DOI: 10.6977/IJoSI.202209_7(3).0004

Detection of CAD using optimization approach with machine learning classification techniques

Savita[1], Geeta Rani [1*] and Apeksha Mittal[2]
[1] Department of Computer Science, GD Goenka University, Gurugram
[2] Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, India
Department of Engineering and Sciences, GD Goenka University Gurgaon, India
[*] Corresponding author E-mail: geetachhikara@gmail.com

(Received 3 February 2022; Final version received 4 August 2022; Accepted 23 August 2022)

Abstract

Coronary Artery Disease is one of the most serious ailments among cardiovascular disorders. High expenses of its' detection using conventional tools like angiography are the key concerns for patients. Supervised machine learning models for automatic classification of the Coronary Artery Diseases are highly accurate, optimal, and cost effective ways. Accuracy of machine learning techniques is dependent on amount and type of data used for training. A large amount of data generated in healthcare organizations help in assisting re-searchers in order to quickly and accurately diagnose the problem using machine learning techniques. Alt-hough, these techniques are effective in diagnosing coronary artery diseases, requires optimization for im-proving the accuracy. In this research, the researchers the authors propose an optimum machine learning based classification approach. They integrate Independent Component Analysis and Principal Component Analysis algorithms for feature extraction, and hybrid of Particle Swarm Optimization and Firefly Algorithm for feature optimization. The performance evaluation, and comparative analysis of the proposed approach and state-of-the-art techniques prove its' superiority.

Keywords: Cardiovascular Diseases, Machine Learning, Optimization Techniques, Coronary Artery Diseases.

1. Introduction

Coronary Artery Disease (CAD) is one of the most life threatening diseases. In this disease, Plaque, a waxy material, gets accumulated inside the coronary arteries as shown in figure 1. It hinders the primary function of arteries and decreases the blood and oxygen supply to the heart muscles. Blood clots can obstruct the flow of blood in the arteries, making it one of the deadliest illnesses. The heart muscles begin to deteriorate. It causes the arm discomfort, stomach ache, nausea, dizziness, tiredness, and sweating. The plaque might burst or harden over time and lead to a heart attack. This may cause death if care is not received immediately. CAD is responsible for more than 30% of all fatalities worldwide. Therefore, it is necessary to detect and cure it at an early stage. Health care therapy for the treatment of Plaque is a lengthy procedure that might takes many years. If CAD is identified at an early stage, the mortality rate can be reduced.

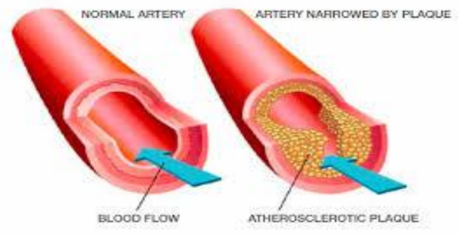


Figure 1. Artery types

The potential of Machine Learning (ML) techniques can be used for the identification of CAD. The growing volume of data may provide opportunities to employ ML techniques to discover important information, and to improve diagnosis accuracy (Latha, 2019).

Putting an effort in this domain, a system is developed for efficient prediction using ML techniques that learn from the previous data (Arora S., Agarwal M., Mongia S, 2021). These techniques are superior in accuracy, are inexpensive in computation, and adaptable to different datasets (Zipes, 2018). Also, ML techniques have the potential to identify the patterns in medical images (Nitesh Pradhan and Geeta Rani, 2020), categorise them into numerous categories (Rani.G and Oza, Da), recreate medical images without contorting its quality (N. Pradhan and V.S Dhaka, 2020), and improving the quality of image without losing more information (N.Kundu and G. Rani, Set). These techniques are capable in data analysis and making decisions according to the trends attainable in the dataset (G.Rani and Oza, 2021). Therefore, ML techniques can be easily accepted.

For the identification of coronary artery disorders, ML based classification approaches such as K Nearest Neighbour (S. Zhang, 2018), Decision Trees (M. Jaworski, 2018) Naive Bayes (C.-z. Gao, 2018) Support Vector Machine (H. Yan, 2018), and Artificial Neural Network (ANN) (X. Liu, 2017) have been employed. But, few studies focused on optimization and hybridization of optimization methods with ML based classification algorithms. These methods deliver excellent results and have been successfully used to predict CAD. Therefore, these methods must be prioritized. In this manuscript, the authors work on the above identified challenges, and propose an integration of optimization and ML based classification approaches for precisely detecting CAD. The paper's main contributions are the following:

1) To minimize redundancy by employing integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for Feature Extraction.

2) To improve accuracy of CAD detection using optimized ML technique.

3) To present a comparison of ML techniques without employing optimization techniques, and after employing ML techniques.

3) To identify best suitable classifier for CAD diseases among SVM, DT, RF, and ANN

Paper is structured as follows: Introduction in mentioned in section 1. Section 2 describes the similar research work which puts light on the several valuable researches done in the diagnosis of CAD. Section 3 covers the proposed methodology and algorithm employed to build the proposed solution. Section 4 discusses the result and discussion of the proposed work and section 5 presents the conclusion and future scope.

2. Related work

On medical datasets, different experiments are done with different classifiers and feature selection algorithms but very little researches have done on the classification of cardiovascular disease that shows good accuracy of classification (M. Fatima and M. Pasha, 2017). For example, the work proposed in reference (K. C. Tan, 2009) propose a hybrid of two ML algorithms SVM, and GA. They used WEKA and LIBSVM tools to analyse the results with 5 different datasets and achieved the accuracy of more than 75% for all 5 datasets. In reference (A. F. Otoom, 2015) SVM, Naïve Bayes, and functional algorithms reported the accuracy of 85.1%, 84.5%, and 84.5% respectively for detection of CAD using Cleveland dataset. Implementation is done in the WEKA tool. 7 best features were selected with the help of the best first selection algorithms.

