An Optimized Machine Learning Approach for Coronary Artery Disease Detection

Savita[1], Geeta Rani[2,*], and Apeksha Mittal[3]

[1] Department of Computer Science, GD Goenka University, Gurgaon, India; Email: savitadagar9@gmail.com
[2] Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, India
[3] Department of Engineering and Sciences, GD Goenka University, Gurgaon, India; Email: apekshamittal3@gmail.com
*Correspondence: geetachhikara@gmail.com

***Abstract*—Rising number of fatalities caused by Coronary Artery Disease is a major concern for the public as well as the health industry. Furthermore, diagnostic methods like angiography are expensive and unaffordable for those who are not well-off. Also, angiography is bothersome for the patient due to allergic responses, renal damage, and bleeding where the catheter is inserted. The researchers in literature proposed the machine learning-based approaches for the detection of Coronary Artery Disease. But, these techniques have low accuracy. Thus, there is a scope for optimization of these techniques. The objective of this paper is to develop a machine learning system for the early detection of Coronary Artery Disease early. Also, it employs optimization methods viz. Particle Swarm Optimization, and Firefly Algorithm with Principle Component Analysis based feature extraction and decision tree-based classification. The proposed technique reports an accuracy of 95.3%. Thus, the technological solution may be used as an automatic diagnostic aid.**

***Keywords*—CAD, data engineering, machine learning, medical diagnosis**

I. Introduction

Coronary Artery Disease (CAD) is one of the most life-threatening diseases worldwide. The number of patients, as well as the mortality rate of these diseases, are increasing across the globe [1]. This disease is more common in males than in females, and children [2]. Due to CAD, approximately 17.9 million deaths take place every year globally [3, 4]. Large amount of data is helpful in providing hidden patterns [5]. The number of deaths is nearly one-third of the total number of deaths across the globe. Further, it is claimed that 50% of all patients diagnosed with heart diseases die within 1-2 years. Also, 3% of the total healthcare expenditure is spent on treating this disease [6, 7]. Experts of CAD are in high demand in healthcare organizations. In developing countries lack of experts, poorly developed technology, and lack of resources may result in misdiagnosis of cardiovascular disorders. Furthermore, the rising volume of data may provide an opportunity to learn the new information or hidden patterns from the data [8].

Also, it is difficult to diagnose the chronic condition of the disease early enough for a patient to recuperate without suffering a heart attack. Technologies like angiography used to diagnose CAD are extremely expensive. Moreover, it is a time-consuming procedure. Thus, it may prove life-threatening for a person reaching a chronic degree of the disease. It also requires several pre-tests such as blood sugar, cholesterol, and an Electrocardiogram (ECG). As the number of people diagnosed with coronary artery disease rises, so does the mortality rate. The ineffectiveness of traditional approaches in early diagnosis, as well as diagnostic mistakes, contribute to the higher fatality rate. Traditional methods, such as angiography, are also renowned for their high cost and adverse effects, including allergic reactions and renal damage. Because of the high mortality rate, it is critical to identify CAD at an early stage to save people's lives. Thus, the demand for an alternative and intelligent solution arises to minimize the delay in diagnosis. Machine Learning (ML)-based automated systems may be effective tools for early identification of CAD at a lower cost and with the fewest diagnostic mistakes. So far, the suggested study focuses on using ML-based approaches for cardiovascular disease prediction.

Also, ML techniques have the potential to recognize patterns in medical images [9], classify them into different categories [10], reconstruct medical images without distorting their quality [11], and enhance the quality of images with the minimum loss of information [12]. ML techniques are equally competent in data interpretation and making decisions using the trends available in the dataset [13]. Thus, these techniques can be easily adapted for the early diagnosis of CAD, knowing its most significant causes and symptoms.

ML models provide a better and faster approach to solve the CAD problems early and at an acceptable cost. Traditionally, the files of each patient were looked after and eventually stored in a relational database with a table-column structure. As the number of patients grows, the traditional techniques of records management are not viable. Furthermore, the rising volume of data may provide an opportunity to learn new information or hidden patterns from the data. ML has been a burgeoning area in recent

Manuscript received December 14, 2021; revised July 9, 2022; accepted July 12, 2022; published February 14, 2023.

years as a result of the vast volume of data available. This data may be used to identify patterns or knowledge. Applying ML models aid health care providers and patients in detecting CAD at an early stage at a low cost, so that even a low-income individual may afford it.

Because of their versatility, optimization techniques are frequently utilized in a variety of fields such as image enhancement [14], and performance improvement of an ML model. These approaches have a lot of versatility when it comes to dealing with complicated nonlinear issues. Particle Swarm Optimization (PSO) and Firefly algorithm optimization techniques have been used to diagnose cardiac problems, however, hybridization of these two optimization algorithms has not been employed yet. The PSO algorithm is well-known for its ease of use and versatility. However, it is readily trapped in the local optima, resulting in low accuracy. Also, there is a huge scope to improve its performance in dealing with complicated issues such as illness diagnosis, and fraud detection.

Some key findings are addressed in this manuscript to create an efficient model. The authors in this manuscript offer an improved method for the early diagnosis of CAD. To extract the essential characteristics, the system uses the PCA algorithm in the first phase to remove unnecessary and duplicated data. The suggested model's accuracy was enhanced in the second stage using a combination of optimization approaches, including PSO and FA. Rather than hybridization of optimization approaches, researchers concentrated only on distinct ML techniques or optimization strategies. Hybridization of optimization approaches can greatly improve the efficiency of our study. In the last phase, the 'Decision Tree' classification approach is employed to classify coronary artery disease.

The main contributions of this paper are as follows:

- Remove the redundant or irrelevant features by employing Principal Component Analysis (PCA) based feature extraction.
- Optimization by employing a hybrid model of Particle Swarm Optimization (PSO) and Firefly Algorithms.
- Comparative analysis of the performance of model without optimization, using a single optimization technique (FA), and integration of 'FA and PSO' optimization techniques.

The remaining paper is organized as follows. Section II discusses the existing literature in the domain of CAD diagnosis. Section III demonstrates the methodology. Section IV presents the experiments conducted using various ML techniques. Finally, Section V concludes and discusses the future directions.

II. Related Work

This section covers research works done in CAD classifications and their performance evaluations.

