
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 6, pp. 5624–5637, DOI 10.12694/scpe.v25i6.3345

HYBRIDIZATION OF MACHINE LEARNING MODEL WITH BEE COLONY BASED FEATURE SELECTION FOR MEDICAL DATA CLASSIFICATION

R. RAJA* AND B. ASHOK†

Abstract. Nowadays, an important count of biomedical data is created continuously in several biomedical equipment and experiments because of quick technical enhancements in biomedical science. The study of clinical and health data is vital to enhance the analysis precision, prevention, and treatment. Initial analysis and treatment are extremely important approaches for preventing deaths in many diseases. Accordingly, the data mining and machine learning (ML) approaches are helpful tools for utilizing minimization error and for providing helpful data for analysis. But the data obtained in digital machines takes higher dimensionality, and not every data attained in digital machines is significant to specific diseases. This article develops an artificial bee colony-based feature selection with optimal hybrid ML model for medical data classification (ABCFS-OHML) technique. The presented ABCFS-OHML technique mainly aims to identify and classify the presence of disease using medical data. To attain this, the presented ABCFS-OHML technique initially pre-processes the input data in two ways namely null value removal and data transformation. Furthermore, the presented ABCFS-OHML technique uses ABCFS model for the choice of effectual subset of features. At last, root means square propagation with convolutional neural network-Hop field neural network (CNN-HFNN) model for classification purposes. The usage of RMSProp optimizer assists in attaining optimal hyperparameter selection of the CNN-HFNN method. The performance validation of the ABCFS-OHML technique takes place using three medical datasets. The comparison study reported that the ABCFS-OHML technique has accurately classified the medical data over other recent approaches.

Key words: Medical data classification; Machine learning; Deep learning; Feature selection; Hyperparameter tuning

1. Introduction. Healthcare sector generates enormous volume of data and Data Science methods act as a supporting factor for extracting hidden knowledge. It allows innovations and opportunities for enhancing health of people by addressing distinct perspectives firstly descriptive, to identify what happened [4]; diagnostic, to detect the cause why it happened predictive, to analyze what will occur and prescriptive, to find how we can make it happen [18]. Data analytics technology renders more effectual apparatuses that aid to present advanced treatment of chronic disease through early detection, home care, accurate medicine, population health and advanced treatment of communicable diseases, and lifestyle support [2, 16]. For last two decades, the authors have modelled many innovative ML approaches for predictive data analysis. Such useful methods were enforced in several data-intensive research zones namely biology, astronomy to mine hidden patterns, and healthcare [12].

ML grants a huge opportunity in this context firstly assisting medical practitioners, physicians, and geneticists to enhance the analysis of large medical data [3], secondly minimizing the health error risk, and lastly enhancing prognostic and diagnostic procedure harmonization. The authors have used ML for enhancing healthcare techniques by learning bio-medical data [19]. ML is a new and intellectual technique that automatically aids to study of particular issues and rises efficiency without explicitly programming. It could automatically find patterns in data and take decisions with minimum human input. In recent years, the expansion of many ML techniques like clustering, and categorization of data, disease prediction had an important effect on the decision-making procedure [20]. Classification can be referred to as a supervised technique of learning in real-time difficulties. It constitutes a method that precisely estimates the targeted class from data collected at numerous classification stages [10]. Feature selection (FS) techniques can be typically enforced to enhance the performance of the method. It minimizes the computing cost through elimination of irrelevant features. Therefore, this makes the diagnosis procedure very comprehensible and accurate [22].

*Department of Computer and Information Science, Annamalai University, Annamalai Nagar - 608 002, Tamil Nadu, India (rajamanira2000@gmail.com)

†Department of Computer Science, PSPT MGR Govt. Arts and Science College, Sirkali, Tamil Nadu, India (ashok.au@gmail.com)

This article develops an artificial bee colony-based FS with optimal hybrid ML model for medical data classification (ABCFS-OHML) technique. The presented ABCFS-OHML technique initially pre-processes the input data in two ways namely null value removal and data transformation. Furthermore, the presented ABCFS-OHML technique uses ABCFS model for the choice of effectual subset of features. At last, root means square propagation with convolutional neural network-Hop field neural network (CNN-HFNN) model for classification purposes. The experimental validation of the ABCFS-OHML approach was executed utilizing three medical datasets.

2. Related works. Bhukya and Manchala [6] introduce a recent metaheuristic rough set-based FS with rule-oriented medical data classification (MRSFS-RMDC) model on MapReduce architecture. The projected method develops a butterfly optimization technique for minimum rough set selection. Furthermore, Hadoop MapReduce was employed for processing large amounts of information. In addition, a rule-oriented classification model called repeated incremental pruning for error reduction (RIPPER) was utilized by adding a set of conditional rules. Sun et al. [21] developed an AFS-DF for COVID19 categorization related to chest CT images. Next, for capturing the higher-level representation of this feature with the comparatively small-scale dataset, the author leverages deep forest models for learning higher level representations of the feature. Furthermore, the author proposed an FS model related to the trained deep forest mechanism for reducing the feature redundancy, whereby the FC has incorporated adaptively with the COVID19 classification method.

Chen et al. [8] designed a confidence-based and cost-effective FS (CCFS) model based on BPSO for improving the efficiency of healthcare data. Especially, CCFS enhances search efficacy by designing a novel updating model which develops feature confidence for considering the fine-grained effect of all the dimensions in the particles on the classifier accuracy. The author in [11], developed a novel algorithm called ensemble embedded FS (EEFS) for handling multilabel bioinformatics data learning problems in an efficient and effective manner. The EEFS doesn't explicitly discover the correlations amongst labels, however, it could sufficiently make use of the label correlation through multilabel classifier and evaluation measure. Moreover, it reduces the accumulated error of information itself by using an ensemble model. Chen et al. [7] suggested confidence related and cost-effective FS model based on binary PSO, CCFS. Firstly, CCFS enhances search efficacy by designing a novel updating model, where confidence of all the features is considered which includes the correlation among categories and features, and historically selected frequency of every feature.

Nagarajan et al. [17] formulate a hybrid GA-ABC that denotes a genetic related ABC method for classification and feature-selection by utilizing classifier ensemble methods. The ensemble classifier has 4 techniques namely DT, SVM, NB, and RF. Karlekar and Gomathi [14] introduce a technique for healthcare data classification utilizing a new ontology and whale optimization-oriented SVM (OW-SVM) method. Primarily, privacy-preserved data can be formulated by implementing Kronecker product bat method, and then, ontology can be constructed for selecting features. The OW-SVM was then modelled through integration of ontology and whale optimization method into SVM where ontology and whale optimization method has been employed for selecting the kernel parameters feasibly.