In paper (G. Parthiban and S. K. Srivatsa, 2012) SVM and naïve bayes algorithms were used to diagnose cardiovascular disease in diabetic patients. WEKA tool was used to implement the algorithms. Data of 500 patients were collected from research Chennai institute. 142 patients had the disease and 358 patients do not. The accuracy of 74% was reported. Similarly, iIn paper (V. Chaurasia and S. Pal, 2014), data mining approaches were used to detect heart disease. 11 attributes were considered in the prediction from a list of 76 attributes. The accuracy of 84.35% was reported by J48, 82.31% by Naïve Bayes, and 85% by Bagging.

Further, in paper (K. Vembandasamy, 2015) datasets consists of 500 patients' records were collected from the Chennai diabetes research institute. Naïve Bayes algorithm is applied on this dataset and accuracy of 86% was reported. Now, In the paper (X. Liu, 2017) heart disease is detected with the help of a hybrid classification system of the ReliefF and Rough Set (RFRS) method. Accuracy of approximately 92.59% was reported. It is evident from the above discussion that a few works focused on optimizing the ML based classifiers for the detection of CAD. Moreover, a few works integrated the feature reduction, classification, and optimizing techniques for prediction of CAD at an early stage.

3. Proposed methodology

Hybridization of PSO and FA algorithms is employed for optimization objectives using decision trees, random forest, SVM, and ANN classification approaches in this proposed methodology section. The decision tree is widely used because of its hierarchical structure, which allows it to solve nonlinear complex problems. Random forest, on the other hand, works better on high-dimensional data and has a faster computation speed on training data than other algorithms, with less data pre-processing. The proposed model uses PCA and ICA for feature extraction, as well as hybridization of PSO and FA with classification approaches such as decision tree, artificial neural network, random forest, and support vector machine. PSO and FA optimization algorithms are used to select best combination of characteristics or the best solution out of a number of options. The proposed methodology is depicted in the figure 2.

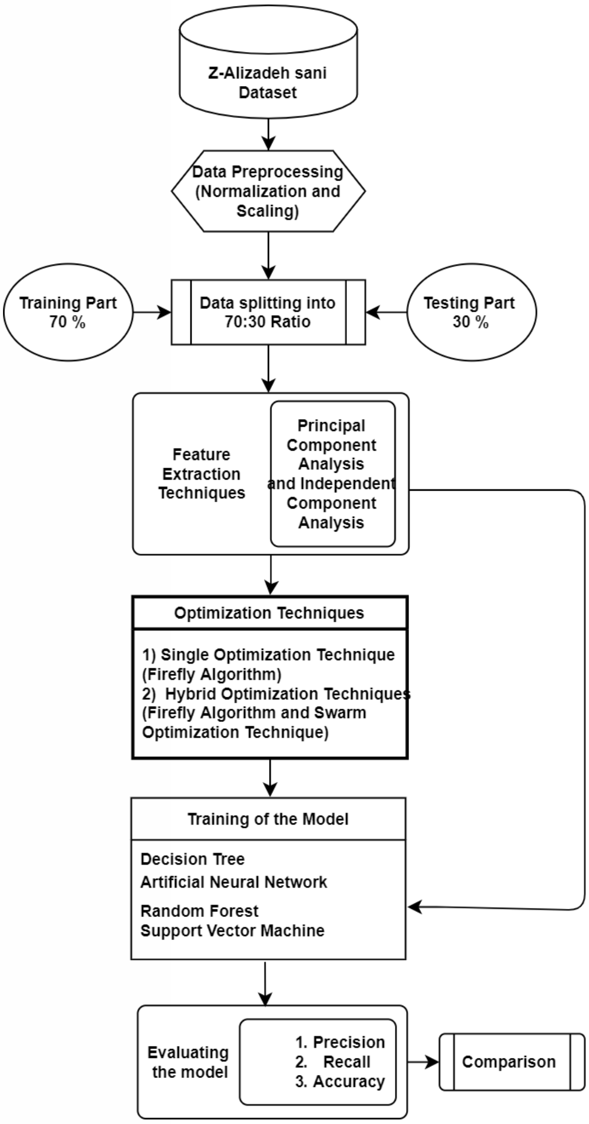

Z-Alizadeh sani Dataset
Data Preprocessing (Normalization and Scaling)
Training Part 70 %
Data splitting into 70:30 Ratio
Testing Part 30 %
Feature Extraction Techniques
Principal Component Analysis and Independent Component Analysis
Optimization Techniques
1) Single Optimization Technique (Firefly Algorithm)
2) Hybrid Optimization Techniques (Firefly Algorithm and Swarm Optimization Technique)
Training of the Model
Decision Tree
Artificial Neural Network
Random Forest
Support Vector Machine
Evaluating the model
1. Precision
2. Recall
3. Accuracy
Comparison

Figure 2: Proposed architecture

In the proposed approach, PCA and ICA are used as a feature extraction technique. Artificial Neural Network (ANN), random forest, SVM, and decision tree algorithms for classification. From the dataset, features are extracted and optimized with the help of PSO and FA for the improved results.

3.1 Dataset

Z Alizadeh Sani Dataset (Set, 2017) consists of records of 303 patients with 56 attributes. All features are divided into four parts demographic, symptom and examination, ECG, and laboratory and echo features. Patients are divided into two categories which are CAD patients or normal. This dataset consists of a large number of attributes and is free from missing values.

3.2 Data preprocessing

In this stage, data is homogenized in the domain of 0 and 1 using the min-max function. Data processing is done in order to make the proposed model machine friendly. Normalization and scaling is performed to convert the string into integers and also convert all the values in the same range.

