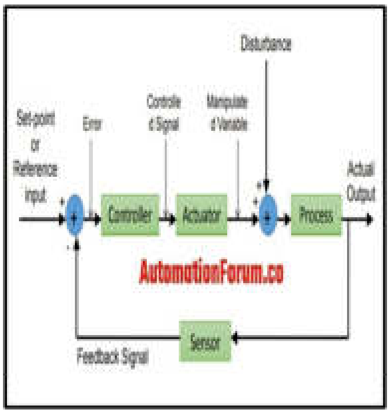


Fig. 3.1: Flow diagram of the Automatic Controlling System

2. Objectives.

1. To know the details of cigarette production line flow based on the automatic control system
2. To analyses the impact of PLC technology on the automatic production system of cigarette
3. To evaluate the challenges that occurred in the implication process of PLC technology and software in the production line flow of cigarette
4. To suggest the best possible way to enhance the production process through the utilization of PLC technology and an automatic production system.

This research provides a comprehensive examination of automation and control technologies in the context of tobacco manufacturing. It offers a detailed exploration of how these technologies are utilized to enhance precision, meet production targets, and ensure product quality in the industry. The paper delves into the role of Programming Logic Controllers, shedding light on their significance as central components of control systems in cigarette production. It provides insights into how PLCs are employed for managing and optimizing production processes.

3. Methodology. This study has been prepared based on secondary sources of information. Secondary sources include those sources which are already used by other organizations. The secondary sources are more easily available than the primary sources. The details of the PLC and the automatic controlling system of production have been understood from secondary sources. The PLC technology assists the employees to continue the production process and this decreases the risk in the business [2]. The development of precise quality of production could be possible through applying digital-based technology as this reduces the number of defaults in the production system [6].

3.1. Effect of automatic control system on the production of cigarette. The application of an automatic controlling system in the manufacturing process of tobacco-containing cigarettes is considered one of the major revolutionary procedures. Therefore with the help of this controlling system, the concerned procedures of production can be controlled [19].Under the cigarette manufacturing process, different types of functions have been used such as examination of the technical standard, investigation of the equipment and machinery that are used for production, method of processing the huge contents of tobacco, and the surrounding environment of the factory where the production process has been performed. The controlling system is capable of monitoring the layering capacity of the machinery that is used in the production process [2]. With this controlling ability, the capability of all the machines and the equipment used in the production process can be monitored and controlled by the excessive production of the cigarette as well.

The shown figure 3.1 contains the workflow of different equipment that is used in the automatic controlling system. First, the user needs to insert an input in the controlling system [10]. After that the input would automatically be sensed by the censoring machines. Once the input has been sensed then it is converted into the next phase where the presence of any kinds of errors has been detected in between the input object [5].

Then the object is further transferred between the controller and the actuator as well. After that by performing the final distribution the object has been undergone into the processed system and the outcome has been performed [11]. This whole process is performed automatically.

It has been observed that with the help of this automatic controlling system along with lowering the production rate the percentage of tobacco filling in each cigarette can also be controlled as well. This technical transformation is also associated with the development of fully automatic control over the excessive production of factories [26].When the tobacco filling becomes low then the adverse impact developed due to the tobacco filling of the particular amount gradually becomes low as well. This controlling nature-filling aspect is helpful in terms of lowering the problems of production [23]. The usual flow of machines that are used for manufacturing tobacco can also be measured and monitored by the automatic controlling system. When the autonomic system can be able to maintain the flow then it also creates an impact on the working consistency of the machines [8]. The flow-maintaining process helps detect any kind of warriors in the machine and not detect the errors but also mitigates the errors and maintains the flow and working consistency is also performed by the automatic controlling system. The production sytem increases the probability of the self-dependency nature of the organization [17]. Digitalized technology mainly controlled through an integrated and specialized system. This mode of production is more systematic in comparison to the other traditional processes [3]. Thus the development of this kind of technology helps to increase the production of cigarettes.

3.2. The meaning of PLC technology and its contribution to the automatic production system. Programmable logic controller is considered as an industrial computer which focuses on manufacturing processes. This technology has given importance to robotic devices, assembly lines and other activities which represents high reliability as well as process fault diagnosis [27]. In PLC technology data came from inputs and operational instructions can be sent through outputs. This technology came from automobile manufacturing industry which provides rugged programmable controllers for replacing hard real-time systems [27]. It also controls the other devices and makes decisions based on getting signals in input as well as it controls outputs. PLC consists of a microprocessor which has to be programmed with the help of a computer and this is written in computer software and then loaded into the PLC [21]. It is a digital computer that is used to control the work equipment in various factories.

PLC technology has been used in chemical industries, food industries, automated industries and biomedical industries. This technology has given a contribution on the automatic production of cigarettes and this machine consists of a control unit as well as successive workstations [24]. Each work stations have operating devices and control units connected to the workstations for controlling operating devices.

Cigarette production depends on the mixed proportion of five tobaccos and the blending process of tobaccos depends on the PLC technology. This technology helps to bring the accuracy of the blending process and it enhances the quality of the final output [15]. There are basically five scales present such as the leaf scale and the other four are called the auxiliary scale. The upper PC of center control sets the blending ratio based on each scale and then sends it to PLC before production [19]. The nuclear scale receives data from PLC and controls the motor speed. Blending accuracy came through this process and it increases the quality as well as automatic production of cigarettes.

Traditionally the manufacturing of the cigarettes have been effectively controlling and driven mechanically [22]. The variety of the rotating and the motions is necessary or important in the machines that are driven through the efficient gear boxes from the drive that are mainly used [16].

3.3. Challenges that occur during the implication of PLC technology in the automatic production line flow. The PLC technology is the form modular solid form of the computer followed the particular instruction. This system helps the organization to increase production [9]. This followed the customized instruction to continue the task. PLCs are industrial computers and are used to monitor different types of industrial activities [28]. The main problems that arrived from the PLC are related to the frequency interface, grounding and power. This means there is a need for a continuous power supply for the correct manipulation of the PLC technology [25]. The problems also arrived from the frequency interface as this frequency varies. The other challenges faced by the organization during the implementation of PLC in the production system are the high cost of maintenance, high cost of implementation [18]. The accounting of the PLC technology is complex, risk of data leakage and so on. The automatic controlling production may increase accidental issues. The

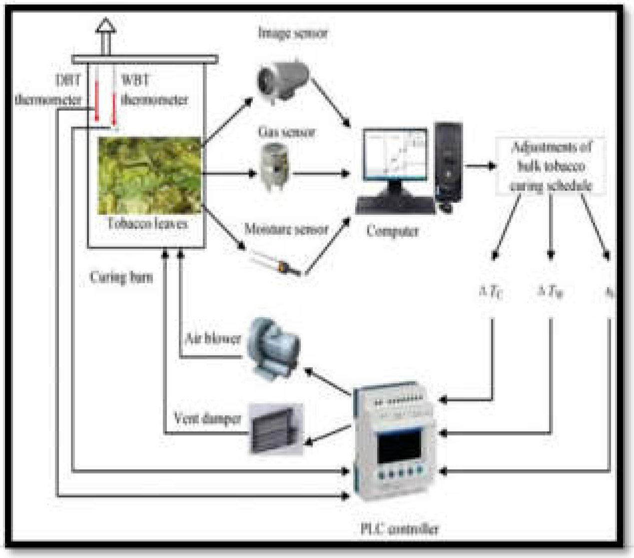


Fig. 4.1: Implication system of PLC for the development of bulk production unit of tobacco

maintenance cost and repetitive maintenance are another problems of automatic control production systems [1]. These are the drawbacks of the PLC and automated production system that may face by the organization during adaptation.

4. Result. The modern technology of cigarette production becomes more advanced and it increases the total quantity of the different cigarette-producing organizations. The automatic control system controls all the tasks related to production [22]. The automatic production system is digitalized technology in which all the production-related tasks continued through the computerized order. The implementation of the PLC technology increases the quality of the production and manages the wastage quantity of the raw materials [12]. This helps in the process of manufacturing superior quality products. The automatic controlling system helps in determining the quality of raw materials used in cigarettes. The tobacco cuts have been used as the raw materials of cigarettes. The development of the organizational production system could be increased through the implication and adaptation of the automatic machinery [7]. The PLC technology includes industrial computers to assist the program process. Digitalized technology increases the production quantity of all organizations. This brings revolution in the business and increases the production quantity [23].In modern times the production system is gone through the implication of the software. The software-based production system increases the preciseness in the production system. PLC technology helps in the checking process of tobacco quality and detects the faults easily [4]. This included a special alarm system to inform the default in the raw materials.

The figure 4.1 represents the steps followed in the cigarette production unit which is based on the PLC technology. All the things like moisture control, gas sensor, and temperature of the room are controlled by the PLC controller [14]. PLC controller is a computer based technology which helps the instructor to control all the tasks related to tobacco production. Along with the machinery layer, the rate of production in each of the layers is also under the monitoring radar of the controlling system of the automation process [11]. This process helps to lower the production rate in reach of the layers of the machines that are used in the cigarette machinery system.

The figure 4.2 represents the demand of the PLC will be high in the upcoming years. The market size of PLC was 11.21 billion US dollars and that increased to 12.51 billion US dollars in 2022. This demarcates that the demand for this technology is increasing. In the modern times, the industrial section becomes self-dependent with the development of technology. PLC helps the origination to earn more revenue from the businesses and it positively impact the organizational growth [20]. The PLC technology is an integrated system of processing

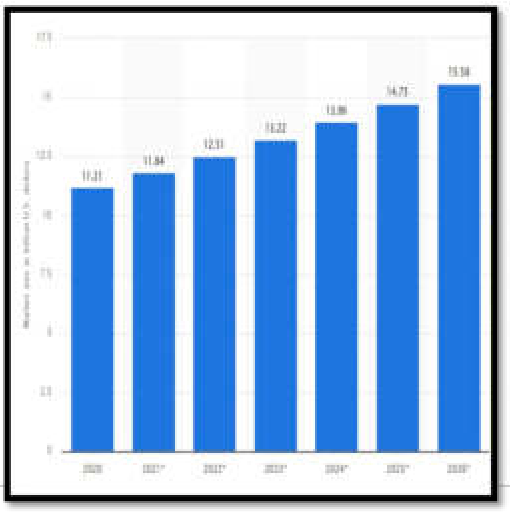


Fig. 4.2: PLC market size from 2020 to 2026

the production task of the industries. This helps in manipulating the temperature, dryness and the moisture [13]. The computerized technique shows all the incidents through the display. This monitoring system helps the organization to observe all the things that happening in the production unit [16]. Thus, the implication of this type of modernized technology is important for the organizational development.

5. Conclusion. In conclusion, this research has illuminated the profound impact of modern automation and technology on the production processes of the tobacco industry, particularly in cigarette manufacturing. The adoption of automatic control systems, driven by Programmable Logic Controller (PLC) technology, has revolutionized the production landscape. This transformative shift has not only increased the overall production quantities but has also ushered in a new era of precision and quality control, surpassing traditional production methods. The influence of software-based technology in conjunction with PLC systems has been instrumental in enhancing the efficiency and effectiveness of the production process. It significantly minimizes errors and discrepancies in the production line while making use of advanced tools such as radar sensors and cameras for real-time quality assessment. The integrated control system ensures that every aspect of production is closely monitored and managed, allowing organizations to achieve higher production volumes and superior product quality. The implementation of these modern technologies represents a pivotal step forward for the tobacco industry, streamlining production processes, reducing raw material wastage, and improving the quality of the end products. As we look to the future, the continued growth and demand for PLC technology further underscore its importance in enhancing production, paving the way for further innovation and growth in this vital industry.

REFERENCES

[1] M. Andronie, G. Lăzăroiu, M. Iatagan, I. Hurloiu, and I. Dijmărescu, *Sustainable cyber-physical production systems in big data-driven smart urban economy: a systematic literature review*, Sustainability, 13 (2021), p. 751.
[2] M. Andronie, G. Lăzăroiu, M. Iatagan, C. Uță, R. Ștefănescu, and M. Cocoșatu, *Artificial intelligence-based decision-making algorithms, internet of things sensing networks, and deep learning-assisted smart process management in cyber-physical production systems*, Electronics, 10 (2021), p. 2497.
[3] M. Bakhshi, S. Hashemi, and H. Dezhdar, *Applying mathematical modeling to create job rotation for improve workforce performance in semi-automatic systems*, International journal of research in industrial engineering, 9 (2020), pp. 318–327.
[4] M. Chattal, V. Bhan, H. Madiha, and S. A. Shaikh, *Industrial automation & control trough plc and labview*, in 2019 2nd international Conference on Computing, Mathematics and Engineering technologies (ICoMET), IEEE, 2019, pp. 1–5.
[5] C. Cimini, F. Pirola, R. Pinto, and S. Cavalieri, *A human-in-the-loop manufacturing control architecture for the next generation of production systems*, Journal of manufacturing systems, 54 (2020), pp. 258–271.

[6] K. Ding, F. T. Chan, X. Zhang, G. Zhou, and F. Zhang, *Defining a digital twin-based cyber-physical production system for autonomous manufacturing in smart shop floors*, International Journal of Production Research, 57 (2019), pp. 6315–6334.
[7] A. El Hammoumi, *Contribution to the optimization and monitoring of photovoltaic systems using iot and embedded technologies: Solar trackers and floating panels*, Available at SSRN 4262688, (2022).
[8] A. Jha and P. Tripathy, *Recent advancements in design, application, and simulation studies of hybrid solar drying technology*, Food Engineering Reviews, 13 (2021), pp. 375–410.
[9] D. S. Kim, B. J. Chung, and Y. M. Chung, *Statistical learning for service quality estimation in broadband plc ami*, Energies, 12 (2019), p. 684.
[10] M. A. Kulkarni et al., *Automatic agriculture crop yield production maintenance system based on remote monitoring techniques in cloud environment*, Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12 (2021), pp. 4409–4416.
[11] E. C. A. Lavopa and L. Zagato, *Advanced digital technologies and industrial resilience during the covid-19 pandemic: A firm-level perspective*, (2022).
[12] G. Lopez, J. Matanza, D. De La Vega, M. Castro, A. Arrinda, J. I. Moreno, and A. Sendin, *The role of power line communications in the smart grid revisited: Applications, challenges, and research initiatives*, IEEE access, 7 (2019), pp. 117346–117368.
[13] E. Lorenz and E. Kraemer-Mbula, *Background study prepared for the unctad contract on preparing a firm-level survey on frontier technology adoption*.
[14] A. K. Memon and K. X. Chen, *Recent advances in mode converters for a mode division multiplex transmission system*, Opto-Electronics Review, 29 (2021).
[15] O. Mendsaikhan, H. Hasegawa, Y. Yamaguchi, and H. Shimada, *Automatic mapping of vulnerability information to adversary techniques*, in The Fourteenth International Conference on Emerging Security Information, Systems and Technologies SECUREWARE2020, 2020.
[16] M. P. MINGSAKUL, J. Jongwanich, et al., *Impact of technology advancements and trade liberalization on productivity: evidence from Thai manufacturing sector*, PhD thesis, Thammasat University, 2020.
[17] N. Mira-Gebauer, C. Rahmann, R. Álvarez-Malebrán, and V. Vittal, *Review of wide-area controllers for supporting power system stability*, IEEE Access, 11 (2023), pp. 8073–8095.
[18] X. Pan, Z. Wang, and Y. Sun, *Review of plc security issues in industrial control system*, Journal of Cybersecurity, 2 (2020), p. 69.
[19] M. T. Pepito, *Revolutionizing industrial electronics & automation: A plc-based instructional package for enhanced learning and efficiency*, European Journal Of Innovation In Nonformal Education, 3 (2023), pp. 61–78.
[20] Z. Sembiring, *Stuxnet threat analysis in scada (supervisory control and data acquisition) and plc (programmable logic controller) systems*, Journal of Computer Science, Information Technology and Telecommunication Engineering, 1 (2020), pp. 96–103.
[21] C. Shao, C. Li, et al., *Application and realization of teaching thinking based on obe in the course of electrical control and plc technology*, Advances in Educational Technology and Psychology, 5 (2021), pp. 175–179.
[22] R. D. Trevizan, J. Obert, V. De Angelis, T. A. Nguyen, V. S. Rao, and B. R. Chalamala, *Cyberphysical security of grid battery energy storage systems*, IEEE Access, 10 (2022), pp. 59675–59722.
[23] A. Tuitoek et al., *A Framework for Adoption of Software Defined Wide Area Networks (Sd-wan) Within the Enterprises in Kenya.*, PhD thesis, University of Nairobi, 2022.
[24] L. Wang, Y. Liu, Q. Wang, J. Wang, Z. Yang, et al., *Plc course teaching method based on obe teaching concept*, Advances in Educational Technology and Psychology, 4 (2020), pp. 101–109.
[25] Y. Wang, Y.-H. Xie, Q.-H. Jiang, H.-T. Chen, R.-H. Ma, Z.-J. Wang, M.-Z. Yin, J. Shen, and S. Yan, *Efficient polymer-mediated delivery system for thiocyclam: Nanometerization remarkably improves the bioactivity toward green peach aphids*, Insect science, 30 (2023), pp. 2–14.
[26] L. Yang, X. Yan, S. Li, D. B. da Costa, and M.-S. Alouini, *Performance analysis of dual-hop mixed plc/rf communication systems*, IEEE Systems Journal, 16 (2021), pp. 2867–2878.
[27] L. Zhipeng, L. Xuejiao, and M. Jiwei, *Application of plc technology in electrical automatic control*, in Journal of Physics: Conference Series, vol. 1684, IOP Publishing, 2020, p. 012132.
[28] S. Zhu and C. Hu, *Design of a grid-connected control system for distributed photovoltaic power generation based on plc*, in Journal of Physics: Conference Series, vol. 2399, IOP Publishing, 2022, p. 012024.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 26, 2023
Accepted: Oct 29, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1191–1201, DOI 10.12694/scpe.v24i4.2550

SMART PARKING SYSTEM: OPTIMIZED ENSEMBLE DEEP LEARNING MODEL WITH INTERNET OF THINGS FOR SMART CITIES

SUDHA PRATHYUSHA JAKKALADIKI*, PETRA POULOVÁ†, PAVEL PRAŽÁK‡, AND BARBORA TESAŘOVÁ§

Abstract. In the recent era of smart city ecosystems and the Internet of Things (IoT), innovative, intelligent parking systems must make cities more sustainable. Every year, the increasing number of city vehicles requires more time to search for parking slots. In large cities, 10% of the traffic congestion occurs because of cruising; drivers spend almost 20 minutes searching for free space to park their vehicles. The passing time of waiting for parking in the traffic leads the issues such as energy, pollution, and stress. There needs to be more than the developed solutions. Therefore, the necessary to create a parking slot availability detection system that informs the drivers in advance about the free parking slot based on location. This paper introduces an enhanced ensemble Deep Learning (DL) model designed to forecast parking slot availability through the integration of IoT, cloud technology, and sensor networks. The devised model, known as Ensemble CNN-Boosted Graph LSTM (ECNN-BGLSTM), is optimized using a Genetic Algorithm (GA) framework. The model's performance is rigorously evaluated using a dataset from Europe, and various metrics, including Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE), are employed for assessment. The experimental findings demonstrate the superior performance of the proposed model compared to existing state-of-the-art approaches.

Key words: Smart parking, Internet of Things (IoT), Optimization, Deep learning (DL), Convolution neural network (CNN), Graph Long short-term memory (LSTM), Genetic algorithm (GA).

1. Introduction. The population of the world is migrating from rural to urban places, which increases the population density of large cities. Based on the United Nations Population Division, the world's population is expected to move to cities by 2050. In practice, the global urban scenario requires advanced technologies to meet the demands of the smart city. The advancement in sensor networks and sensor technologies leads new models to build, deploy and promote sustainable development systems to escalate the challenges in urbanization [12, 1]. Urban mobility sustainability and traffic congestion reduction are critical challenges in urbanization where parking space is limited [22].

Cities are getting smarter with the increased utilization of the Internet of Things (IoT), and the applications have risen rapidly in various areas of cities, including transportation systems, hospitals, shopping malls, airports, etc. In IoT-based smart cities, the smart parking system is the most significant topic due to the increasing number of vehicles. In recent years, parking issues, while new vehicles join during traffic, have attracted many researchers [19]. With growing technologies, the IoT concept and deep learning (DL) are utilized in smart city planning that gradually tackles mobility issues and provides citizens with sustainable infrastructure for ecological, economic, and social scenarios [3]. Recently, most intelligent systems have been in the form of mobile-based solutions which helps drivers report traffic jams, accidents, road conditions, and alternative routes. However, parking is still challenging because of the more significant number of vehicles moving across the roads.

The primary motivations behind the research are:

1. With increasing urbanization and population growth, cities face challenges related to traffic congestion

*Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic (sudha.jakkaladiki@uhk.cz)
†Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic
‡Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic
§Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic

and parking shortages. Intelligent parking solutions help address these issues by making better use of existing parking spaces.

2. Many cities strive to reduce their carbon footprint and improve air quality. Smart parking can contribute to this goal by reducing traffic congestion and lowering vehicle emissions.
3. Smart parking aligns with the broader goal of enhancing urban mobility and creating a connected transportation network. It encourages the use of public transportation and other eco-friendly commuting options.

