IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 2, April, 2024, pp. 643-653
Journal homepage: http://iieta.org/journals/ria

Cardiovascular Disease Prediction: Employing Extra Tree Classifier-Based Feature Selection and Optimized RNN with Artificial Bee Colony

Yaso Omkari Daddala, Kareemulla Shaik*

School of Computer Science and Engineering VIT-AP University, Amaravati 522237, Andhra Pradesh, India

Corresponding Author Email: kareemulla.shaik@vitap.ac.in

https://doi.org/10.18280/ria.380228

Received: 17 September 2023
Revised: 11 December 2023
Accepted: 18 January 2024
Available online: 24 April 2024

Keywords:
cardiovascular disease, recurrent neural network, feature selection, artificial bee colony optimization, heart disease classification, extra tree classifier

ABSTRACT

Cardiovascular disease (CVD) stands as the most widespread severe illness impacting human health on a global scale. Forecasting CVDs in advance becomes more and more crucial as CVDs increase exponentially every day. Deep Learning (DL) algorithms are self-adaptive to recognize patterns and analyze data more effectively in CVD prediction. Over the past few decades, many researchers and practitioners have examined different predictive algorithms, but most of those studies are based on small-sized datasets like less than 10,000 patient records. The major shortcomings of earlier research lie in its reliance on small-sized datasets, elevating the risk of overfitting. In contrast, our study addresses this limitation by utilizing Kaggle's cardiac dataset encompassing 70,000 patients and 11 features. The primary objective of this study is to minimize the risk of overfitting and accurately predict CVD by showcasing the effectiveness of using comprehensive datasets. This paper proposes a hybrid DL methodology by utilizing a Extra Tree Classifier with Artificial Bee Colony optimized Recurrent Neural Network (ETC-ABC-RNN) for accurate classification of CVDs with 96% accuracy. By measuring accuracy, precision, recall, and F1, the efficiency of the system is demonstrated. The outcomes demonstrated that the suggested methodology surpassed various methods in predicting heart disease.

1. INTRODUCTION

The heart, being the most essential muscular organ, performs the crucial task of sending oxygen and nutrients to the tissues while simultaneously removing carbon dioxide and other waste products. A complex connection of blood vessels called the cardiovascular system is comprised of arteries, veins, and capillaries throughout the body. The heart, a muscular pump with four chambers that propels oxygenated blood to the body's tissues through arteries and returns deoxygenated blood to the lungs via veins. Capillaries, tiny vessels, facilitate the exchange of oxygen, nutrients, and waste products. Blood, comprising red and white blood cells, platelets, and plasma, carries vital substances to cells and aids in waste removal. This intricate system ensures oxygen delivery, nutrient transport, waste removal, and hormone distribution, playing a fundamental role in maintaining overall health. Any abnormality in normal blood circulation can lead to serious heart complications [1].

Cardiovascular diseases (CVD) can be of several types, including coronary artery disease, congenital heart disease, and arrhythmia [2]. Alarming statistics from the World Health Organization (WHO) reveal that CVDs annually claim an estimated 17 million lives globally, surpassing other causes of mortality. Despite the potential for addressing these conditions through lifestyle adjustments, there is a concerning upward trend in CVD incidence. Multiple WHO reports state that approximately 31% of all global deaths are attributed to CVDs. Addressing key risk factors such as tobacco use, an unhealthy diet, and physical inactivity could result in an 80% reduction in premature deaths from heart attacks and strokes, major contributors to global mortality [3].

There are several factors that trigger heart disease, including inadequate blood supply to the body's parts. There are various symptoms associated with CVDs, such as an abnormal heartbeat, difficulty breathing, chest pain, feeling faint, impulse to vomit, and swelling of the feet. The most important risk factors for heart diseases are mainly dependent on human lifestyle such as work style, stress, genetic mutations, alcohol consumption, cholesterol levels, obesity, smoking [4, 5]. Mortality can be reduced by noticing early signs and making lifestyle changes, such as exercising, avoiding smoking, and seeking appropriate medical care. Genetic predispositions play a crucial role in CVDs, influencing factors such as susceptibility and disease progression. Unlike lifestyle-related risk factors, these genetic factors may not be as easily modified through behavioral changes. Understanding and identifying these genetic components is essential for personalized approaches to prevention and treatment in cardiovascular health.

To make accurate diagnoses, hidden patterns must be discovered in the immense volume of medical data. Artificial Intelligence (AI) is expanding into every facet of life, including medicine. Furthermore, machine learning (ML), which is part of AI, aids in the classification of genes as well as diseases like diabetes [6, 7], cancer [8, 9], liver [10, 11],

thyroid [12], brain stroke [13], prostate [14], and skin [15]. As a result, clinicians have been recommending enhanced detection using ML-based predictive models to minimize the death rate and improve clinical decision-making. AI is pivotal in transforming healthcare, enhancing diagnostics, personalizing treatments, and streamlining processes. Its integration into predictive analytics allows for early disease detection and risk assessment, significantly impacting patient outcomes. Overall, AI revolutionizes medical practices and improves healthcare efficiency. Researchers continue to face challenges in identifying the most pertinent factors that increase the risk of heart disease in order to achieve a high level of prediction performance. While these methods prove effective in dealing with a limited number of traditional CAD risk factors, including age, smoking, diabetes, blood pressure, weight, and cholesterol, they may not cover all potential factors.

In recent years, many prediction models have been developed using a few publicly available datasets on heart disease. Many studies analyzed data from the UCI repository and Kaggle. The three most well-known datasets used to detect coronary artery disease (CAD) are: Cleveland dataset with size of 303 instances [16], UCI heart disease dataset with 1025 instances [17], and Z-Alizadeh Sani dataset with 303 instances [18]. A complex pattern and large data often degrade the performance of the ML model. The major limitation in earlier research stems from limited datasets, increasing the risk of overfitting, with most datasets ranging from 300 to 10,000. This constraint may hinder the model's ability to generalize, especially when confronted with larger datasets. To comprehend hidden patterns within diverse data, ML algorithms require exposure to various samples from the dataset. In contrast, our method involves utilizing an extensive cardiovascular disease dataset featuring 70,000 patients and 11 features, effectively minimizing the risk of overfitting.

The primary goal in this research is to employing extensive and diversified datasets, thereby emphasizing the efficacy of utilizing substantial data for improved accuracy and to design an expertise system which supports diagnosis. This research introduces a novel approach to tackle this challenge, which involves using an Extra Tree Feature Set based Recurrent Neural Network combined with an Artificial Bee Colony Optimization model. The aim is to achieve more precise predictions of cardiovascular diseases (CVDs).

1.1 Our contributions

In this paper, our contributions are the following:

(1) The analysis of Deep Learning (DL) classification technique is carried out using a large dataset that consists of 70,000 instances collected from Kaggle.