3.3 Feature extraction

The availability of thousands of features in a dataset may cause inconsistencies. Thus, it is necessary to remove redundant data, and extract highly correlated features in order to improve the accuracy of the model. The complex structure of the dataset increases computation time and makes data processing slow. So, there is a need for feature extraction algorithms to get rid of these problems. PCA is the most widely used dimension reduction techniques. It is one of the that helps to identify correlations and patterns in a dataset so that it can be transformed into a new dataset that has significant low dimensions without loss of much information. Narrowing down a couple of variables from the original dataset. Various steps are involved here 1) Standardization of data: scaling of the data is done in such a manner that variables and their values lie within a similar range z=variable value-mean/standard deviation 2) Computing covariance matrix-it express 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-these are computed from the covariance matrix to determine principal components of the data set 4) Computing principal components-

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.

The second feature extraction process with the combination of machine learning used is the independent component analysis. ICA describes a reproductive model for the experimental multivariate records, which is usually a large samples dataset. In a machine learning model, the variables used as input are supposed to be linear combinations of some new suppressed variables. The suppressed variables are supposed nongaussian and commonly self-determining which are well known as the independent components based on the observed information. ICA has nearly correlated to PCA i.e. principal component analysis. ICA is a much extra controlling procedure, however, proficient in finding the essential parameters used for the feature extraction process. The information analyzed using the ICA process could originate from several application areas, including images, databases, financial methods, and psychometric dimensions. In many belongings, the dimensions are set of time series data where the blind source split-up is used to illustrate the defined problem.

3.4 Feature optimization

In feature optimization hybridization of swarm particle optimization and firefly nature-inspired algorithm is used for the selection of instances and to identify the best solution among different possible solutions for classification.

3.4.1 Firefly algorithm

Various assumptions of fireflies are the following like attraction towards each other, attraction is directly proportional to their brightness, in case of same brightness of two fireflies, they move randomly and from random walk new solutions are identified.

The steps of firefly algorithms are given below:

1) Initialization of parameters is done.

2) Population of n fireflies is generated.

3) Fitness value of each firefly is evaluated.

4) Condition is checked, accordingly position and intensity of light are updated.

5) Best solution is reported.

Firefly algorithm is widely used for its numerous advantages as compared to other optimization techniques. Those advantages are 1) Population is divided into several parts automatically 2) It deals with multi-model optimization 3) In solutions, there is high randomness.

3.4.2 Particle swarm optimization

Particle swarm optimization algorithm is inspired by the social behavior of birds flocking and solves hard complex problems. This algorithm is a population-based algorithm. Members of the population are called particles and the population is called a swarm. There are various advantages of PSO over other optimization techniques. PSO is applied on coronary artery diseases successfully and it has given fantabulous results. In PSO objective function tries to maximize or minimize the values which need to optimize. Various advantages of PSO are following. 1) It is very easy to implement and has very few parameters to work with 2) Higher chances of finding out the global optima 3) Takes a short time of computation. 4) Robust and converge fast.

The steps of the particle swarm optimization are the following:

1) Population and parameters are initialized.

2) The fitness value of each particle is evaluated and the best particle is chosen.

3) The velocity and position of every particle are calculated.

4) Find the current best (Gbest).

5) Update iteration t.

6) Output is Gbest.

3.5 Classification techniques

In the classification technique, data is divided into numerous categories or several classes. In the classification approach when the test data is applied and the training model gets loaded, the predictions will be made based on the supervised learning models and at that stage it can be noticed through the class labels that whether the person is having disease or not. The main objective of using these classification techniques is to find out the category/class of the new data belongs to. In the early prediction case it depends on the type of input and the processing done on the input based on the early prediction characteristics which can reduce the quantitative risks for the same. In the proposed approach main concentration is given to the end to end prediction.

3.5.1 Decision tree algorithm (DT)

Decision tree is one of the most popular algorithms used for classification. Based on the condition

Advantage of Decision Tree approach:-

1) The decision tree generates easy and understandable rules.

2) With much computation, classification is performed in decision trees.

3) Continuous and categorical variables are easy handled by decision trees.

4) It gives a crystal clear idea about the prominent fields for the purpose to predict and classify.

3.5.2 Random forest (RF)

Random forest is one of the widest algorithms used in machine learning. Random forests are bagged decision trees, which are split on a features subsets basis. RF can manage categorical, binary, and numerical features. Very little pre-processing is required as there is no requirement for rescaling and transformation. There are various advantages of the Random forest classification technique over other classification techniques.

1) On high dimensions data, it works efficiently.

2) It can manage lots of inputs variables without deletion of variables.

3) It gives a clear idea about which variable is more valuable.

4) Most accurate learning algorithm.

5) Parallelizable (Process can be split into multiple machines to run).

3.5.3 Support vector machine (SVM)

SVM is one of the significant supervised learning systems, used for sorting and regression complications. But, mainly, it is used for ordering difficulties in data science and machine learning concepts. The objective of this supervised process is to generate the superlative decision boundary that can separate n-dimensional space into modules that are called class labels so that the efficient prediction will be made on the new dataset in the real-world scenario. This top decision borderline is known as a hyperplane. SVM elects the extreme data vectors that help in generating the decision boundary. These extreme points are well known by support vectors as shown in figure 3.

of the internal node, the tree is divided into different edges and the last decision is taken on that branch, which stops splitting more, and that will be the last decision of the classification process.