Researchers have developed different approaches using various types of collected data in the literature to solve the issues. Most researchers used machine learning (ML) and time series models to compute parking places' occupancy in the existing literature. Due to the increasing number of sensor data, the conventional decision support systems are not providing efficient performance like deep neural networks since the deep learning (DL) approaches help estimate linear and nonlinear samples. Comparatively, deep learning (DL) methods help predict occupancy, specifically feed-forward networks. Yet, simple deep neural networks still need to be integrated with temporal domain data to predict parking solutions [16]. Hanzl [10] conducted a study to direct drivers to find parking spaces in urban areas. Static and dynamic routing approaches are analyzed with their advantages and disadvantages. Canli and Toklu [7] developed LSTM based model to predict parking space availability. The mobile application-based dynamic access is provided with the LSTM model and the occupancy rates of the parks in the desired locations are displayed in the applications using the relevant parameters. Intending to develop a smart parking system using the DL model, this paper contributes the following:

1. Developed a smart car parking space availability prediction system using IoT and ensemble DL approaches.
2. Using the considered European market data, the available car parking space is predicted with the features extracted by the Ensemble CNN and Boosted Graph LSTM model. A Genetic algorithm optimizes the ECNN-BGLSTM to improve the model performance.
3. The model is evaluated based on the metrics such as Root mean square error (RMSE), Mean square error (MSE), and Mean absolute error (MAE). The experimental analysis results show that the proposed model outperforms the state-of-the-art approaches.

The remaining section of this paper is as follows: section 2 discusses the related works of literature on the smart parking system. Section 3 introduced the ensemble DL model for predicting the available parking space. Section 4 discusses the experimented results and compares them with state-of-the-art approaches. Section 5 concludes the proposed model with its merits and future directions.

2. Related work. This section discusses the literature related to predicting smart car parking systems. Siddiqui et al., [20] accomplished the empty car parking services through sensors, and artificial neural networks (ANN) have been used for the prediction. Aizenberg et al., utilized the DL-based approaches initially based on ANN. The DL approach is differed from ANN by adding more hidden layers where the successive layers are taken from the output of the previous layers based on learning and data representation [6, 5]. Ghulam Ali et al., [1] proposed deep, long-term memory (DLSTM) by integrating IoT, cloud, and sensor technologies. Using the Birmingham parking dataset, three kinds of experiments are conducted to predict the availability of free parking space based on location, days of the week, and hours of the day. The result analysis shows that the DLSTM model outperforms other state-of-the-art approaches.

Liu et al., [14] developed a combination of Graph CNN and Recurrent neural network (RNN) to predict the real-time parking slot occupancy. Graph neural network has been utilized to extract the spatial data of the traffic flow, and RNN has been used to extract the temporal data of the traffic flow. The demonstration shows that the integration model performs better than other approaches. Carli et al., [8] proposed LSTM with a gated Recurrent unit (GRU) to predict the free parking slot availability, and the demonstration proves that GRU performs better than the LSTM model. Bock et al., [4] investigated the taxi fleets applicability to detect the on-street parking availability. The detection of free parking spots is analyzed using the taxi transit frequencies and the availability of parking spaces from vehicles equipped with GPS sensors. Tekouabou et al., [21] developed a combination of IoT and ensemble approaches to predict the free parking availability in the smart city. Using the Birmingham parking dataset, the obtained prediction accuracy was 94% using the bagging ensemble model. Kothai et al., [13] proposed a hybrid DL model using Boosted LSTM and CNN to predict the dynamic behavior of vehicle congestion in the road. From the traffic images, the CNN extracts the features, and BLSTM strengthens the weak classifiers. The real traffic scenario is simulated using OMNeT++ and SUMO with TensorFlow Python libraries. The experimental results show that the proposed model secured an accuracy of 98%.

Hébert et al., [11] created a high-resolution prediction model to prognosticate the accident circumstances on the roads by intersecting via big data analytics. Big data analysis is the approach that permits researchers to extract significant data from a larger amount of heterogeneous data. The authors employed balanced RF to sample the imbalanced data with various ML approaches such as decision trees, ANN, and Bayesian networks. Exploiting the features, including the attributes of weather, arterial segment, date, and time, road accidents are successfully prognosticated. Shengdong et al., [9] proposed a hybrid Multi-Model DL approach to forecasting the traffic flow by incorporating GRU and 1DNN to attain the features. The features of spatial local features have long dependencies, and correlations are processed by the end-to-end multimodel. The CNN-GRU model solves the issues with the long temporal dependencies of spatial-temporal correlations.

Moses and Parvathi [15] employed support vector regression to map the input using nonlinear mapping with M-dimensional features. The MSE estimates the performance by scaling the average error squares. The linear regression approach erects the scalar and independent response, and the DT approach reckons the information gain. Sellami and Alaya [18] developed Self-adaptive Multi kernel clustering for VANET to determine the load balance and other attainable resources. This model depicts the ambivalent density of deceleration, conveyance nodes, and bounded radio ranges using three stages such as cluster initialization, cluster adaption, and fusion stage. Compared to other approaches, this model poses the resultant to predict the traffic in the urban area, and the algorithm's complexity is optimized with the optimization algorithm. Ranjan et al., [17] integrate the CNN, CNN-transpose, and LSTM to redirect the congestion level. The encoder of the convolution process encodes the input spatial extraction features into a low-resolution latent state. The recurrent network with LSTM ascertains the time series data. The PredNet and ConvLSTM attain improved accuracy in predicting traffic congestion by associating the spatial and temporal features.

The literature mentioned above primarily discusses the proposed models in various ML and DL models. A research gap exists in validating these models using real-world traffic data from urban areas. Real-world validation will help assess the models' performance under diverse traffic conditions and validate their practical applicability. Addressing these research gaps would contribute to advancing traffic prediction and resource management in smart cities, leading to more efficient and reliable urban mobility solutions.

3. Proposed method. The developed framework is based on the API for finding parking, an Optimized DL model, a cloud-based parking server, and an open data portal. The API receives the parking request from the applications and returns the parking list, which is all suitable. The car parks within 2 km of the destination are listed as a choice for a parking slot. The direct request is made to the cloud-based smart parking server for each parking slot and analyzed through the developed DL approach. The cloud-based server receives the detected available parking slot in real-time and if the request is for a future date, then it is directed to the DL-based service. The ECNN-BGLSTM detects the parking slots in that service and returns the result. The data collection from the open data portal is described in the data preparation stage. These details are illustrated in the proposed model in Fig 3.1.

3.1. Smart car parking system (SCPS). The SCPS system informs the drivers regarding the parking slot availability of various parking locations. The checking of parking slot availability is a time-dependent problem. This paper developed a real-time response through a cloud server and optimized a DL-based approach to predict the available parking slot accurately. The developed SCPS highly depends on the optimized hybrid DL model, which provides accurate data on the availability of parking slots to the drivers. Based on the query requested by the driver, the SCPS analyzed the results and fetched the available parking slot response based on the request's specific time slot or a specific day. These fetched data are processed by the DL model and predict the free parking availability and transfer the data to the cloud car parking server to inform the driver regarding the availability.

3.2. Decision support system using optimized ensemble DL model. The proposed DSS for parking slot availability prediction receives input from sensors and predicts the availability of various parking locations

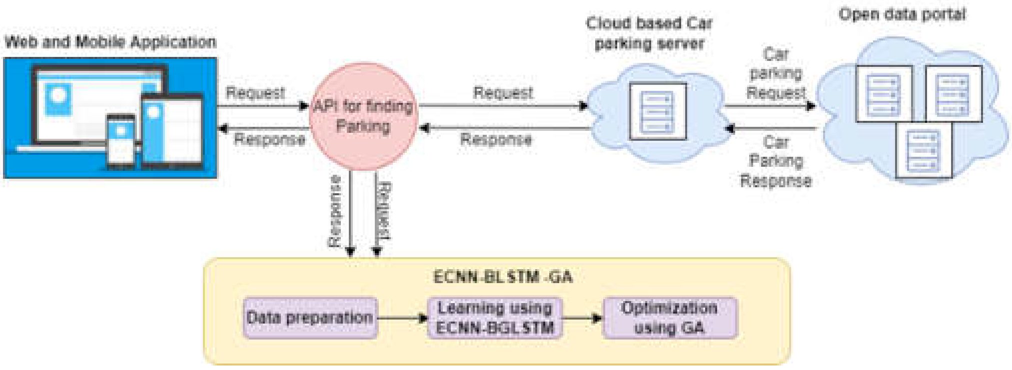


Fig. 3.1: Overview of the proposed smart parking detection system

at a specific time. Once the data of the request is obtained, applied ECNN-BLSTM model for prediction. Initially, CNN was employed to extract the features from the input car traffic images and has been trained by the BGLSTM model.

3.2.1. CNN. CNN has the powered ability to represent the features of the input image through neuron local connectivity and weight sharing. The layers include the convolution layer to learn the feature representation and the pooling layer to accomplish the invariance. In the convolution layer, the inputs are received by the neuron from the previous layer. The different features are learned by the kernels to convolute from the previous layer. The convolution process is denoted in Eqn 3.1

$$Y_F^l = \sigma\left(\sum_{i=1}^{F_{l-1}} Y_K^{l-1}, w_{KF}^l + b_F^l\right), \; F \in [1, F_l] \tag{3.1}$$

where F is the filter, l is the layer, Fth activation map of lth layer is denoted as Y_F^l, Kth activation map of (l-1)th layer is denoted as Y_K^{l-1}, W and b connects the fth activation map of lth layer at Kth position and bias respectively. The filters of layer 'l' are denoted as $F_l$and σ is the nonlinear activation function [17].

The pooling operations subdued the activation map spatial size as denoted in Eqn 3.2

$$Y_F^l(i,j) = \sigma\left(\sum_{a=1}^{f_{l-1}}\sum_{b=0}^{m-1}\sum_{c=0}^{n-1}\left(w_{KF}^l(b,c)\bigotimes Y_K^{l-1}(i+b, j+c) + c_F^l\right)\right) F \in [1, F_l] \tag{3.2}$$

where m, n is the size of filter in the convolution process, $Y_F^l(i,j)$ obtained from the output of the previous layer, b, and c are the kernel function. The convolution layer is entailed by the location of the l+1th pooling layers fth activation function by learning the output of the previous layer with the filter size (2,2). The update function is computed as in Eqn 3.3

$$Y_F^{l+1}(i,j) = \sigma\left(\sum_{a=1}^{f_{l-1}}\sum_{b=0}^{m-1}\sum_{c=0}^{n-1}\left(w_{KF}^{l+1}(b,c)\bigotimes Y_K^{l-1}(2i+b, j+c) + c_F^{l+1}\right)\right) F \in [1, F_{l+1}] \tag{3.3}$$

3.2.2. GLSTM. The traditional LSTM is based on the backpropagation model which consists of three gates including input, forget, and output gate. The graph LSTM is differentiated from the standard LSTM in terms of the graph structure where each tree node declares LSTM unit. The forward direction of the node captured the history and the backward direction characterizes the response. This mode has five layers including one input, three hidden, and one output layers. The input layer consists of a recurrent neural network (RNN) which denotes the extracted features from CNN. Each feature of the input layer is declared by the neuron.

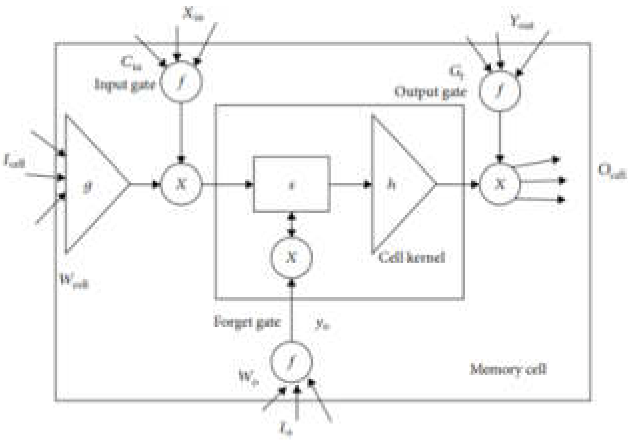


Fig. 3.2: GLSTM Cell structure [2]

The usage of a batch normalization layer reduces the overfitting issue which standardizes the prediction of the hidden layer. For the iteration t, the Graph LSTM computed the forward hidden sequence $\overrightarrow{h}$ and Backward hidden sequence $\overleftarrow{h}$ and the output sequence Y as in Eqn 3.4 and 3.5

$$\overrightarrow{h} = H(w_{F\overrightarrow{h}}\mathcal{F}_t + w_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1}\mathcal{F}_t + b_{\overrightarrow{h}}) \tag{3.4}$$

$$\overleftarrow{h} = H(w_{F\overleftarrow{h}}\mathcal{F}_t + w_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1}\mathcal{F}_t + b_{\overleftarrow{h}}) \tag{3.5}$$

$$Y_t = w_{\overrightarrow{h}Y}\overrightarrow{h}_t + w_{\overleftarrow{h}}\overleftarrow{h}_t + b_Y \tag{3.6}$$

where H is the hidden layer function, w is the weight matrices and b is the bias vector of each feature F. For each feature, the GLSTM cell is denoted in Fig 3.2.

For the particular iteration t, the input 'I', forget '$\mathcal{F}$', output 'O' gate and cell state '$\mathbb{C}$' with the activation function σ is updated as in Eqn 3.7 to 3.11

$$\mathcal{F}_t = \sigma(w_{F\mathcal{F}}F_t + w_{\overrightarrow{h}F}\overrightarrow{h}_{t-1} + w_{C\mathcal{F}}\mathbb{C}_{t-1} + b_{\mathcal{F}}) \tag{3.7}$$

$$I_t = \sigma(w_{F_I}F_t + w_{\overrightarrow{h}i}\overrightarrow{h}_{t-1} + w_{C_I}\mathbb{C}_{t-1} + b_I) \tag{3.8}$$

$$O_t = \sigma(w_{F_O}F_t + w_{\overrightarrow{h}o}\overrightarrow{h}_{t-1} + w_{C_O}\mathbb{C}_{t-1} + b_O) \tag{3.9}$$

$$\mathbb{C}_t = \mathcal{F}_t\mathbb{C}_{t-1} + I_t tanh(w_{F_\mathbb{C}}F_t + w_{\overrightarrow{h}\mathbb{C}}\overrightarrow{h}_{t-1} + b_\mathbb{C}) \tag{3.10}$$

$$\overrightarrow{h}_t = O_t tanh\mathbb{C}_t \tag{3.11}$$

For backward G-LSTM, the parameters are updated as follows in Equation from 3.12 to 3.18.

$$\mathcal{F}_t = \sigma(w_{F\mathcal{F}}F_t + w_{\overleftarrow{h}F}\overleftarrow{h}_{t-1} + w_{c_F}\mathbb{C}_{t-1} + b_F) \tag{3.12}$$

$$I_t = \sigma(w_{f_I} f_t + w_{\overleftarrow{h}_I} \overleftarrow{h}_{t-1} + w_{c_I} \mathbb{C}_{t-1} + b_I) \tag{3.13}$$

$$O_t = \sigma(w_{f_O} f_t + w_{\overleftarrow{h}_O} \overleftarrow{h}_{t-1} + w_{c_O} \mathbb{C}_{t-1} + b_O) \tag{3.14}$$

$$\mathbb{C}_t = F_t \mathbb{C}_{t-1} + I_t tanh(w_{f_{\mathbb{C}}} f_t + w_{\overleftarrow{h}_C} \overleftarrow{h}_{t-1} + b_{\mathbb{C}}) \tag{3.15}$$

$$\overleftarrow{h}_t = O_t tanh \mathbb{C}_t \tag{3.16}$$

3.2.3. Ensemble CNN-BGLSTM. The ensemble model aggregates the BGLSTM and CNN to overcome the overfitting issue of the prediction system. The GLSTM model trained with the G_k^i over the training data of iteration k. Initially, G(i) is equally set as $G_1(i) = 1/n$. The modified output cell of Eqn 3.9 and Eqn 3.14 is declared as, (Forward) $O_k = \sum_{k=0}^{n} \left(\sigma(w_{F_k} F_k + w_{\overrightarrow{h}_k} \overrightarrow{h}_{k-1} + w_{C_k} \mathbb{C}_{k-1} + b_k) \right)$ 3.9

(Backward)$O_k = \sum_{k=0}^{n} (\sigma(w_{f_k} f_k + w_{\overleftarrow{h}_k} \overleftarrow{h}_{k-1} + w_{c_k} \mathbb{C}_{k-1} + b_k))$ 3.14

The error function defined by the user is identified as the boosting outputs expressed mathematically as,

$$E_k = (O_{acutal} - O_k) \tag{3.17}$$

For the expression, the network parameter a_k is computed as,

$$a_k = 0.5 \left\{ ln \frac{(1 - E_k)}{E_k} \right\} \tag{3.18}$$

E_k is the final boosted ensemble output which is computed while the error is declared as zero and the mathematical expression is declared as in Eqnn 3.19

$$Y_k = sign \left(\sum a_k . O_k \right) / a_k \tag{3.19}$$

3.2.4. Optimization using GA. The genetic algorithm (GA) is a global optimization approach that mimics biological evolution in nature. It obtains the optimal solution through continuous evolution, based on the survival of the fittest. GA involves three operations: selection, crossover, and mutation. Crossover is the primary method for generating offspring and forming a new generation.

The GA process includes an initial population, parameter coding, designing a fitness function, genetic operations, and control parameter settings. The main ideas behind GA are as follows: (i) Generating the initial population randomly and conducting classification. The next population is generated through GA selection, crossover, and variations, serving as the first generation. (ii) Merging the parental population with the offspring population, generating the second generation, and conducting non-inferior classification. Selecting the appropriate individuals to create the new parental population based on each individual's focusing distance. (iii) Continuing the generation of offspring until the condition is satisfied.

GA-based optimization has several advantages over other approaches. It operates directly on decision variable coding and structural objects like sets, matrices, sequences, graphs, and trees. GA is not only a simulation of biological gene processes and genetic evolution but also facilitates the application of genetic operators. It has been successfully applied in various fields such as automatic function, number optimization, production scheduling, machine learning, and image processing. GA utilizes objective function values for searchability and fitness function to measure individual fitness, providing an advantage when dealing with difficult-to-derive objective functions. Another advantage of GA is its global search ability, starting from an initial population with diverse individuals to avoid getting stuck at local optima. GA follows probability rules rather than deterministic rules, offering flexible searching and reducing parameter influences. Its strong scalability allows it to be combined with various technologies.

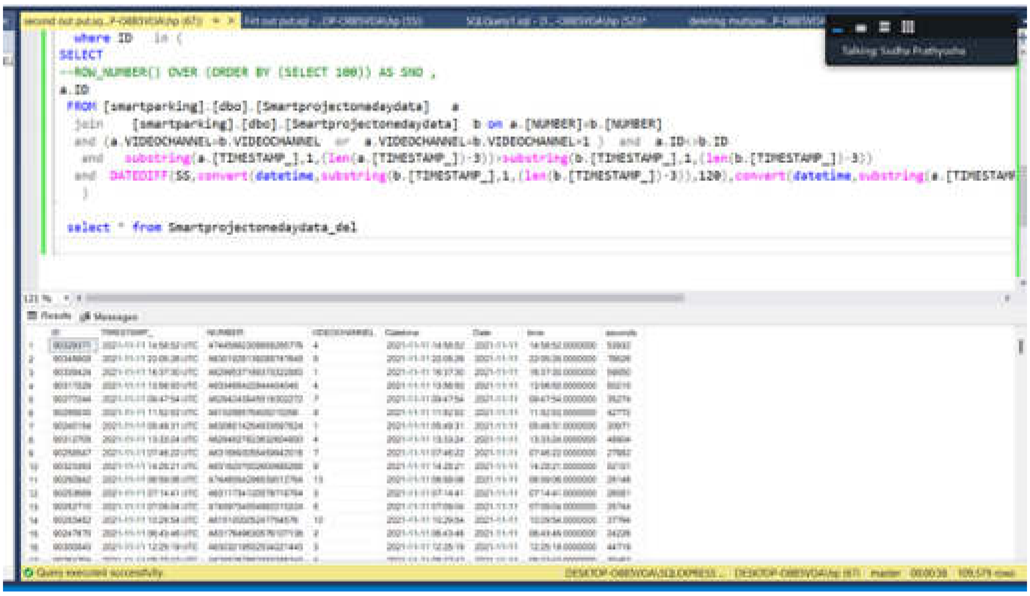


Fig. 4.1: Sample data

4. Results and Discussions. This section presents the experimental results and discussions using Keras, a neural network library. The Jupiter notebook was chosen for compiling the program. Geographic data from the European market was collected for processing. Initially, the dataset consisted of eight parameters: ID, timestamp, number, video_channel, date_time, date, time, and seconds, with 109,736 records. However, the fixed parameter values for each car park did not significantly affect the model's performance. After recording the data, abnormal entries were identified and removed through preprocessing to eliminate duplicate entries, missing values, and Langrange interpolation. The data was then standardized, resulting in 109,579 records. Fig 4.1 illustrates a sample of the processed data.

4.1. Metrics used. The predictive model performance is evaluated using the metrics such as Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy values.