3. The Proposed Model. In this article, a new ABCFS-OHML method was formulated for medical data classification. At the preliminary stage, the presented ABCFS-OHML method initially pre-processes the input data in two ways namely null value removal and data transformation. Moreover, the presented ABCFS-OHML technique uses ABCFS model for the choice of effectual subset of features. Finally, the RMSProp optimizer with CNN-HFNN model for classification purposes. Fig. 3.1 defines the overall process of ABCFS-OHML system.

3.1. FS using ABC Algorithm. Once the medical data is pre-processed, the next level is to choose an optimal feature subset. ABC is the optimization technique stimulated by honeybee nectar gathering behaviors and proposed on the basis of random population [9]. Due to its simplicity, global search capability, and efficiency robustness, it is demonstrated to be nature-inspired algorithm for managing constrained and unconstrained multi-or-single-objective global optimization issues. In this work, each bee was classified into scouts, employed bees, and onlookers. The three types of honeybee use nectar sources via separation of labour and cooperation and continually upgrade the position of nectar source by sharing and marking for detecting the optimum nectar source. The position of nectar sources matches with possible solution to optimization problems, and quality of nectar sources are evaluated by the fitness values of optimization issue. The steps for ABC algorithms are

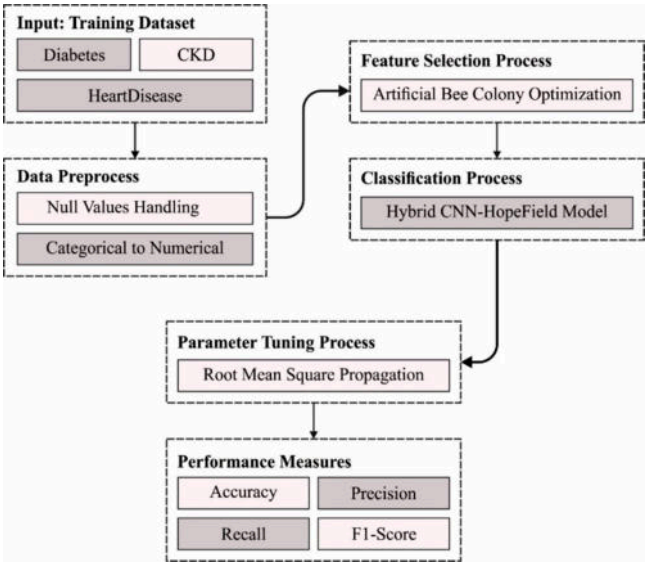

Fig. 3.1: Overall process of ABCFS-OHML system

given below.

1. Initialized phase: Set the number of maximum iterations, bee colonies, optimized range, and dimensionality.
   In this study, ABC technique was utilized to enhance duty cycle D, the optimized range was set to (0.85, 1), the primary amount of the onlooker and employed bees are set as 30 and 50 correspondingly, dimension was set as 1, and the maximal amount of iterations was set as 10.
2. *Employed bee phase:* They search for a novel food source nearby the existing food source.
   To generate candidate food location from the older one, the novel solution was related to the present solution, and fitness can be evaluated based on the subsequent equation as follows:

$$V_{ij}=X_{ij}+\phi_{ij}\,(X_{ij}-X_{kj}) \tag{3.1}$$

   Now, y_{ij} represents the novel position of food sources; X_{ij} indicates the existing food resource; $i\in\{$
3. *Onlooker bee stage:* The onlooker chooses food resources afterward sharing data of employed bees and define the amount of nectar. The best solution is chosen based on the probability that is evaluated as follows:

$$p_i=\frac{f_i}{\sum_{i=1}^{s_N} f_i} \tag{3.2}$$

From the expression, f_i indicates the fitness function (FF) values of j^{th} solution.

4. *Scout bee stage:* In all the iterations, the scout bee monitors the variations of all the solutions in swarm. Once the food source could not be upgraded by means of predefined cycle, it is detached from population, and employed bees of food resource turns out be scouts and utilize subsequent formula for finding a novel random food source position:

$$X_{ij}=X_{\min\ j}+rand\,[0,1]\,(X_{\max\ j}-X_{\min\ j}) \tag{3.3}$$

The FF leveraged in this presented technique was modelled to maintain a balance between the number of selected features in all solutions (minimal) and the classifier accuracy (maximal) achieved with the use of these selected features, Eq. (3.4) signifies the FF for evaluating solutions.

$$Fitness=\alpha\gamma_R\,(D)+\beta\frac{|R|}{|C|} \tag{3.4}$$

whereas $\gamma_R(D)$ indicates the classifier error rate of a given classifier was utilized here). $|R|$ represents the cardinality of the subset which is selected and $|C|$ refers to the total number of attributes in the datasets, β andα were 2 parameters respective to the significance of subset length and classification quality. $\in$ [1, 0] and β= 1−α.

3.2. Data Classification using CNN-HFNN Model. To detect and classify medical data, the CNN-HFNN model is exploited in this study. CNN has robust representation learning abilities by extracting and learning features automatically from inputs. CNN methods were commonly made up of fully connected (FC), convolutional layers, and pooling layers in classification applications. In a chain-oriented DNN, the FC layers have many parameters belongs to the network, which influences the computational complexity and memory occupancy. For several real time issues, accelerating inference period becomes a significant matter due to the hardware design implications. To handle this issue, the replacement of the FC layers in addition to Hopfield neural networks (HNNs) was proposed [15]. This presented structure will combine an HNN and a CNN: A pretrained CNN technique was employed for extracting features, subsequently, an HNN, which will be assumed as an associative memory that stores every feature constituted by the CNNs. The HNN architecture has interconnected powerful features and neurons of content addressable memory that are crucial for solving numerous optimization and combinatorial tasks. The HNN technique involves organized neurons. In bipolar detection, the neuron from discrete HNN is employed; 1 is implemented to represent the true state, and the falsification can be determined as 1. The fundamental analysis of neuron state activation from HNN is characterized as follows.