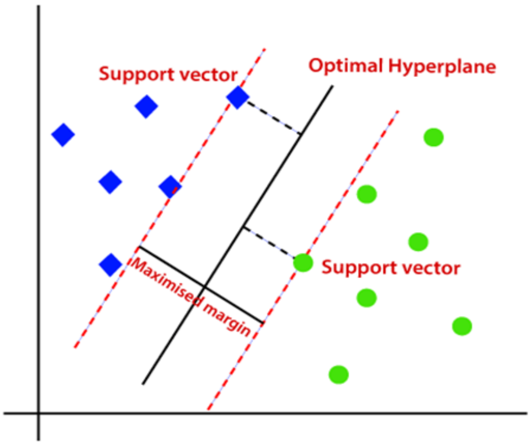


Figure 3. Artificial neural network

The working of ANN is like the structure of a human understanding. Comparable to the human mind that has processing elements called neurons interrelated to the other neurons in the network, artificial neural networks also have neurons that in a similar way the ANN structure are interconnected nodes to other nodes organized in the form of layers. ANN structures are arranged as per the sequence of the layers. There are mainly three layers in ANN i.e. Input layer, Hidden Layer, and Output Layer as shown in figure 4. It regulates weighted input passed to an activation utility to produce the relationships. The activation function decides the life of the node in the ANN that whether it should be accepted or not between the layers. There are unique activation functions accessible that can be functional on the type of requirement that needs to be implemented.

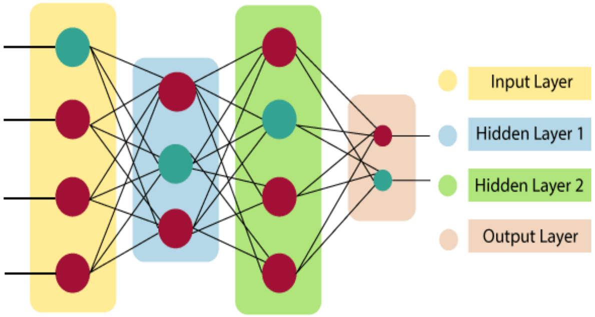


Figure 4: Typical structure of an N

3.6 Classification task

In the classification technique, data is divided into numerous categories or several classes. In the classification approach when the test data is applied and the

stage it can be noticed through the class labels that whether the person is having disease or not. The main objective of using these classification techniques is to find out the category/class of the new data belongs to. In the early prediction case it depends on the type of input and the processing done on the input based on the early

prediction characteristics which can reduce the quantitative risks for the same. In the proposed approach main concentration is given to the end to end prediction.

3.7 Proposed algorithm

Step 1: Input Records such that R = R1, R2… RN as data and perform the framing of the data to process in a structured format.

Step 2: Scaling of the Records.

Step 3: Extract the feature vector F(x) & transform it to produce the covariance.

Step 4: If PCA performs Eigenvector extraction to extract meaningful understandings for the transformation to map the data and generate a new vector space.

If ICA Perform transformations on the data to get the independent Gaussian components.

Step 5: Perform instance selection using optimization methods firefly algorithm and particle swarm optimization for the dimensionality reduction process.

Step 6: Generate splitting of the data into different ratios 70% is training data & 30% is the test unknown data.

Step 7: Generate the Ensemble Learning.

Step 8: Upload unknown samples.

Step 9: Call the trained model and implement classification on Test dataset.

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

4. Experiments and results

The proposed work is implemented in python using the Pycharm editor by creating the virtual environment. A virtual environment is created to reduce the dependencies among the different libraries installed for the smooth functioning of the implementation. Pycharm is training model gets loaded, the predictions will be made based on the supervised learning models and at that

widely used to perform complex technical computations and is handier to run the simulation environment.

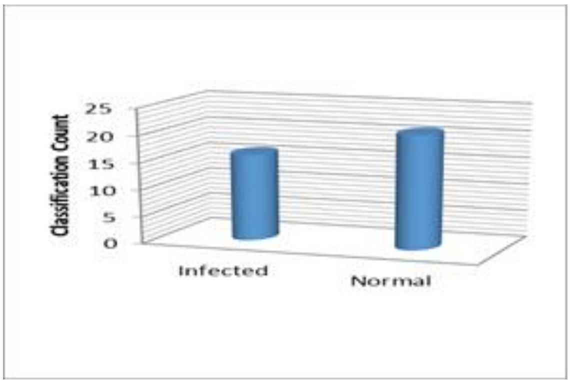


Figure 5: Classification

Fig 5. Shows the classification counts from the test data and is unknown data and performs predictive modeling using Random Forest ensemble classification. This is a necessary step because these predictive counts will be matched with the unknown test labels to perform the validations that the predicted output is a truly positive and true negative or not.

4.1 Performance evaluation using PCA in the combination of optimization and classification

The below discussed results are evaluated using extraction done with PCA with the hybrid optimization and classifications. The results are evaluated using Decision Tree, SVM, Ensemble Learning, and Neural Network. It can be seen that the neural network is achieving high performance because it is capable of reducing the training and validation losses by updating the weights which will reduce the biasing error in the information transfer between the layers.

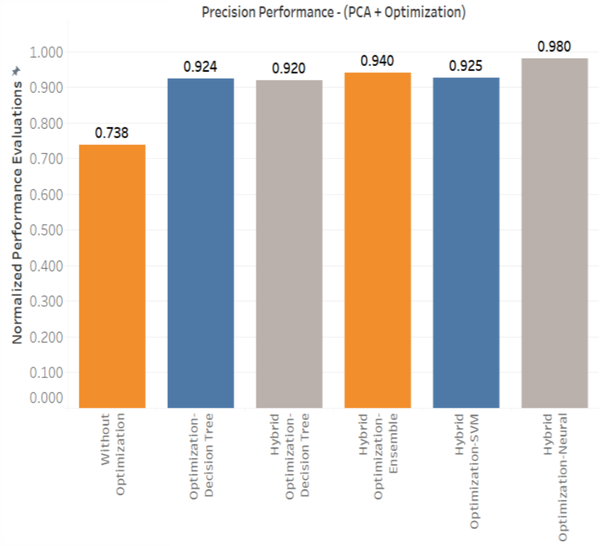


Figure 6: Precision comparison

Fig 6 shows the precision performance in which it can be seen clearly that the ANN using hybrid optimization is achieving high precision in the classification of the true predicted instances for the CAD than ensemble learning, SVM, and decision tree which are closely related to the precise performance. It shows that how many true relevant instances are predicted efficiently as a model performance. If the precision is high the true positive rates concerning the false rejection and false acceptance increases which increases the accuracy of the model. The precision is evalu-ated using the given expression below in equation 1.

$$P = \mathrm{X(p)}/(X(p) + Y(p)) \qquad (1)$$

Where P is the precision of the model and X (p) is the true positive rate and Y (p) is the false positive rate.