1. RMSE has been used to find the prediction model deviation which takes the values from 0 to $+\infty$. The closer the value of 0, the lower the prediction model deviation. The mathematical expression is stated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - Y)^2} \tag{4.1}$$

2. MSE: It depicts how the regression curse is closer to the set of points and it is generally a regression problem and it is stated as,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y)^2 \tag{4.2}$$

3. MAE: Average error magnitude in the predictions and the low value indicates the accurate prediction.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - Y| \tag{4.3}$$

Table 4.1: Error assessment

Train-Test set	Models	MAE	MSE	RMSE	MdAE	MSLE
70-30%	SVM	2.5	8.5	2.91	0.032	0.014
	RF	2.43	7.12	2.66	0.033	0.017
	LSTM	0.97	2.53	1.58	0.023	0.008
	Proposed ECNN-BGLSTM	0.88	1.83	1.35	0.018	0.002
80-20%	SVM	2.71	8.62	2.93	0.035	0.015
	RF	2.53	7.81	2.79	0.036	0.019
	LSTM	1.02	3.12	1.77	0.034	0.010
	Proposed ECNN-BGLSTM	0.91	2.03	1.42	0.021	0.004
90-10%	SVM	2.83	9.12	3.01	0.057	0.017
	RF	2.87	8.37	2.89	0.043	0.021
	LSTM	1.11	3.18	1.78	0.045	0.011
	Proposed ECNN-BGLSTM	0.98	2.43	1.55	0.035	0.005

4. Median Absolute Error (MdAE):

$$MdAE = median\left(\sum_{i=1}^{N} |Y_i - Y|\right) \tag{4.4}$$

5. Mean Squared Log Error (MSLE): It is declared as,

$$MSLE = \frac{1}{N}\sum_{i=1}^{N} \left(\log\left((Y_{\,i}) \ +1\right) - \log\left((Y) \ +1\right)\right)^2 \tag{4.5}$$

where Y_i is the actual available parking space and Y is the prediction of the available parking space predicted by the ECNN-BGLSTM.

4.2. Result analysis and comparison. The efficiency of the proposed model is compared with standard Deep Learning (DL) approaches, including SVM, Random Forest, and LSTM [7]. The model's quality is evaluated under three different cross-validation scenarios: 70:30, 80:20, and 90:10, where the first term represents the percentage of training data and the second term represents the percentage of testing data. The models were executed for 100 epochs, and their performance is presented in Table 4.1. Among the scenarios, the proposed model achieved an improved prediction rate with a reduced error rate compared to other approaches, specifically in the 70:30% scenario.

Fig 4.2 displays the actual and predicted values of both the proposed and existing approaches. The green line represents the actual prediction results, while the red line denotes the predicted results of the approaches. Specifically, Fig 4.2 (d) illustrates the efficient prediction of the parking occupancy rate using the proposed model, outperforming the other approaches. In contrast, Fig 4.2 (a) for SVM, Fig 4.2 (b) for RF, and Fig 4.2 (c) for LSTM show good predictions in linear time series but worse predictions at maximum and minimum points.

The accuracy comparison in terms of the number of epochs for predicting available parking spaces is depicted in Fig 4.3. As the number of epochs increases, the accuracy also improves. At 100 epochs, the proposed model achieved an enhanced accuracy of 98.5% for predicting parking space availability. In comparison, other approaches, such as SVM, RF, and LSTM, achieved accuracies of 90%, 91%, and 95.3%, respectively. Fig 4.4 shows the training loss comparison. As the number of epochs increases, the loss value decreases. Notably, the proposed ECNN-BGLSTM model outperforms other approaches by reducing the training loss to 0.032, whereas state-of-the-art approaches like SVM (0.09), RF (0.08), and LSTM (0.072) exhibit higher loss values.

5. Conclusion. IoT enables the intelligent interconnection between things, the cloud, and individuals, facilitating efficient action plans. This smart exchange of information is processed and sent to the cloud

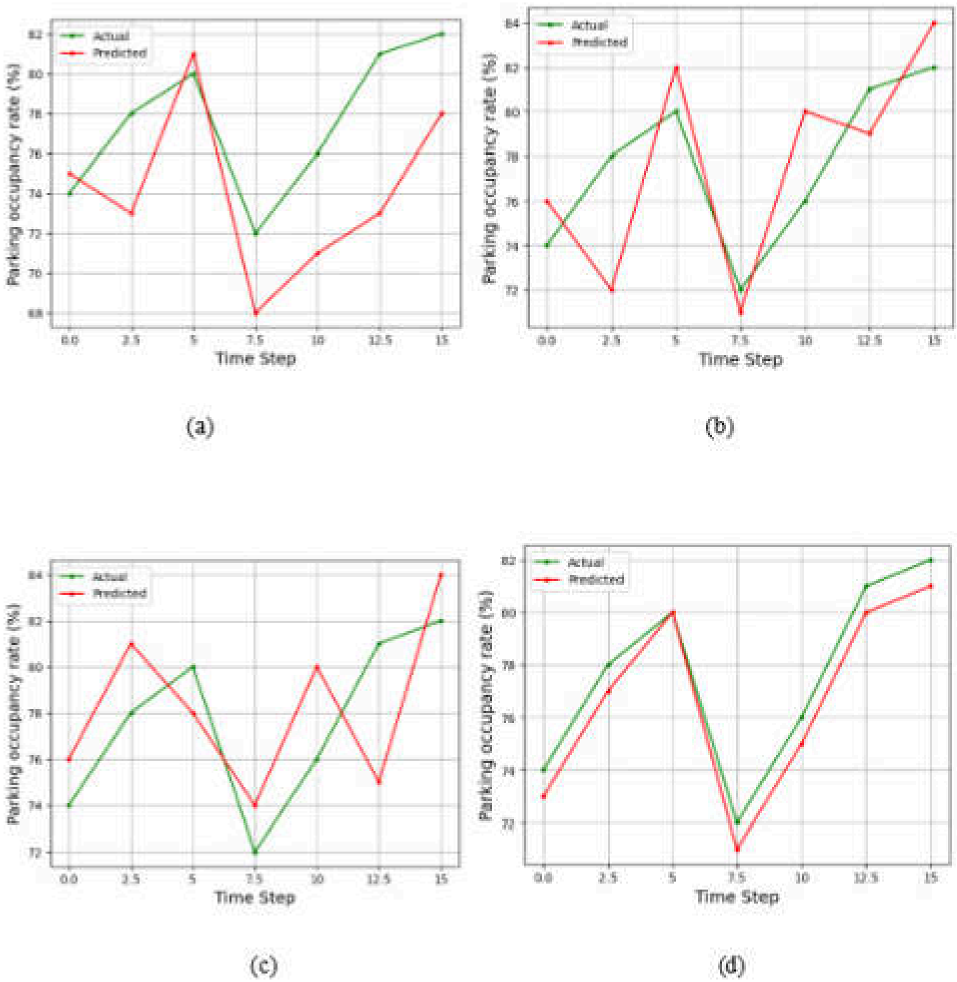


Fig. 4.2: Prediction results (a) SVM (b) Random Forest (c) LSTM (d) Proposed ECNN-BGLSTM

server via the web, where it is utilized for car parking convenience. The challenge of finding available parking spaces in the European market often results in wrong parking, leading to vehicle congestion and traffic jams. Therefore, the development of a smart parking system that notifies users about available spaces is crucial for Europeans. In this paper, we present an intelligent smart car parking space availability prediction system that utilizes ensemble Deep Learning (DL) and IoT. Using data sourced from the European market, the prediction of available parking spaces is carried out by the Ensemble CNN and Boosted Graph LSTM model, utilizing feature extraction. To enhance the model's performance, Genetic Algorithm optimization is employed on the ECNN-BGLSTM. Evaluation of the model encompasses metrics such as Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). The experimental analysis underscores the superior performance of the proposed model, achieving an enhanced accuracy of 98.5% with notable reductions in error rates and losses. Subsequent research will delve into examining traffic density with diverse parameters, alongside the implementation of novel strategies to efficiently address parking space availability and further extend the smart parking system to encompass smart charging, thus contributing to the realm of smart city applications.

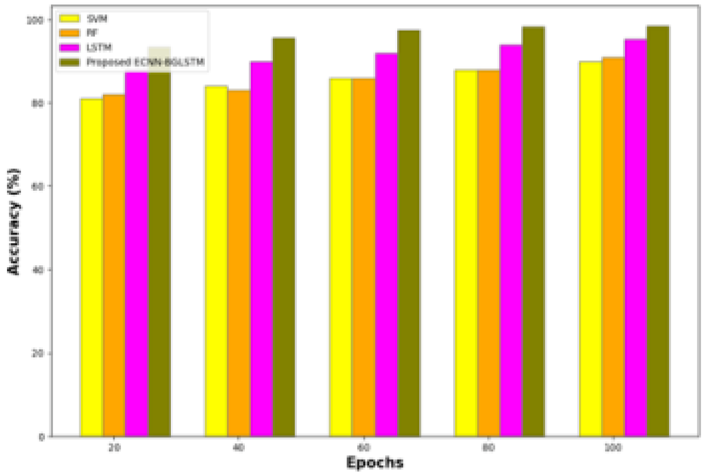


Fig. 4.3: Accuracy Comparison

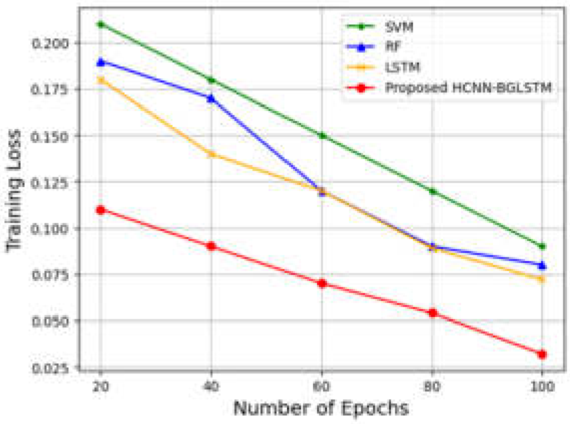


Fig. 4.4: Training Loss Comparison

Acknowledgment. The intention of the article was expressed and proposed within the framework of Smart parking project execution at the University of Hradec Kralove, Czech Republic and the study was supported by the SPEV project, run by the faculty of informatics and management, at the University of Hradec Kralove, Czech Republic.

REFERENCES

[1] G. Ali, T. Ali, M. Irfan, U. Draz, M. Sohail, A. Glowacz, M. Sulowicz, R. Mielnik, Z. B. Faheem, and C. Martis, *Iot based smart parking system using deep long short memory network*, Electronics, 9 (2020), p. 1696.

[2] S. Amudha and M. Murali, *Deep learning based energy efficient novel scheduling algorithms for body-fog-cloud in smart hospital*, Journal of Ambient Intelligence and Humanized Computing, 12 (2021), pp. 7441–7460.

[3] J. Bélissent et al., *Getting clever about smart cities: New opportunities require new business models*, Forrester Research, 3 (2010), pp. 1–31.

[4] F. Bock, S. Di Martino, and A. Origlia, *Smart parking: Using a crowd of taxis to sense on-street parking space availability*, IEEE Transactions on Intelligent Transportation Systems, 21 (2019), pp. 496–508.

[5] O. Brun, Y. Yin, J. Augusto-gonzalez, M. Ramos, O. Brun, Y. Yin, J. Augustogonzalez, M. Ramos, and E. Gelenbe, *Iot attack detection with deep learning to cite this version?: Hal id?: Hal-02062091 iot attack detection with deep learning*,

in Proc. ISC Secur. Working, 2019.

[6] S. Cakir, S. Toklu, and N. Yalcin, *Rpl attack detection and prevention in the internet of things networks using a gru based deep learning*, IEEE Access, 8 (2020), pp. 183678–183689.

[7] H. Canli and S. Toklu, *Deep learning-based mobile application design for smart parking*, IEEE Access, 9 (2021), pp. 61171–61183.

[8] R. Carli, G. Cavone, S. Ben Othman, and M. Dotoli, *Iot based architecture for model predictive control of hvac systems in smart buildings*, Sensors, 20 (2020), p. 781.

[9] S. Du, T. Li, X. Gong, and S.-J. Horng, *A hybrid method for traffic flow forecasting using multimodal deep learning*, arXiv preprint arXiv:1803.02099, (2018).

[10] J. Hanzl, *Parking information guidance systems and smart technologies application used in urban areas and multi-storey car parks*, Transportation Research Procedia, 44 (2020), pp. 361–368.

[11] A. Hébert, T. Guédon, T. Glatard, and B. Jaumard, *High-resolution road vehicle collision prediction for the city of montreal*, in 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 1804–1813.

[12] A. Hussain, U. Draz, T. Ali, S. Tariq, M. Irfan, A. Glowacz, J. A. Antonino Daviu, S. Yasin, and S. Rahman, *Waste management and prediction of air pollutants using iot and machine learning approach*, Energies, 13 (2020), p. 3930.

[13] G. Kothai, E. Poovammal, G. Dhiman, K. Ramana, A. Sharma, M. A. AlZain, G. S. Gaba, and M. Masud, *A new hybrid deep learning algorithm for prediction of wide traffic congestion in smart cities*, Wireless Communications and Mobile Computing, 2021 (2021), pp. 1–13.

[14] J. Liu, J. Wu, and L. Sun, *Control method of urban intelligent parking guidance system based on internet of things*, Computer Communications, 153 (2020), pp. 279–285.

[15] A. Moses and R. Parvathi, *Vehicular traffic analysis and prediction using machine learning algorithms*, in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE, 2020, pp. 1–4.

[16] C. Pengzi, Y. Jingshuai, Z. Li, G. Chong, and S. Jian, *Service data analyze for the available parking spaces in different car parks and their forecast problem*, in Proceedings of the 2017 International Conference on Management Engineering, Software Engineering and Service Sciences, 2017, pp. 85–89.

[17] N. Ranjan, S. Bhandari, H. P. Zhao, H. Kim, and P. Khan, *City-wide traffic congestion prediction based on cnn, lstm and transpose cnn*, IEEE Access, 8 (2020), pp. 81606–81620.

[18] L. Sellami and B. Alaya, *Samnet: Self-adaptative multi-kernel clustering algorithm for urban vanets*, Vehicular Communications, 29 (2021), p. 100332.

[19] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani, *Selection of effective machine learning algorithm and bot-iot attacks traffic identification for internet of things in smart city*, Future Generation Computer Systems, 107 (2020), pp. 433–442.

[20] S. Y. Siddiqui, M. A. Khan, S. Abbas, and F. Khan, *Smart occupancy detection for road traffic parking using deep extreme learning machine*, Journal of King Saud University-Computer and Information Sciences, 34 (2022), pp. 727–733.

[21] S. C. K. Tekouabou, W. Cherif, H. Silkan, et al., *Improving parking availability prediction in smart cities with iot and ensemble-based model*, Journal of King Saud University-Computer and Information Sciences, 34 (2022), pp. 687–697.

[22] E. I. Vlahogianni, K. Kepaptsoglou, V. Tsetsos, and M. G. Karlaftis, *A real-time parking prediction system for smart cities*, Journal of Intelligent Transportation Systems, 20 (2016), pp. 192–204.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 26, 2023
Accepted: Oct 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1203–1213, DOI 10.12694/scpe.v24i4.2557

EVALUATION OF REGIONAL ROAD TRANSPORT SAFETY SERVICE LEVEL WITH EDGE COMPUTING IN SCALABLE SENSOR AND ACTUATOR NETWORKS

XIAOXU DANG*, GUOYU WANG†, XIAODONG ZHOU‡, AND SHIHUI WANG§

Abstract. In contemporary society, road transport has assumed a pivotal role in the daily lives of individuals, with virtually everyone being a road user in one capacity or another. Unfortunately, a significant portion of road accidents stem from user negligence and a lack of awareness regarding road safety principles and regulations. The majority of these accidents are fundamentally a result of human errors, errors that could potentially be averted through the comprehensive utilization of technology. The integration of edge computing within actuator networks, combined with the scalability of sensors, holds the potential to proactively detect and prevent impending collisions. This cutting-edge approach not only accumulates and processes sensor-generated data but also employs it to mitigate the occurrence of road accidents. This discussion delves into the potential advantages offered by smart sensors and advanced technology, focusing on the enhancement of safety within transportation networks. The core attributes of these intelligent technologies facilitate the establishment of robust connectivity and the effective implementation of processes, ushering in an innovative era in the realm of road safety through smart technology.

Key words: Road transport, Safety services, Edge computer, Scalable sensor, Actuar networks

1. Introduction. In contemporary society, road transport has seamlessly integrated itself into the fabric of daily human existence, impacting the lives of individuals from all walks of life. Whether it's the daily commute to work, the school run, or leisurely travel, road transport plays a crucial role. However, this convenience comes at a cost, as road accidents and mishaps continue to pose significant threats to public safety. The root causes of these accidents often stem from human factors, primarily the carelessness of road users or their inadequate knowledge of road safety principles and guidelines. The consequential human errors leading to road accidents underscore the imperative need for a paradigm shift in road safety strategies. It is within the boundless potential of technology that the solutions to these persistent challenges may lie. In particular, the emerging field of edge computing, in synergy with actuator networks and scalable sensor technology, holds the promise of transforming road safety by proactively detecting and preventing collisions. This innovative approach involves the gathering and processing of sensor-generated data, utilizing it to predict and avert potential accidents.

The Internet in the modern era is a great revolution of the advances technology has been thus far. This revolution of the digital world has brought a huge change in the lifestyle, economies, and societies. Technology in the digital era has transformed the lives of the human, which can only be possible to imagine just a few centuries ago.

Although the rise in road accidents due to the errors of drivers is alarming, it is very important to address that and use edge-computing technology to decrease the maximum loss possible [2]. The aim of this study is to understand the safety services of regional road transport with edge computing in the sensor that is scalable and an actuator network.

This research discussion delves into the potential advantages offered by smart sensors and cutting-edge technological advancements in the realm of transportation safety. By focusing on the enhancement of safety within transportation networks, this research aims to explore the transformative role that technology can play in revolutionizing road safety measures. The core features of these intelligent technologies facilitate the creation of robust and stable connectivity, thus ushering in an innovative era in the domain of road safety through the implementation of smart technology. This study endeavors to shed light on the evolving landscape of road safety,

*Institute of Management, Xi'an University of Science and Technology, Xi'an, 710054, China (xiaoxudangres21@outlook.com)
†Institute of Management, Xi'an University of Science and Technology, Xi'an, 710054, China
‡Department of Transportation of Shaanxi Province, Xi'an, 710075, China
§Institute of Management, Xi'an University of Science and Technology, Xi'an, 710054, China

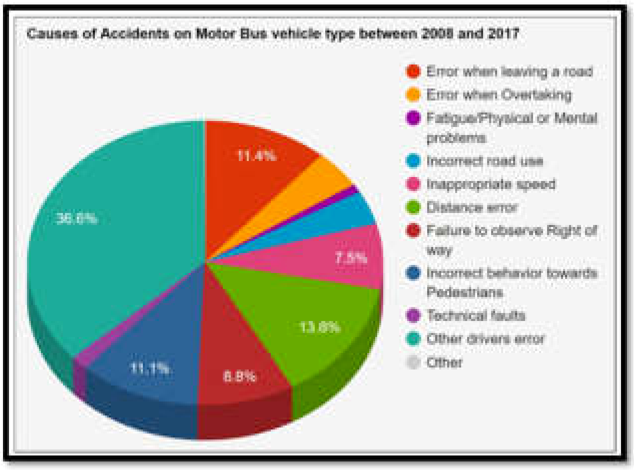


Fig. 1.1: Causes of motor bus vehicle-type accidents, 2008 to 2017

offering insights into the synergy between technology and transportation safety, with a particular emphasis on the possibilities and advantages of smart sensors and advanced mechanisms.

2. Objectives. In this study, some of the key objectives of the research have been identified and evaluated. This includes the basic concept of regional rod safety services, edge computing, and other related things. The objectives also include minute details of the relations between the key terms of this study. The identified key objectives are as followed:

RO 1: To identify the basic concept of safety services of regional road transport

RO 2: To evaluate the idea of edge computing in scalable sensor and actuator network

RO 3: To determine the importance of safety services in the regional road transport

RO 4: To investigate the impact of edge computing in scalable sensor and actuator networks in terms of safety services in the regional road transport

RO 5: To analyze the challenges faced by safety services in regional road transport while implying edge computing in scalable sensor and actuator networks

RO 6: To evaluate the recommended strategies in order to mitigate these challenges faced by safety services in regional road transport while implying edge computing in scalable sensor and actuator networks

The motivation for undertaking this research is rooted in the pressing need to address the persistently high rate of road accidents and mishaps that threaten the safety and well-being of road users worldwide. Road transport has become an integral part of modern human life, providing unparalleled convenience, yet it also presents inherent risks, primarily stemming from human errors, negligence, and a lack of awareness about road safety. These accidents not only result in a loss of lives and property but also place immense burdens on healthcare systems and infrastructure.The research is driven by the realization that technology, particularly in the form of edge computing and smart sensors, has the potential to revolutionize road safety by providing real-time monitoring, data analysis, and proactive collision prevention measures. As the world becomes more interconnected and data-driven, harnessing technology's full potential in enhancing road safety is not just a promising endeavor but a necessity.