(2) This study focuses on analyzing factors that increase the risk while predicting heart disease. An expanded dataset is created based on the study, adding new features such as mean arterial pressure (MAP), pulse pressure (PP), body mass index (BMI).

(3) Feature selection is performed using the Extra Tree Classifier (ETC).

(4) The focus of this research is on creating a prediction model for classification of heart disease utilizing Artificial Bee Colony Optimization and Recurrent Neural Networks.

(5) New features boosted performance, including Mean Atrial Pressure and Pulse Pressure, and the proposed hybrid methodology yielded the maximum accuracy for the dataset.

The remaining portion of the paper is structured in the following manner: Related studies relevant to our problem statement on heart disease prediction are explored in Section II. The dataset description, feature optimization and ranking, DL technique with optimization technique employed for heart disease prediction are in Section III. Section IV discusses the proposed ETC Feature Set-based Recurrent Neural Network with Artificial Bee Colony Optimization model's framework. Section V presents the various metrics utilized for the system evaluation and the analysis of experimental findings obtained from this study, and last, Section VI outlines the observations inferred from the outcomes.

2. RELATED WORK

Even though medical cardiology is critical, AI has already made tremendous strides. Several researchers used various AI techniques, such as ML and DL, for foreseeing cardiac disease. It has been widely accepted that these findings have had a substantial impact on medical science. In general, DL-based approaches can effectively extract meaningful patterns from large and complex data sets and use this knowledge and experience to improve performance. Various datasets have been used by researchers for cardiac disease classification. Largest portion datasets are UCI heart disease dataset, Cleveland dataset, and Z-Alideshani dataset. Most CVD prediction models use publicly available datasets, with a limited number of patient records.

Several feature extraction algorithms like minimal-redundancy, relief, and maximal relevance are examined and proved that relief-based feature selection improves the performance of ML algorithms [19]. Fisher score feature selection algorithm [20], majority voting [21], chi-square [22] are applied and improved the performance of the ML model drastically. KNN Algorithm [23], hybrid random forest [24], and ANN based classification model [25] with hyper parameter tunning, optimized Gradient Boosted Decision Tree [26] obtained good classification results. In all the studies described above, small datasets were used, and magnificent results were obtained with various ML and DL models.

A significant reduction has been observed in accuracy when the data size is increased. Some researchers demonstrated this by using two different dataset sizes. Padmanabhan et al. [27] conducted research on auto-ML. Two datasets were used in the study: the UCI dataset and the cardiovascular disease dataset. Considering the size of the cardiovascular dataset, training has taken more time. Feature significance is determined using PCA. Another study by Hagan et al. [28] used Kaggle's dataset alongside with UCI's arrhythmia dataset. According to the authors, ensemble algorithms performed well on small datasets when compared with large datasets. With gradient boosting and RF, Kaggle's dataset yields the highest accuracy. A comparison of ML models on small and large datasets was performed by Ouf and ElSeddawy [29]. With the assistance of diverse cross-validation methods, the performance of several classifiers is analyzed. A neural network with holdout cv achieved the maximum results as 71.82%. Jubier Ali et al. [30] applied the feature selection algorithm to large datasets to enhance efficiency. In this investigation, the performance of the ML algorithm is improved with lowered features. SVM, with the most significant features, achieved the great accuracy.

Most significant feature selection is the essential part in a model's performance. A variety of feature selection methods

were analyzed by Hasan and Bao [31] for feature extraction. A few classification algorithms are applied with ANN, including SVM, KNNs, Naive Bayes, and XG Boost. According to research, the highest accuracy was obtained by integrating wrapper methods with XG Boost. An analysis of the dataset's performance is carried out by applying various machine learning algorithms. The gradient boosting algorithm with recursive feature elimination outperformed the other ML algorithms in terms of accuracy. Prajwal et al. [32] examined different risk factors for heart disease prediction. By examining the correlation between the variables, features are selected. In comparison to the other ML algorithms, SVM showed the highest accuracy.

Many studies have shown that ensemble approaches are extremely effective at predicting CAD. According to literature [33], Shorewala stacked models are effective in categorizing disease classification than bagging and boosting models. The stacked model of KNN, RF, and SVM outperforms bagging and boosting methods. The LASSO method is applied for the extraction of features. Cheekati et al. [34] applied LASSO feature selection on bagging model and achieved 75% accuracy.

Martins et al. [35] analyzed various data mining techniques and models using RapidMiner. For performance analysis, WEKA software is used. Weights are used to extract attributes, and some attributes with a 0 value are removed. In this paper, the authors analyze the performance of cross-validation and split-validation methods. A split-validation was found to be more effective than a cross-validation. According to this analysis, optimized decision trees produced the highest accuracy at 73%. Bhatt et al. [36] demonstrated that MLP models have a best performance of 87%. Using K-mode clustering, classification accuracy is improved. The hyperparameters are tuned through grid search to achieve the best results.

After studying the earlier research conducted by various researchers as shown in Table 1, it has been observed that many researchers consider datasets with fewer than 1000 patient records. In this study, 70000 patient records were taken from a cardiovascular dataset to examine this issue. Numerous techniques for heart disease classification with ML and DL are applied for classification among all Neural networks with parameter optimization got better results on dataset. For feature selection filter, wrapper, and embedded methods achieved best accuracy. After analyzing the sources mentioned earlier, it becomes evident that feature extraction and optimization plays essential part in boosting the model's ability. The vast exploration space poses challenges for feature selection. Consequently, a robust and comprehensive search technique is essential to effectively address this issue.

Table 1. Literature review analysis

Methodology	Advantages	Future Work
Auto ML [27]	An analysis is conducted to compare AutoML performance using two datasets. The AutoML requires fewer lines of code in contrast to the graduate student's code, which consists of several hundred lines.	The focus was solely on AutoML functionality, and there is a need to enhance accuracy through diverse parameter optimization techniques.
Random forest [28]	The classification was conducted using two datasets, and the highest accuracy was achieved in the smaller-sized dataset with 452 ECG records.	The model needs to focus on feature analysis to identify the most effective attributes for classification improvement.
Neural network with hold out [29]	A comprehensive study was conducted on four cross-validation techniques, namely hold-out, 10-fold, repeated random, and stratified 10-fold.	There is a need to focus on feature selection and optimize neural networks to enhance the overall results.
Random Forest [30]	Comparisons were made among the results of various ML algorithms under different configurations for feature sets, namely 3, 8, and 12.	While the model performs well with the top three features, it neglects important data that significantly influences overall performance. It is crucial to consider a sufficient number of features from the dataset. Additionally, neural networks tend to yield optimal results when applied to diverse and extensive datasets.
Bagged Decision Tree [31]	A detailed analysis, and evaluation of the filter, wrapper, and embedded feature selection methods has been conducted.	Performance enhancement can be achieved by leveraging neural networks in conjunction with optimization algorithms.
SVM [32]	To mitigate overfitting in the models, synthetic data was generated, and a new feature, BMI, was introduced.	Model performance can be improved through feature selection techniques.
Neural network [33]	A comparative classification analysis was conducted on, ensemble methods such as bagging, boosting.	Neural networks, coupled with optimization algorithms, can enhance performance.
Optimized Decision tree [34]	Utilizing LASSO feature selection on a bagging model resulted in an accuracy of 75%.	Further exploration of optimizing methods is necessary for enhancing performance.
Neural networks [35]	This study proved split-validation was found to be more effective than a cross-validation.	Minimizing the number of false positives is crucial to significantly enhance the model performance.
MLP classifier [36]	Classification accuracy is enhanced by leveraging K-mode clustering, with fine-tuned hyperparameters through grid search for optimal results.	The evaluation did not assess the interpretability of the algorithm's capability to explain the formed clusters.