Mohan *et al*. [15] proposed a hybrid model by integrating Random Forest and a Linear Model (HRFLM). This model is applied to the UCI Cleveland dataset for improving the performance accuracy of CAD diagnosis. After selecting 13 prominent attributes like age, sex, chest pain, resting blood pressure, cholesterols, Fasting Blood Sugar, Res Electrocardiographic, Max Heart Rate, Exercise-Induced, old peak, slope, Major Vessels, the model reported an accuracy of 88.7% with R Studio Rattle tool which is used for classification purpose.

Latha *et al*. [16] proposed an ensemble classification model to improve the accuracy of weak classifiers by combining multiple classifiers and predicting heart diseases at an early stage as well. Improvement of weak classifiers was achieved up to 7% after employing ensemble methods like bagging, boosting, stacking, and majority voting. Techniques like Support Vector Machine, C4.5, Bayes Net, Naïve Bayes, Multilayer perceptron, and PART with ensemble methods have been employed. These classifiers are implemented on the Cleveland dataset from the UCI machine learning repository. Among all the classifiers, the Logistic Regression reported the highest accuracy of 79.20%. The performance of the work was further improved with a feature selection implementation.

Khateeb *et al*. [17] used different classification techniques such as Naïve Bayes, K- Nearest Neighbor, J48, and Bagging classifiers. They used heart diseases dataset comprising 303 records and 14 attributes. The authors reported an accuracy of approximately 79.20% by using the Re-sampling Weka option.

Pereira *et al*. [18] worked on a Logistic Regression algorithm for the classification of CAD. The algorithm shows an accuracy of 84.15%. The author also applied Deep Learning (DL) processes to the same dataset and observed that DL techniques improved the accuracy and reported the highest accuracy of 91.67%. Also, these techniques reduced the error of classification by 8.33%. An improvement in the value of the Kappa indicator *i.e*. 0.831, indicates an excellent agreement. Attributes like chest pain type, ST depression induced by exercise relative, and thalassemia have a great impact on cardiovascular disease.

Saboji *et al*. [19] proposed a model which diagnose the heart diseases with less number of attributes like ECR, cholesterol, chest pain, fasting sugar, Maximum Heart Rate (MHR), etc. from the Cleveland dataset. The authors applied the random forest algorithm for predicting heart disease and demonstrated that the model reported an accuracy of 98% on the dataset of 600 records. Apache Spark and Hadoop platform is used implementation. In this approach, better accuracy was observed with the limited number of attributes and comparison of proposed model is done against Naïve Bayes.

Uyara *et al*. [20] proposed a Genetic Algorithm (GA) based on Recurrent Fuzzy Neural Networks (RFNN) The objective of their research is to assist health care professionals in early diagnosis of heart diseases. Cleveland dataset comprising 303 samples is used. The dataset is divided into training and testing datasets comprising 252 and 45 samples respectively. 6 samples were ignored from 303 samples due to presence of missing values in them. Among all the employed techniques, Naïve Bayes and SVM showed the best accuracy of 96.63% and 97.78% respectively. Author also compared their proposed model with ANN-Fuzzy_AHP approach which

shows 91.1% accuracy. An improvement of 6.68% proves the supremacy of the model.

Feizi-Derakhshi *et al*. [21] reduced the number of features with enhanced feature selection techniques. The authors employed the KNN algorithm and Imperialist Competitive Algorithm along with a meta-heuristic approach. They designed an automatic heart disease detection system that is suitable for health care organizations for categorizing patients. They used the dataset of heart disease collected from Tehran Shahid Rajaei hospital. Comparison of the proposed model is done with SMO, Neural Network, Naïve Bayes, and Bagging SMO, and proved the supremacy of their work. In the future, the feature selection methods can be employed for handling the missing data.

Khourdifi *et al*. [22] detected cardiovascular diseases using hybridization of optimization techniques (Ant Colony Optimization algorithm (ACO) and Particle Swarm Optimization (PSO)) to improve the accuracy of the proposed model. Fast Correlation-Based Feature Selection (FCBF) is used for the filtration of redundant features. The overall accuracy of the proposed model was 99.65%. KNN, Random Forest, Artificial Neural Network (ANN), SVM, and Naïve Bayes classification techniques have been used.

Abdar *et al*. [23] tested ten traditional ML algorithms followed by SVM-based algorithms on the Z-Alizadeh-Sani dataset. The objective of the study is to detect the CAD with a better performance compared to classical machine learning techniques. GA and PSO methods coupled with 10-fold cross-validation. Their proposed algorithm N2Genetic-nuSVM shows the highest accuracy of 93.08% among these techniques. Next, preprocessing approaches such as standardization, or mean removal and variance scaling, can be applied to improve the performance of the employed techniques instead of normalization along with other evolutionary techniques.

Hassannataj *et al*. [24] applied various classification techniques like SVM, Chi-Square Automatic Interaction Detection (CHAID), Decision Tree (C5.0), and Random Forest algorithm (RF) on the Z-Alizadeh Sani dataset. They considered chest pain as the most prominent predictor. Among all, random forest outperforms with an accuracy of 91.40 %. However, this research increased the accuracy of CAD diagnosis by selecting significant predictive features, fuzzy intelligent systems can be used in combination with AI models to diagnose CAD.

Next, the authors in reference [25] employed six ML algorithms namely Logistic Regression, Support Vector Machine, Gaussian Naive Bayes, Decision Tree, K-Nearest-Neighbor and Neural Networks to predict heart disease. They used Cleveland Heart Disease dataset taken from the UCI repository and containing 24 attributes. The Logistic Regression reported the highest accuracy of 86.8%. Its low accuracy is a limitation in adopting this technique for real life. Thus, there is a huge scope to improve the accuracy of prediction of CAD.

Shah [26] applied Naïve Bayes, K-NN, Decision tree, and Random forest algorithms which reported the accuracy of 88.157%, 90.789%, 80.263%, and 86.84% respectively on the Cleveland dataset. The dataset consists of 300 records and 14 attributes. The approach implemented in is limited to calculation of accuracy for evaluating the performance of the models.

Padmajaa [27] *et al*. developed a model to reduce death rates by early prediction of cardiovascular diseases using numerous classification techniques like Logistic Regression, Random Forest, Gaussian Naïve Bayes, Gradient boosting, K-NN algorithm, Multinomial neighbors, SVM, and Decision trees. The authors used feature selection algorithm in order to reduce execution time and improve performance of classifiers. Among all the above mentioned algorithms, random forest gives the best accuracy on the Cleveland dataset.