The advent of smart sensors and cutting-edge mechanisms offers an unprecedented opportunity to advance road safety by mitigating human errors, predicting potential collisions, and enabling timely interventions. The motivation for this research lies in exploring how these innovative technologies can be effectively integrated into transportation networks to not only reduce road accidents but also to foster more efficient and sustainable road transport systems. By comprehensively examining the potential advantages and impacts of smart sensor technology and advanced mechanisms in road safety, this research seeks to contribute to the broader goal of making road transport safer, more reliable, and environmentally sustainable. Ultimately, the motivation is to

Table 3.1: Tips for road safety

Tips for Road Safety	Do not use mobile phones while driving or riding Walk carefully on the footpaths and cross the road when the traffic light is red using the zebra crossing Never drink and drive Be careful of the pedestrians, children, and all the citizens Wear helmets and seatbelts accordingly Never go above the speed limit

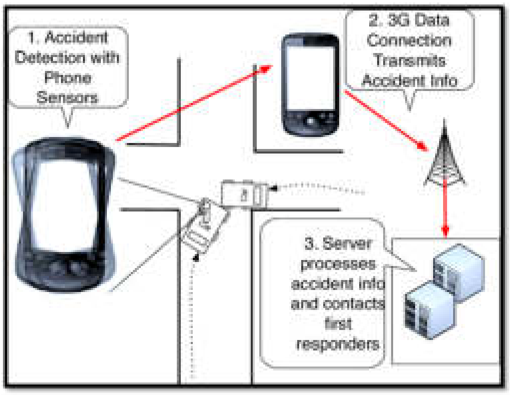


Fig. 3.1: Cooperation of multiple vehicles and avoiding collision in automated vehicles

pave the way for innovative solutions that can significantly reduce the toll of road accidents and enhance the quality of life for all road users.

3. Methodology. This research has chosen the secondary qualitative method for conducting this research. In this section, this work represents all the data and information that are collected. The collected data and information helps to understand the current situation of the safety services in the regional transport system, specifically on the road [1]. This work also examines the possibilities of edge computing in scalable sensor and actuator networks and how it can be used to minimize human error. This study evaluates the challenges that can be seen in road safety services during implying edge computing in scalable sensor and actuator networks [5]. The data also reveals a few suggestions for minimizing the challenges of regional road safety services as much as possible with the help of edge computing in actuator networks and scalable sensors. This data and information gained from the research will help the regional road safety services to take proper measurements and precautions for the best safety [6].

3.1. Identifying the basic concept of the safety services of regional road transport. In the modern era, road safety has become an important part of human evolution. Everybody uses the road in one way or the other. The current system of road transportation has been able to minimize the distance but it has increased the number of road accidents in an alarming number [10]. The reports show that almost in every country numerous numbers of road accidents happen, and this not only results in the loss of human and animal lives in a huge number, but also damages the lives of other humans and cause harm to other resources. For instance, in India, a report says, almost eighty thousand people are killed in different types of road accidents, statistically saying which is on and about 13 percent of the total fatality that happens all across the globe [8]. These accidents basically happen due to the lack of awareness of the road safety in the regional area or the carelessness of the drivers or the other users of the roads.

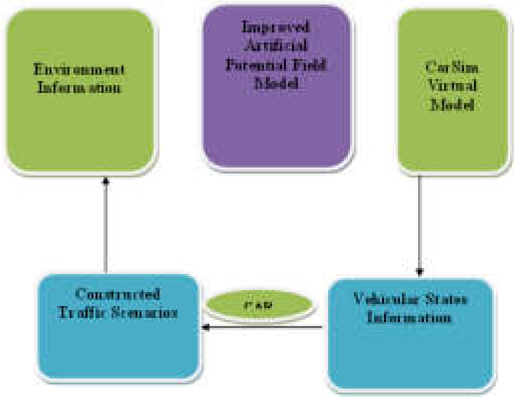


Fig. 3.2: Co-simulation for avoiding obstacles in road

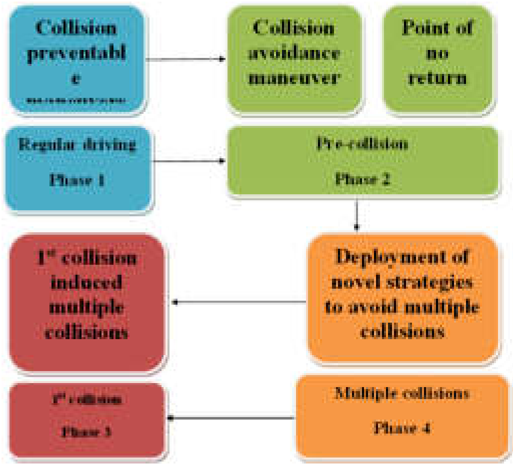


Fig. 3.3: Factors of interaction while driving

3.2. Evaluating the idea of edge computing in actuator networks and scalable sensor. The above picture shows that if the cars become AI enabled that will help the vehicles to sense other vehicles from distance and avoid possible collision [11]. It can also cooperate with multiple vehicles at the same time and the automated vehicle helps to reduce the human errors in times of mishaps.

Edge computing refers to an information technology that is distributed and an architecture where the data of the client is being processed at the periphery of the selected network [12]. This happens very close to the source of the data where the data originated in the first place.

In the above figure, the scenario is constructed on the hypothesis, the vehicle that can be possible obstacle is moving on a straight road and the speed limit is 15 km per hour. Another vehicle is running on the same lane at the speed of 80 km per hour. In order to avoid any kind of collision the above architecture is made.

On the other hand, scalability refers to the basic ability of the off computing in order to handle all the resources that are growing as per the requirements [13]. These requirements can include the new cases that are being used and on boarded, as well as measuring the already used cases that are present all across the current facility.

Figure 3.3 shows different Edge computing which allows all the devices that are connected to the network in different remote locations to process and analyze all the data that are presented in the 'edge' of the specific network. This data processing and analysis happened with the help of a local server or by the device [15].

Table 3.2: Road safety rules for cars and bikes

Rules of road safety for Cars on the Road	Wear seatbelts Avoid being distracted while driving Respect the mentioned speed limits Maintain the condition of the car and keep servicing regularly
Rules of road safety for Bikes on Road	Wear ISI-certified helmets Avoid swerving between lanes while riding

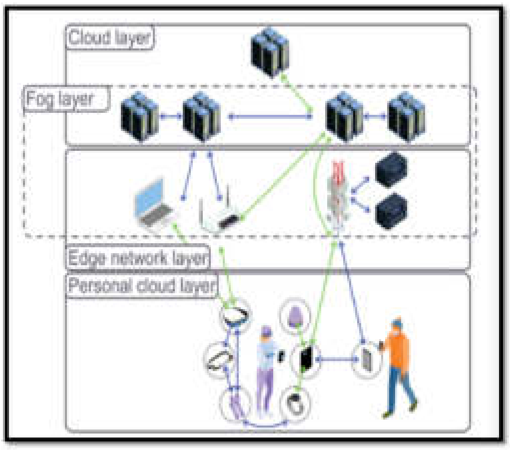


Fig. 3.4: Architecture of edge computing for safety services

When the data needs to be analyzed and processed in the data entered at the central system, the server only transmits the important ones, which can help to minimize the latency of the data.

3.3. Identifying the importance of safety services in the regional road transport system. According to the Bureau of National Crime Records in 2019 almost 467171 road accidents happened in a particular country. When the number of mishaps is so high, it is very important that these issues need to be addressed as shown in table 3.2. The importance of road safety lost its meaning of it when the public do not maintain the basic road safety rule [14]. The basic rules follow crossing the road in the zebra crossing when the traffic light is red; the black and white strips are there for a reason. In addition, the traffic light should be followed while driving to ensure the safety of both pedestrians and others including the safety of the driver. Another important thing to keep in mind while driving or riding is to avoid any kind of drugs or alcohol [17]. Drinking and driving not only puts the driver in a risky situation, but it can also harm others on the road too.

3.4. Discussion of the impact of edge computing in scalable sensor and actuator networks in terms of safety services in the regional road transport. The sector of information security regarding the technical process of safety and monitoring through cloud computing is forming an advanced ecosystem. The conventional structure of edge computing in terms of safety services in the local transportation system has to go through some significant layers of understanding.

Figure 3.4 represents the steps of cutting-edge cloud computing through its mechanical process [3]. The global report of the World Bank states that more than 5% of the individual country's economy is affected by road accidents. Apart from the digital cities, the regional transformation has faced the rapid issue of inconveniences of road trouble.

The lack of qualified workers, from certified drivers to experienced specialists, has made it difficult to fill positions in the transportation sector [4]. The usage of the cloud layers of cutting-edge technology and scalable sensor enable improved workforce allocation with a constrained workforce. Technology has also contributed to

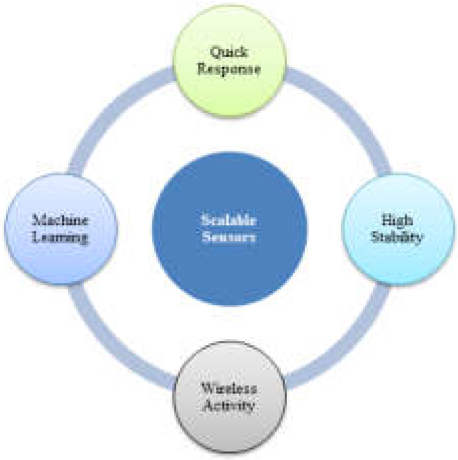


Fig. 3.5: Key feature of Scalable sensors in road safety

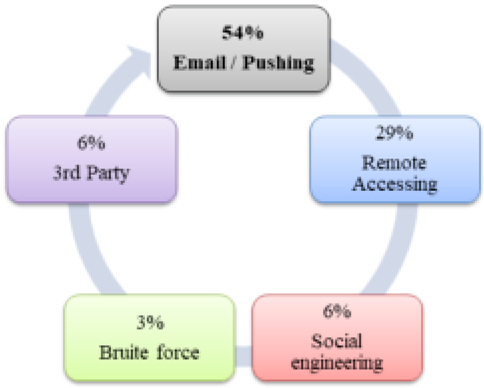


Fig. 3.6: Key feature of Scalable sensors in road safety

the development of motion sensing and monitoring of the whole road network even in regional areas through the cloud connectivity of certain technology which easily monitors driver's road skills.

Figure 3.5 highlights the key components of the scalable sensors that have laid out an innovative approach to prevent probable accidents [7]. The features consist of transmission of quicker response between receiver and sender, optimization of machine algorithm, and wireless connectivity for boundless operation. The cloud network-based computation process is much more stable, and durable, and provides real-time performance. This modern technology can assist the regional road transportation system in the section of improving operator skills and reducing potentially high-cost expenses by identifying behaviors that lead to wear and tear, such as forceful braking and acceleration, monitoring fuel economy, and guaranteeing on-time deliveries [9]. Direct delivery of notifications and alerts from the front end to the drivers makes them aware of the probable occurrence beforehand.

3.5. Evaluation of the challenges faced by safety services in regional road transport during the implementation of edge computing in sclable sensor and actuator networks. The diversity of this ecosystem comes with certain limitations, particularly boundaries and security challenges. This systematic discussion includes the identify similarities, differences, situational attacks, and cyber hacks in the various layers of the computing process. The significant challenges point out the essential security and privacy threats.

Table 3.3: Probable challenges of sensor components

Sensor components	**Internal challenges**
Framework design	The systematic transmission will fail if the sensors have to perform a wider network and complex algorithms. The power consumption and maintaining accuracy with the swift flow of transmission can face misbalance.
Signal processing capability	The circuit framework of smart sensors is the key point for the monitoring and transmission of signals. The key components of the sensor's placement are too dense to transmit a stable and noise-free signal to the receiver.
Reliability maintenance	The reliability and ease of accessing functions hold the whole operation of practical usage of cloud cutting-edge programs in smart sensors. A huge maintenance burden can be faced for the repair and maintenance of the smart sensors if breaks down to technical error.

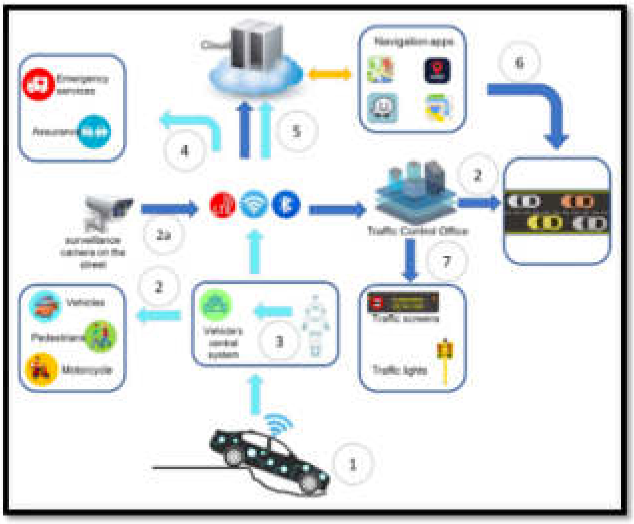


Fig. 3.7: Recommendable Strategies for Improving the Scalable sensor technology

Figure 3.5 focuses on the types of cyber-attacks in percentile status on the implementation of sensors and edge technology[16]. Around 54% of email or phishing attack on the technology is established through the criminals followed by the remote accessing technology that holds 29% of the cyber bullying types. Brute forcing, parameter blocking, and transmission preventing are some other important functions of cyber-attacks on the activity of sensors in edge computer networking [19].

Table 3.3 discusses the significant problems and internal challenges of the sensor components [20]. Coalition of transmission loss ransomware and networking failures are the prominent hindrance to the cutting-edge technology in smart sensing that needs to be checked and monitored properly.

3.6. Assesment of the recommended strategies in order to mitigate these challenges faced vy safety services in regional road transport in scalable sensor and actuator networking. The problems regarding the implementation of sensor technology in road transportation in the rural area require some strategic changes to improve the networking activity and the better development the sender-receiver transmission process.

Figure 3.7 describes the probable recommendation of updating for improving the technical features of cutting-edge technology in the smart sensor networking process [21]. To decrease the number of traffic inconveniences and control better road accession a thorough network connection has to be improved in terms of

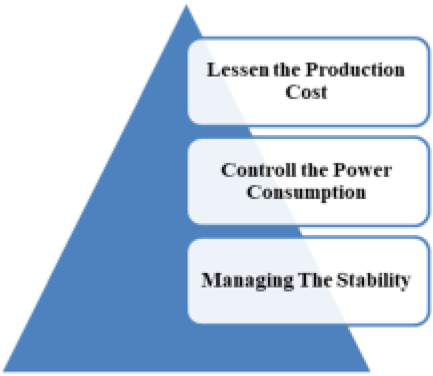


Fig. 3.8: Suggestion for Machinery advancement

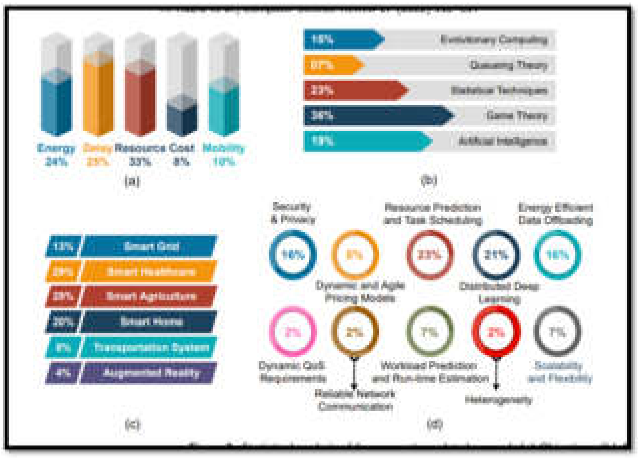


Fig. 4.1: Networking solutions depend on the IoT in the data computation

the vehicle's core system, and differentiation technology for identifying pedestrians, bikes, and light and heavy vehicles. The motion radar system can help to detect traffic symptoms automatically. The continuous existence of surveillance cameras in the street help to cover road occurrence [22].

Figure 3.8 highlights the machinery updating suggestion apart from the other key features [24]. It deals with the management of lessening the production cost and can focus into expanding the business requirement. The practical usage of controlling power consumption is a significant aspect of improving the network strategy. Advancement of wireless connectivity and proper coverage of network tower can enhance the coverage of controlling devices and helps to increase the durability of the network function.

4. Result. In the modern data of world, the advancement of technologies and the systematic implementation of fog technology make the development of road safety. This includes the application of sensor technologies for increasing safety. The computation of Fog technology helps to obtain the working performance of the sensors. It depends on the challenges created by IoT in computer and network sensing technology. The major portions of the road safety technologies of networking solutions depend on the IoT in the data computational. It makes the 57 per cent of the challengers as per a survey from 2018 to 2021. As the technologies are based on Artificial intelligence techniques so this adopts and makes creative changes within the technology as per the required change.

These changes in the computing technologies of the sensors impact the 43 per cent safety of transporting

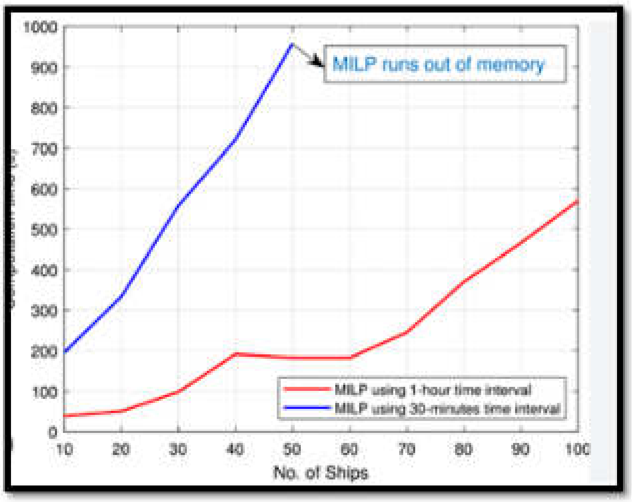


Fig. 4.2: Developments in MILP

vehicles as shown in Figure 4.1 [8]. The improvements of the technologies include the decreasing of data transformation to cloud networks by decreasing latency. Thus in the situation of analyzing the traffic gets easily managed and the time consumption in the transportation also decreases. Therefore these technologies include collecting data from the sensor networks to cloud storing technologies and they use those for making certain predictions to develop the safety and smooth running of transportation on the Road. This technology basically makes digital communication between the networking server and drivers, traffic management, and Police by the application of internet servers.

The application of IoT technologies in data storing of the regional networking servers makes the provision of information through cloud servers. This process includes the performance of FOG computation that directly collects the information as the resources and operates them in the cloud networking server. The technologies include the performance of operating directly through the cloud and this helps in regular monitoring of traffic control. The technologies of the Fog computation make the development and the transformation speed of the information collected by the traffic sensors to a cloud server. This helps in the transformation of the information about the regular changes in the traffic situation to all users quickly. These technologies in the cloud storing and transformation policies impact 39 per cent [18]. The programming technologies of the server are based on algorithmic language programming so they automatically adapt themselves with the adaptive changes. Thus the procedures of FOG computation with a low rate of latency make the major preference of the transporting bodies like the drivers.

This preference and the application benefits of algorithmic programming help in attracting the communication preference and reduce the time of responses. This response is based on the result of communication between the Fog server and the consumer. This also makes the reduction of consumption energy of the server and the drivers get actual and proper time-to-time information about the road circumstances. The cloud storage of the fog computation server includes storing data in the Databus increases the reliability band security of the data [18]. This implementation of the Fog computation helps to make the performance sharing, securing and storing of data smooth. This workflow and the smooth running of the sensor networking technology decrease the whole rate latency and consumption of energy. Therefore these technologies include collecting data from the sensor networks to cloud storing technologies and they use those for making certain predictions to develop the safety and smooth running of transportation on the Road. This infrastructure of the FOG computation makes the development of the traffic system in the regional division as well as in the different locations.

5. Problem Statement. People often avoid a much concerning issue while driving is their health issue. It is suggested that if a person has some health issues that can distract the person while driving, the person should avoid driving. Every country has its own kind of rules and regulations that must be followed while driving, riding, or even walking on the roads there [18]. Most of countries use different types of sign boards to signal the slowdown, schools, and hospitals ahead, different turns and no turning corners, different hand gestures, even the usage of horns and all that should be followed to be safe on the road. These safety concerns do not only focus on the drivers, most of the time when a road accident happens but there are also two sides to that accident [23]. On one side, there is the driver who is injured and harmed, while on the other side, there can be a pedestrian, another vehicle, or anybody standing safely can be harmed. Therefore, it is very important to abide by all the rules so and be safe while on the roads to avoid any kind of mishaps.

6. Conclusion. In concluding part it can be stated that the advancement of sensor technology in rural areas can now improve the radar facility with the progression of wireless networking and Bluetooth connectivity. Updated navigation maps and cloud networking helps the drivers to have all the necessary information regarding the position of individual vehicles and the present traffic updates. Smart sensors can also be employed in new domains thanks to machine learning and artificial intelligence applications developed in recent years, computation, and model training, It surely helps to the enhance processing power of embedded systems. These technological improvements are characterized by reduced data transmission to cloud networks, leading to decreased latency in traffic analysis and management. The integration of sensor data collected from networks into cloud storage systems facilitates predictive capabilities, thereby promoting road safety and traffic efficiency. Notably, this technology fosters digital communication among networking servers, drivers, traffic management, and law enforcement agencies through internet-based applications.