$$S_i = \begin{cases} 1, & if \quad \sum_j W_{ij} S_i > \psi, \\ -1, & Otherwise \end{cases} \tag{3.5}$$

whereas W_{ij} denotes the synaptic weighted vector of HNN-RANkSAT derived from *j-th* to *i-th* neuons. S_i indicates the state of *i-th* neurons from HNN, and ψ represents the existing values. The value ψ= 0 is to verify that the network energy is reduced to 0. The synaptic weighted connection from discrete HNN has no connection with itself, and the synaptic connected in one neuron to others is 0 (W_{iii}=W_{jjj}=W_{kkk} and W_{ii}=W_{jj}=W_{kk}). Consequently, HNN has symmetrical features regarding structure. The HNN method is similar and intricate fact to Ising technique of magnetism. In bipolar expression, the neuron state is termed as spin point executes the magnetic field trajectory. Each neuron is compelled for flipping still it achieves a stable equilibrium state based on the following equation.

$$S_i \rightarrow sgn \ [h_i(t)] \tag{3.6}$$

The local field vector connects all the neurons from HNN is defined by h_i. The sum of field is caused by each neuron state in the following:

$$h_i = \sum_{k}^{N} \sum_{j}^{N} W_{ijk} S_j S_k + \sum_{j}^{N} W_{ij} S_j + W_i \tag{3.7}$$

The task of local field is to evaluate the final state of neuron and generate each possible 3-SAT-induced logic accomplished in the final state of neuron. The predominant feature of HNN network is the detail that it converges continuously as follows [1]:

$$E_{F_{RANkAT}} = \sum_{i=1}^{NN} \prod_{j=1}^{V} T_{ijk} \tag{3.8}$$

in which V and NN imply the number of variables and neurons created from $F_{RANkSAT}$ correspondingly. The inconsistency of $F_{RANkSAT}$ demonstration as:

$$T_{ij} = \begin{cases} \frac{1}{2}(1-S_\rho), \ if \ \neg\rho \\ \frac{1}{2}(1+S_\rho) \ , \ otherwise \end{cases} \tag{3.9}$$

The value $F_{RANkSAT}$ is proportionate to value of inconsistency in the logical clause as follows:

$$S_i\,(t{+}1) = \begin{cases} 1, & h_i = \sum_K^N \sum_J^N W_{ijk} S_j S_k + \sum_J^N W_{ij} S_j + W_i \geq 0 \\ -1, & h_i = \sum_K^N \sum_J^N W_{ijk} S_j S_k + \sum_J^N W_{ij} S_j + W_i < 0 \end{cases} \tag{3.10}$$

Eq. (3.10) describes the Lyapunov energy function from the HNN.

$$H_{F_{RINNAT}} = -\frac{1}{3} \sum_{i=1,i\neq j,j\neq kj=1}^{N} \sum_{i\neq j,j\neq ik=1}^{N} \sum_{i\neq j,k\neq i}^{N} W_{ijk} S_i S_j S_k - \frac{1}{2} \sum_{i=1,i\neq jj=1}^{N} \sum_{i\neq j}^{N} W_{ij} S_i S_j - \sum_{i=1}^{N} W_i S_i \tag{3.11}$$

Eq. (3.11) is used to classify when the solution acquires global/local minimal energy. HNN makes the optimal allocation when the induced neuron state acquires global minimum energy. Restricted analyses are incorporated with HNN and ACO as a single computation network. Consequently, the robustness of ACO improves the trained process from HNN as follows:

$$\left| H_{F_{RANKSAT}} - H_{F_{RANKSAT}}^{\min} \right| \leq \xi \tag{3.12}$$

In Eq. (3.12), ξ indicates a tolerance value. The value $\xi = 0.001$. When the $F_{RANkSAT}$ logical representation embedding from HNN does not fulfill the condition, afterward that the neurons are surrounded in the incorrect pattern from the final state.

3.3. Hyperparameter Tuning Employing RMSProp Technique. To adjust the hyperparameters of the CNN-HFNN technique, the RMSProp optimizer is used. The RMSprop optimizer restricts the oscillation in the vertical direction [5]. Thus, learning rate can be increased and the presented model could take large step in the horizontal direction which converges fast. The RMSprop calculation is demonstrated as follows. The momentum value is represented as beta and is generally fixed as 0.9.

$$vdw = \beta \cdot vdw + (1-\beta) \cdot dw^2$$

$$vdb = \beta \cdot vdw + ^{(1-\beta)\cdot db^2}$$

$$W = W - \alpha \cdot \frac{dw}{\sqrt{vdw} + \varepsilon}$$

$$b = b - \alpha \cdot \frac{db}{\sqrt{vdb} + \varepsilon}$$

In backpropagation model, dW and db are used for updating W and b parameters:

$$W = W - learning\ rate * dW$$

$$b = b - learning\ rate * db$$

In RMSprop, before utilizing dW and db individually for all the epochs, exponentially weighted average of square of dW and db has been considered:

$$S_{dW} = \beta^* S_{dW} + (1-\beta)^* dW^2$$

$$S_{db} = \beta^* S_{db} + (1-\beta)^* db^2$$

whereas β beta is additional hyperparameter and takes value from zero to one. The newly weighted average is made by weights, average of prior and present value square. Afterward computing exponential weighted average, the parameter has been updated.

$$W = W - learning\ rate * dW/sqrt(S)$$

$$b = b - learning\ rate * db/sqrt(S)$$

S_{dW} is quite lesser such that are splitting it by dW. While S_{db} is quite larger such that splitting db with relatively large value reduces the update on vertical dimension.

Table 4.1: Dataset description

Description	CKD	Diabetes	Heart Disease
Number of Instances	400	768	270
Number of Attributes	24	8	13
Number of Class	2	2	2
Number of Positive Samples	250	268	120
Number of Negative Samples	150	500	150
Data source	[23]	[13]	[24]