Khamparia *et al*. developed a model with the combination of ML and DL in order to predict the heart diseases [28]. Classification techniques have been applied on various datasets such as Cleveland, Hungary, Switzerland, and Long Beach. Lasso algorithm is used for feature selection and choosing important features. The integration of DL with Logistic regression reported the highest accuracy of predicting CAD. The comparison in the performance of models used in literatures is shown in Table I, and the comparative analysis of techniques proposed in literature is discussed in Table II. It is evident from the results shown in Table I, and Table II that deep learning techniques report the highest accuracy of 94.2%. Thus, there is a scope of improving the accuracy of CAD prediction. Moreover, none of the existing techniques focus on proving automatic CAD detection for patients with renal impairment and catheter implanted.

TABLE I. COMPARISON IN PERFORMANCE OF STATE-OF-THE-ART MODELS

Algorithm Used	Accuracy (%)	Specificity (%)	Sensitivity (%)
Logistic Regression	83.3	82.3	86.3.
KNN	84.8	77.7	85.0
Random forest	80.3	78.7	78.2
Decision Tree	82.3	78.9	78.5
SVM	83.2	78.7	78.2
Deep Learning	94.2	83.1	82.3

TABLE II. COMPARATIVE ANALYSIS OF TECHNIQUES EMPLOYED FOR CAD DETECTION

Author(s)	Techniques	Objectives/Advantage of the research	Dataset	Performance	Future scope
Mohan *et al*. [15]	'HRFLM', a hybrid of random forest and linear method.	Developing a hybrid model of Random Forest and a Linear Model (HRFLM) for the better prediction of heart diseases.	Cleveland dataset with R Studio Rattle	Accuracy of 88.7%.	Integration of ML techniques with feature selection methods may be used to improve accuracy of CAD diagnosis.

Latha *et al*. [16]	SVM, C4.5, Bayes Net, Naïve Bayes, Multilayer Perceptron, and PART with Ensemble Methods.	Using ensemble classification models for improvement of the accuracy of weak classifiers in order to detect heart diseases at a premature stage.	Cleveland heart dataset	Logistic Regression reported highest accuracy of 79.20%.	Employing different ensemble algorithms for early diagnosis of CAD.
Khateeb [17]	Naïve Bayes, IBK K-Nearest Neighbor, J48, and Bagging classifiers /ML Techniques	Reviewed multiple classifiers implemented on the dataset and found quite good accuracy from different feature reduction methods.	Heart Disease Dataset.	Highest accuracy of 79.20%	Employing feature reduction methods before classification.
Pereira *et al*. [18]	Logistic regression, Naïve Bayes, Random Forests, k-NN, SVM, Area under the ROC curve, Deep Learning.	Classify data to reduce the bias of the estimation, which improves heart prediction accuracy.	Heart Disease UCI Data Set	Logistic Regression reported the highest accuracy of 84.15%.	Increasing the number of relevant attributes for further improvement in performance.
Saboji *et al*. [19]	Random Forest algorithm implemented on Spark framework.	Predict heart diseases with a limited number of attributes and good accuracy	Data collected from Cleveland, Switzerland, and Hungary.	Accuracy of 98%	Scope of improving the accuracy.
Uyara *et al*. [20]	Genetic algorithm based trained recurrent fuzzy neural networks	Precise prediction of heart diseases with better results.	Cleveland heart disease dataset	Accuracy of 97.78%	Scope for inclusion of pathology history of patients for reliable prediction.
Feizi-Derakhshi *et al*. [21]	Imperialist Competitive algorithm with Meta-heuristic approach used to optimize feature selection and K-nearest Neighbor algorithm.	To develop an automatic system to diagnose heart disease, and classify the patients with less numbers of attributes.	Cleveland dataset.	Accuracy of 91.53%.	Feature selection method can be used for dealing with incomplete and missing data.
Khourdifi and Bahaj [22]	Fast Correlation based feature selection, and SVM, Naïve Bayes, Random Forest, KNN classifiers, integrated with optimization techniques namely ant colony optimization and particle swarm optimization.	Heart disease detection using classification algorithms, and optimization techniques PSO and ACO.	Heart Disease Dataset	Accuracy of 99.65%.	Scope to reduce the computation efficiency of proposed methods.
Abdar *et al*. [23]	N2Genetic optimizer based on the fusion of 10-fold cross-validation with GA or PSO is employed and SVM for classification.	Optimizing feature selection and for predicting CAD.	Z-Alizadeh-Sani dataset.	Highest accuracy of 93.08%.	Preprocessing approaches such as standardization, mean removal, and variance scaling can be applied for improving accuracy of prediction.
Fernandes and Freitas [25]	Logistic Regression, Support Vector Machine, Gaussian Naive Bayes, Decision Tree, K-Nearest-Neighbor, and Neural Networks for CAD prediction	Increasing accuracy of heart disease prediction.	Cleveland Heart Disease dataset.	Highest accuracy of-86.8% by Logistic Regression.	Scope to improve accuracy.
Shah, Patel, and Kumar [26]	Naïve Bayes, Decision Tree, K-Nearest Neighbour, and Random Forest algorithm.	To predict heart diseases efficiently and accurately with few attributes.	Cleveland dataset	Highest accuracy of 90.789% by K-NN classifier.	Integrating models for improving the accuracy.
Bharti *et al*. [27]	K nearest neighbour, SVM, Random forest, Decision tree and DL	Heart disease prediction with using ML techniques	Cleveland, Hungary, Switzerland, and Long Beach.	Accuracy of 94.2% using Deep Learning technique	Scope to improve accuracy by fine-tuning the hyperparameters, and embedding optimization techniques.

III. Methodology

This section includes the details about the proposed classification system. In this research, the authors employ the decision tree for classification because it is an efficient hierarchical tree structure for handling nonlinear problems. Also, it reports low classification error rates. Further, the hybridization of PSO and FA algorithms has been employed for the optimum feature selection. The workflow of the proposed model is shown in Fig. 1.

Fig. 1 shows the flow diagram of the proposed work. It shows the role of processing and normalization of the data before performing the classification process. The data needs to be sorted efficiently therefore normalized feature vector will be extracted, Feature optimization is done to extract the relevant information to be utilized by the classifier in decisions making. From this, it can be noticed

how much performance can be optimized to predict CAD and model validation. Once the CAD and non-CAD are predicted successfully, the performance is evaluated in terms of precision, recall, and accuracy.