Additionally, the utilization of IoT technologies in regional networking server data storage facilitates information provision via cloud servers. This process involves the execution of FOG computing, which efficiently collects, processes, and operates information in cloud networking servers. Consequently, the performance and speed of information transformation from traffic sensors to cloud servers significantly impact traffic monitoring and control, contributing to 39% of the overall effects. The server programming technologies, characterized by algorithmic language programming, seamlessly adapt to dynamic changes. FOG computing procedures, with low latency rates, are preferred by transportation stakeholders, notably drivers, due to their effective communication and timely response capabilities. This not only reduces server energy consumption but also ensures drivers receive accurate real-time information about road conditions. Furthermore, the cloud storage of the FOG computation server, along with data storage in the Databus, enhances data reliability and security.

The implementation of FOG computation substantially enhances the flow of sensor networking technologies, resulting in decreased latency and energy consumption. This infrastructure facilitates the development of traffic systems across regional divisions and diverse locations. The collective impact of these innovations is a promising step towards more efficient, secure, and reliable road transportation

REFERENCES

[1] O. Ali, M. K. Ishak, M. K. L. Bhatti, I. Khan, and K.-I. Kim, *A comprehensive review of internet of things: Technology stack, middlewares, and fog/edge computing interface*, Sensors, 22 (2022), p. 995.

[2] A. Alwarafy, K. A. Al-Thelaya, M. Abdallah, J. Schneider, and M. Hamdi, *A survey on security and privacy issues in edge-computing-assisted internet of things*, IEEE Internet of Things Journal, 8 (2020), pp. 4004–4022.

[3] A. Arena, P. Perazzo, C. Vallati, G. Dini, and G. Anastasi, *Evaluating and improving the scalability of rpl security in the internet of things*, Computer Communications, 151 (2020), pp. 119–132.

[4] A. T. Atieh, *The next generation cloud technologies: a review on distributed cloud, fog and edge computing and their opportunities and challenges*, ResearchBerg Review of Science and Technology, 1 (2021), pp. 1–15.

[5] A. Bourechak, O. Zedadra, M. N. Kouahla, A. Guerrieri, H. Seridi, and G. Fortino, *At the confluence of artificial intelligence and edge computing in iot-based applications: A review and new perspectives*, Sensors, 23 (2023), p. 1639.

[6] B. K. Dar, M. A. Shah, S. U. Islam, C. Maple, S. Mussadiq, and S. Khan, *Delay-aware accident detection and response system using fog computing*, Ieee Access, 7 (2019), pp. 70975–70985.

[7] K. Z. Ghafoor, L. Kong, S. Zeadally, A. S. Sadiq, G. Epiphaniou, M. Hammoudeh, A. K. Bashir, and S. Mumtaz, *Millimeter-wave communication for internet of vehicles: status, challenges, and perspectives*, IEEE Internet of Things Journal, 7 (2020), pp. 8525–8546.

[8] A. Hazra, P. Rana, M. Adhikari, and T. Amgoth, *Fog computing for next-generation internet of things: fundamental, state-of-the-art and research challenges*, Computer Science Review, 48 (2023), p. 100549.

[9] S. Islam, S. Badsha, S. Sengupta, H. La, I. Khalil, and M. Atiquzzaman, *Blockchain-enabled intelligent vehicular edge computing*, IEEE Network, 35 (2021), pp. 125–131.

[10] S. S. Jazaeri, S. Jabbehdari, P. Asghari, and H. Haj Seyyed Javadi, *Edge computing in sdn-iot networks: a systematic review of issues, challenges and solutions*, Cluster Computing, (2021), pp. 1–42.

[11] M. H. Kashani and E. Mahdipour, *Load balancing algorithms in fog computing*, IEEE Transactions on Services Computing, 16 (2022), pp. 1505–1521.

[12] A. W. Kiwelekar, P. Patil, L. D. Netak, and S. U. Waikar, *Blockchain-based security services for fog computing*, Fog/Edge Computing For Security, Privacy, and Applications, (2021), pp. 271–290.

[13] M. Laroui, B. Nour, H. Moungla, M. A. Cherif, H. Afifi, and M. Guizani, *Edge and fog computing for iot: A survey on current research activities & future directions*, Computer Communications, 180 (2021), pp. 210–231.

[14] C. P. Lee, F. T. J. Leng, R. A. A. Habeeb, M. A. Amanullah, and M. H. ur Rehman, *Edge computing-enabled secure and energy-efficient smart parking: A review*, Microprocessors and Microsystems, (2022), p. 104612.

[15] F. Liu, G. Tang, Y. Li, Z. Cai, X. Zhang, and T. Zhou, *A survey on edge computing systems and tools*, Proceedings of the IEEE, 107 (2019), pp. 1537–1562.

[16] L. Liu, C. Chen, Q. Pei, S. Maharjan, and Y. Zhang, *Vehicular edge computing and networking: A survey*, Mobile networks and applications, 26 (2021), pp. 1145–1168.

[17] H. T. Malazi, S. R. Chaudhry, A. Kazmi, A. Palade, C. Cabrera, G. White, and S. Clarke, *Dynamic service placement in multi-access edge computing: A systematic literature review*, IEEE Access, 10 (2022), pp. 32639–32688.

[18] I. Martinez, A. S. Hafid, and A. Jarray, *Design, resource management, and evaluation of fog computing systems: a survey*, IEEE Internet of Things Journal, 8 (2020), pp. 2494–2516.

[19] A. Meslin, N. Rodriguez, and M. Endler, *Scalable mobile sensing for smart cities: The musanet experience*, IEEE Internet of Things Journal, 7 (2020), pp. 5202–5209.

[20] G. Peserico, A. Morato, F. Tramarin, and S. Vitturi, *Functional safety networks and protocols in the industrial internet of things era*, Sensors, 21 (2021), p. 6073.

[21] M. Peyman, P. J. Copado, R. D. Tordecilla, L. d. C. Martins, F. Xhafa, and A. A. Juan, *Edge computing and iot analytics for agile optimization in intelligent transportation systems*, Energies, 14 (2021), p. 6309.

[22] M. A. Rahman, A. T. Asyhari, I. F. Kurniawan, M. J. Ali, M. Rahman, and M. Karim, *A scalable hybrid mac strategy for traffic-differentiated iot-enabled intra-vehicular networks*, Computer Communications, 157 (2020), pp. 320–328.

[23] M. Songhorabadi, M. Rahimi, A. MoghadamFarid, and M. H. Kashani, *Fog computing approaches in iot-enabled smart cities*, Journal of Network and Computer Applications, 211 (2023), p. 103557.

[24] F. Tonini, B. M. Khorsandi, E. Amato, and C. Raffaelli, *Scalable edge computing deployment for reliable service provisioning in vehicular networks*, Journal of Sensor and Actuator Networks, 8 (2019), p. 51.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 29, 2023
Accepted: Oct 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org

Volume 24, Issues 4, pp. 1215–1222, DOI 10.12694/scpe.v24i4.2568

A STUDY OF MENTAL TRAINING AND SKILL ENHANCEMENT IN PHYSICAL EDUCATION TEACHING COMBINED WITH DEEP LEARNING ALGORITHMS

SHENGFEI HU*AND ZIYAO GAO†

Abstract. The study sheds light on teaching physical education, which aids in improving students overall physical and mental well-being. Individuals benefit from physical education instruction through developing their talents and minds. Teaching physical education is crucial since a person's whole well-being mostly depends on their physical and mental health. Deep learning algorithms are used in physical education training to improve quality and help students become more mentally and skillfully fit. The deep learning method is a machine learning component that aids the healthcare and e-commerce sectors by comprehending how the human brain functions. Comprehending the physical education teaching process and how it can be enhanced also aids individuals. This study contributes to our understanding of the value of physical education instruction and how it affects a person's physical and mental health.

Key words: Mental training, deep learning algorithm, machine learning

1. Introduction. Physical education teaching helps us to grow the physical and mental health of an individual. Physical education teaching also helps individuals by providing mental training and skills development. Physical education teaching is always important as both mental and physical health are the main factors for good well-being in a person. Physical education teaching provides more quality with the use of deep learning algorithms and helps to increase mental fitness and skill enhancement in an individual. The deep learning process is a part of machine learning, which understands the workings of the human brain and helps in the healthcare and e-commerce industries. This also helps individuals by understanding the physical education teaching process and how it can be improved. This study helps to understand the importance of physical education teaching and its impact on mental and physical health in an individual. The study provides proper data to analyze the topic of the study. In the methodology section of the study, a secondary qualitative method is used for data collection which can enrich the quality of the result.

The deep learning and the knowledge of the branch machine languages that are attempted in order to models high-level abstraction of the data and the information by the usage of the multi-layer neurons. The usage of the multi-level neurons is composed in the complex structured or the transformations that are non-linear. This help in the increase in the volume of the data and information for the computing of the power and the complex networks in the areas and the fields that are taken into considerations.

The processes of the physical education and the traditional method or the techniques that are related to the physical educations are usually limited by the professional or the teachers who are at professional's level. Moreover, the traditional education related to physical and improvement of the skills required quality as well as effects of the teachings. The quality and the effects that are related to the teaching are often very difficult to guarantee. In addition the traditional methods of the teaching or learning on the basis of the physical education as well as enhancement of the skills also require certain venue or space having a strict rules and regulations and more training time.

Physical education plays a pivotal role in promoting holistic well-being and skill development among students. While traditional teaching methods have been effective, the integration of advanced technologies offers new opportunities to enhance the learning experience. This study delves into the synergistic relationship between mental training, skill enhancement, and deep learning algorithms in the context of physical education

*Physical Education college of South-Central Minzu University, Wuhan, 430074, China, email: shengfeiresea@outlook.com
†Physical Education college of South-Central Minzu University, Wuhan, 430074, China

teaching. By combining the principles of pedagogy with cutting-edge technology, we aim to explore innovative strategies that can revolutionize physical education and contribute to the overall development of students.

Main contribution of this research is,

1. This study delves into the sphere of physical education, illuminating its pivotal role in enhancing the holistic well-being of students, encompassing physical and mental dimensions.
2. By harmoniously integrating mental training and skill enhancement strategies, this research underscores the multifaceted advantages of physical education, positioning it as a conduit for developing not only physical prowess but also cognitive abilities.
3. Leveraging deep learning algorithms, typically employed in healthcare and e-commerce sectors, within the realm of physical education, this study advances our understanding of how these advanced technologies can amplify the quality of teaching and bolster students' mental and physical fitness.

2. Objectives.

1. To understand the role of physical education teaching in improving mental fitness in individuals.
2. To know the role of physical education teaching on the skill enhancement in an individual.
3. To evaluate how the deep learning algorithm integrates with physical education teaching to provide the best results.
4. To suggest the best possible way to utilize the skill development and deep learning algorithm

3. Methodology. The secondary qualitative method is used for the collection of the data in the study. The secondary data are collected from sources that have already been collected for certain uses. The sources such as government websites, books, and documents from government libraries and newspapers are used for secondary data collection methods [19]. In this study secondary data is used as this has certain advantages, the first advantage is that this method is both time effective as well as cost-effective.

The next advantage is that this process of data collection is quite simple as any individual can collect the data. There are no hard skills required in the collection of secondary data and this makes the method simple. The next advantage is that the secondary data sources are easily accessible to individuals as compared to the primary data collection methods.

3.1. Physical education teaching helps individuals to improve their mental and physical health. Physical education teaching is about strategic procedure in workout techniques that helps to improve the health and mind of an individual. Physical and mental health plays a significant role in improving the mental health and physical health of an individual [4]. Physical education teaching is guided by experts who have deep knowledge and information about the physical and mental health of a human being. Physical and mental health is important for better and for staying away from any type of health disorder in the long run [22]. There are many advantages of physical training for physical health in a person. The first advantage is physical growth, the physical training helps to develop muscles and make the bone strong in an individual [23].

The next benefit of physical training is the mitigation of stress levels in individuals. In recent times stress has become one of the main problems in a person. Stress leads to these various types of problems such as heart disease, mental health issues, and many types of health problems in a person [16]. Physical training helps a person to be fresh and free from many types of diseases in the long run.

The next advantage of physical training is that it helps to be disciplined and patient. Physical training helps to develop confidence and self-discipline in a person. Physical training also helps with focus and concentration skills in a person [8]. The focus is very important for the growth and development of a person which can be improved by physical training. The sleep cycle which is very important for a healthy lifestyle in a person can be improved by the physical training process in person [24]. The physical training process also helps in character development which is important to be a part of society. In the same way, physical training helps to improve the mental health of a person. Mental health and peace is also important factor for the good well-being of a person. Physical activity helps to develop a good mood in a person; physical training helps to increase the serotonin level in the human brain which helps individuals to be in a good mood and stress-free [2]. The release of serotonin in the human body is important to help protect an individual's mental and physical health issues.

Physical training also helps to avoid depression in a person, the depression which is the cause of many

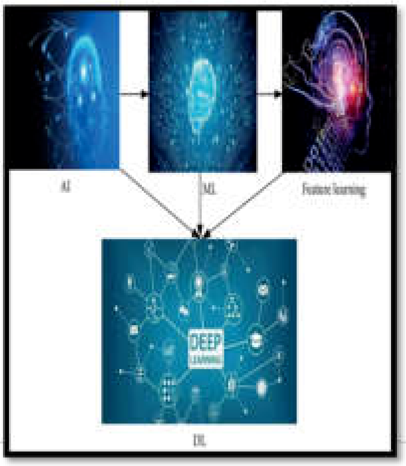


Fig. 3.1: The physical education training by the use of deep learning algorithm

types of health problems in a human being. Physical training and education help in developing self-esteem in a person [11].

3.2. The role of physical education teaching in improving the skills of individuals. Physical education teaching helps to improve performance in a person. Through consistent physical activities, a person can improve certain skills which are important for living. Physical education improves the mental peace and mental sharpness in a person which helps to acquire any skills easily [3]. Any type of physical training such as swimming, running, and walking make a positive impact on the human brain that helps in the easy learning process by a person. The physical training process helps A person to think strategically and also helps to make better decisions. Physical education teaching also helps a person to develop social skills. Social skills are very important for living in a society and to gain respect from society [5]. Physical education teaching helps to develop better communication skills and better management skills to individuals. These teamwork and management skills are important for a person to live a better life with family as well as in society. Physical training can help a person develop management skills and learn teamwork [7].

Physical education also helps in the self-improvement process and better character development in a person. The physical training process is important for the student and the younger generation in a country. Physical training also helps in getting enough sleep in an individual. Proper sleep plays a significant role in the performance and skill development of a person [18]. It became easy for a person to learn any new skills when the person got proper sleep. Physical training also helps in developing leadership quality and better communication skills in a person. Leadership qualities and better communication are important in a person who is living in society.

The physical training is a effective and the efficient procedures for the treatment of the mental heaths and as well as anti-anxiety. The physical training help in relieves stress and tensions that is bale to boost or developed the mental as well as the physical energy. The good physical healths also help in the enhancement of the well-being through the release of the endorphins. It is not just the about the capacity of the aerobic and the size of the muscle, but it help in the improvement of the mental heaths developing overall body personality.

The individual who do physical training regularly because it give them an enormous sense of the well-being and be more energetic in comparison with the individual that do not do any physical training. The physical training has many benefits that will make the growth in the overall health of the individual. There are many advantages of the physical training that includes better sleeps at night, feeling energetic during the day, growth in the power of the memory, and it also make the individual relaxed and has the positive approach about themselves and their life.

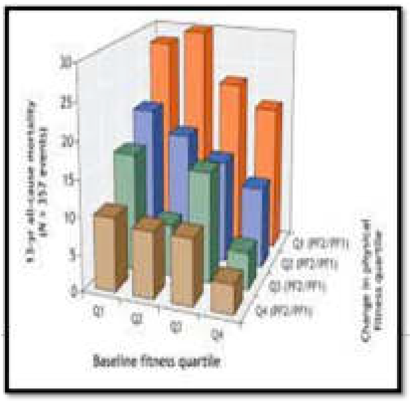


Fig. 3.2: The advantages of physical health on overall health

Fig. 3.3: The health of the students and the service of counseling

3.3. Discussion of the process of integration of deep learning algorithm with physical education teaching to generate the best results. The deep learning algorithm is part of the machine learning process that helps to understand the human brain structure and this helps to make any activity easier for human use. The deep learning algorithm process uses the data and analyzes the data to help an individual in any process [12]. In the context of this study, the deep learning algorithm helps to improve the physical education teaching process [29]. Physical training education merged with deep learning algorithms helps in getting better and more effective results. The deep learning algorithm makes any process more effective and strategic and gets better results [29].

The deep learning algorithm uses the data and analyzes this data to understand the role of physical education teaching on the human brain and mental health in an individual [27]. By understanding the human brain technology can make the process better and more effective. In the digital age, digital technology is used in all fields for better performance and better results [21].

The deep learning algorithm is also used in different fields such as healthcare, automobile industries, and the entertainment process. There are certain advantages of deep learning algorithms in physical education teaching [15]. The first advantage is the automatic process of learning; the deep learning algorithm automatically learns the characteristics of the data that help to make the process more simple and effective. The next advantage is that the deep learning algorithm helps in managing large and complex data [25]. The deep learning algorithm can manage the complicated dataset in a process. The next advantage is that the deep learning algorithm

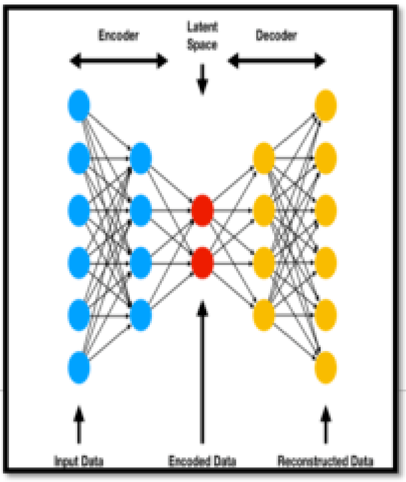


Fig. 3.4: The deep learning algorithm

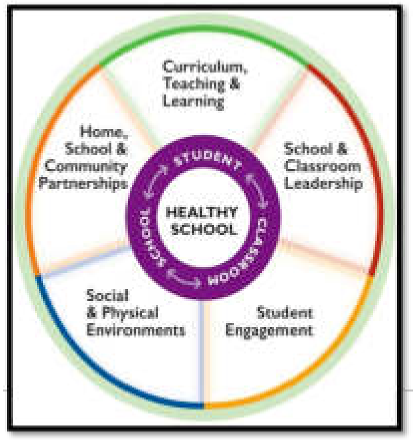


Fig. 3.5: Representation of the healthy school

also helps in increasing the performance of the process [13]. The deep learning algorithm helps to understand complex issues and helps to solve the issues by analyzing the data [17]. The next advantage is that the deep learning algorithm helps to manage the structured and unstructured data, file such as image, text, and audio can be easily managed and analyzed by this algorithm technology [10].

The first crucial step is selecting the appropriate deep learning algorithms for the given context. This decision depends on the specific objectives of the physical education program. Deep learning models are trained on the preprocessed data. This training phase involves feeding the models with labeled data (e.g., performance ratings, emotional states) to learn patterns and correlations. The models adapt and improve their performance over time.

4. Results. The study helps to understand the importance of physical education teaching in improving the physical and mental health of an individual. As mental and physical health is the primary determinants of a person's well-being, physical education instruction is always crucial [1]. By utilizing deep learning algorithms, physical education instruction is of higher quality and contributes to a person's improvement in both mental health and skill-building. The deep learning method, which is a component of machine learning, assists the

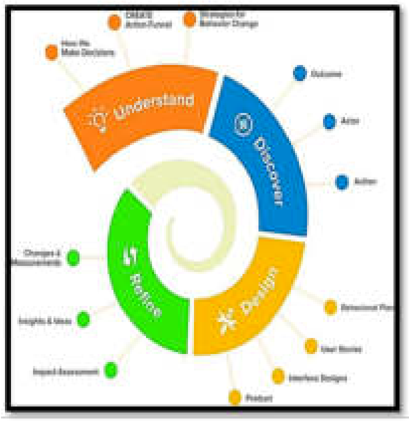


Fig. 4.1: Physical education system using multimedia technology

healthcare and e-commerce sectors by understanding how the human brain functions [28].

The deep learning algorithm procedure analyzes data to aid a person in any venture. In the context of this study, the deep learning algorithm aids in improving the training process of physical education. Combining deep learning algorithms with physical education helps provide better and more productive results. Any procedure may be made more strategic and effective to achieve better results using the deep learning algorithm [14].

A person's mental clarity and peace are improved through physical education, which makes it easier for them to pick up new abilities. Any physical exercise, including swimming, running, and walking, has a favorable effect on the brain and facilitates a person's ability to learn [6]. A person can think strategically and make better decisions with the aid of physical exercise. A person can strengthen their social skills with the use of physical education teaching, for one to function in society and to be respected by others, social skills are crucial. Physical exercise has various benefits for a person's physical health [9].

The first benefit is the individual's physical development; exercise helps build muscles and strengthen bones. The reduction of stress levels in people is the next advantage of physical exercise. In recent years, stress has emerged as one of a person's major issues. Stress causes a variety of concerns in a person's health, including heart disease, mental health conditions, and many others. As a result, technological development is very important as it can make a process more effective and flexible [20]. The deep learning algorithm which is part of machine learning technology helps in physical education teaching to make it more effective and flexible for human use [26]. In recent times all fields are influenced by technology, hence physical education teaching is also been developed by the use of deep learning algorithms. In recent times there are many institutions that use the deep learning algorithm for improving physical education teaching [30]. Many big companies in recent times are investing in deep learning algorithms and physical training education for the better health of individuals.