3. MATERIALS AND METHODS

An overview of the various classification models utilized in predicting heart disease, datasets utilized, and feature ranking algorithms are described in the below section.

3.1 Dataset description

The data utilized in this research was collected from Kaggle online data repository. This dataset consists of 70,000 patient records with 12 attributes. During a medical examination, this information was collected. These attributes determine a person's risk of cardiac problem. A total of three categories were identified such as: objective, examination, and subjective. Patient data such as age, height, weight, and gender are included in the objective type. An examination type signifies medical information obtained from a physical checkup. Patients provide information about their habits under the subjective feature type. To meet the above three requirements, 12 features have been included. According to the collected data, there were two categories: inclusion or exclusion of heart disease with various factors that pose a risk. There are several features that are not directly related to cardiovascular disease prediction but may provide significant insights. A thorough and detailed narration of every feature and type of the value can be found in Table 2.

Table 2. Description of heart disease dataset

Feature Name	Feature Category	Description and Value
age	Objective type	Age (in days)
height	Objective type	Height (in cms)
weight	Objective type	Weight (in kgs)
gender	Objective type	Men (1)/Women (2)
ap-hi	Examination type	Systolic BP (integer values)
ap-lo	Examination type	diastolic BP (integer values)
cholesterol	Examination type	Cholesterol level (Level 1,2,3)
gluc	Examination type	Glucose level (Level 1,2,3)
smoke	Subjective type	Smoking habit (0/1)
alco	Subjective type	Alcohol intake (0/1)
active	Subjective type	Physical activity (0/1)

3.2 Deep learning algorithms

Following a thorough review of the literature to tackle overfitting in datasets, the Extra Trees Classifier is chosen for its ensemble approach, constructing decision trees with random feature subsets. LSTM, a type of RNN, is selected for classification tasks on large and diversified datasets, leveraging its capability to capture long-term dependencies and effectively handle diverse data. Its adeptness in overcoming gradient-related challenges enhances its suitability for tasks involving complex patterns and varied information. Artificial Bee Colony (ABC) optimization is chosen for classification tasks, offering an effective nature-inspired algorithm for global optimization. By leveraging the foraging behaviour of honeybees, ABC facilitates efficient exploration of the solution space, enhancing the fine-tuning of parameters for improved classification performance.

3.2.1 Extra Tree Classifier (ETC) for feature selection:

A crucial component of machine learning is feature-selection, which aims to eliminate the irrelevant and irrelevant attributes from the model in order to enhance its quality and efficiency. An embedded method for extracting pertinent features was demonstrated through the use of the extra-tree algorithm in our study.

Extra-trees classifier (ETC) operates by generating a substantial quantity of decision trees using the entire training dataset. Initially, the algorithm picks a split rule based on a partially arbitrary cut point and a random subset of features (K). It proceeds to choose a random split that creates two random child nodes from the parent node.

This process repeats in each child node until a leaf node, which has no further child nodes, is reached. The final prediction is established by combining the predictions of all trees through a majority vote. In the process of constructing the forest, ETC perform feature selection by considering the Gini importance of every feature from the dataset [37].

The Gini importance of each feature is calculated from the dataset, and are then sorted in descending order on their respective importance score values. Our investigation focused on identifying the significant attributes for input into the classification model. Ultimately, we identified the top ten features that optimize model accuracies, including age, BMI, weight, height, MAP, systolic, PP, cholesterol, diastolic, and glucose as described in Table 3.

Table 3. Feature rankings given by ETC method

Feature Name	Feature Rankings
Age	0.234112
BMI	0.155018
Weight	0.136884
Height	0.132956
MAP	0.084757
Systolic	0.076705
PP	0.047813
Cholesterol	0.041836
Diastolic	0.041122
Glucose	0.015046

3.2.2 Recurrent neural networks (long short-term memory)

Recurrent neural networks (RNNs) are modified to create long-short-term memory (LSTM) networks for retaining past information. LSTM is influenced by the learning output of RNN, and the vanishing gradient problem of RNN is solved. Learning long-term dependencies of variables can be done with the LSTM method.

A neural network algorithm can boost gradient explosions and gradient disappearances by calculating long-term time series using the LSTM algorithm. Three gates are installed in the LSTM unit, namely forget gate, input gate, and output gates. Memory values are modified by the input gate. As the forget gate determines what details are to be discarded from a block, the output gate determines what parts of a cell's state should be output [38]. The formulation of LSTM can be given in the following formulas 1 to 6.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f \quad (1)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i \quad (2)$$

$$\tilde{C}_t = \tanh(W_C.[h_{t-1}, x_t] + b_C \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

where, σ is the sigmoid function is used to form three gates in the memory cell. Tanh is the hyperbolic tangent function used

to scale up the output of a memory cell. The input gate, forget gate, output gate, memory cell content, and new memory cell content are represented by i, f, o, and C, respectively.

3.2.3 Artificial Bee Colony (ABC*)*

Optimizing numerical problems with the use of a swarm intelligence system called the Artificial Bee Colony (ABC) algorithm. It was conceived after observing the resourceful actions of honey bees during their foraging. Foragers and food sources round out the model. In this research, artificial bee colony algorithm works on feature subset generation that contains most relevant features for Cardiovascular Disease Detection. An array of workable answers is generated by iteratively altering groups of variables throughout the employed phase.

The observer phase is when these solutions are sorted from best to worst based on the values assigned to their properties. In the scout phase, a solution is given up if it is not improved upon after a set number of iterations. These steps are repeated until a solution that is very close to optimal is found by the algorithm. The location of a food source in ABC, a population-based algorithm, stands in for a possible alternative to the optimization issue, while the quantity of nectar it provides stands in for the quality called fitness of the linked solution. There are as many working bees as there are total population solutions. In the first step, P percent of the population is randomly sprayed across the solution space, and then their fitness is calculated using the parameters FV1, FV2, ..., FVn.