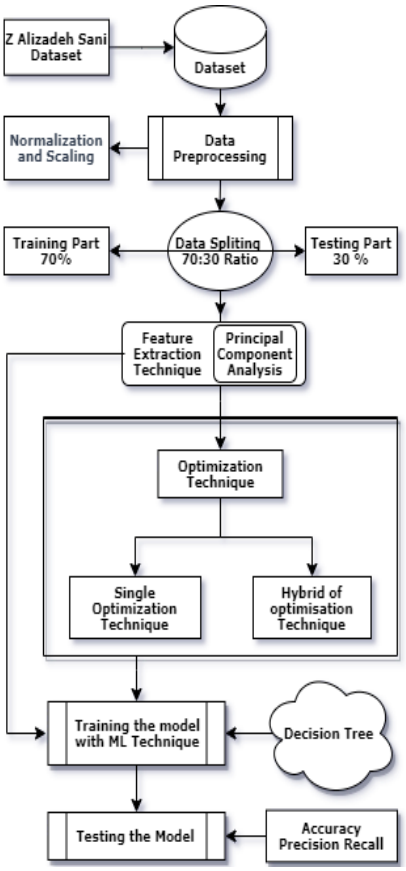


Figure 1. Overview of the proposed approach.

A. Data Collection

Dataset is taken from the UCI Machine Learning Repository site. Z-Alizadeh Sani dataset is used for training and evaluating the performance of the developed model. This dataset consists of 303 records comprising 216 patients with CAD, and 88 Normal samples. More than 50% narrowing of artery represents a CAD patient whereas the remaining samples are labeled as Normal. The dataset includes 56 attributes of each sample. The values in Z-Alizadeh-Sani dataset are numeric and string as shown in Table III.

TABLE III. SAMPLE DATASET

Length	Sex	BMI	DM	HTN	Current_ Smoker	EX_ Smoker	FH	Obesity	CRF	CVA
163	Female	27.099251	0	0	0	0	0	Y	N	N
145	Female	25.2080856	0	1	0	0	0	Y	N	N
167	Female	28.6851447	0	0	0	0	0	Y	N	N
165	Male	23.8751148	0	0	1	0	0	N	N	N
164	Male	22.3881499	0	0	0	1	0	Y	N	N
152	Female	32.4619114	1	1	0	0	0	Y	N	N
150	Female	31.5555556	1	1	0	0	0	Y	N	N
160	Female	25.390625	0	1	0	0	0	Y	N	N
152	Female	28.566482	0	0	0	0	0	Y	N	N
160	Female	22.65625	0	1	0	0	0	N	N	N
160	Female	28.90625	0	1	0	0	0	Y	N	N

B. Data Pre-processing

Pre-processing involves data cleaning and normalization. In the above-stated dataset, no missing values were found. Therefore, only normalization and scaling were performed. Features are scaled in the interval [0, 1]. After performing normalization, strings are converted into numeric values ‘0’ and ‘1’. For example, the male and female labels have been transformed into 0 and 1 respectively. Similarly, the yes for the value of a feature is labeled as ‘1’ and the no labeled as ‘0’.

C. Feature Extraction

There are thousands of features in the dataset that results in inconsistencies and redundancies. Also, redundant features increase the computation time. Thus, these redundant features are removed and highly correlated features are coupled together to improve the accuracy of the model. So, there is a need for feature extraction algorithms to get rid of these problems. PCA is used for feature extraction. It is one of the efficient feature extraction techniques that help to identify correlations and patterns in a dataset. It can transform the input dataset into a new dataset that has significantly low dimensions without the loss of such information. Narrowing down a couple of variables from the original dataset. Various steps are involved such as 1) Standardization of data: scaling of the data is done in such a manner that variables and their values lie within a similar range of variable value-mean/standard deviation 2) computing covariance matrix which expresses the correlation between different variables in a dataset. It is necessary to remove the dependent feature because it contains biased and redundant information which reduces the performance of the proposed model. Covariance can be negative or positive. 3) Calculating the eigenvectors and eigenvalues which are computed from the covariance matrix to determine the principal components of the data set. 4) Computing principal components mean that the highest eigenvalue has the most significant feature and forms the first principal component. 5) Last step is to rearrange the original data with the final principal components which represent the maximum and most significant information of the dataset.

D. Feature Optimization

This step deals with the combination of Swarm Particle Optimization and Firefly Nature-Inspired Algorithm for the instance selection which selects the relevant feature vector and finds the best among different possible solutions for the classification step.

1) Firefly algorithm

The firefly algorithm is inspired by the flashing behavior of fireflies. Various assumptions are there for this algorithm like the attraction of fireflies towards each other, attraction is proportional to brightness, less bright is attracted to the brighter fireflies, if the brightness for both is the same, fireflies move randomly and new solutions are generated by a random walk and attraction of fireflies. FA is applied in nonlinear problems, dynamic problems, feature selection, fault detection many more.

There are numerous advantages of FA over other optimization algorithms. 1) Automatic subdivision of the

whole population into subgroups. 2) Natural capability of dealing with multi-model optimization. 3) High randomness in the solutions.

Basic steps of Firefly Algorithm

- Initialize the parameters.
- Generate the population of n fireflies
- Calculate the fitness value of each firefly
- Check if (t=1 to max t)
- Update the position and light intensity for each firefly.

2) Particle Swarm Optimization (PSO)

It is a metaheuristic algorithm that solves hard optimization problems based on complex problems. This algorithm is inspired by the social behavior of flocking birds. It is a population-based stochastic search algorithm. PSO solves the problems by having a population of candidate solutions (particles). Each member of the population is called a particle and the population is called a swarm. Each particle has velocities. There are various advantages of PSO due to which it is quite popular such as easy to implement, few usages of parameters that can be used as optimal functions, ANN training, fuzzy control system. PSO is successfully applied to the heart disease data showing outstanding results. The objective function is used in metaheuristic algorithms to maximize or minimize the values that you are trying to optimize.

Basic steps of PSO

(1) Initialization (Initialize parameters and population).
(2) Evaluate the fitness value of each particle and select the best particle among them.
(3) Find velocity and position for each particle.
(4) Evaluate fitness (Find the current best).
(5) Update t=t+1.