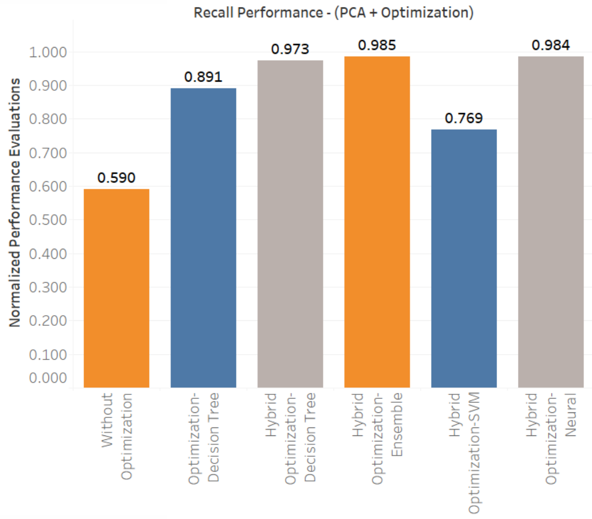


Figure 7: Recall comparison

Fig 7: Recall Comparison shows the recall performance in which it can be seen clearly that the ANN and en-semble learning is achieving nearly the same and high recalling of the relevant instance selections using hybrid optimization in the classification of the true predicted instances for the CAD. Also, it can be seen the SVM is not achieving a high recall rate because of high non-linearity but the decision tree is matching with the performance of the ANN recall rate. Recall shows the relevancy of the prediction is performance based on the training data which should be high. If the relevancy of the prediction is high then there will be high information retrieval and it will increase the high sensitivity of the model. The recall is evaluated using the given expression below in equation 2.

$$R = \mathrm{X(p)}/(X(p) + Y(n)) \qquad (2)$$

Where R is the recall of the model and X (p) is the true positive rate and Y (n) is the false-negative rate.

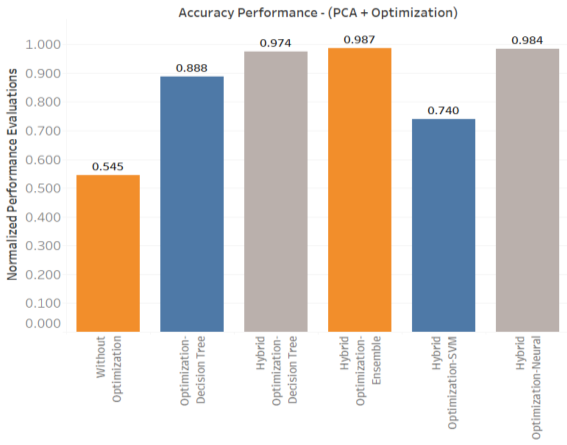


Figure 8: Accuracy comparison

Figure 8: Accuracy ComparisonFig. 8 shows the accuracy performance in which it can be seen clearly that the ANN and ensemble learning using hybrid optimization is achieving high accuracy in classification of the true pre-dicted instances for the CAD and having a very minute difference in the precision value. On the other hand it SVM is achieving low accuracy due to high non-linearity as a result of which the weights become variable and is achieving high variances in the classification rates whereas the decision tree is also achieving good accuracy in the prediction as it reduces the entropy of the data and improves the decision making for the classifications. Accuracy shows the prediction is achieved closely to true values presented in the dataset which should be high for true posi-tive and

true negative rates. If the accuracy is high then classification error rates will

be low for true prediction. The accuracy is evaluated using the given expression in equation 3.

$$A = \frac{X(p)}{(X(p) + X(n) + Y(p) + Y(n))(1 + x)^n} \qquad (3)$$

Where A is the accuracy of the model and X (p) is the true positive rate and Y (n) is the false-negative rate, X (n) are the true negative rate and Y (n) false-negative rate.

Table 1: Confusion matrix (Decision tree)

Classes	**Actual(CAD =Yes)**	**Actual (CAD= No)**
Predicted (CAD= Yes)	TP (202)	FP (15)
Predicted (CAD= No)	FN (11)	TN (75)

Table 2: Confusion matrix (Ensemble learning)

Classes	**Actual (CAD =Yes)**	**Actual (CAD= No)**
Predicted (CAD= Yes)	TP (204)	FP (12)
Predicted (CAD = No)	FN (8)	TN (79)

Table 3: Confusion matrix (SVM)

Classes	**Actual (CAD=Yes)**	**Actual (CAD = No)**
Predicted (CAD = Yes)	TP (150)	FP (48)
Predicted (CAD = No)	FN (49)	TN (56)

Table 4: Confusion matrix (ANN)

Classes	**Actual (CAD =Yes)**	**Actual (CAD = No)**
Predicted (CAD = Yes)	TP (206)	FP (9)
Predicted (CAD = No)	FN (7)	TN (81)