These populations become the onlooker bees once they are introduced to the solution space [39, 40]. The probability of a search in a given population is defined as

$$PS(p) = \sum_{i=1}^{M} \frac{G(T_i) * G(T|S_p)}{\sum_{i=i+1}^{L} \frac{G(T|S_p)}{G(T_i)}} + \max(P, P+i) \quad (7)$$

where, G is the probability function, p is the probability search value for each record, S, T are the current and next records in the dataset. The fitness value is arrived at after determining the last possible bound, using the formula.

$$FitV(PS(p)) = \frac{\sum_{i=i+1}^{L} \frac{G(T|S_p)}{G(T_i)}}{1 + \frac{1}{PS}} \quad (8)$$

$$FV(p) = \begin{cases} \frac{1}{1+FitV(p)} + PS & if\ FitV \geq 0 \\ FitV + PS & if\ FitV < 0 \end{cases} \quad (9)$$

4. SYSTEM MODEL

An overview of the model's workflow is given in this section. Figure 1 describes the architecture of the proposed ETC Feature Set based Recurrent Neural Network with Artificial Bee Colony Optimization (ETC-RNN-ABC) model.

•Data pre-processing

•Feature analysis – new features BMI, MAP, PP are added to the dataset

•Extraction of features –Extra Tree Classifier algorithm is applied for most important feature extraction

•Deep Learning – Recurrent Neural Network- LSTM

•Cross-validation – Two cases are applied (1) 7-fold cross validation (2) 10-fold cross validation

• Optimization algorithm – Artificial Bee Colony Optimization

• Evaluation of results – Based on confusion matrix (accuracy, precision, recall, f1-score)

•Classification – Binary classification (0-absent, 1-present)

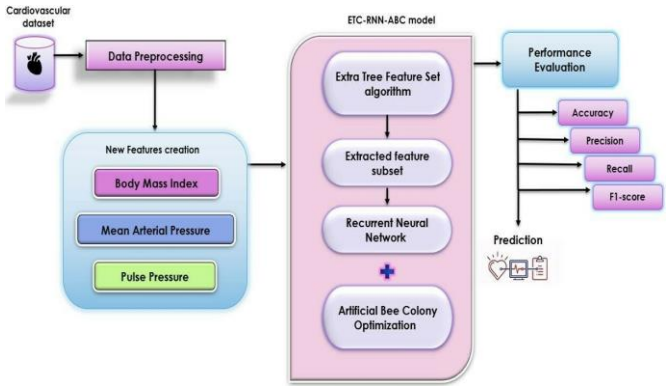

Figure 1. Basic workflow of proposed methodology

5. RESULTS AND COMPARATIVE ANALYSIS

In this section, feature engineering, data preprocessing, performance evaluation metrics and results are described in detailed manner.

5.1 Feature engineering

After a detailed analysis of the data from the cardiovascular dataset, some new features were calculated to enhance the current features. BMI is often regarded as a useful measure for assessing general obesity because it provides insight into the deposition of visceral fat and its impact on metabolic health. A wealth of studies has investigated how central obesity affects cardiac function. According to studies, being overweight can increase the risk of cardiovascular disease by up to 45% [41]. BMI can be calculated using the following formula 10:

$$BMI = \frac{weight}{height^2} \quad (10)$$

A person's mean arterial pressure (MAP) is known as the average blood pressure during a cardiac cycle. A significant direct correlation exists between peripheral resistance and cardiac output measured by MAP. MAP is a significant risk factor for people with type 2 diabetes and is also associated with a higher rate of CVD hospitalization [42]. MAP is calculated by using systolic and diastolic pressure values as shown in the following Eq. (11):

$$MAP = \frac{systolic\ pressure + (2 * diastolic\ pressure)}{3} \quad (11)$$

In recent years, high pulse pressure has been considered a risk factor for cardiovascular disease. Further, since PP is affected by both systolic and diastolic blood fluctuations, it may provide better cardiovascular risk predictions than blood pressure (BP). The PP can indicate an increased stiffness of

large arteries because of the pulsatile component of BP [43]. Pulse pressure can be calculated as described in Eq. (12).

$$PP = systolic - diastolic \quad (12)$$

The new feature BMI, MAP, and PP are added to the cardiovascular dataset and the total 15 features are considered in our study.

5.2 Pre-processing

The effectiveness of the ML model is contingent upon the quality of the data preprocessing. We preprocessed the data before analyzing it to account for missing values and outliers. Due to their impact on overall performance, missing values should be identified and eliminated. None of the fifty features in the cardiovascular data set had missing values. During data analysis, non-categorical numeric variables between 0 and 1 were standardized to ensure uniform distributions in the dataset. As evidenced by its class distribution, this dataset has balanced classes. The rate of people without cardiac issue is 49.5%, and the people with cardiac issue is 50.4%.

During the dataset analysis, some contrasts were found. Systolic values should be higher than diastolic values, and these are removed from the dataset. Some people have heights less than 125 or greater than 210. These values are also eliminated from the height feature in the dataset. Patterns can be revealed through exploratory data analysis by examining data through graphical representations as described in Figures 2-9.