E. Classification Technique (Decision Tree)

A decision tree is an efficient hierarchical tree structure used to perform various decision-making procedures. It includes the condition on the internal node, as a result of which the tree is divided into edges. So, the decision is made on the branch that stops splitting anymore which is the final decision of the decision tree classification process. In this technique, all extracted features are measured and dissimilar split arguments are practiced and tested with a cost function. The selection is made on the splitting as per the best cost i.e. lowest cost. In the proposed work we have used the pruning process for the decision making which can increase the performance of a tree. The proposed algorithm is shown in section G. It comprises removing the unnecessary decisions on the classification by reducing the branches or having low importance. This will reduce the complexity, and thus cumulative its predictive influence by decreasing the overfitting problem which most of the time machine learning model encounters. The pruning is done at the leave's node and eliminates each node by most standard classes. It is also known as error reduction. More refined pruning approaches can be recycled by introducing a learning constraint which is used to consider whether participating nodes can be detached based on the magnitude of the sub-tree. Also, the entropy is evaluated before and after splitting so that the disorder among the data points and nodes decision will be evaluated and is calculated as given in Eq. (1).

$$E(x) = -\sum_{k=i}^{n} X(i) log2\,[X(i)] \tag{1}$$

where *n* is the number of groups and *X(i)* is the belonging group probability.

Information Gain is one of the crucial parts of the decision-making process in the decision tree which is evaluated as shown in Eq. (2).

$$I(g) = E(Pn) - E(Cn) \tag{2}$$

where $I(g)$ is the information gain, $E(Pn)$ is the entropy in the parent node and $E(Cn)$ is the entropy evaluated for the child node based on the weighted average.

F. Estimating The Performance: K-Fold Cross-Validation Method evaluation

Coronary artery disease detection is seen as a classification problem. Decision tree algorithms have been used for classification purposes. Performance analysis is done with one optimization technique and with the hybridization of two nature-inspired optimization algorithms. It is determined that DT (Decision Tree) is performing outstanding with the hybridization of optimization methods. The most widely classified techniques used for CAD detection are ANN, DTs, and SVM which are applied to most the cardiovascular disease datasets. ANN, DTs, and SVM are considered the top three methods that are showing outstanding results and are easy to use along with the low computation burden that is reported in the literature. There are numerous advantages of decision trees like 1) Dealing with missing data 2) Dealing with irrelevant data 3) Fast in computations.

G. Performance Measure Indices

The model's performance is now assessed using performance matrices like accuracy, precision, and recall, as well as the 10-fold cross-validation technique. For the implementation of these approaches, Pycharm editing techniques were utilized.

Performance can be evaluated in terms of accuracy, precision, and recall. In the case of positive classification, a person is classified as a CAD patient, similarly, in the case of negative classification, a person is classified as a non CAD patient.

TP (True Positive)-The number of instances correctly classified to the CAD class.

TN (True Negative)-The number of instances correctly classified to the non-CAD class.

FP (False Positive)-The number of instances from the Non-CAD class is incorrectly classified as the CAD class. For example, identifying non CAD patients as CAD patients.

FN (False Negative)-The number of instances from the CAD class is incorrectly classified as the Non-CAD class.

For example, identifying CAD patients as normal patients.

Accuracy-It shows the observations which are correctly identified from all prospective out of all the observations as mentioned in Eq. (3).

The high accuracy signifies the ratio of true positive and true negative to the summation of all instances of the population.

Precision- The precision is the ratio of the true positive concerning to the summation of the true positive and false negative as mentioned in Eq. (4).

Recall- Recall shows the ratio of the true positive concerning the summation of the true positive and false negative as mentioned in Eq. (5).

$$Accuracy\ (Acc) = \frac{(TP + TN)}{(TP\ +\ TN\ +\ FP\ +\ FN)} \quad (3)$$

$$Precision\ = \frac{TP}{(TP + FP)} \quad (4)$$

$$Recall\ or\ Sensitivity\ = \frac{TP}{(TP\ +\ FN)} \quad (5)$$

H. Proposed Algorithm

Step 1: Input Data such that D = D_1, D_2… D_N as data and perform the data framing to process data smoothly.

Step 2: Normalize & scale the data.

Step 3: Generate the feature extraction F(x) & transform to generate the covariance process.

Step 4: Extract the Eigenvalues and vectors to extract meaningful insights for the transformation T = X W for the data mapping (M) to generate a new space vector E(x).

Step 5: Optimize the feature extraction using optimization methods (Firefly Algorithm and Particle Swarm Optimization for the dimensionality reduction process.

Step 6: Generate splitting of the data into 70% (Training data) and 30% (Testing data) ratios.

Step 7: Generate the Classification Decision Model.

Step 8: Upload test samples.

Step 9: Load the trained model and perform classification.

Step 10: Estimate the performance of the model in terms of Precision, Recall, and Accuracy.

IV. Experiments and Results

The proposed work is simulated using a virtual environment in the python language. Pycharm editor tool is used for the simulation of the real-time environment by taking care of all dependencies and without affecting the simulation environments. For the classification of the CAD, the decision tree classifier is used. It performs the bagging of the data to obtain high classification rates.

The demonstration of CAD detection using a single optimization method, hybridization of optimization methods, and without optimization techniques using a decision tree classifier is done.

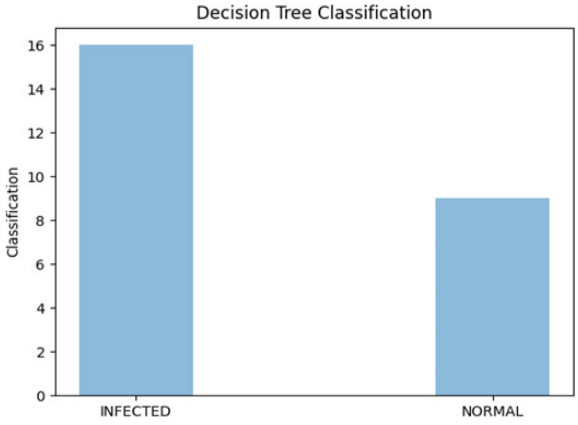


Figure 2. Classification.

Fig. 2 shows the classification count in which it can be seen the categorization of the normal and infected people. The y-axis in the above figure are showing the total count of infected and non-infected from the test set. The normal people are the ones who have no CAD (Coronary Artery Disease) and the infected are the ones having CAD. This is automatically classified using a training and testing process in which the data is trained with some feature extraction and optimization process and the decisions are made using a tree structure in the hierarchical form which gives the classified result and is helpful in the medical diagnosis.

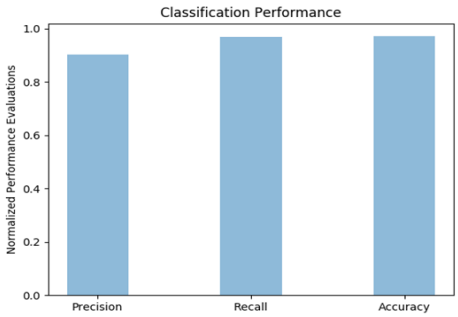


Figure 3. Performance evaluations.