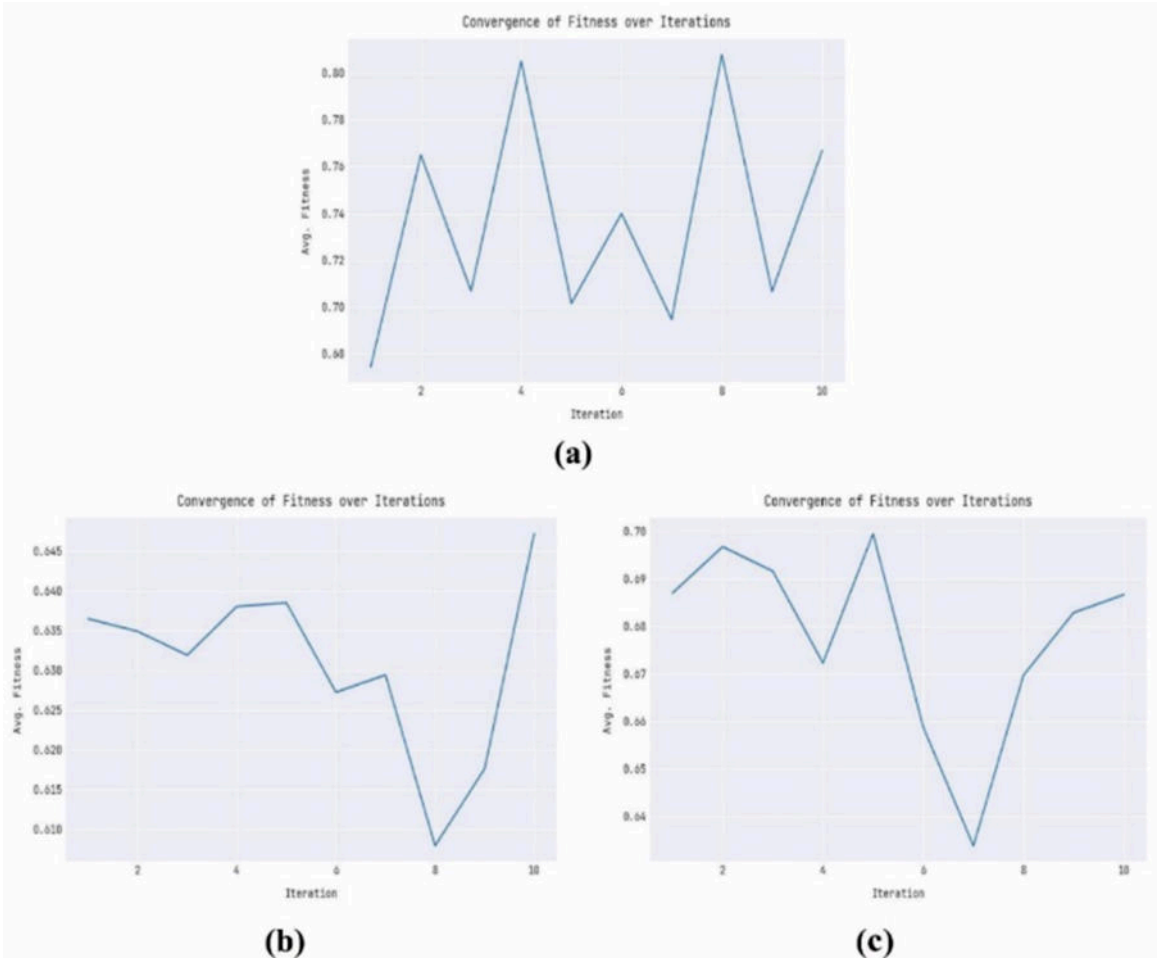

Fig. 4.1: Datasets (a) CKD (b) Diabetes (c) Heart disease

4. Results and Discussion. The experimental validation of the ABCFS-OHML method is tested using three medical datasets namely CKD, Diabetes, and HD. Table 4.1 represents the detailed description of three medical datasets.

Fig. 4.1 shows the convergence study of the ABCFS-OHML model on the applied datasets. On the CKD dataset, the ABCFS technique has chosen the following features: sg, al, su, rbc, pcc, sc, sod, pot, pcv, rbcc, htn, appet, pe, and ane. Besides, on diabetes dataset, the ABCFS technique has elected preg, plas, mass, pedi, and age features. Finally, on HD dataset, the chosen features are sex, chest, resting_blood_pressure,

Table 4.2: Result analysis of ABCFS-OHML technique with various measures under three datasets

Measures	**CKD Dataset**	**Diabetes Dataset**	**Heart Disease Dataset**
Accuracy	99.00	96.74	98.15
Precision	99.20	94.18	97.52
Recall	99.20	96.64	98.33
F1-Score	99.20	95.40	97.93

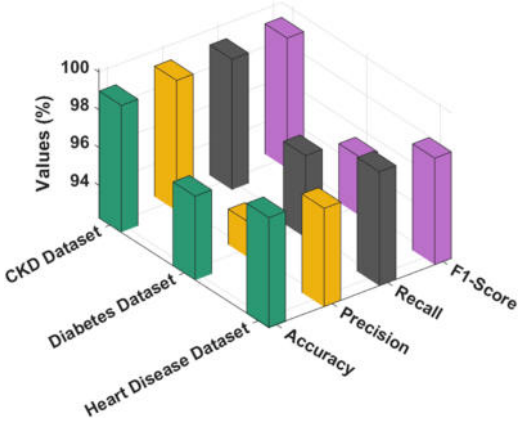

Fig. 4.2: Result analysis of ABCFS-OHML technique under three datasets

maximum_heart_rate_achieved, exercise_induced_angina, oldpeak, slope, and thal.

Table 4.2 and Fig. 4.2 report the outcomes offered by the ABCFS-OHML technique. The outcomes inferred the ABCFS-OHML technique has gained effectual performance on every dataset. For instance, on CKD dataset, the ABCFS-OHML technique has offered $accu_y$ of 99%, $reca_l$ of 99.20%, $prec_n$ of 99.20%, and $F1_{score}$ of 99.20%. Meanwhile, on diabetes dataset, the ABCFS-OHML technique has rendered $accu_y$ of 96.74%, $reca_l$ of 94.18%, $prec_n$ of 96.64%, and $F1_{score}$ of 95.40%. Eventually, on heart disease dataset, the ABCFS-OHML approach presented $accu_y$ of 98.15%, $reca_l$ of 97.52%, $prec_n$ of 98.33%, and $F1_{score}$ of 97.93%.

Fig. 4.3 establishes the classifier results of the ABCFS-OHML approach under CKD dataset. Fig. 4.3 a shows the confusion matrix presented by the ABCFS-OHML method. The figure highlighted the ABCFS-OHML technique has identified 148 instances under notckd and 248 instances under ckd. Also, Fig. 4.3 b illustrates the precision-recall study of the ABCFS-OHML technique. The figures stated the ABCFS-OHML approach has gained maximal precision-recall performance in every class. At last, Fig. 4.3 c exemplifies the ROC study of the ABCFS-OHML method. The figure depicted the ABCFS-OHML approach has resulted in proficient results with higher ROC values in every different class label.

Fig. 4.4 portrays the classifier results of the ABCFS-OHML method under diabetes dataset. Fig. 4.4 a represents the confusion matrix provided by the ABCFS-OHML technique. The figure displayed the ABCFS-OHML approach has identified 484 instances under notckd and 259 instances under ckd. Likewise, Fig. 4.4 b illustrates the precision-recall analysis of the ABCFS-OHML algorithm. The figures stated the ABCFS-OHML technique has gained maximal precision-recall performance in every class. Finally, Fig. 4.4 c displays the ROC study of the ABCFS-OHML algorithm. The figure represented the ABCFS-OHML methodology has resulted in proficient outcomes with maximal ROC values in different class labels.