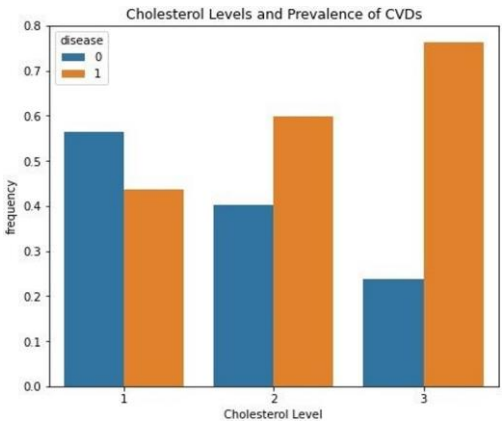

Figure 2. Analysis graph of cholesterol levels

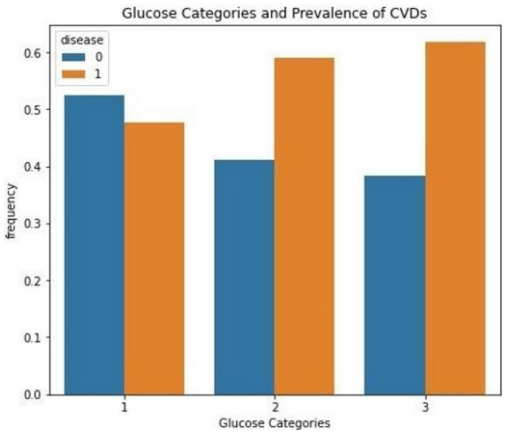

Figure 3. Analysis graph of glucose levels

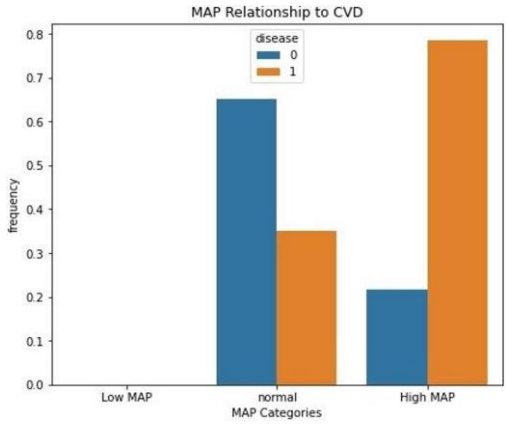

Figure 4. Analysis graph of MAP feature

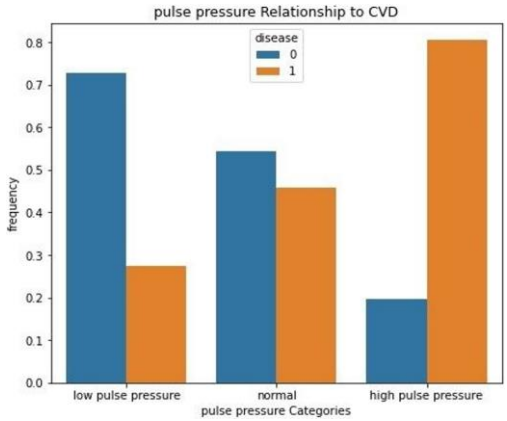

Figure 5. Analysis graph of PP feature

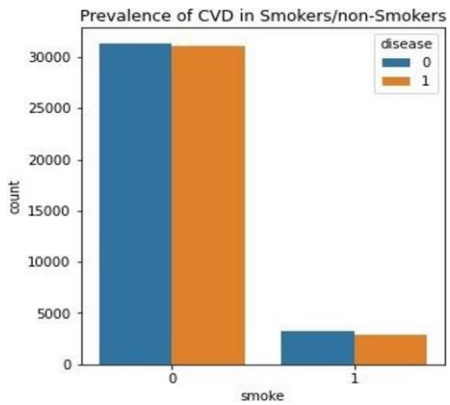

Figure 6. Analysis graph of smoking feature

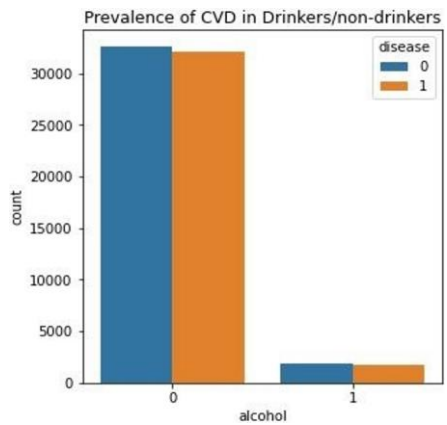

Figure 7. Analysis graph of alcohol feature

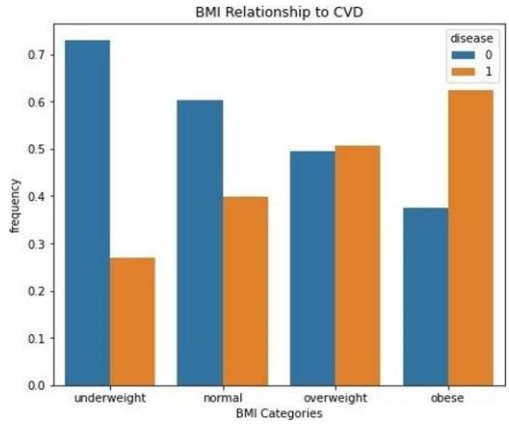

Figure 8. Analysis graph of BMI feature

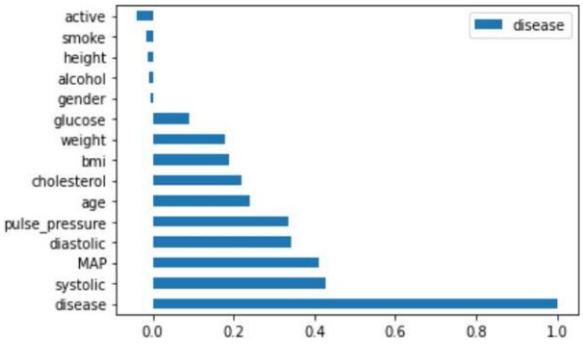

Figure 9. Importance of total 15 features

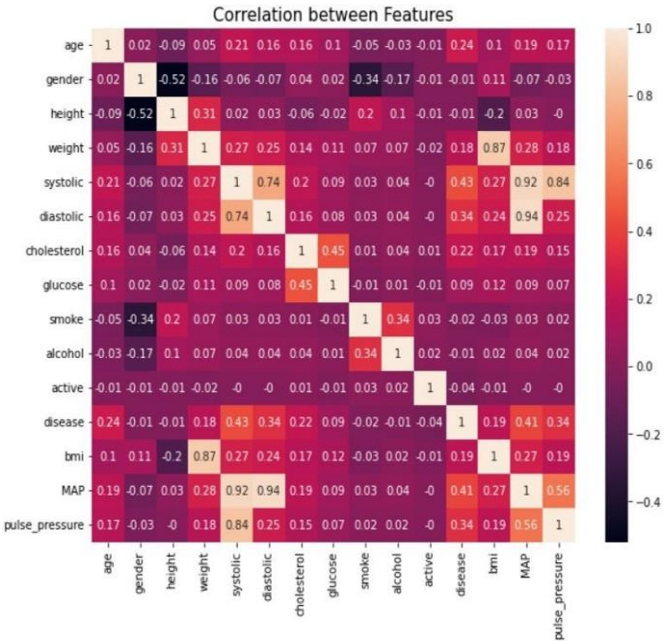

Figure 10. Correlation matrix with all 15 features

The BMI graph indicates that people with a high or obese BMI are more prone to experience a CVD than with the people. When the MAP value exceeds 100, it is a sign that the arteries are under a lot of pressure. According to the MAP graph, most people with high MAP are more likely to develop heart disease than those with low MAP. According to the PP graph, heart disease is more likely to occur among people with high pulse pressure. As shown in the cholesterol graph, it seems that if a person has a cholesterol level of 2 or 3, they may be more at risk of CVD than a person with a cholesterol level of 1. A person with glucose level 2, 3 are having more risk while comparing with the lower level as shown in glucose graph. Through feature importance graph, we see that age, BMI, weight, height, MAP, systolic, PP, cholesterol level, diastolic, and glucose are all significant when predicting if someone may have a CVD. Smoking and drinking features still require more comprehensive information to be collected, like how much they are consuming daily. The correlation matrix for all the features is represented in the below Figure 10.

5.3 Performance evaluation metrics

The competence of model is evaluated by a confusion matrix, which is a combination of four values. Figure 11 describes the confusion matrix and its values. TP and TN represent correct predictions, while FP and FN represent incorrect predictions. Statistical measures such as accuracy, precision, F1-score, and recall have been used to validate our algorithms. Here are the corresponding formulas, represented in the following equations. Accuracy refers to the proportion of correct predictions among both positive and negative classes. Precision measures the percentage of accurately predicted positive classes out of all predicted positives. Recall, on the other hand, represents the ratio of correctly predicted positive instances to all instances in the actual positive class. The F1-score is derived from precision and recall to provide a combined measure of a model's performance.