Fig. 3 shows the performance evaluations using precision, recall, and accuracy. It can be seen from the classification results that the precision and recall are giving appropriate output for the classified results and based on which the accuracy for the true positive rates also increases. Precision and recall of the proposed classification are high to get low false positive and negative rates which should be low for low classification error rates and increases the authenticity of the system.

Fig. 4 gives the accuracy performance which is divided into three different phases. The very first is the without optimization phase in which it can be seen that the performance is not giving high rates for the classified results for the extracted features. The second phase is the optimization phase in which one optimization process is used for the classification. It can be seen that with one optimization approach the accuracy results got improved and this is further improved using the hybrid optimize the

approach to get the best possible solutions in terms of the instance selections for the true positive rates. The accuracy achieved with the hybrid optimization is 95.87% which is an appropriate outcome for our proposed work.

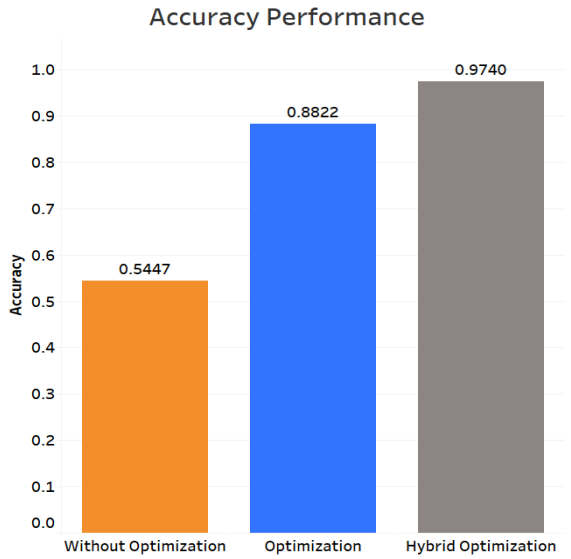


Figure 4. Accuracy comparison.

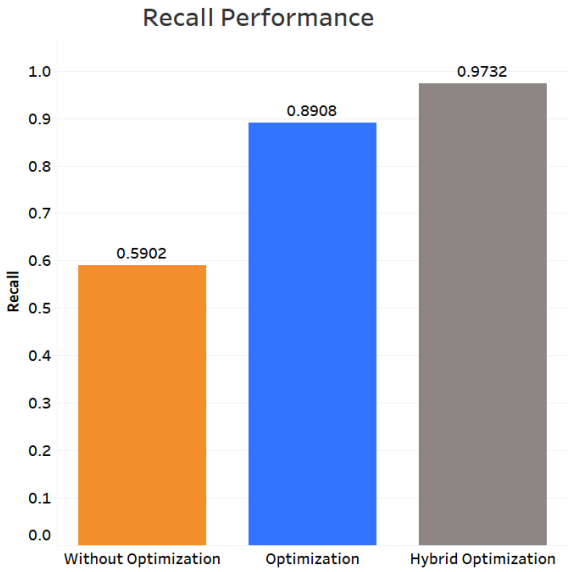


Figure 5. Recall comparison.

Fig. 5 presents the recall comparison of the three phases. It can be seen that the recall is coming high which means that the proposed approach can retrieve efficient information from the training process with low classification error rates to achieve low false positive and negative rates. It is also necessary for accurate information retrieval because the testing phase is completely dependent on the training and validation process and if the training error rate increases it means the model is under fitting and is not able to recall relevant information which must be reduced for the high gaining of the information. It can be seen from the figure that the recall performance of the hybrid optimization is 0.93 in a normalized form which is an appropriate outcome for our proposed work. It explains the relevance of the classification task.

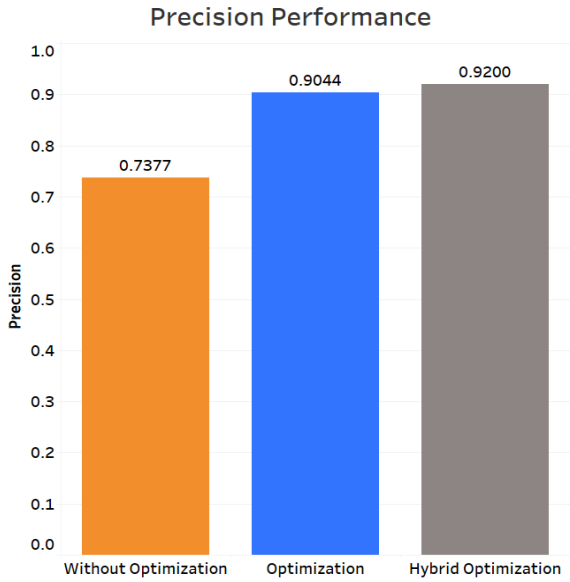


Figure 6. Precision comparison.

Fig. 6 demonstrates the precision comparison of the three phases. It can be seen that the precision is high for the proposed approach. This indicates the quality performance in terms of the selection of the instances. The precision must be high for the high accuracy and low classification error rates. Precision is deeply dependent on the type-1 error which should be reduced and controlled through the regularization methods. It will give evaluations for the case if the model will go in the overfitting process. The proposed hybrid approach achieves a precision of 0.94. Precision provides the relevance of the classes from all observations. It shows the belongings of the predicted positive classes which belong to the positive classes in the actual data.

The confusion matrix gives the visualization of the performance of the algorithm as given in Table IV.