Fig. 4.5 exhibits the classifier results of the ABCFS-OHML approach under heart disease dataset. Fig. 4.5 a depicts the confusion matrix presented by the ABCFS-OHML method. The figure stated the ABCFS-OHML method has identified 148 instances under notckd and 248 instances under ckd. Also, Fig. 4.5 b establishes

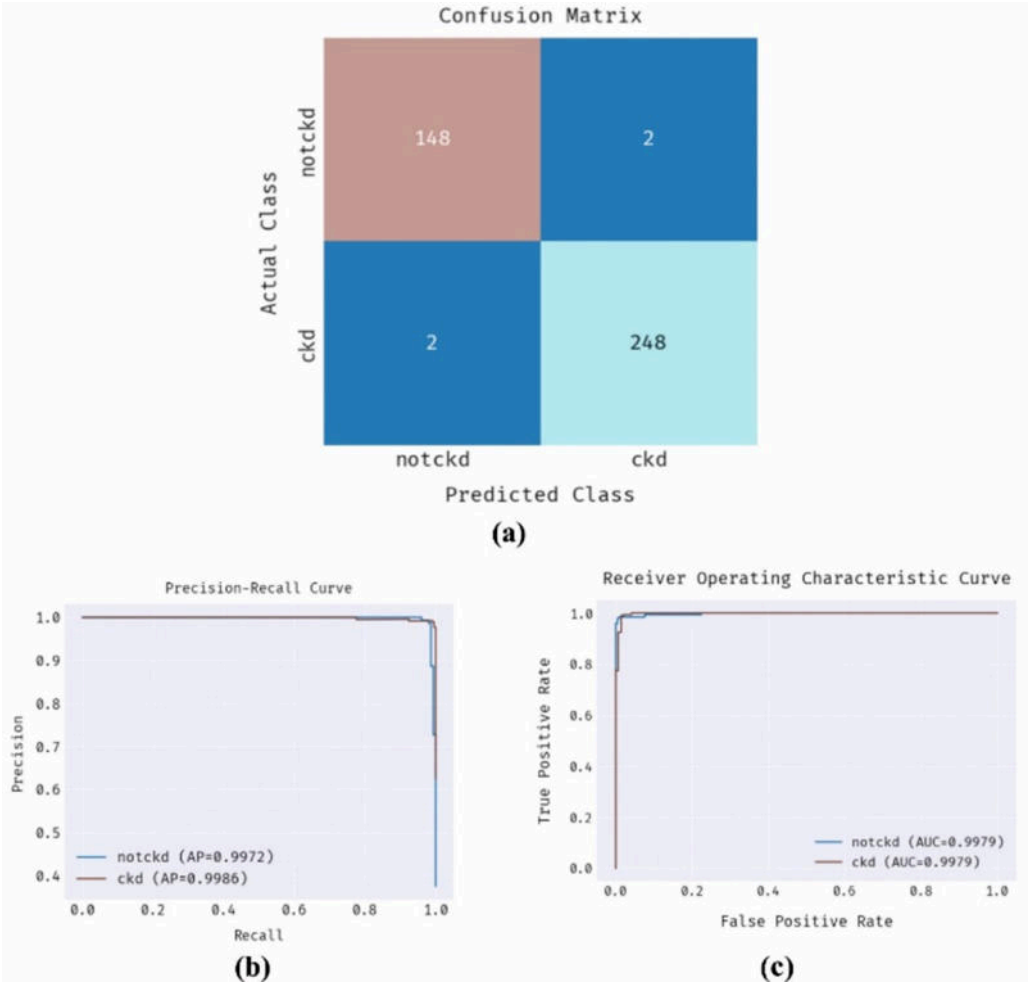

Fig. 4.3: CKD dataset (a) Confusion Matrix (b-c) PR and ROC curves

the precision-recall investigation of the ABCFS-OHML approach. The figures highlighted the ABCFS-OHML technique has reached maximum precision-recall performance under all classes. Lastly, Fig. 4.5 c exhibits the ROC study of the ABCFS-OHML method. The figure exhibited the ABCFS-OHML approach has resulted in proficient outcomes with maximal ROC values in different class labels.

Fig. 4.6 delivers the accuracy and loss graph analysis of the ABCFS-OHML method in three datasets. The outcomes exhibited accuracy value seems to be higher and loss value seems to lower with an increase in epoch count. It is noted that the training loss is low and validation accuracy is high on the test three datasets.

For reassuring the improved performance of the ABCFS-OHML method, the detailed comparative review of CKD dataset is given in Table 4.3 and Fig. 4.7. The outcomes represented the OlexGA model has resulted in least $accu_y$ of 75%. Then, the XGBoost and LR models have resulted in slightly improvised $accu_y$ of 83% and 82% whereas the PSO algorithm has reached even improved $accu_y$ of 95%. Although the DT and ACO models have accomplished reasonable $accu_y$ of 90% and 87.50%, the ABCFS-OHML model has shown maximum $accu_y$ of 99%.

For reassuring the enhanced performance of the ABCFS-OHML technique, a brief comparative study on Diabetes dataset is given in Table 4.4 and Fig. 4.8. The outcomes denoted the Voted Perceptron approach has resulted in least $accu_y$ of 66.79%. Then, the DT algorithm resulted in slightly improvised $accu_y$ of 73.82%