Table 5: Recall performance

Test No.	**Decision tree**	**Ensemble Learning**	**SVM**	**ANN**
1	0.973	0.984	0.783	0.983
2	0.952	0.984	0.763	0.981
3	0.954	0.982	0.736	0.977
4	0.973	0.976	0.763	0.983
5	0.967	0.977	0.757	0.985
6	0.963	0.985	0.738	0.972
7	0.954	0.972	0.778	0.987
8	0.957	0.987	0.772	0.983
9	0.968	0.978	0.764	0.982
10	0.962	0.962	0.746	0.975

Table 6: Precision performance

Test No.	**Decision tree**	**Ensemble Learning**	**SVM**	**ANN**
1	0.926	0.940	0.925	0.970
2	0.921	0.945	0.933	0.969
3	0.930	0.944	0.926	0.961
4	0.925	0.932	0.913	0.967
5	0.922	0.936	0.921	0.976
6	0.935	0.943	0.913	0.971
7	0.928	0.948	0.929	0.957
8	0.936	0.934	0.914	0.961
9	0.924	0.936	0.932	0.956
10	0.937	0.933	0.911	0.973

Table 7: Accuracy performance

Test No.	**Decision Tree**	**Ensemble Learning**	**SVM**	**ANN**
1	0.971	0.986	0.739	0.983
2	0.966	0.973	0.756	0.981
3	0.973	0.983	0.769	0.986
4	0.977	0.981	0.775	0.973
5	0.979	0.977	0.747	0.988
6	0.965	0.984	0.749	0.981
7	0.966	0.986	0.743	0.983
8	0.972	0.983	0.769	0.985
9	0.969	0.985	0.776	0.978
10	0.975	0.979	0.725	0.983

Table 1, 2, 3, 4 shows confusion matrix using Decision Tree, Random Forest, SVM, ANN. Table 5, 6, 7 shows the comparison of the different test sets run on the CAD classification. The above evaluation shows that the proposed approach achieved high-performance analysis and attained low false acceptance and false rejection rates. It can also be seen that the specificity and sensitivity of the model are also increased and as a result of which the accuracy also increases. Eventually, the proposed ensemble learning using boosting evaluation and ANN is achieving high accuracy rates with low classification error rates.

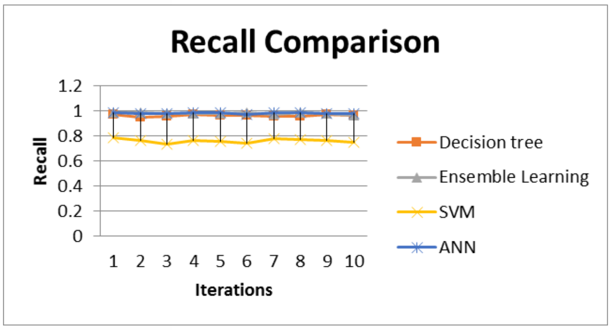
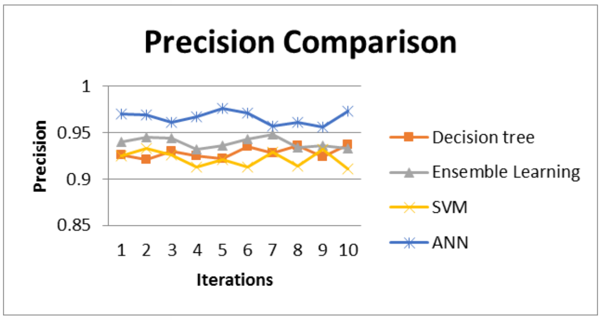
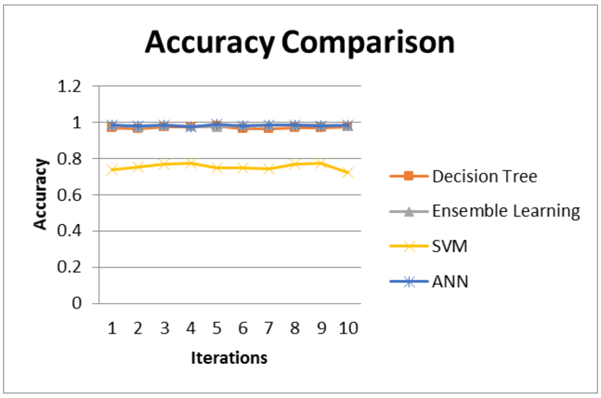


Figure 9(a) Recall performance analysis



9(b) Precision Performance Analysis



9(c) Accuracy Performance Analysis

Figure 9: (a) Recall, (b) Precision (c) Accuracy test evaluations

Fig. 9 shows the performance evaluations of various test cases run on different test samples and performance is evaluated in the same. It can be seen that the performance is achieving high accuracy, precision, and recall for the various test samples and the proposed supervised learning machine learning comparison is achieved. It can be seen that the ANN and ensemble learning is almost similar in performance which shows that the modelling of the data in such a way that the proposed models can achieve high true positive and negative rates.