TABLE IV. CONFUSION MATRIX

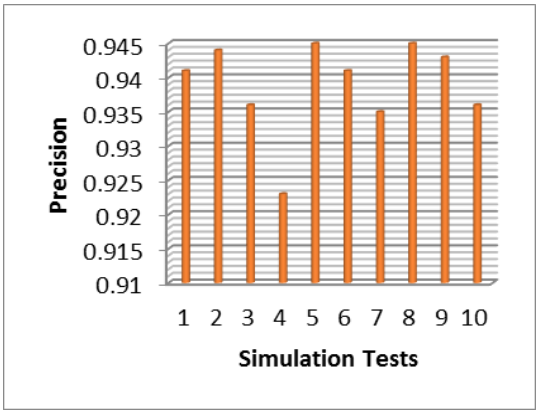
Classes	Actual (Heart diseases=Yes)	Actual (Heart Diseases = No)
Predicted (Heart Diseases= Yes)	TP (210)	FP (4)
Predicted (Heart Diseases= No)	FN (7)	TN (84)

TABLE V. RECALL, PRECISION AND ACCURACY PERFORMANCE

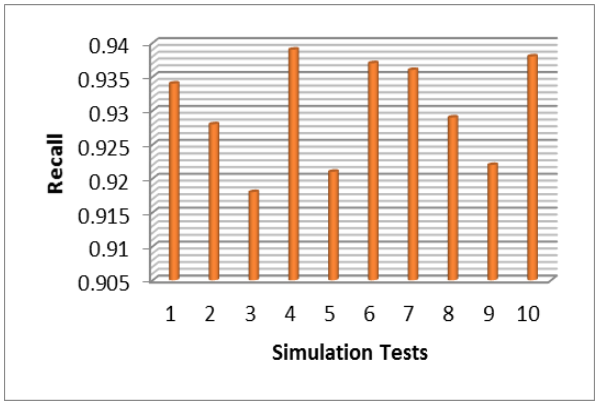
Test No.	Recall	Test No.	Precision	Test No.	Accuracy
1	0.934	1	0.941	1	0.941
2	0.928	2	0.944	2	0.955
3	0.918	3	0.936	3	0.953
4	0.939	4	0.923	4	0.943
5	0.921	5	0.945	5	0.957
6	0.937	6	0.941	6	0.958
7	0.936	7	0.935	7	0.932
8	0.929	8	0.945	8	0.958
9	0.922	9	0.943	9	0.951
10	0.938	10	0.936	10	0.949

Table V presents the performance analysis on different simulation runs on the PYCHARM editor for the machine

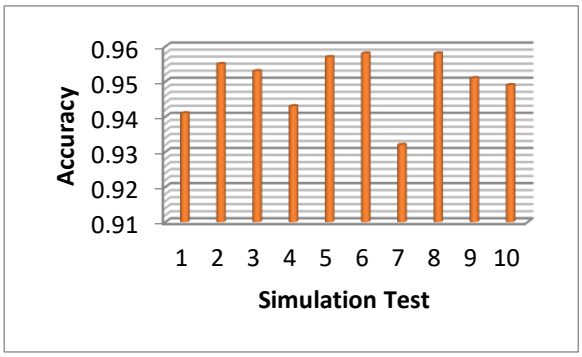
learning model to classify the CAD diseases. It can be seen that there is a very minute difference in the performance in different runs. The model achieves low false positive rates and false rejection rates to attain high true positive and negative rates with low classification error rates.



(a)



(b)



(c)

Figure 7. Precision (b) recall (c) accuracy.

Fig. 7 illustrates the performance evaluations on various tests simulations done for the proposed classification model. It can be seen from the figure that simulations are run on the Pycharm editor tool to check the validity of the data on different iterations. The results show very minute difference among all the different runs in terms of precision, accuracy, and recall. This indicates that the model variances are very less and are having low standard deviations. The comparative analysis of the results is shown in Table VI.

TABLE VI. PERFORMANCE COMPARISON

Parameter	Accuracy
Pereira [28]	82.46%
Wang *et al*. [29]	92.59%
Derakhshi [30]	94.43%
Pamungkas [31]	86.67%
Proposed	95.87%

The proposed model comparison is shown in Fig. 8 below.

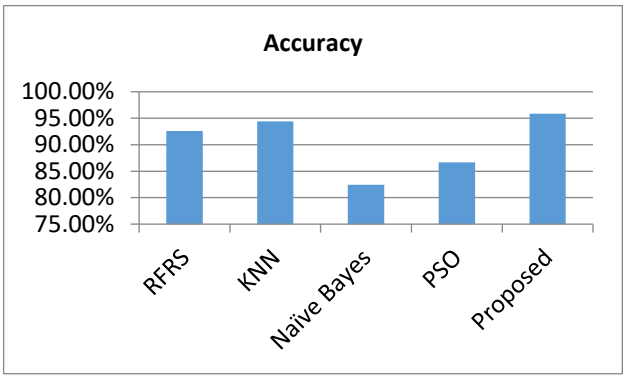


Figure 8. Proposed model comparison.

V. CONCLUSION AND FUTURE WORK

This research proposes an enhanced machine-learning approach for the early detection of CAD. The performance of ML approaches used for CAD detection is compared 'without utilizing optimization techniques', 'using a single optimization algorithm', and using a 'hybrid of more than one optimization technique'. The results show that combining a decision tree-based classification technique with a hybrid of Particle Swarm Optimization and Firefly Algorithm with Principle Component Analysis predicts CAD with a 95.87% accuracy. The classification accuracy is higher than the maximum accuracy reported in state-of-the-art models, which is 94.43 percent. The approach provides a low-cost CAD detection solution. It is also beneficial to people who have renal impairment or bleeding where the catheter is implanted. Thus, the technique can be adopted to develop an automatic and assisting tool for CAD detection. As a result, it can be used to create an automatic and assisted CAD detecting tool. Furthermore, there is a lot of potential in the methodologies given for CAD identification, such as using multiple ML classifiers, as well as alternative feature extraction techniques and optimization techniques, to improve classification accuracy and reduce computation time.

CONFLICT OF INTEREST

The authors declare no conflicts of interest associated with this manuscript.

AUTHOR CONTRIBUTIONS

Savita mainly conducted this research with the implementation of the proposed algorithm; Geeta

Chhikara and Apeksha Mittal helped in refining the paper; all authors finally approved the final version.

REFERENCES

[1] C. Trevisan, G. Sergi, S. J. B. Maggi, *et al.*, "Gender differences in brain-heart connection," in *Brain and Heart Dynamics*, Cham, Switzerland: Springer, 2020, p. 937.

[2] M. S. Oh and M. H. Jeong, "Sex differences in cardiovascular disease risk factors among Korean adults," *Korean J. Med*., vol. 95, no. 4, pp. 266-275, Aug. 2020.

[3] World Health Organization and J. Dostupno. (2016). Cardiovascular diseases: Key facts. [Online]. 13(2016), p. 6. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[4] K. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Comput. Sci*., vol. 120, pp. 588-593, Jan. 2017.

[5] S. Arora, M. Agarwal, and S. Mongia, "Comparative analysis of educational job performance parameters for organizational success: A review," in *Proc. the International Conference on Paradigms of Computing, Communication and Data Sciences. Algorithms for Intelligent Systems*, 2021.