Figure 4.2 represents that the youths of different countries are not involved with the physical activities. The adolescence of 94.2% of South Korea is not involved with the physical activities. The other countries like Italy, France, Russia, and China, Germany's adolescence of 88.6%, 87%, 84.5%, 84.3%, and 83.7% are not involved with the physical activities respectively.

Thus, the development of the mental training and the skill development are essential for the development of the physical strength of the people. The deep learning technology assists the human being to practice all the physical activities and other activities in an appropriate way.

The physical training helped the individual to learn the meaning of the confidence by boosting and enhancing the skills through making new fosters and friends as well as the sprits creating the healthy competition environment. This environment of the healthy competitions helps the individual to take participant in the variety of sports. The physical training or education not only increased the physical health of the individual but also help in the developed in the mental health.

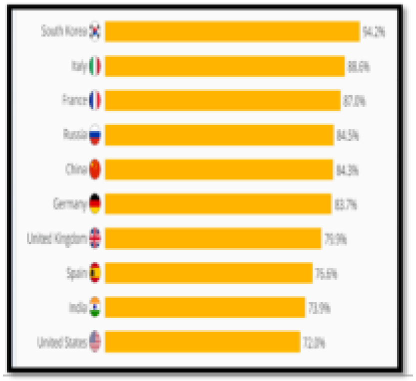


Fig. 4.2: Percentage of youths not engaged in physical activity

5. Conclusion. It can be concluded that physical education teaching is significant in improving the physical as well as mental health of a human being. The deep learning algorithm is a kind of machine learning process that helps to understand the human brain framework and functioning of the human brain. The physical education teaching process becomes more effective when it is integrated with the deep learning algorithm. The study has helped to understand the impact of physical education teaching on an individual and how the technology can be used positively to make a process more effective. The study has discussed the topic with the help of data which is collected by the use of a secondary qualitative method. By harnessing the power of deep learning algorithms, the quality of physical education instruction is elevated, contributing to substantial improvements in an individual's mental and skill development. Moreover, physical education is instrumental in honing social skills, a vital ingredient for an individual's societal integration and acceptance by peers. in future Deep neural network will be implemented and tested for accuracy.

REFERENCES

[1] N. Bakhmat, B. Maksymchuk, O. Voloshyna, V. Kuzmenko, T. Matviichuk, A. Kovalchuk, and E. Manzhos, *Designing cloud-oriented university environment in teacher training of future physical education teachers*, Journal of Physical Education and Sport®(JPES).-2019.-Vol 19 (Supplement issue 4), Art 192.-P. 1323-1332.-Publisher: University of Pitesti Country of publisher: Romania., (2019).

[2] M. Behnke, M. Tomczak, L. D. Kaczmarek, M. Komar, and J. Gracz, *The sport mental training questionnaire: Development and validation*, Current Psychology, 38 (2019), pp. 504–516.

[3] T. M. Bozkurt and M. A. Öztürk, *The effect of mental training levels of athletes on attitudes towards playing digital games*, Revista de Psicología del Deporte (Journal of Sport Psychology), 32 (2023), pp. 320–330.

[4] O. B. Çelik and N. B. Güngör, *The effects of the mental training skills on the prediction of the sports sciences faculty students'anxiety levels.*, International Journal of Eurasian Education & Culture, 5 (2020).

[5] I. Demchenko, B. Maksymchuk, V. Bilan, I. Maksymchuk, and I. Kalynovska, *Training future physical education teachers for professional activities under the conditions of inclusive education*, BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 12 (2021), pp. 191–213.

[6] J. Fernandez-Rio, E. de las Heras, T. González, V. Trillo, and J. Palomares, *Gamification and physical education. viability and preliminary views from students and teachers*, Physical education and sport pedagogy, 25 (2020), pp. 509–524.

[7] G. Fyall and M. W. Metzler, *Aligning critical physical education teacher education and models-based practice*, (2019).

[8] R. Guidotti, C. Del Gratta, M. G. Perrucci, G. L. Romani, and A. Raffone, *Neuroplasticity within and between functional brain networks in mental training based on long-term meditation*, Brain Sciences, 11 (2021), p. 1086.

[9] H.-C. Jeong and W.-Y. So, *Difficulties of online physical education classes in middle and high school and an efficient operation plan to address them*, International journal of environmental research and public health, 17 (2020), p. 7279.

[10] H. S. Lee and J. Lee, *Applying artificial intelligence in physical education and future perspectives*, Sustainability, 13 (2021), p. 351.

[11] F. M. Leo, A. Mouratidis, J. Pulido, M. López-Gajardo, and D. Sánchez-Oliva, *Perceived teachers' behavior and students' engagement in physical education: The mediating role of basic psychological needs and self-determined motivation*, Physical Education and Sport Pedagogy, 27 (2022), pp. 59–76.

[12] T. Liu, D. Wilczyńska, M. Lipowski, and Z. Zhao, *Optimization of a sports activity development model using artificial*

intelligence under new curriculum reform, International Journal of Environmental Research and Public Health, 18 (2021), p. 9049.

[13] M. S. Mahmood, Y.-S. Kim, R. Herbert, and W.-H. Yeo, *Universal brain-machine interfaces enabled by flexible scalp electronics and deep convolutional neural networks*, in Electrochemical Society Meeting Abstracts 236, no. 53, The Electrochemical Society, Inc., 2019, pp. 2307–2307.

[14] L. S. Marques, C. Gresse von Wangenheim, and J. C. Hauck, *Teaching machine learning in school: A systematic mapping of the state of the art*, Informatics in Education, 19 (2020), pp. 283–321.

[15] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, and A. Mosavi, *Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis*, IEEE Access, 8 (2020), pp. 150199–150212.

[16] J. H. Ohrt, J. D. Deaton, K. Linich, J. D. Guest, B. Wymer, and B. Sandonato, *Teacher training in k–12 student mental health: A systematic review*, Psychology in the Schools, 57 (2020), pp. 833–846.

[17] Y. K. Ooi and H. Ibrahim, *Deep learning algorithms for single image super-resolution: a systematic review*, Electronics, 10 (2021), p. 867.

[18] W. O'Brien, M. Adamakis, N. O'Brien, M. Onofre, J. Martins, A. Dania, K. Makopoulou, F. Herold, K. Ng, and J. Costa, *Implications for european physical education teacher education during the covid-19 pandemic: a cross-institutional swot analysis*, European Journal of Teacher Education, 43 (2020), pp. 503–522.

[19] M. R. Roller et al., *A quality approach to qualitative content analysis: Similarities and differences compared to other qualitative methods*, in Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, vol. 20, 2019.

[20] U. Sanjar and H. Doston, *Innovative methods of improving the professional skills of a teacher of physical culture*, INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, IT, ENGINEERING AND SOCIAL SCIENCES ISSN: 2349-7793 Impact Factor: 6.876, 16 (2022), pp. 20–23.

[21] A. Shrestha and A. Mahmood, *Review of deep learning algorithms and architectures*, IEEE access, 7 (2019), pp. 53040–53065.

[22] I. Solhaug, M. de Vibe, O. Friborg, T. Sørlie, R. Tyssen, A. Bjørndal, and J. H. Rosenvinge, *Long-term mental health effects of mindfulness training: a 4-year follow-up study*, Mindfulness, 10 (2019), pp. 1661–1672.

[23] M. Turgut and O. M. Yasar, *Mental training of college student elite athletes.*, Journal of Education and Learning, 9 (2020), pp. 51–59.

[24] V. Varea, G. Gonzalez-Calvo, and A. García-Monge, *Exploring the changes of physical education in the age of covid-19*, Physical Education and Sport Pedagogy, 27 (2022), pp. 32–42.

[25] D. Řadovanović and S. Dukanović, *Image-based plant disease detection: a comparison of deep learning and classical machine learning algorithms*, in 2020 24th International conference on information technology (IT), IEEE, 2020, pp. 1–4.

[26] R. Wang, Y. Liu, Y. Lu, J. Zhang, P. Liu, Y. Yao, and G. Grekousis, *Perceptions of built environment and health outcomes for older chinese in beijing: A big data approach with street view images and deep learning technique*, Computers, Environment and Urban Systems, 78 (2019), p. 101386.

[27] D. Yang, E.-S. Oh, and Y. Wang, *Hybrid physical education teaching and curriculum design based on a voice interactive artificial intelligence educational robot*, Sustainability, 12 (2020), p. 8000.

[28] Y. Zeng, *Evaluation of physical education teaching quality in colleges based on the hybrid technology of data mining and hidden markov model*, International Journal of Emerging Technologies in Learning (iJET), 15 (2020), pp. 4–15.

[29] W. Zhang, H. Li, Y. Li, H. Liu, Y. Chen, and X. Ding, *Application of deep learning algorithms in geotechnical engineering: a short critical review*, Artificial Intelligence Review, (2021), pp. 1–41.

[30] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan, and X. Chen, *Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study*, ISA transactions, 107 (2020), pp. 224–255.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 30, 2023
Accepted: Oct 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1223–1229, DOI 10.12694/scpe.v24i4.2569

A STUDY ON OPTIMIZING ERROR DETECTION AND CORRECTION STRATEGIES IN PHYSICAL EDUCATION AND SPORT TEACHING USING DATA MINING ALGORITHMS

ZIYAO GAO*, SHENGFEI HU†, GUO YU‡, AND YINHUI LI§

Abstract. In the fiercely competitive realm of sports and physical education, the application of data mining algorithms has emerged as a vital solution. Machine learning has streamlined processes, offering a seamless means of elevating the quality of education and training provided to students, particularly in the context of sports. This technological support empowers the sports education system to make more informed decisions pertaining to the physical development of aspiring athletes. In this comprehensive study, a blended approach of qualitative methods has been leveraged to gather intricate insights, enriching the overall understanding of the subject. Additionally, an in-depth exploration of articles and journals has been undertaken to scrutinize the practical implementation of data algorithm techniques geared towards enhancing physical training. The resultant findings underscore a substantial and tangible nexus between data algorithms and the domain of sports education. Of paramount significance is the central role played by data mining algorithms in augmenting performance. Notably, the National Sports Board (NSB) has extensively harnessed this technology to meticulously monitor players' on-field performance, ultimately leading to a granular comprehension of each player's capabilities. This paper emphasizes the methods of optimizing mistake detection and its joining systems for increasing the punishment in the operational procedures.

Key words: Optimizing Error Detection, Data mining algorithm, sports teaching, physical education

1. Introduction. Optimization is a methodological act and process of developing a functional and fully effective source. The increased implementation of technology has developed the internet-based education system and increased the growth of physical education. According to, Alamiedy et al. (2020), the development of network technology and information technology helps to achieve high convenience in performing the algorithms of data mining. An algorithm of machine learning (ML) or data mining is a set of calculations and heuristics that develops a data model. Furthermore, to develop the algorithm must scrutinize the provided data after analyzing the trends and patterns.

In the area of information, the knowledge and skills of people have been enhanced rapidly with data mining algorithms (DMA). Nowadays, the requirement of technology has been in high demand in sports teaching (SP) and physical education (PE). DMA tackles the challenge of flossing information in the education sector. In recent times, the importance and function of data mining have been highly released mainly in the sports field.

This paper highlights the processes of optimizing error detection and its associated strategies for enhancing the correction in the operational processes. These processes are generally implicated in physical education and sports teaching utilizing the algorithms of data mining approaches. In addition to this, it helps to implicate data mining tools to direct sports training and related physical sports. Moreover, optimizing error detection and strategic Eros detection techniques promotes physical education and boosts the tactic and strategic analysts. Furthermore, the tools of the logarithms of data mining help to optimize error detection in the field of education and sports. As per the view of Alamiedy et al. (2020), these tools are sent to be used in competitive sports to mine out the informational data from the massive datasets directing the sports trainers to focus on physical education growth and development.

The lack of authentic and more informative information about the data Ming algorithm is the limitation

*Physical Education college of South-Central Minzu University, Wuhan, 430074, China
†Physical Education college of South-Central Minzu University, Wuhan, 430074, China
‡School of Physical Education Wuhan University of Technology Wuhan China 430070
§Physical Education department, Zhongnan University of Economics and Law, Wuhan, 430073, China (yinnhuilireser@outlook.com)

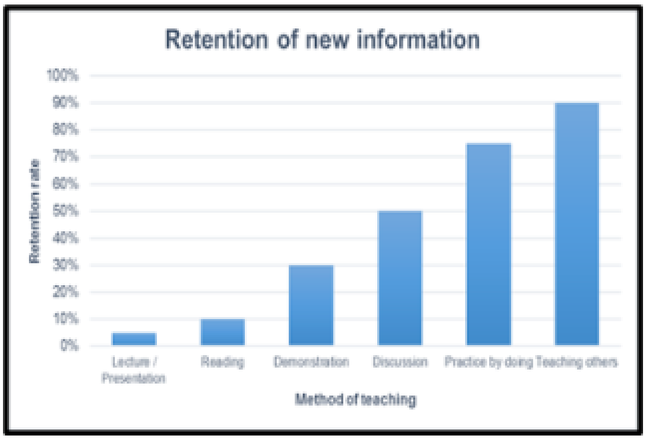


Fig. 1.1: Rate of using data mining algorithm in physical education

of the study. Further, all aspects of the technology of data mining have not been evaluated in the study. In addition, this study consumes lots of time to collect more informative data on physical education development.

1.1. Motivation . In the modern era, physical education and sports hold a pivotal place in individual development and overall well-being. However, to maximize the benefits of physical education, it is imperative to embrace technological advancements. This study is motivated by the realization that data mining algorithms and machine learning can revolutionize the way physical education is administered and enhance the outcomes for students and athletes. The motivation behind this research is to explore the synergies between data mining algorithms and physical education, harnessing the power of technology to optimize training, skill enhancement, and mental well-being.

The research contributes,

1. By integrating data mining algorithms, physical education programs can be tailored to individual needs, optimizing training regimens for each student or athlete.
2. The study sheds light on how data-driven approaches can be used for mental training, helping individuals build resilience, focus, and emotional well-being alongside physical prowess.

2. Optimizing Error Detection and Correction . Error identification and correction are strategies used to recognize and precise the faults that arise at the time of storage and transmission of data. Hence, some algorithms are designed to identify the error because of interference and the factors that influence data corruption. As suggested by Shahid et al. [8], Error detection is the process of identifying the errors at the time of transmitting and to the transmitters. Furthermore, error correction is the reconstruction of the error-free and original data.

In today's highly interconnected and data-driven world, the need for robust error detection and correction methods is paramount. From critical systems like financial transactions to everyday tasks such as communication, errors can have far-reaching consequences. The motivation behind this research is to harness the power of advanced technologies to optimize error detection and correction processes, ensuring the integrity and reliability of data and systems.

3. Strategies of optimization error correction or detection. It includes several techniques such as simple parity check, checksum, hamming code, and cyclic.

Parity check: It is the method the network develops to analyze the errors and evaluate the data received integrity at the side of the receiver. Of At the time of event parity, if the 1s number is even, the value of the partial bit is 0. Further, if the other number is odd then 1 is the value of the party

Checksum: It is the method that helps to detect the data which has been transmitted. The procedure includes the data dividing into even segment sizes.

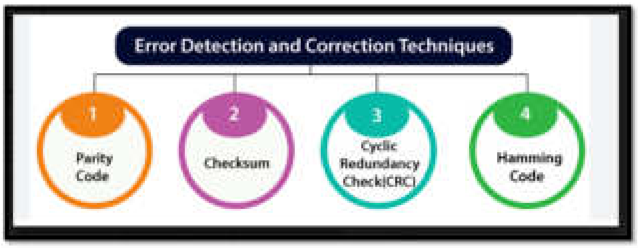


Fig. 3.1: Strategies of optimization error correction or detection

Hamming code: It is a system that identifies and detects errors at the time of storing data or transmitting. It is highly used in the ECC RAM and enhances the processing.

Cyclic Redundancy Check (CRC): It is the method of knowing the details of the transmitted data in the network.

3.1. Impact of Data mining algorithm on sports and physical education. As per the comment of Wang et al.[**?**], the implication of data mining has improved sports functions and other physical education activities. Moreover, it is one of the effective approaches in the field of educational systems that help in the improvement of the physical health of the students and players in the education system. Physical education involves a series of theoretical as well as practical knowledge to improve the efficiency of sports. A big database can be achieved through the implication of data mining algorithms.

Data mining algorithms allow for in-depth performance analysis in sports. They enable coaches and athletes to dissect various aspects of performance, such as speed, agility, endurance, and accuracy. By analyzing data from wearable devices, video footage, and performance metrics, patterns and trends can be identified, leading to targeted improvements.

3.2. Application of data mining algorithm on physical education and sports teaching. The sports sectors and physical education are required to fulfill the qualification, informatization, and automation based on big data performance. As per Haoxiang & Smys [2], For sport and education data mining relies on a multimedia base of data of high value and importance. For instance, in physical education, the reliable data collected by the professor is very big and the authentic vital detail included in the data cannot be easily gotten as the noise of the students throughout the class. Further, the management of sports systems optimizes and obtains the resources of data. Besides this, the mining of data of different aspects influences the knowledge of the students and evaluates the overall result of sports teaching.

Evaluation of student's performance

Data mining tools refer to predictive modeling that has largely helped to predict the performance of students. To develop this many processes have been fulfilled including the regression and classification of different data. In addition, major classification has been used to estimate the performance of the students. SPSS, Rapid Miner, KEEL, and Orange are some tools that help to evaluate the performance.

Providing training and guiding the students

The data of the competitors can be evaluated by associating and putting rules with the features of the ball to identify the interconnection between the technical action and the players. It is a helpful tool for giving training and guiding through the physical process of physical Nationalon. The National Board Association (NBA) has recently implemented a data mining tool in conjunction with the recording of the image of the game of basketball.

Comparing the performance of various candidates

Different students' performance can be evaluated with various tools and the monitoring process can also be performed with the development of strategies to mitigate the challenges. According to Gunawan,[1], There have been various tools that help to develop a high insight into the performance of the students. Different situation views can be perceived with these tools that help in the future enhancement of the performance.

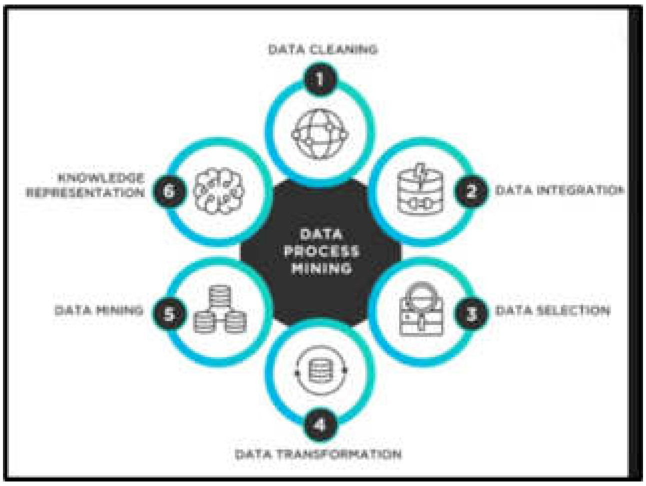


Fig. 3.2: Theory of data mining technology

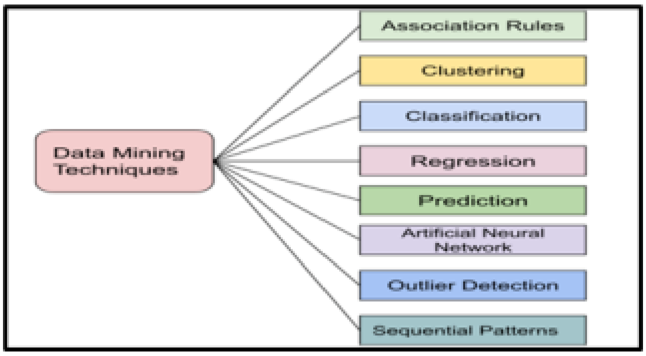


Fig. 4.1: Technologies of data mining

4. Techniques of Data Mining. There have been Various techniques have been used in data mining which involve correlation analysis; classification, clustering Outlier detection and the induction of the decision tree have been analyzed below:

Correlation analysis

Correlation analysis is an extension of the rules of the association, often the confidence and support parameters can be analyzed and the pattern can be detected. It is the method of statistical support to evaluate the linear linking between the variables and identify the relations. Sports and education teaching data mining helps to optimize the optimize.

Induction of decision tree

Decision-making is the method of learning the prospects of decision administration from the training sets. As stated by Lee, [4], a decision tree is used in the classification of data mining which creates a classification of regression and tree structure. The goal of the decision tree is to build a model that can function the databases effectively and help in splitting the data into distinctive classes. The scalability of the database are measured in the detection of inductive reasoning.

Outlier Detection

The local outlier detection in the algorithm is analyzed to compute the viability of the local density practices with respect to the substantial practices. An outlier detection has the significance of identifying the bad data and the good data that probably helps in meeting the central distribution. As per the critical analysis by

Maximov & Weslye, [5], sorting out the values administers progress towards checking interquartile takeaways depending on the respective data. The four major techniques are the Z-score, Numeric outlier, DBSCAN and the isolation forest respectively. Outlier detection has enabled the course of algorithmic actions to take place in prerequisites of Python detection.