4.2 Performance evaluation using ICA in a combination of optimization and classification

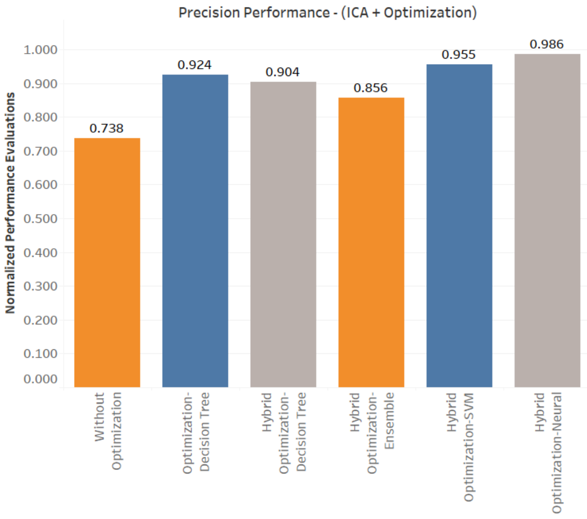


Figure 10: Precision comparison

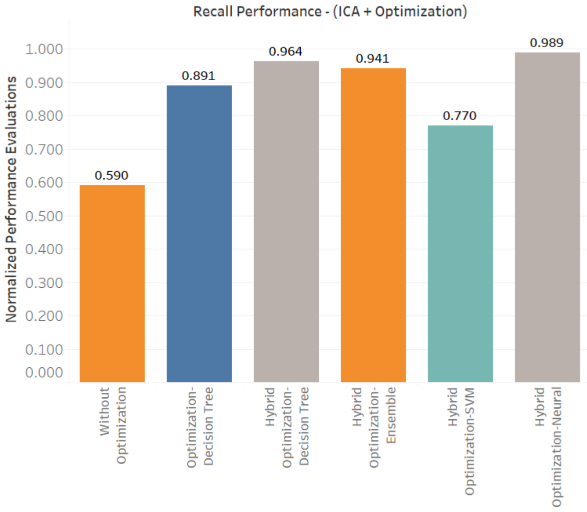


Figure 11: Recall comparison

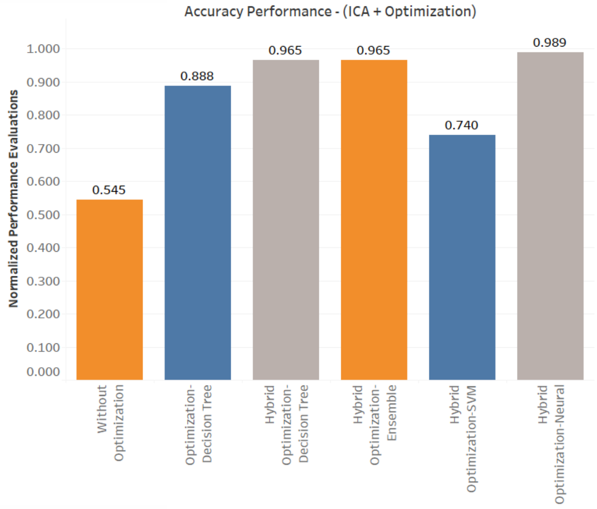


Figure 12: Accuracy comparison

Table 8: Confusion Matrix (Decision Tree)

Classes	Actual(CAD =Yes)	Actual(CAD= No)
Predicted(CAD = Yes)	**TP (204)**	**FP (13)**
Predicted(CAD = No)	**FN (9)**	**TN (77)**

Table 9: Confusion matrix (Ensemble learning)

Classes	Actual (CAD=Yes)	Actual(CAD= No)
Predicted(CAD = Yes)	**TP (206)**	**FP (9)**
Predicted(CAD = No)	**FN (7)**	**TN (81)**

Table 10: Confusion matrix (SVM)

Classes	Actual (CAD =Yes)	Actual(CAD=No)
Predicted (CAD= Yes)	**TP (152)**	**FP (45)**
Predicted (CAD = No)	**FN (51)**	**TN (55)**

Table 11: Confusion matrix (ANN)

Classes	Actual (CAD=Yes)	Actual (CAD = No)
Predicted(CAD= Yes)	**TP (209)**	**FP (6)**
Predicted(CAD= No)	**FN (3)**	**TN (85)**

Table 12: Recall performance

Test No.	Decision tree	Ensemble Learning	SVM	ANN
1	0.964	0.944	0.770	0.988
2	0.954	0.956	0.775	0.989
3	0.966	0.956	0.757	0.988
4	0.967	0.968	0.755	0.986
5	0.968	0.968	0.768	0.978
6	0.955	0.957	0.774	0.976
7	0.969	0.954	0.769	0.989
8	0.958	0.949	0.759	0.986
9	0.969	0.945	0.743	0.989
10	0.954	0.964	0.769	0.987

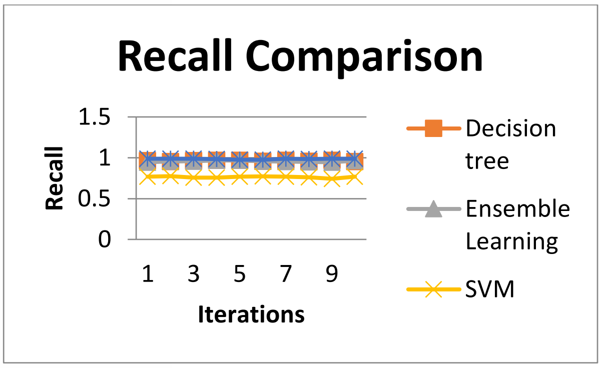
Table 13: Precision performance

Test No.	Decision tree	Ensemble Learning	SVM	ANN
1	0.926	0.940	0.925	0.970
2	0.921	0.945	0.933	0.969
3	0.930	0.944	0.926	0.961
4	0.925	0.932	0.913	0.967
5	0.922	0.936	0.921	0.976
6	0.935	0.943	0.913	0.971
7	0.928	0.948	0.929	0.957
8	0.936	0.934	0.914	0.961
9	0.924	0.936	0.932	0.956
10	0.937	0.933	0.911	0.973