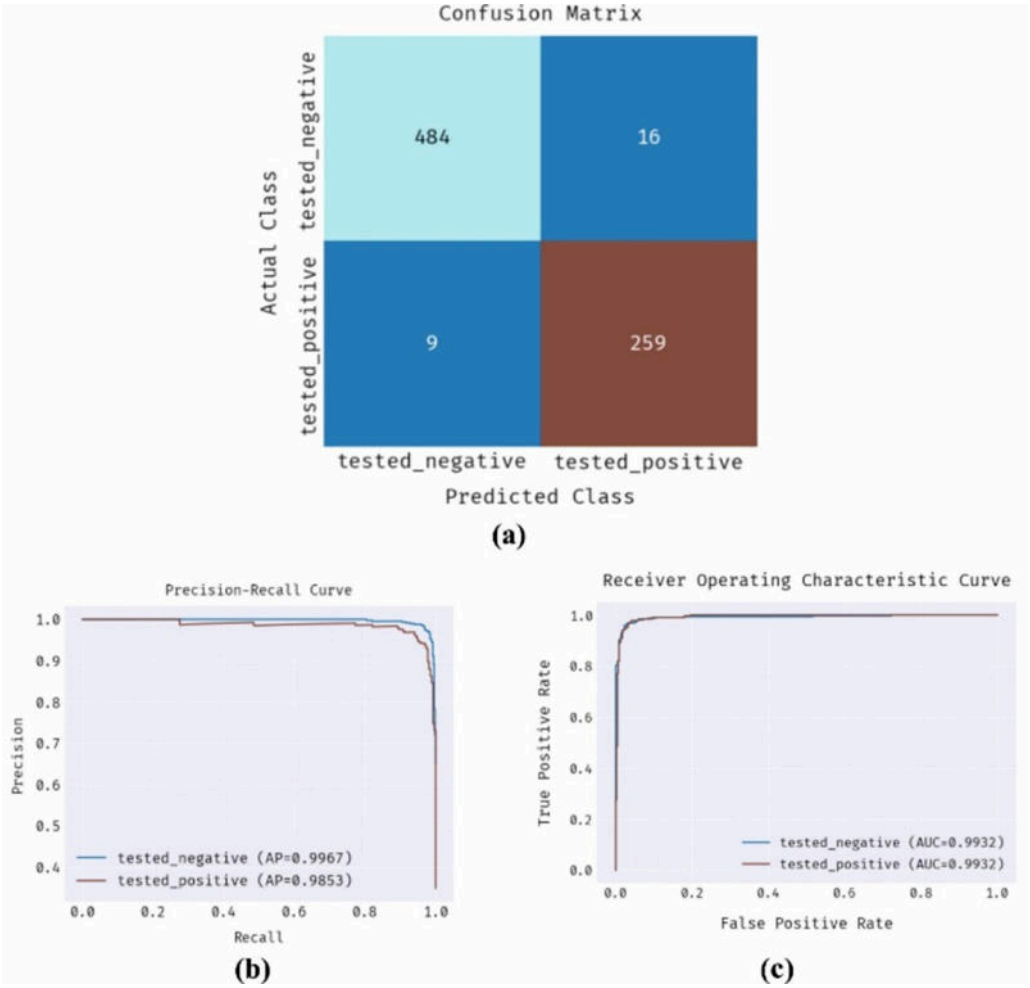

Fig. 4.4: Diabetes dataset (a) Confusion Matrix (b-c) PR and ROC curves

Table 4.3: ***Accu$_y$*** analysis of ABCFS-OHML technique with existing approaches under CKD dataset

Methods	**Accuracy**
ABCFS-OHML	99.00
Decision Tree	90.00
ACO	87.50
PSO	85.00
XGBoost	83.00
Logistic Regression	82.00
OlexGA	75.00

whereas the LogitBoost algorithm has reached even improved $accu_y$ of 74.08%. Though the LR and GBT techniques have established reasonable $accu_y$ of 77.21% and 88.67%, the ABCFS-OHML method has shown maximum $accu_y$ of 96.74%.

For reassuring the enhanced performance of the ABCFS-OHML technique, a detailed comparative study on heart disease dataset is given in Table 4.5 and Fig. 4.9. The results showed the RT algorithm has resulted

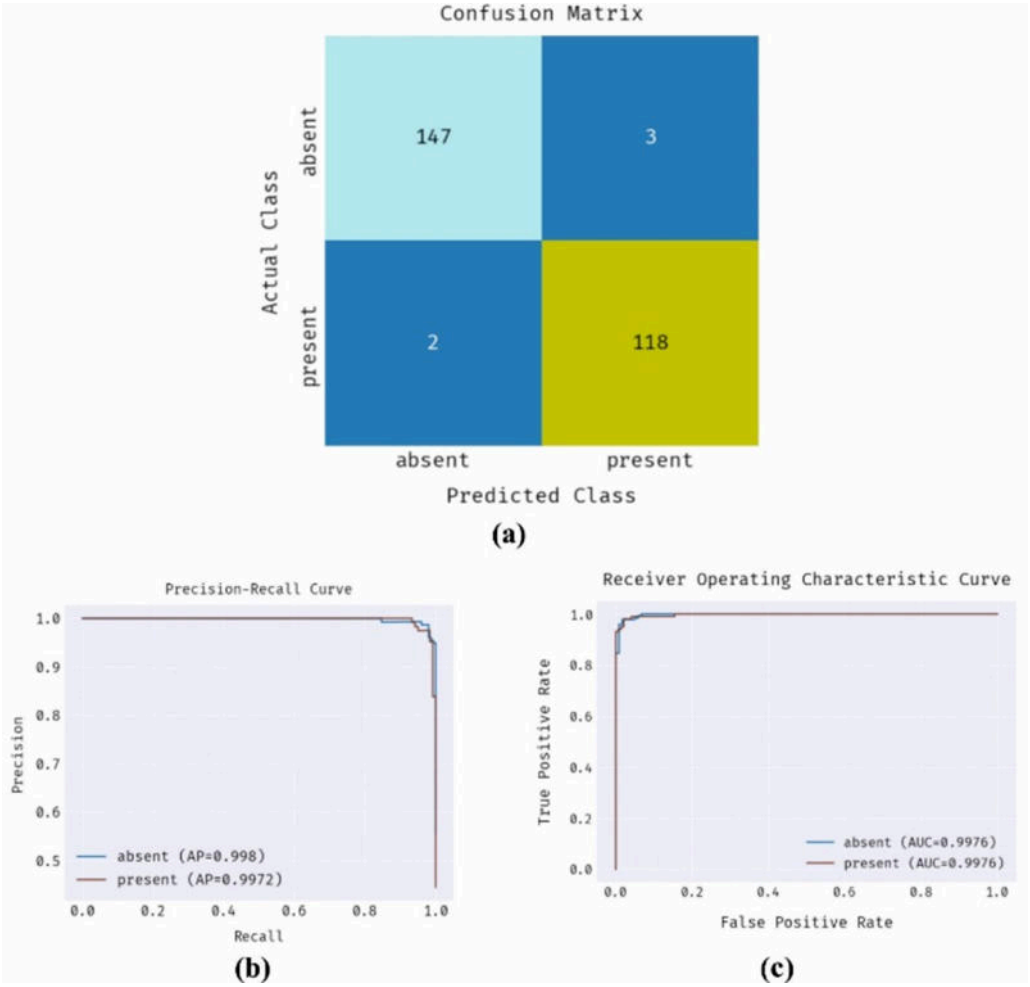

Fig. 4.5: Heart Disease dataset (a) Confusion Matrix (b-c) PR and ROC curves

Table 4.4: $\boldsymbol{Accu_y}$ analysis of ABCFS-OHML technique with existing

Methods	**Accuracy**
ABCFS-OHML	96.74
GBT	88.67
LR	77.21
Voted Perceptron	66.79
LogitBoost	74.08
DT	73.82

in least $accu_y$ of 76.29%. Then, the J48 approach resulted to slightly improvise $accu_y$ of 76.66% whereas the NBTree method has reached even improved $accu_y$ of 80.37%. Although the RF and RBFNetwork techniques have exhibited reasonable $accu_y$ of 81.85% and 84.07%, the ABCFS-OHML approach has displayed maximum $accu_y$ of 98.15%.