Regression

Regression analysis is the concept of determining the course of actions corresponding to independent variables. Building a valid algorithm is a mandate that helps in functioning the training practices based on valid testing techniques. As opined by Ranjan & Avasthi, [7], after diving the data into testing and training, the model is learnt by following various samples that are relative to the linear regression object. In addition to that, plotting and visualizing the predicted data can contribute to accomplishing certain factors that are imported from both the simple and multiple variations respectively. As a result of this, the corresponding modules are associated with numeric values.

Clustering Analysis

Cluster analysis in data mining mainly refers to the group of objects that are associated with different grouping elements in data analytics. The clustering process is a multivariate data mining that is accustomed to measuring the user's attributes and characteristics in a significant manner. As per the critical analysis by Križanić, [3], identifying discrete and sales transactions, it is important to measure the resources that are available in an ample amount. Clustered analyses are applied in the form of image processing and pattern recognition. The strategic resources are put forward to configure the clustered methods that are related to machine learning.

Sequential pattern

Sequential patterns are the frequent patterns that are prosecuted to several input devices to clarify the courses of algorithm and data mining. Initially, the user needs to specify the parameter called minimum support threshold to acknowledge the application of subsequent networks. As mentioned by Pavithra, [6], an instance of the sequential pattern is that it helps in modifying the occurrence of networking practices thereby administering the successive outlets to deliver data effectively. The generalized sequential pattern is associated with learning possibilities to identify the pattern from a large set of databases.

Interconnection between the sports teaching and data mining algorithm

Physical training has been likely to improve with the algorithms of data mining. Scores of the student least in the database of the teaching has been reflected in the understanding of the situation (Zeng, et al. 2020). Further, the overall plans and strategies can be implemented to improve the advanced technologies. In addition, the record of each student's performance can be easily accessed with the algorithms of data mining.

5. Methodology. Secondary qualitative method has been follow in the study to collect detail about the data mining tools in the physical training and sport. The data has been collected through the articles, books and magazines. It is more reliable source to gather the data and enhance the more information about the reseaechr topic.

6. Result. According to Shakya, [9], different students' learning capacity can be identified and the tendencies to sustain in the long term have been evaluated. The data mining of education aims at building the techniques for explaining virus data kinds occurring in the context of education.

The above figure states that the sport and education sector has been highly using the data mining tools to increase the efficiency level.Data mining efficiently enhances the team performance by matching players to some particular situation. Furthermore, each player's contribution can be analyzed and the opposition's tendency in the match has been evaluated to exploit the weak point of the players. It involves different actions such as knowing the extraction and function at the ground of the match.

The above figure states that there are different correction strategies and error detection optimization processes in the area of physical education den sports teaching or training in which the algorithms of data mining can be used. The techniques of data have the potential to recognize the performance levels of the players (Lopes, et al. 2020). Application of the training sets comprises class labels and attributes that are implemented in the algorithm

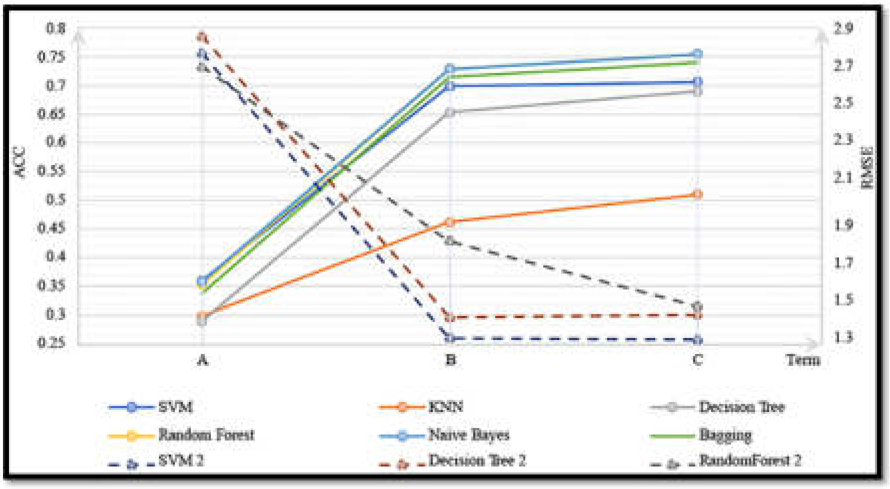


Fig. 6.1: Data driven mining impact on physical education

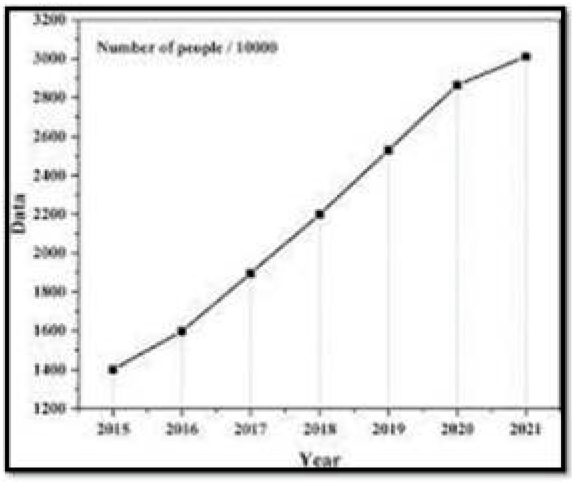


Fig. 6.2: Growth due to Data mining techniques in the educational systems

7. Conclusion. The overall research concludes that there is a positive impact of the data mining method on the sports and teaching sector. Furthermore, several algorithms have been used in this process to increase the potential impact. Nowadays, internet technology and computer technology have a higher influence on the physical fitness of students. Big data analysis, statistics, algorithms, and some advanced technologies are linked to developing the mode of information sports tactics and techniques in the education sector. With the help of tools of mining valuation detail can be extracted from large data such efficient and fast methods can assist the teacher as well as the students to fulfill the academic goals. The objective talent identification facilitated by these methods ensures that deserving individuals receive recognition and opportunities irrespective of biases or preconceptions. The implementation of data mining in strategic game planning elevates the competitiveness of teams, enhancing their chances of success. As the demand for sports analytics professionals grows, future research can focus on developing and enhancing educational programs in this field. This would help meet the increasing need for experts who can effectively leverage data for sports performance and strategy.

REFERENCES

[1] D. Gunawan, *Classification of privacy preserving data mining algorithms: a review*, Jurnal Elektronika dan Telekomunikasi, 20 (2020), pp. 36–46.

[2] W. Haoxiang, S. Smys, et al., *Big data analysis and perturbation using data mining algorithm*, Journal of Soft Computing Paradigm (JSCP), 3 (2021), pp. 19–28.

[3] S. Križanić, *Educational data mining using cluster analysis and decision tree technique: A case study*, International Journal of Engineering Business Management, 12 (2020), p. 1847979020908675.

[4] H. S. Lee and J. Lee, *Applying artificial intelligence in physical education and future perspectives*, Sustainability, 13 (2021), p. 351.

[5] I. I. Maximov, D. Alnæs, and L. T. Westlye, *Towards an optimised processing pipeline for diffusion magnetic resonance imaging data: Effects of artefact corrections on diffusion metrics and their age associations in uk biobank*, Human Brain Mapping, 40 (2019), pp. 4146–4162.

[6] M. Pavithra, A. Sindhana, T. Subajanaki, and S. Mahalakshmi, *Effective heart disease prediction systems using data mining techniques*, Annals of the Romanian Society for Cell Biology, (2021), pp. 6566–6571.

[7] R. Ranjan and D. Avasthi, *Enhanced edge detection technique in digital images using optimised fuzzy operation*, Webology, 19 (2022), pp. 5402–5416.

[8] H. Shahid, H. Ashraf, H. Javed, M. Humayun, N. Jhanjhi, and M. A. AlZain, *Energy optimised security against wormhole attack in iot-based wireless sensor networks*, Comput. Mater. Contin, 68 (2021), pp. 1967–81.

[9] S. Shakya, *A self monitoring and analyzing system for solar power station using iot and data mining algorithms*, Journal of Soft Computing Paradigm, 3 (2021), pp. 96–109.

Edited by: Sathishkumar V E
Special issue on: Scalability and Sustainability in Distributed Sensor Networks
Received: Aug 30, 2023
Accepted: Oct 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 1231–1240, DOI 10.12694/scpe.v24i4.2016

PREDICTIVE MAINTENANCE SYSTEM FOR ROTATING MACHINERY ONBOARD SHIPS FOR DETECTING PERFORMANCE DEGRADATION

BHARAT JAYASWAL,* SMITA AGRAWAL,† SWATI JAIN,‡ RAVI SINGH,§ KASHYAP,¶ AND PRAKRUT CHAUHAN‖

Abstract. Maintenance of rotating machinery is crucial for extending the lifespan and increasing the reliability of equipment onboard ships. Presently, breakdown and preventive methodologies are used for the maintenance of equipment. Further, dataloggers collect critical machinery parameters, and parameter data is used for real-time parameter monitoring. The availability of such extensive monitoring data has also led to the adoption predictive maintenance methodologies in the industry, wherein machine learning-based analysis of recorded data is used to predict impending defects and prompt required maintenance. In this paper, we propose a predictive maintenance system that records data through a network of sensors installed over multiple electrical motor pump sets onboard the ship and uses statistical analysis to detect equipment degradation. Our system has been deployed onboard a ship to undertake real-time predictive maintenance of electrical motor pump sets used in firemain, AC plants, stabilizers, steering pumps and other auxiliary engine room machinery.

Key words: Predictive maintenance system, vibration and current analysis, Naval ships, Unsupervised learning, Detecting performance degradation, Electrical motors

1. Introduction. The rotating machinery is the lifeline of ships, and specifically, electrical motor pump sets constitute a significant component of critical systems, namely, firemain, air conditioning, steering, stabilizers, propulsion, power generation (alternators) and various auxiliary pumps of circuits of fresh water, seawater and chilled water lines. The ship-board machinery is prone to defects due to exposure to sea climate, moisture and vibrations from the roll/pitch of the ship. Presently, Condition-based Monitoring (CBM) and Planned Preventive Maintenance (PPM) strategies are majorly adopted for the maintenance of equipment. However, these maintenance strategies are affected by more downtime (due to refit and maintenance periods) and are expensive (due to replacement of components) [22]. Predictive maintenance [13] (PdM) harnesses the potential of AI towards increasing the life span of equipment and enhances the reliability of machinery [14] PdM techniques analyze recorded sensor data of machinery to develop cognition for impending failures, and thereby, prompt required maintenance. Predictive maintenance within the shipping industry is in its early stages [5]. Recent research has been undertaken using supervised learning approaches over labelled data from simulated defects on machinery. Vibration analysis can detect abnormal vibration patterns, which may indicate potential machinery defects. PdM techniques using vibration analysis are helpful towards the identification of deviations from normal patterns and resolving issues before major breakdowns occur[18]. Study of vibration analysis [15] using Fast Fourier transform (FFT) has been undertaken by inducing defects in electrical motors. Vibration analysis of gearboxes [10, 2] of electrical motors has been undertaken using SVMs with experimental set-up and inducing defects using faulty bearings. Other than vibration analysis, PdM techniques may be implemented for ship-borne machinery using analysis of running current [1], thermography[4] and oil analysis [8]. In the absence of labelled (labels indicating operational and defective life), real-time datasets of a lifetime of motors and unsupervised learning approaches have also been attempted. Wescoat et al. describe vibration analysis of a paint dosing pump used for Condition Monitoring (CM) based maintenance using unsupervised learning [21]. The work majorly presents techniques to organize data using unsupervised approaches. In both supervised

*Center of Excellence for AI and Big Data, INS Valsura, Indian Navy, India (bharat.jayaswal85@navy.gov.in)
†CSE Department, Institute of Technology, Nirma University, India (smita.agrawal@nirmauni.ac.in)
‡CSE Department, Institute of Technology, Nirma University, India (swati.jain@nirmauni.ac.in)
§Center of Excellence for AI and Big Data, INS Valsura, Indian Navy, India (ravisingh@navy.gov.in)
¶CSE Department, Institute of Technology, Nirma University, India (16bit089@nirmauni.ac.in)
‖CSE Department, Institute of Technology, Nirma University, India (16bce044@nirmauni.ac.in)

and unsupervised approaches, vibration analysis [16, 12] has been found of key importance for undertaking predictive maintenance of rotating machinery. Olesen et al. [3] review state-of-the-art techniques for applying PdM in thermal power plants and pump systems. The review describes various challenges concerning applying PdM for rotating machinery viz unavailability of run-to-failure labelled data set and restriction of state of the art techniques mostly to vibration analysis. Our work utilizes analysis of current parameters in addition to vibration analysis.

During our study, we developed and deployed an AI-based predictive maintenance system onboard a ship to predict the degradation of electrical motor pump sets using a comparative approach. Our contributions are as follows:

- **Data Collection and Exploratory Data Analysis (EDA)** – Dataloggers were installed over one ship to collect data on 16 motors of different ship-borne systems. EDA techniques were used to get insights into patterns of machinery usage and understand the data for further application towards predictive maintenance.
- **Pre-Processing Techniques** – Real-time data received from dataloggers installed on electrical motors may have erroneous values, view accidental grounding of sensor equipment, malfunctioning of sensors or power fluctuations. Our work proposes pre-processing techniques to learn machinery exploitation patterns and segregate erroneous readings. We have used a decision tree-based classifier [7] in our model for filtering erroneous readings.
- **Prediction of Performance Degradation** – To undertake a regression-based analysis of Remaining Useful Life (RUL) [23], we require lifetime data of a motor, i.e. from the installation of a new motor till such time, the motor gets faulty. However, data collection for such rule-based analysis requires recording the parameters of motors for over two years. Our approach circumvents the need to record the parameters of an electrical motor for its lifetime, as we present a methodology for statistical and comparative analysis of motors of similar types, which may be in different stages of their lifetime. Our model determines the Gaussian distribution of data, compares the Gaussian distribution of similar machines and uses empirical rules to predict the degradation of machinery.

Lastly, we demonstrate the efficacy of our system on real-world deployment onboard Naval Ships. We propose future work that can be undertaken to enhance our solution.

2. Application Overview. The application architecture is shown in Fig. 2.1. The system consists mainly of four modules: real-time data collection, pre-processing, data analytics and user interface. During the development of the application, firstly, we used a data collection module to collate data. After that, we used EDA techniques to understand the data. Based on the results of EDA, we developed pre-processing and comparative analytics modules (used for degradation prediction). Subsequently, we deployed the complete application and user interface onboard the ship for real-time predictive maintenance of electrical motor pump sets. In this paper, we briefly explain the data collection module, followed by the results of our EDA techniques. Further, we explain techniques used for pre-processing and prediction of degradation.

3. Data Collection Module. This module comprises a network of sensors (voltage, vibration and current) installed over electrical motor pump sets, a common data bus, and a central polling computer. This rugged system takes data feed from 16 machineries at an interval of one minute. Data were collated for three weeks to design and develop other system modules. The data snapshot is shown in Fig. 3.1 and Table 3.1. The recorded parameters consist of voltage and current in three power supply phases to the machinery and RMS values of vibration (mm/Sec) taken at one-minute intervals. Since information about the state of the machinery, i.e. Degraded/ Operational was not available, our system has been designed to work on unlabelled datasets.

4. Exploratory Data Analysis (EDA). The techniques used for EDA and the results obtained are explained in subsequent paragraphs.

4.1. Analysis using DBSCAN. Density-Based Spatial Clustering Application with Noise (DBSCAN) [17] algorithm and parameters' timeline charts were used over the dataset to undertake EDA. The results of DBSCAN clustering over PCA [11] transformed dataset for one motor are shown in Fig. 4.1 and explained as follows:

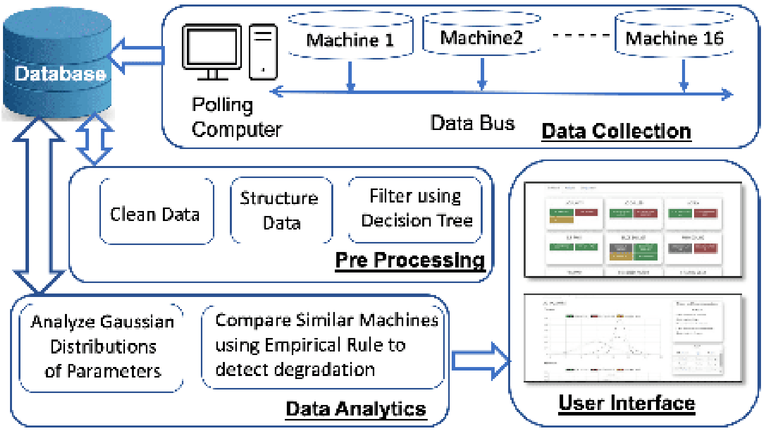


Fig. 2.1: Architecture of Predictive Maintenance System

Table 3.1: Summary of total Data of 16 Machineries

No. of Motors	**No. of days**	**No of Samples per motor**	**Total No of Samples**	**Remarks**
16	21	31449	5,03,184	One sample per minute 24X7 for 21 days

- The red and blue clusters indicate data points considered normal values by the DBSCAN algorithm, and black dots correspond to anomalies detected by DBSCAN.
- The presence of two clusters was further analyzed in the dataset using timeline plots of current and vibration readings, as shown in Fig. 4.2. Further, the colours (red and blue) of vibration values and current in Fig. 4.2 were mapped with colours of data points in Fig. 4.1 for appreciation of affiche acy of DBSCAN clustering. It was found that zero-valued current readings were grouped in blue clusters, and data points with non-zero values were largely clustered in red clusters. As the motor does not draw current in switched OFF state, it was observed that the red cluster consisted of data points in the ON state of machinery, and the blue cluster indicated the OFF state of machinery.
- Further, it was also observed that anomalies observed by the DBSCAN algorithm were about exceptionally high current values during state transition from OFF state to ON state. It is essential to mention that the datalogger recorded current values at a sampling rate of one datapoint per minute, and therefore, exceptionally high current values were starting current values captured during state transition. It was inferred that though the DBSCAN algorithm detects these points as anomalies, these data points were not anomalous as starting current is consistently higher than the running current during regular operation of the motors. It was also inferred that DBSCAN could not detect graceful degradation, as during a graceful degradation, values will iteratively increase slowly to higher values; therefore, all these density-reachable values would be clustered together.

4.2. Analysis of the running status of the motor. The running status of motors within the group of similar machinery was also studied using histogram analysis. The points with zero current values were labelled OFF state, and non-zero current values were labelled ON state. The histograms of current values concerning two motors are shown in Fig 4.3. This analysis was later used to identify the instances of readings when the current value was zero and filter out readings, as done in Section 5.3.

5. Pre-Processing Techniques. Since the data was recorded from the real-time running of the motors onboard one ship, multiple factors contributed to erroneous values in the data. For example, during the time

Fig. 3.1: Snapshot of Data from one machinery

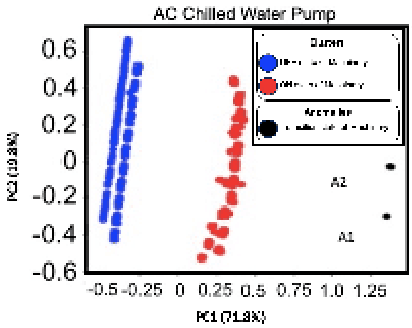


Fig. 4.1: Results of DBSCAN on data from AC sea water pump -1 - (The plot has principal eigenvectors on the X and Y axis, and therefore, no units are indicated)

series analysis of the readings, it was seen that current values were dropping to zero for three to four consecutive readings due to the accidental grounding of the sensor. Also, there were Direct Current (DC) offsets observed in the readings due to power supply fluctuations [19]. Therefore, the following pre-processing techniques were employed in the AI-based predictive maintenance system (as indicated earlier in fig. 2.1).

5.1. Cleaning of Data. The real-time data generated from dataloggers required data cleaning, including timestamps formatting and removing data files, wherein all readings, viz. current, voltage and vibration, were zero-valued.

5.2. Structuring of Data. The data points were provided with a machine identification number for each machinery. After that, time series data for each machinery was analyzed and observed to be a series of ON-OFF states. It was found that current values remained zero consecutively in the time series when the machinery was OFF and remained non-zero consecutively when running. Therefore, data about each machinery was structured as running instances by scanning the current values in increasing order of time series. The data points about OFF instances of machinery were discarded for further processing.