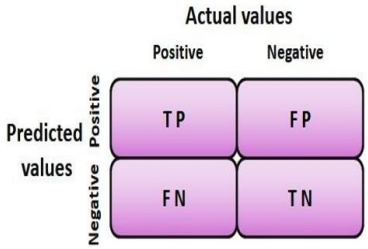

Figure 11. Confusion matrix

The below Eqs. (13) to (16) can be utilized to express these metrics.

$$Accuracy = \frac{(True\ posi + True\ neg)}{(Total)} * 100 \quad (13)$$

$$Precision = \frac{True\ posi}{(True\ posi + Flase\ posi)} * 100 \quad (14)$$

$$Recall = \frac{True\ posi}{(True\ posi + Flase\ neg)} * 100 \quad (15)$$

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (16)$$

5.4 Discussion

This research utilized an MSI computer with Intel (R) Core (TM) i5-10500H processor and 16 GB of RAM. The model is implemented in Python on Google Colab. There are 15 features in the dataset after feature engineering, along with three additional features: BMI, MAP, and PP. ETC algorithm is used to obtain the most crucial attributes after preprocessing the data. After applying the ETC algorithm rankings, top 10

features were selected from total features, including age, BMI, weight, height, MAP, systolic, PP, cholesterol level, diastolic, and glucose.

Table 4. Results with 7-fold and 10-fold cross validation

Cross validation	Accuracy	Precision	Recall	F1-score
Case 1: Without BMI, MAP, PP Features				
RNN-LSTM (Old 12 Features)				
7-fold cross validation	87.1	85.9	89.5	87.7
10-fold cross validation	88.5	89.2	89.3	89.3
RNN-LSTM (Top 10 features by ETC)				
7-fold cross validation	90.5	90.5	91.9	91.2
10-fold cross validation	91.1	91.5	91.9	91.7
Case 2: With newly inserted features - BMI, MAP, PP				
RNN-LSTM (Old 12+new 3=Total 15 features)				
7-fold cross validation	92.2	92.4	92.8	92.6
10-fold cross validation	92.7	93.5	92.9	93.2
RNN-LSTM (Top 10 features among 15 by ETC)				
7-fold cross validation	95.2	96.1	95.0	95.5
10-fold cross validation	96	96	97	96

As shown in Table 4, total of two cases were examined in this study (i) old features 12 and (ii) with newly added features BMI, MAP, PP. According to the Table 3, the model is tested using 5-fold and 10-fold cross validation, and the 10-fold cross validation yielded the most accurate results. In both cases again model is evaluated with and without feature selection. Among all the cases 10-fold cross validation yielded the best results while comparing with the 7-fold cross validation. The results highlight a significant improvement of almost 5% with the addition of new features. Initially, the best accuracy of 91.1% was achieved with old features and 10-fold cross-validation. After introducing BMI, MAP, and PP, a substantial increase was observed, reaching an accuracy of 96%. This underscores the pivotal role of BMI, MAP, and PP in enhancing the model's performance. According to the figure below, case (ii) top 10 features obtained excellent results by applying ETC with 10-fold cross validation compared to case (i).

The total data has been separated into two segments: 70% of the data used to train and 30% used to test the proposed model. RNN-LSTM is used to build the model with ABC optimizer. Epochs measure how often weights are updated when the training set is selected once. Models are better at generalizing their learning with raised epochs. The model performance here is analyzed over 100 epochs. Dropout is applied which is a regularization technique used to avoid the overfitting problem and ensure the model. The model has the Adam optimizer, ReLU activation function, and mean squared error loss function. The training and validation loss represents the performance of the proposed model. The proposed model training and testing loss is very minimum that represents the performance of the model as high. Figure 12 represents the training and validation loss levels of the ETC-RNN-ABC model.

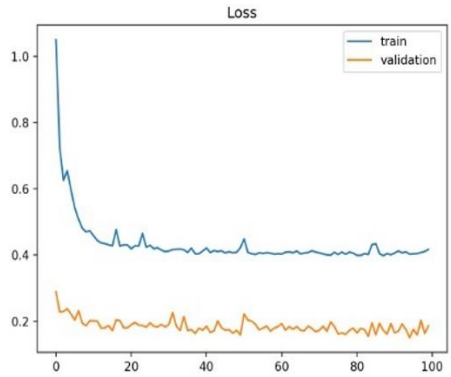

Figure 12. Loss graph of model for 100 epochs

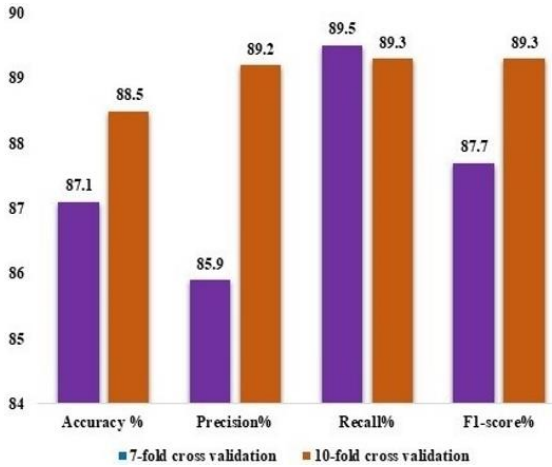

Figure 13. Cross validation comparison of old 12 features

After applying the ABC algorithm for model optimization, the model has shown excellent results while comparing with the many studies presented in Table 5.

Table 5. Comparison with literature survey studies

Methodology	Accuracy%	Precision%	Recall%	F1-score%
Auto ML [27]	74	68	-	-
Random forest [28]	72.69	74.85	69.48	-
Neural network with hold out [29]	71.8	72	72	72
Random Forest [30]	72.69	74.8	69.4	-
Bagged Decision Tree [31]	74.8	76.2	67.4	71.5
SVM [32]	72.6	-	-	-
Neural network [33]	74.9	76.2	68.2	73
Optimized Decision tree [34]	73.14	75.35	69.22	77.63
Neural networks [35]	71.82	72	72	72
MLP classifier [36]	87.23	88.7	84.8	86.7
Proposed ETC-RNN-ABC model	**96**	**96**	**97**	**96**