These results concluded that the ABCFS-OHML model can effectually classify medical data.

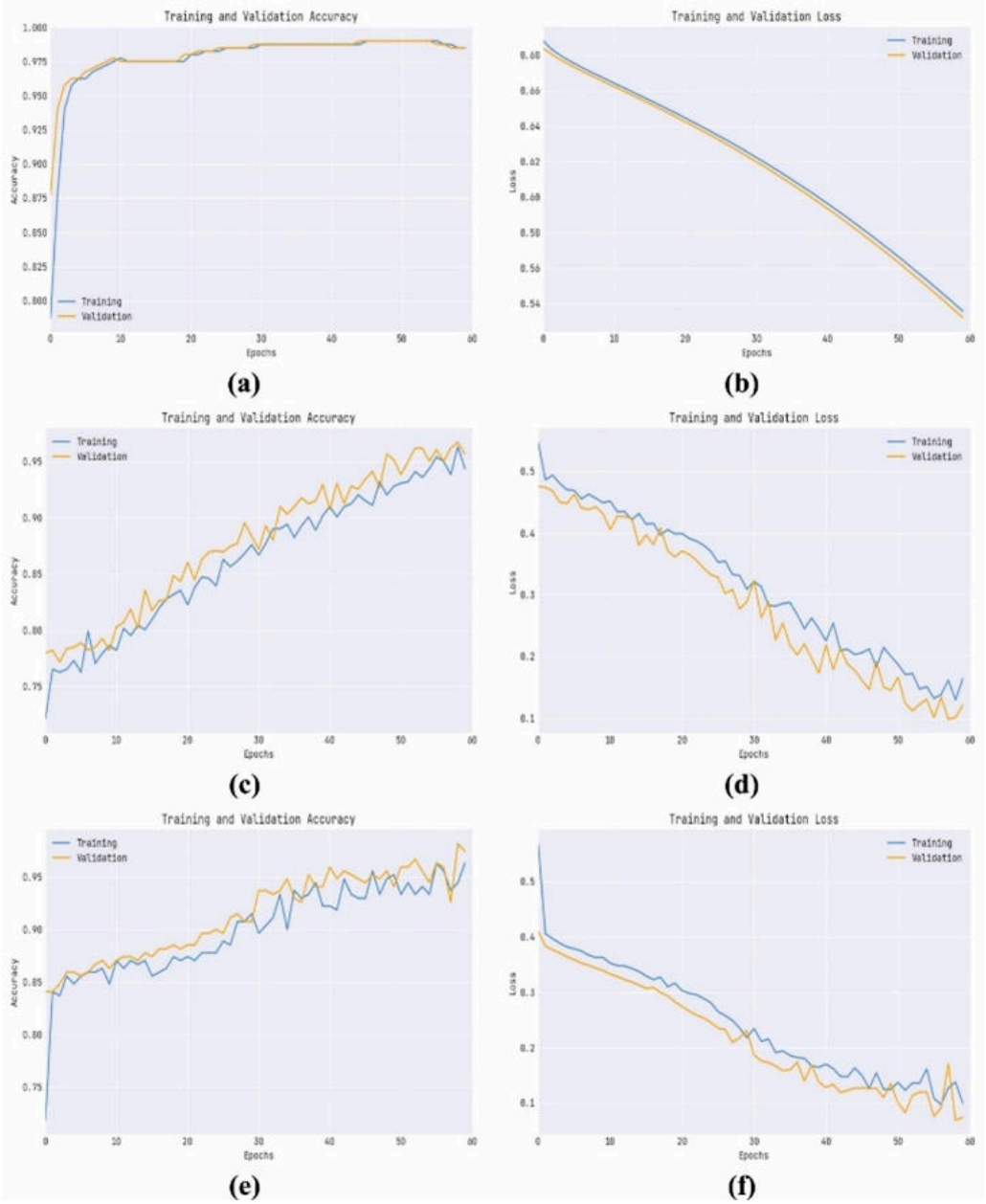

Fig. 4.6: Accuracy and loss analysis datasets (a and b) CKD (c and d) Diabetes (e and f) Heart disease

5. Conclusion. In this article, a new ABCFS-OHML approach was devised for medical data classification. At the preliminary stage, the presented ABCFS-OHML approach initially pre-processes the input data in two ways namely null value removal and data transformation. Additionally, the presented ABCFS-OHML technique uses ABCFS model for the choice of effectual subset of features. Finally, the RMSProp optimizer with CNN-HFNN model for classification purposes. The usage of RMSProp optimizer assists in attaining optimal hyperparameter selection of the CNN-HFNN method. The performance validation of the ABCFS-OHML technique takes place using three medical datasets. The comparison study reported that the ABCFS-OHML technique has accurately classified the medical data over other recent approaches. Thus, the presented

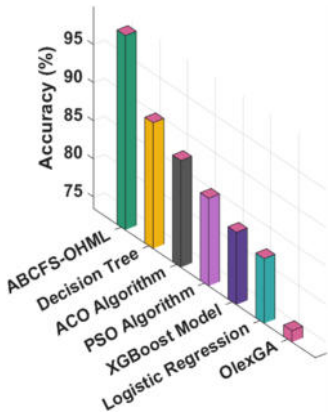

Fig. 4.7: $\boldsymbol{Accu_y}$ analysis of ABCFS-OHML technique under CKD dataset

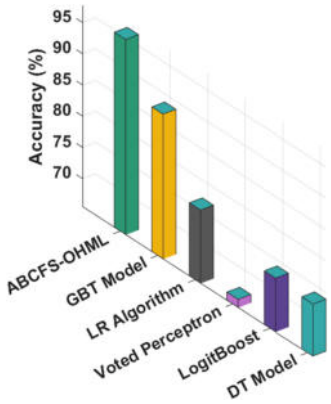

Fig. 4.8: $\boldsymbol{Accu_y}$ analysis of ABCFS-OHML technique under Diabetes dataset

Table 4.5: $\boldsymbol{Accu_y}$ analysis of ABCFS-OHML technique with existing approaches under heart disease dataset

Methods	**Accuracy**
ABCFS-OHML	98.15
J48	76.66
Random Tree	76.29
RBFNetwork	84.07
NBTree	80.37
Random Forest	81.85

ABCFS-OHML technique can be employed for accurate medical data classification. The limitations of the ABCFS-OHML technique includes the restricted scalability for handling extremely large datasets and potential overfitting. In future, outlier removal approaches will be employed to boost the classification performance of the ABCFS-OHML technique.