5.3. Filtering of erroneous readings. During the visualization of timeline charts of vibration and current data for each running instance, it was observed that a few readings needed to be revised and, therefore, required to be discarded. We show an example of a timeline plot of vibration readings of one running instance of one machine in Fig. 5.1. We observed in this example that the vibration readings coming from Sensors were rising intermittently and, at times, continually dropping to Zero. This behaviour may be due to some instantaneous sensor grounding or intermittent contact. We categorized such readings as bad readings. Further, we have shown an example of good reading from another motor in Fig. 5.2. In this example, it can be seen that the vibration occasionally drops to zero, and also, there are DC offsets in readings a few times. However, overall the reading is good compared to readings provided in Fig. 5.1. Therefore, we have applied pre-processing techniques in our work, wherein data points of each running instance were passed through a decision tree-based classifier to accept good readings and reject bad readings. A total of 360 readings were labelled as good and bad readings based on non-linearity and experience of domain experts. The decision tree provided 94%

6. Prediction of Degradation. Our algorithm for the degradation prediction is based on a comparative analysis of Gaussian distributions of current and vibration readings of similar types of motors. This module consists of the following two parts:

- Determining the univariate Gaussian distribution of current as well as vibration.
- Comparing the univariate distribution of current and vibration of similar types of motors based on Empirical rule and providing machine status labels as follows:

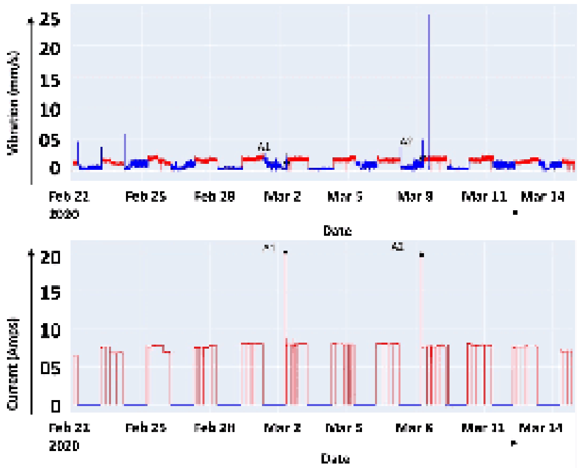


Fig. 4.2: Timeline of vibration and current values of AC chilled water Pump - 1

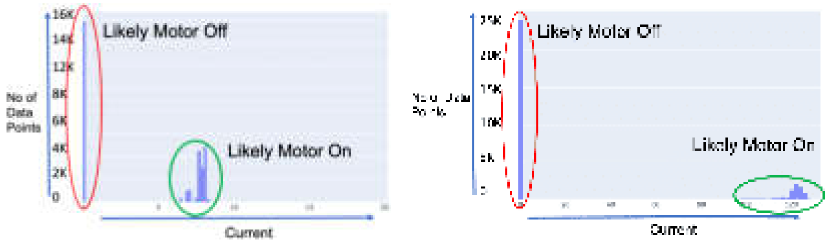


Fig. 4.3: Histogram of current readings – AC Sea Water Pump 1 & AC Plant 2

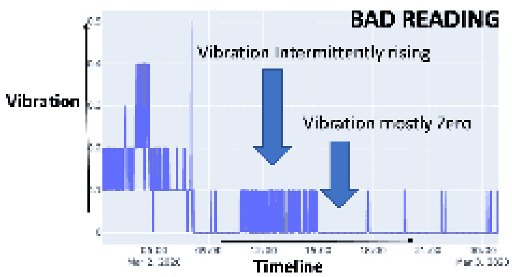


Fig. 5.1: Example of Bad readings from Vibration Sensor

- **Satisfactory, Green :** If the Gaussian distribution of vibration and current values of machinery is similar to that of machinery in the same group.
- **Degradation Likely, Amber :** If the Gaussian distribution of vibration and machinery's current values is higher than that of machinery in the same group.
- **Degradation Imminent, Red :** If the Gaussian distribution of vibration and machinery's current values is very high compared to the Gaussian distribution of machinery in the same group.

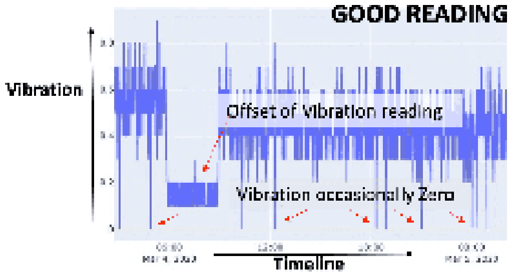


Fig. 5.2: Example of Good readings from Vibration Sensor

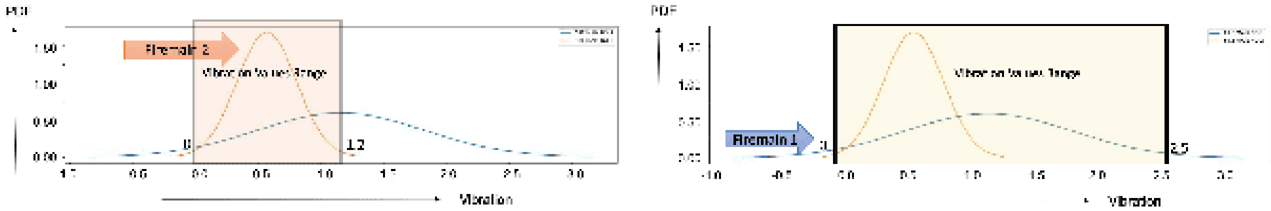


Fig. 6.1: Gaussian distribution on vibration data of Firemain 1 and Firemain 2

6.1. Examples of functioning of the application. The algorithm's functioning is shown with examples of firemain motors, AC motors and stabilizer motors.

- **Firemain Motors:**There were two Fireman motors in the dataset, i.e., Firemain 1 and 2. Fig. 6.1 shows the probabilistic distribution of vibration readings of both motors, with the Y axis showing the Probability Density Function (PDF). Our algorithm learns that 99.7% of the vibration values from Firemain motor two are spread between 0 and 1.2. Similarly, the algorithm learns that 99.7% of vibration values from Firemain motor 1 are spread between 0 and 2.5. Ideally, the vibration values should be the same for the same type of motor. Our algorithm undertakes a similar analysis of the probabilistic distribution of current readings of both motors. Distributions of current readings are shown in Fig. 6.2. It was observed that the current readings of Firemain 1 were higher than Firemain 2.
- **AC Plants:**The dataset consisted of three AC plants, namely AC plants 1, 2 and 3. The vibration and current analysis of AC plants are shown in Fig. 6.3. It was inferred that the vibration readings of all AC plants were similar, but the current readings of AC plant-1 and AC plant-2 were higher than AC plant-3.
- **Stabiliser Motors:** Our dataset contained data of two stabilizer motors, i.e. port and stabilizer. The vibration and current analysis of stabilizer motors is shown in Fig. 6.4. It was observed that the vibration readings of both stabilizer motors were similar, but the current readings of the port stabilizer were higher than Stbd Stabiliser.

6.2. Evaluation of the Results. In our study, we focused on predicting the performance degradation of rotating machinery using unsupervised techniques. Since labelled data for performance degradation of rotating machinery onboard ships were unavailable, employing traditional measured model quality metrics was challenging. We adopted a practical approach to overcome this limitation, validating our model based on real-world examples and gathering feedback from onboard users. It allowed us to assess the model's performance effectively in real operating conditions. Additionally, we conducted a comparative analysis, contrasting our PdM model with established CBM algorithms. The insights from this comparison offer a comprehensive evaluation of our model's effectiveness and practical utility for predicting performance degradation in rotating machinery.

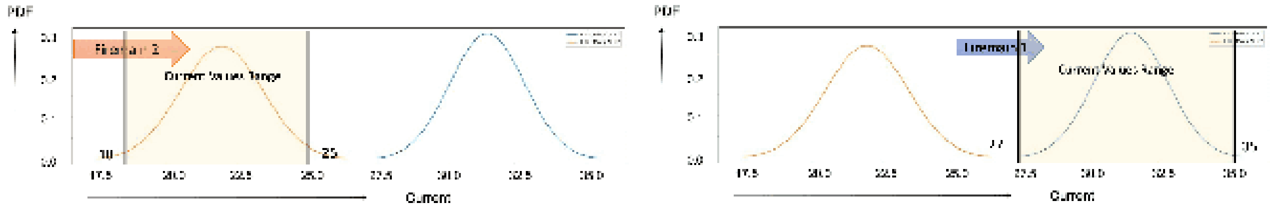


Fig. 6.2: Gaussian distribution on current data of Firemain 1 and Firemain 2

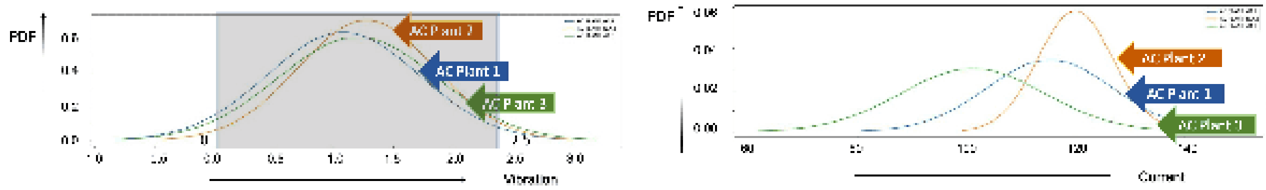


Fig. 6.3: Gaussian distribution on vibration and current data of AC Plants

We explain our evaluation with three examples:

- **Firemain :** For example, vibration as well as current values of both Firemain 1 and Firemain 2 were within threshold values as per CBM and therefore, both machineries were working satisfactorily. However, our PdM model compared parameter values of similar machinery, and even though the values were within the threshold, our model observed that parameter readings of Firemain 1 were trending to higher values. Therefore, our model indicated an impending degradation in the state of machinery well in time to prompt maintenance/ overhauling. Based on the results of our model, Firemain 1 landed at the repair yard, and the health of the motor was assessed. It was found that the motor bearings were degraded, and post-greasing of the motor and replacement of bearings, vibration and current readings were similar to that of Firemain 2.
- **AC Plants :** Similarly, values of current drawn by all three AC Plants were within the threshold as per CBM, and the machinery was working satisfactorily. However, our PdM techniques observed that the current drawn by AC Plants 1 and 2 is higher than that of AC Plant 3 during operation. A higher current indicates either higher torque or higher load. In order to rule out the possibility of a higher load, all three AC Plants were disconnected from Ship's AC system and all three AC compressors were tested in full load condition. The current values of AC Plant 1 and 2 in full load condition were also higher than that of AC Plant 3, indicating likely degradation of machinery. Therefore, all three Plants were recommended to be exploited with careful monitoring for further degradation.
- **Stabiliser Motors :** It was observed that values of vibration drawn by both Stabiliser Motors were within the threshold as per CBM, and the machinery was working satisfactorily. However, our PdM techniques observed that the current drawn by the port stabilizer motor was higher than that of the stbd stabilizer during operation. Further, it was found that the port stabilizer was operating at a higher torque than the stbd stabilizer during the ship's sailing, owing to the inherent stability of the ship. The same led to higher current usage by the port stabilizer motor. Therefore, the port stabilizer motor was recommended to be exploited with careful monitoring for further degradation.

7. User Interface Design. We have developed an intuitive Graphical User Interface (GUI) for data visualization and prompting the required maintenance based on a decision matrix. The GUI has been developed using Python, Django framework and chart JS. The screenshots of the application are shown in Fig. 7.1.

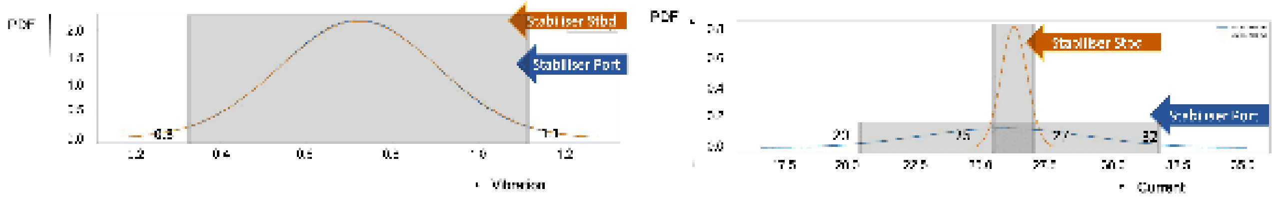


Fig. 6.4: Gaussian distribution on vibration and current data of stabiliser

Table 7.1: Recommendation Matrix

Current → / Vibration ⇓	Red	Amber	Green
Red	• Check winding insulation. • Check input supply. • Undertake routines of contactors of starter panel. • Check connections of cables and lugs. • Carry out alignment checks of motor and load. • Check vibrations and condition of bearings. • Check fan and motor shaft for any wear and tear.	• Check winding insulation. • Carry out alignment checks of motor and load. • Check vibrations and condition of bearing. • Check fan and motor shaft for any wear and tear. • Check screws of motor foundation.	• Carry out alignment checks of motor and load. • Check vibrations and condition of bearings. • Check fan and motor shaft for wear & tear. • Check screws of motor foundation.
Amber	• Check winding insulation. • Check input supply. • Undertake routines of contactors of starter panel. • Check connections of cables and lugs. • Carry out greasing of bearings. • Check fan and motor shaft for any wear & tear. • Check screws of motor foundation.	• Check starting and running current of motor. • Check vibration readings of driving and non-driving end. • Carry out greasing of bearings. • Check screws of motor foundation.	• Check vibration readings of driving and non-driving end. • Carry out greasing. • Check screws of motor foundation.
Green	• Check winding insulation. • Check input supply. • Undertake routines of contactors of starter panel. • Check connections of the cable and lugs. • Check vibration of driving & non-driving end of motor.	• Check starting and running current of motor. • Check connections of the cable and lugs. • Check screws of motor foundation.	• All Parameters are within threshold.

7.1. Explanation of Results to User. Vollert et al. [20] presents a review of publications addressing PdM problems with a focus on model interpretability. Since explainability is one of the major challenges in the PdM model due to the black-box nature of algorithms. Our model explains results to the end users through charts and decision matrices which is crucial to explain to the user how our model arrived at a specific decision. It ensures that the end users understand and appropriately trust the model's recommendations so that the AI-based Predictive Maintenance System can be deployed confidently. Therefore, we devised a matrix for the explanation of the results of our algorithm to the user. Our model explains the observations of our algorithm and prompts required maintenance for the machinery as per the matrix shown in Table 7.1. The screenshots of the application are shown in Fig. 7.2.

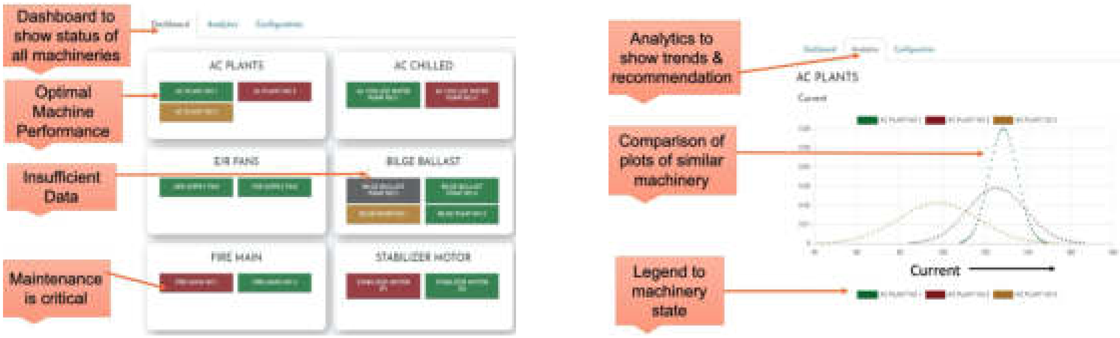


Fig. 7.1: Deployment of the System onboard ship

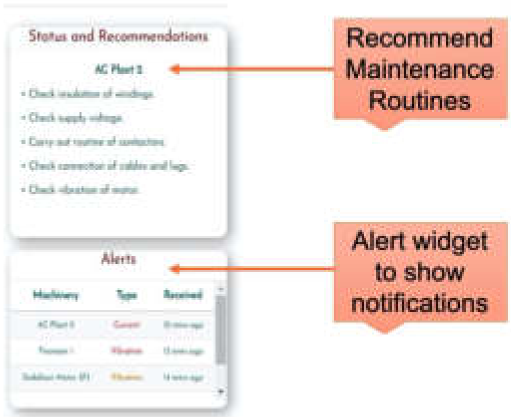


Fig. 7.2: Recommendations to the user

8. Deployment of the Application. The application was successfully deployed onboard a ship, and data feeds from dataloggers were integrated into the application to provide a real-time utility for predictive maintenance to users.

9. Conclusion. We have developed a technique to process sensor data from various electrical motor pump sets fitted onboard ships, recorded by a network of sensors. Our application involves pre-processing techniques for cleaning, structuring and filtering data. EDA has allowed us to understand the dataset for its application for predictive maintenance. Further, we incorporated machine learning techniques and empirical rules to perform comparative analytics of data of similar machines [6, 9]. A complete application along with GUI was developed and deployed onboard a ship by combining pre-processing techniques, EDA, and comparative analysis modules.

10. Future Work. Our application may be enhanced by integrating knowledge of the design parameters of machinery for comparing the design and observed parameters. Further, the application can facilitate the collection of more maintenance data by providing functionalities to record details of maintenance routines, downtime of equipment, spares consumed during maintenance and similar data. The availability of detailed maintenance data would facilitate using other AI and ML algorithms for predictive maintenance.

Acknowledgments. This work was carried out at the Center of Excellence for AI and Big Data, INS Valsura, Indian Navy, Under the MoU with Nirma University and Indian Navy.

REFERENCES

[1] J. Antonino-Daviu, *Electrical monitoring under transient conditions: A new paradigm in electric motors predictive maintenance*, Applied Sciences, 10 (2020).

[2] E. P. Carden and P. Fanning, *Vibration based condition monitoring: a review*, Structural health kmonitoring, 3 (2004), pp. 355–377.

[3] J. Fausing Olesen and H. R. Shaker, *Predictive maintenance for pump systems and thermal power plants: State-of-the-art review, trends and challenges*, Sensors, 20 (2020).

[4] A. N. Huda and S. Taib, *Application of infrared thermography for predictive/preventive maintenance of thermal defect in electrical equipment*, Applied Thermal Engineering, 61 (2013), pp. 220–227.

[5] V. J. Jimenez, N. Bouhmala, and A. H. Gausdal, *Developing a predictive maintenance model for vessel machinery*, Journal of Ocean Engineering and Science, (2020).

[6] R. Kakkar, J. Alzubi, A. Dua, S. Agrawal, S. Tanwar, R. Agrawal, G. Sharma, P. N. Bokoro, and R. Sharma, *Padaav: blockchain-based parking price prediction scheme for sustainable traffic management*, IEEE Access, 10 (2022), pp. 50125–50136.

[7] S. Kaparthi and D. Bumblauskas, *Designing predictive maintenance systems using decision tree-based machine learning techniques*, International Journal of Quality & Reliability Management, (2020).

[8] S. Keartland and T. L. Van Zyl, *Automating predictive maintenance using oil analysis and machine learning*, in 2020 International SAUPEC/RobMech/PRASA Conference, 2020, pp. 1–6.

[9] R. Kothari, R. Kakkar, S. Agrawal, P. Oza, S. Tanwar, B. Jayaswal, R. Sharma, G. Sharma, and P. N. Bokoro, *Selection of best machine learning model to predict delay in passenger airlines*, IEEE Access, (2023), pp. 1–1.

[10] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, *Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis*, Neurocomputing, 168 (2015), pp. 119–127.

[11] G. Li and Y. Hu, *Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis*, Energy and Buildings, 173 (2018), pp. 502–515.

[12] K. I. Masani, P. Oza, and S. Agrawal, *Predictive maintenance and monitoring of industrial machine using machine learning*, Scalable Computing: Practice and Experience, 20 (2019), pp. 663–668.

[13] G. Palem, *Condition-based maintenance using sensor arrays and telematics*, arXiv preprint arXiv:1309.1921, (2013).

[14] V. Pandya, S. Agrawal, S. Jain, and B. Jayaswal, *Early fault detection for rotating machinery onboard ships motor using fuzzy logic and k-means*, in Proceedings of Fourth International Conference on Computing, Communications, and Cyber-Security, S. Tanwar, S. T. Wierzchon, P. K. Singh, M. Ganzha, and G. Epiphaniou, eds., Singapore, 2023, Springer Nature Singapore, pp. 567–580.

[15] S. Patil and J. Gaikwad, *Vibration analysis of electrical rotating machines using fft: A method of predictive maintenance*, in 2013 fourth international conference on computing, communications and networking technologies (ICCCNT), IEEE, 2013, pp. 1–6.

[16] C. Scheffer and P. Girdhar, *Practical machinery vibration analysis and predictive maintenance*, Elsevier, 2004.

[17] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, *Dbscan revisited, revisited: Why and how you should (still) use dbscan*, ACM Trans. Database Syst., 42 (2017).

[18] M. Trivedi, R. Kakkar, R. Gupta, S. Agrawal, S. Tanwar, V.-C. Niculescu, M. S. Raboaca, F. Alqahtani, A. Saad, and A. Tolba, *Blockchain and deep learning-based fault detection framework for electric vehicles*, Mathematics, 10 (2022), p. 3626.

[19] J. Vico and R. Hunt, *Protection principles for electrical motors in the cement industry*, in 2010 IEEE-IAS/PCA 52nd Cement Industry Technical Conference, 2010, pp. 1–13.

[20] S. Vollert, M. Atzmueller, and A. Theissler, *Interpretable machine learning: A brief survey from the predictive maintenance perspective*, in 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2021, pp. 01–08.

[21] E. Wescoat, M. Krugh, A. Henderson, J. Goodnough, and L. Mears, *Vibration analysis utilizing unsupervised learning*, Procedia Manufacturing, 34 (2019), pp. 876–884.

[22] P. Zhang, Y. Du, T. G. Habetler, and B. Lu, *A survey of condition monitoring and protection methods for medium-voltage induction motors*, IEEE Transactions on Industry Applications, 47 (2011), pp. 34–46.

[23] Z. Zhang, X. Si, C. Hu, and Y. Lei, *Degradation data analysis and remaining useful life estimation: A review on wiener-process-based methods*, European Journal of Operational Research, 271 (2018), pp. 775–796.

Edited by: Katarzyna Wasielewska-Michniewska
Research paper
Received: Jun 15, 2023
Accepted: Sep 26, 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_ε 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.