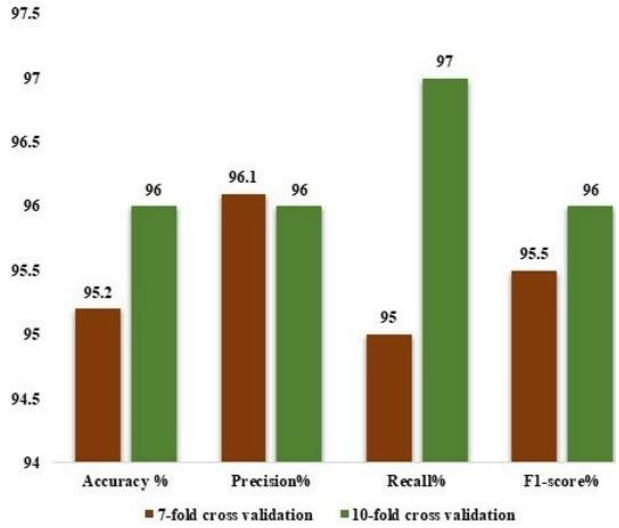

Figure 14. Cross validation comparison of features from ETC

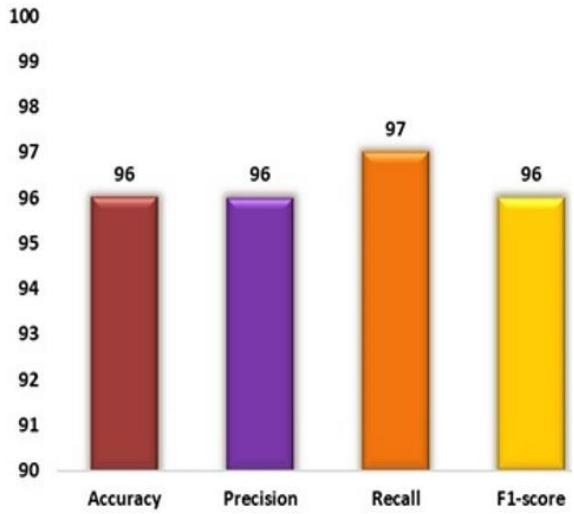

Figure 15. Performance of the proposed methodology

Figures 13 and 14 shows the cross validation comparative analysis of the proposed model. As shown in the Figure 15, the proposed model achieved accuracy as 96%, precision as 96%, recall as 97%, and F1-score as 96%.

6. CONCLUSIONS

As chronic disease epidemics grow, identifying those at high risk of cardiovascular disease plays an increasingly important role in medicine. To this end, many AI-based algorithms for predicting cardiovascular disease risk have been presented. Most of the prognosis prediction approaches currently available use small datasets based on ML and DL. A cardiovascular dataset of 70000 patient records provided by Kaggle is used for model evaluation in this paper. This comprehensive dataset addresses limitations from earlier research on smaller datasets, showcasing the effectiveness of the proposed approach. A hybrid DL approach was proposed to predict cardiovascular disease in this study. Extra tree classifier is used to extract the most significant features as feature selection is crucial in model performance. The artificial bee colony optimization algorithm and the RNN-LSTM are proposed in this study to form a hybrid method. To evaluate a model, performance analysis metrics namely accuracy, precision, recall, and F1-score are used. Improving CVD patient diagnosis accuracy on larger datasets is a primary focus of the proposed ETC-RNN-ABC model. An analysis of the proposed methodology with several studies shows that it is highly accurate in predicting CVDs by 96%. The proposed method attains a robust 96% accuracy, surpassing various studies in the literature survey, and demonstrates effectiveness in minimizing overfitting risks. The proposed model excels in cost-effectiveness and ease of use compared to non-AI methods like angiography testing, utilizing existing data to reduce expenses related to invasive procedures and improve accessibility. While the training time is longer because of the size of the data, future work will focus on tackling the time complexity.

REFERENCES

[1] Amin, R., Al Ghamdi, M.A., Almotiri, S.H., Alruily, M. (2021). Healthcare techniques through deep learning: Issues, challenges and opportunities. IEEE Access, 9: 98523-98541. https://doi.org/10.1109/ACCESS.2021.3095312

[2] Quer, G., Arnaout, R., Henne, M., Arnaout, R. (2021). Machine learning and the future of cardiovascular care: JACC state-of-the-art review. Journal of the American College of Cardiology, 77(3): 300-313. https://doi.org/10.1016/j.jacc.2020.11.030

[3] Alizadehsani, R., Abdar, M., Roshanzamir, M., et al. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. Computers in Biology and Medicine, 111: 103346. https://doi.org/10.1016/j.compbiomed.2019.103346

[4] Gayathri, B., Sujata, S., Thakur, R. (2023). Cardiovascular diseases and ageing in India: A propensity score matching analysis of the effects of various risk factors. Current Problems in Cardiology, 48(5): 101606. https://doi.org/10.1016/j.cpcardiol.2023.101606

[5] Christopher, O., Xiong, Z., Huang, Y., et al. (2023). Risk score for coronary heart disease (CHD-RISK) and hemodynamically significant aortic valve stenosis. Nutrition, Metabolism and Cardiovascular Diseases, 33(5): 1029-1036. https://doi.org/10.1016/j.numecd.2022.12.023

[6] Akbar, S., Midhunchakkaravarthy, D. (2020). A novel filtered segmentation-based Bayesian deep neural network framework on large diabetic retinopathy databases. Revue d'Intelligence Artificielle, 34(6): 683-692. (2020). https://doi.org/10.18280/ria.340602

[7] Birjais, R., Mourya, A.K., Chauhan, R., Kaur, H. (2019). Prediction and diagnosis of future diabetes risk: A machine learning approach. SN Applied Sciences, 1(2019): 1-8. https://doi.org/10.1007/s42452-019-1117-9

[8] Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M.M., Hasan, M., Kabir, M.N. (2020). Breast cancer prediction: A comparative study using machine learning techniques. SN Computer Science, 1: 1-14. https://doi.org/10.1007/s42979-020-00305-w

[9] Kadir, T., Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imaging techniques. Translational lung cancer research, 7(3): 304.