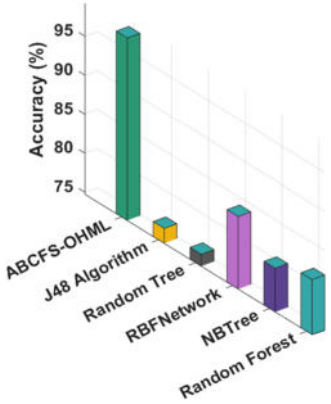

Fig. 4.9: $\boldsymbol{Accu_y}$ analysis of ABCFS-OHML technique under heart disease dataset

REFERENCES

[1] H. Abubakar, A. Muhammad, and S. Bello, *Ants colony optimization algorithm in the hopfield neural network for agricultural soil fertility reverse analysis*, Iraqi Journal For Computer Science and Mathematics, 3 (2022), pp. 32–42.

[2] F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, *A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion*, Information Fusion, 63 (2020), pp. 208–222.

[3] M. Ali and T. Aittokallio, *Machine learning and feature selection for drug response prediction in precision oncology applications*, Biophysical reviews, 11 (2019), pp. 31–39.

[4] S. Amaouche, A. Guezzaz, S. Benkirane, M. Azrour, S. B. A. Khattak, H. Farman, and M. M. Nasralla, *Fscb-ids: Feature selection and minority class balancing for attacks detection in vanets*, Applied sciences, 13 (2023), p. 7488.

[5] D. V. Babu, C. Karthikeyan, A. Kumar, et al., *Performance analysis of cost and accuracy for whale swarm and rmsprop optimizer*, in IOP Conference Series: Materials Science and Engineering, vol. 993, IOP Publishing, 2020, p. 012080.

[6] H. Bhukya and S. Manchala, *Design of metaheuristic rough set-based feature selection and rule-based medical data classification model on mapreduce framework*, Journal of Intelligent Systems, 31 (2022), pp. 1002–1013.

[7] Y. Chen, Y. Wang, L. Cao, and Q. Jin, *An effective feature selection scheme for healthcare data classification using binary particle swarm optimization*, in 2018 9th international conference on information technology in medicine and education (ITME), IEEE, 2018, pp. 703–707.

[8] ———, *Ccfs: a confidence-based cost-effective feature selection scheme for healthcare data classification*, IEEE/ACM transactions on computational biology and bioinformatics, 18 (2019), pp. 902–911.

[9] L. Fan and X. Ma, *Maximum power point tracking of pemfc based on hybrid artificial bee colony algorithm with fuzzy control*, Scientific Reports, 12 (2022), p. 4316.

[10] C. B. Gokulnath and S. Shantharajah, *An optimized feature selection based on genetic approach and support vector machine for heart disease*, Cluster Computing, 22 (2019), pp. 14777–14787.

[11] Y. Guo, F.-L. Chung, G. Li, and L. Zhang, *Multi-label bioinformatics data classification with ensemble embedded feature selection*, IEEE access, 7 (2019), pp. 103863–103875.

[12] D. Jain and V. Singh, *Feature selection and classification systems for chronic disease prediction: A review*, Egyptian Informatics Journal, 19 (2018), pp. 179–189.

[13] Kaggle, *Pima indians diabetes database*. `https://www.kaggle.com/uciml/pima-indians-diabetes-database`. Accessed: 2024-08-10.

[14] N. P. Karlekar and N. Gomathi, *Ow-svm: Ontology and whale optimization-based support vector machine for privacy-preserved medical data classification in cloud*, International Journal of Communication Systems, 31 (2018), p. e3700.

[15] F. E. Keddous and A. Nakib, *Optimal cnn–hopfield network for pattern recognition based on a genetic algorithm*, Algorithms, 15 (2021), p. 11.

[16] N. Mahendran and D. R. V. PM, *A deep learning framework with an embedded-based feature selection approach for the early detection of the alzheimer's disease*, Computers in Biology and Medicine, 141 (2022), p. 105056.

[17] S. M. Nagarajan, V. Muthukumaran, R. Murugesan, R. B. Joseph, and M. Munirathanam, *Feature selection model for healthcare analysis and classification using classifier ensemble technique*, International Journal of System Assurance Engineering and Management, (2021), pp. 1–12.

[18] B. Remeseiro and V. Bolon-Canedo, *A review of feature selection methods in medical applications*, Computers in biology and medicine, 112 (2019), p. 103375.

[19] A. Sharma and P. K. Mishra, *Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis*, International Journal of Information Technology, 14 (2022), pp. 1949–1960.

[20] D. Singh, D. S. Sisodia, and P. Singh, *Multi-objective evolutionary approach for the performance improvement of learners using ensembling feature selection and discretization technique on medical data*, Current medical imaging, 16 (2020), pp. 355–370.
[21] L. Sun, Z. Mo, F. Yan, L. Xia, F. Shan, Z. Ding, B. Song, W. Gao, W. Shao, F. Shi, et al., *Adaptive feature selection guided deep forest for covid-19 classification with chest ct*, IEEE Journal of Biomedical and Health Informatics, 24 (2020), pp. 2798–2805.
[22] R. Tang and X. Zhang, *Cart decision tree combined with boruta feature selection for medical data classification*, in 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), IEEE, 2020, pp. 80–84.
[23] UCI Machine Learning Repository, *Chronic kidney disease dataset*. `https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease`. Accessed: 2024-08-10.
[24] ———, *Statlog (heart) dataset*. `http://archive.ics.uci.edu/ml/datasets/statlog+(heart)`. Accessed: 2024-08-10.

Edited by: Neelakandan Subramani
Special issue on: Transforming Health Informatics: The Impact of Scalable Computing and Advanced AI on Medical Diagnosis
Received: Feb 9, 2024
Accepted: Jun 23, 2024