[6] A. U. Haq, J. P. Li, M. H. Memon, *et al.*, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Inf. Syst*., vol. 2018, pp. 1-21, Dec. 2018.

[7] S. Arora, "An empirical study - The cardinal factors towards recruitment of faculty in higher education institutions using machine learning," in *Proc. 8th International Conference on Signal Processing and Integrated Networks*, 2021, pp. 491-497.

[8] S. Arora and M. Agarwal, "Empowerment through big data: Issues and challenges," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 2, pp. 423-431, 2018.

[9] N. Pradhan, G. Rani, V. S. Dhaka, *et al.*, "Diabetes prediction using artificial neural network," in *Deep Learning Techniques for Biomedical and Health Informatics*, B. Agarwal, V. E. Balas, L. C. Jain, R. C. Poonia, Manisha, Eds., Academic Press, 2020, pp. 327-339. https://doi.org/10.1016/B978-0-12-819061-6.00014-8

[10] G. Rani, M. G. Oza, V. S. Dhaka, *et al.*, "Applying deep learning-based multi-modal for detection of coronavirus," *Multimedia Systems*, 2021. https://doi.org/10.1007/s00530-021-00824-3

[11] N. Pradhan, V. S. Dhaka, G. Rani, *et al.*, "Transforming view of medical images using deep learning," *Neural Comput. & Applic.*, vol. 32, pp. 15043-15054, 2020. https://doi.org/10.1007/s00521-020-04857-z

[12] N. Kundu, G. Rani, and V. S. Dhaka, "Machine learning and IoT based disease predictor and alert generator system," in *Proc. Fourth International Conference on Computing Methodologies and Communication*, 2020, pp. 764-769. doi: 10.1109/ICCMC48092.2020.ICCMC-000142

[13] G. Rani, M. G. Oza, V. S. Dhaka, *et al.*, "Applying deep learning-based multi-modal for detection of coronavirus," *Multimedia Systems*, 2021. https://doi.org/10.1007/s00530-021-00824-3

[14] G. Rani and M. Agarwal, "Contrast enhancement using optimum threshold selection," *IJSI*, vol. 8, no. 3, pp. 96-118, 2020. http://doi.org/10.4018/IJSI.2020070107

[15] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *Special Section on Smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet of Things*, vol. 7, 2019.

[16] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, 100203, 2019.

[17] N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique," in *Proc. the International Conference on Big Data and Internet of Thing*, 2017, pp. 21-26.

[18] N. Pereira, "Using machine learning classification methods to detect the presence of heart disease," master's dissertation, Technological University Dublin, 2019.

[19] R. G Saboji and P. K. Ramesh, "Scalable solution for heart disease prediction using classification mining technique," *International Conference on Energy, Communication, Data Analytics and Soft Computing*, 2017.

[20] K. Uyara and A. İlhana, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," in *Proc. 9th International Conference on Theory and Application of Soft Computing, Computing with 24-25, Computing with Words and Perception*, Budapest, Hungary, August 2017.

[21] J. Nourmohammadi-Khiarak, M. Feizi-Derakhshi, K. Behrouzi, *et al.*, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health and Technology*, vol. 10, no. 3, pp. 667-678, 2020.

[22] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, 2019.

[23] M. Abdar and W. Ksiazek, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 179, 104992, 2019.

[24] J. H.j Joloudari and E. H. Joloudari, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, 2020.

[25] A. R. Fernandes and G. J. L. Freitas, "Heart disease prediction and classification using machine learning," *Foundations of Artificial Intelligence*, vol. 12, no. 1, 2022.

[26] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, p. 345, 2020. https://doi.org/10.1007/s42979-020-00365-y

[27] B. Padmajaa, C. Srinidhib, K. Sindhuc, *et al.*, "Early and accurate prediction of heart disease using machine learning model," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 6, 2021.

[28] R. Bharti, A. Khamparia, M. Shabaz, *et al.*, "Prediction of heart disease using a combination of machine learning and deep learning," *Hindawi Computational Intelligence and Neuroscience*, vol. 2021, 8387680, 2021. https://doi.org/10.1155/2021/8387680

[29] X. Liu, X. Wang, Q. Su, *et al.*, "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Computational and Mathematical Methods in Medicine*, 2017.

[30] J. N. Khiarak, M. F. Derakhshi, K. Behrouzi, *et al*., "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health and Technology*, pp. 1-12, 2019.

[31] S. H. Wijaya, G. T. Pamungkas, M. B. Sulthan, "Improving classifier performance using particle swarm optimization on heart disease detection," in *Proc. International Seminar on Application for Technology of Information and Communication*, 2018, pp. 603-608.

Savita is pursuing Ph.D from GD Goenka University in the area of Artificial Intelligence (Machine Learning). She has completed MCA (Master in Computer Application) from KIIT College of Engineering Gurugram in 2013. Experience of teaching as an adjunct faculty in KR Mangalam University and Gurugram University (Gurugram) and currently working as an asst professor at DPG Degree College. Key research areas are machine learning, Deep Learning, and Artificial Intelligence, Web User Profiling and Recommender System.

Geeta Rani has 12 years' experience in teaching at renowned organizations namely Manipal University Jaipur, NSIT, NIT and GDGU. Proficient in Image Processing, Machine Learning, Web User Profiling and Recommender Systems. Eight patents have been published. Eighteen copyrights are registered for the software works. Expertise in writing and filing IPR, in Govt. of India. Qualified test of women scientist in IPR. Published papers in SCI and SCOPUS indexed Journals. Chief Editor of the book "Disease Prediction using Machine Learning". Published many book chapters in SCOPUS indexed book series. Published twelve articles in SCI and SCOPUS indexed Journals. Conducted conferences and Hackathons. Presented research ideas at many conferences. Delivered talks as distinguished speakers at several workshops, FDPs and conferences.

Apeksha Mittal received doctorate in Computer Science Engineering from University School of Information, Communication & Technology, Guru Gobind Singh Indraprastha University, Delhi, in 2021 in the area of Artificial Neural Networks and received a B.Tech. in Computer Science Engineering from Guru Gobind Singh Indraprastha University, Delhi in 2013 and M.Tech in Computer Science from Banasthali University, Rajasthan in 2015. Currently working as Assistant Professor in GD Goenka University, Gurugram. Her key research areas are Artificial Neural Networks, Deep Learning, Machine Learning and Artificial Intelligence.