https://doi.org/10.21037/tlcr.2018.05.15
[10] Fayaz, S.A., Zaman, M., Kaul, S., Butt, M.A. (2022). How M5 Model Trees (M5-MT) on continuous data are used in rainfall prediction: An experimental evaluation. Revue d'Intelligence Artificielle, 36(3): 409-415. https://doi.org/10.18280/ria.360308
[11] Wu, C.C., Yeh, W.C., Hsu, W.D., et al. (2019). Prediction of fatty liver disease using machine learning algorithms. Computer Methods and Programs in Biomedicine, 170: 23-29. https://doi.org/10.1016/j.cmpb.2018.12.032
[12] Chaubey, G., Bisen, D., Arjaria, S., Yadav, V. (2021). Thyroid disease prediction using machine learning approaches. National Academy Science Letters, 44(3): 233-238. https://doi.org/10.1007/s40009-020-00979-z
[13] Bandi, V., Bhattacharyya, D., Midhunchakkravarthy, D. (2020). Prediction of brain stroke severity using machine learning. Revue d'Intelligence Artificielle, 34(6): 753-761. https://doi.org/10.18280/ria.340609
[14] Barlow, H., Mao, S., Khushi, M. (2019). Predicting high-risk prostate cancer using machine learning methods. Data, 4(3): 129. https://doi.org/10.3390/data4030129
[15] Bascil, M.S. (2019). Convolutional neural network to extract the best treatment way of warts based on data mining. Revue d'Intelligence Artificielle, 33(3): 165-170. https://doi.org/10.18280/ria.330301
[16] Absar, N., Das, E.K., Shoma, S.N., et al. (2022). The efficacy of machine-learning-supported smart system for heart disease prediction. Healthcare, 10(6): 1137. https://doi.org/10.3390/healthcare10061137
[17] Ramesh, T.R., Lilhore, U.K., Poongodi, M., Simaiya, S., Kaur, A., Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. Malaysian Journal of Computer Science, 132-148. https://doi.org/10.22452/mjcs.sp2022no1.10
[18] Gupta, A., Kumar, R., Arora, H.S., Raman, B. (2022). C-CADZ: Computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset. Applied Intelligence, 52(3): 2436-2464. https://doi.org/10.1007/s10489-021-02467-3
[19] Haq, A.U., Li, J.P., Memon, M.H., Nazir, S., Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018: 1-21. https://doi.org/10.1155/2018/3860146
[20] Saqlain, S.M., Sher, M., Shah, F.A., Khan, I., Ashraf, M.U., Awais, M., Ghani, A. (2019). Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowledge and Information Systems, 58: 139-167. https://doi.org/10.1007/s10115-018-1185-y
[21] Latha, C.B.C., Jeeva, S.C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16: 100203. https://doi.org/10.1016/j.imu.2019.100203
[22] Appari, G.D., Borra, S.P.R., Haritha, T., Mandava, V.S.R., Balaji, T., Kalapala, V.S., Kodepogu, K.R. (2023). An improved CHI 2 feature selection based a two-stage prediction of comorbid cancer patient survivability. Revue d'Intelligence Artificielle, 37(1): 83-92. https://doi.org/10.18280/ria.370111
[23] Pawlovsky, A.P. (2018). An ensemble based on distances for a kNN method for heart disease diagnosis. In 2018 International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, pp. 1-4. https://doi.org/10.23919/ELINFOCOM.2018.8330570
[24] Mohan, S., Thirumalai, C., Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7: 81542-81554.. https://doi.org/10.1109/ACCESS.2019.2923707
[25] Yazid, M.H.A., Satria, M.H., Talib, S., Azman, N. (2018). Artificial neural network parameter tuning framework for heart disease classification. In 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics, Malang, Indonesia, pp. 674-679. https://doi.org/10.1109/EECSI.2018.8752821
[26] Shaik, K., Ramesh, J.V.N., Mahdal, M., Rahman, M.Z.U., Khasim, S., Kalita, K. (2023). Big data analytics framework using squirrel search optimized gradient boosted decision tree for heart disease diagnosis. Applied Sciences, 13(9): 5236. https://doi.org/10.3390/app13095236
[27] Padmanabhan, M., Yuan, P., Chada, G., Nguyen, H.V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. Journal of Clinical Medicine, 8(7): 1050. https://doi.org/10.3390/jcm8071050
[28] Hagan, R., Gillan, C.J., Mallett, F. (2021). Comparison of machine learning methods for the classification of cardiovascular disease. Informatics in Medicine Unlocked, 24: 100606. https://doi.org/10.1016/j.imu.2021.100606
[29] Ouf, S., ElSeddawy, A.I.B. (2021). A proposed paradigm for intelligent heart disease prediction system using data mining techniques. Journal of Southwest Jiaotong University, 56(4): 220-240. https://doi.org/10.35741/issn.0258-2724.56.4.19
[30] Jubier Ali, M., Chandra Das, B., Saha, S., Biswas, A. A., Chakraborty, P. (2022). A comparative study of machine learning algorithms to detect cardiovascular disease with feature selection method. In Machine Intelligence and Data Science Applications, pp. 573-586. https://doi.org/10.1007/978-981-19-2347-0_45
[31] Hasan, N., Bao, Y. (2021). Comparing different feature selection algorithms for cardiovascular disease prediction. Health and Technology, 11: 49-62. https://doi.org/10.1007/s12553-020-00499-2
[32] Prajwal, K., Tharun, K., Navaneeth, P. (2022). Cardiovascular disease prediction using machine learning. In 2022 International Conference on Innovative Trends in Information Technology, Kottayam, India, pp. 1-6. https://doi.org/10.1109/ICITIIT54346.2022.9744199
[33] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26: 100655. https://doi.org/10.1016/j.imu.2021.100655
[34] Cheekati, V., Natarajasivan, D., Indraneel, S. (2021). Ensemble approaches can aid in the early detection of coronary heart disease. Natural Volatiles & Essential Oils, 8(5): 12224-12239.
[35] Martins, B., Ferreira, D., Neto, C., Abelha, A., Machado, J. (2021). Data mining for cardiovascular disease prediction. Journal of Medical Systems, 45: 1-8. https://doi.org/10.1007/s10916-020-01682-8
[36] Bhatt, C.M., Patel, P., Ghetia, T., Mazzeo, P.L. (2023).

Effective heart disease prediction using machine learning techniques. Algorithms, 16(2): 88. https://doi.org/10.3390/a16020088

[37] Reddy, B.K., Delen, D. (2018). Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. Computers in Biology and Medicine, 101: 199-209. https://doi.org/10.1016/j.compbiomed.2018.08.029

[38] Hsu, W., Warren, J.R., Riddle, P.J. (2022). Medication adherence prediction through temporal modelling in cardiovascular disease management. BMC Medical Informatics and Decision Making, 22(1): 313. https://doi.org/10.1186/s12911-022-02052-9

[39] Karaboga, D., Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. Journal of global optimization, 39: 459-471. https://doi.org/10.1007/s10898-007-9149-x

[40] Zamee, M.A., Han, D., Won, D. (2023). Integrated grid forming-grid following inverter fractional order controller based on Monte Carlo Artificial Bee Colony Optimization. Energy Reports, 9: 57-72. https://doi.org/10.1016/j.egyr.2022.11.149

[41] Bakhtiyari, M., Kazemian, E., Kabir, K., et al. (2022). Contribution of obesity and cardiometabolic risk factors in developing cardiovascular disease: A population-based cohort study. Scientific Reports, 12(1): 1544. https://doi.org/10.1038/s41598-022-05536-w

[42] Sesso, H.D., Stampfer, M.J., Rosner, B., Hennekens, C.H., Gaziano, J.M., Manson, J.E., Glynn, R.J. (2000). Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. Hypertension, 36(5): 801-807. https://doi.org/10.1161/01.HYP.36.5.801

[43] Mizuhara, R., Mitaki, S., Takamura, M., Abe, S., Onoda, K., Yamaguchi, S., Nagai, A. (2022). Pulse pressure is associated with cognitive performance in Japanese non-demented population: A cross-sectional study. BMC Neurology, 22(1): 137. https://doi.org/10.1186/s12883-022-02666-6