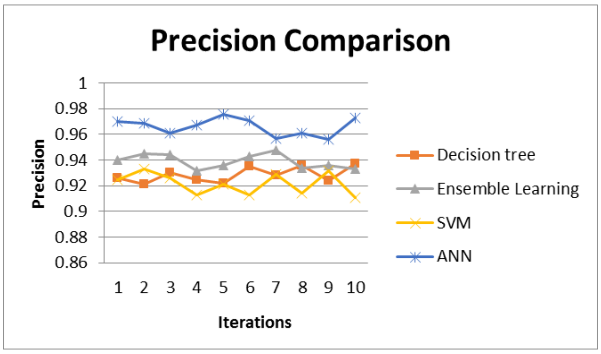
Table 14: Accuracy performance

Test No.	Decision Tree	Ensemble Learning	SVM	ANN
1	0.964	0.963	0.736	0.988
2	0.963	0.971	0.748	0.985
3	0.965	0.975	0.765	0.987
4	0.978	0.974	0.753	0.984
5	0.966	0.969	0.756	0.989
6	0.977	0.971	0.735	0.986
7	0.965	0.973	0.766	0.978
8	0.966	0.989	0.757	0.989
9	0.978	0.961	0.763	0.988
10	0.961	0.973	0.754	0.986

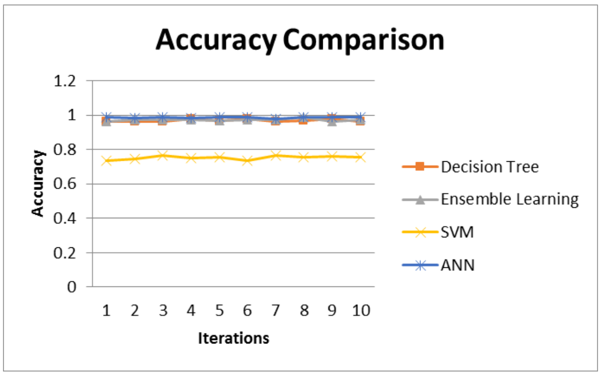
Figure 10, 11, 12 shows precision, recall and accuracy comparison using ICA with optimization and classification techniques. Table 8, 9, 10, 11 shows confusion matrix using Decision tree, Ensemble Learning, SVM and ANN. Table 12, 13, 14 shows the comparison of machine learning approaches after various iterations run on the data and simulated the results. The criteria of overfitting and under fitting are taken care also in the implementation covering all the test cases by using the hyper parameter tuning process. It can be seen that the decision tree, ensemble learning, and ANN is performing well in the detection but SVM is somewhat outfitted the classification because of the non-separable of support vectors on the hyperplane. Also, it can be noticed that the ANN is performing well in terms of the high accuracy, precision, and sensitivity of the model.



13(a) Recall performance



13(b) Precision comparison



13(c) Accuracy comparison

Fig 13: (a) Recall, (b) Precision (c) Accuracy test evaluations

Fig. 13 shows the performance evaluations of various test cases run on different test samples and performance is evaluated in the same. It can be seen that the performance is achieving high accuracy, precision, and recall for the various test samples and the proposed ensemble learning is performing well which is the desired output. The proposed model is tested on various classifiers such as decision tree and ensemble random forest learning process which is achieving high performance in the classification of true positive and true negative rates.

Table 15: Performance comparison

Author's Names	**Accuracy**
X. Liu, X. Wang et al. (X. Liu et al. 2017)	**92.59%**
J. N. Khiarak, M.F.Derakhshi (J. N. Khiarak et al, 2019)	**94.43%**
N. Pereira (N. Pereira, 2019)	**82.46%**
Devansh Shah (Devansh Shah et al, 2021)	**90.78%**
S.H. Wijaya, G.T. Pamungkas (S. H. Wijaya et al, 2018)	**86.67%**
Proposed Model	**98.88%**

Eventually, the classification is performed with the combination of feature extraction and optimization as instance selections. The test cases implemented using PCA and ICA are achieving efficient and nearer performance and ICA is achieving minute high performance than PCA due to its Gaussian independent nature and a high degree of freedom which will reduce the variance and standard deviation among the data points and also reduces the non-linearity's which will increase the model performances. Performance comparison with different authors is shown in table 15.

5. Conclusions

The proposed work does the comparison of CAD with the help of machine learning algorithms without optimization, single optimization, and a combination of optimization techniques with decision tree, random forest, SVM and ANN classification techniques. Pycharm tool is used with the Z-Alizadeh Sani dataset to analyse the developed model. After pre-processing, the PCA algorithm is applied to a dataset to extract the features. After extracting the features, hybridization of optimization techniques is done by using techniques PSO and FA to improve the accuracy of the developed model. Experiments demonstrate the efficiency of the proposed model. In this paper, the performance is evaluated on different machine learning algorithms because the dynamicity of the data changes the working of the model. So it is necessary to evaluate the performance with at least four machine learning processes. In the proposed model accuracy score is 98.4%. Also, the proposed solution can help clinical experts on our solution because in the dataset almost all realistic features are considered from which any patient suffered for the disease. Also based on the feature engineering process, it can help to find the patterns of the diseases which can help doctors to diagnose the diseases. Based on the precision and recall which is the combinations of true and false measures to achieve the relevance of the data and prediction, the clinical experts can make best use of the proposed solution. From the result and discussion, it can be seen that ANN is giving satisfactory results in comparison to the other algorithms and then Ensemble learning i.e bagging-based random forest is achieving good accuracy nearer to the ANN. The implementation of the Deep learning models with transfer learning will be the future scope of the current approach because as the data grows the complexity of the model will also increase which is the drawback of the current machine learning algorithms.

References

Arora S., Agarwal M., & Mongia S. (2021). Comparative Analysis of Educational Job Performance Parameters for Organizational Success: A Review.

Dave M., Garg R., Dua M., & Hussien J. (eds) Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences. Algorithms for Intelligent Systems. Springer, Singapore.

A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, & M. Ashour, (2015). Effective diagnosis and monitoring of heart disease. International Journal of Software Engineering and Its Applications, (9)1,143-156.

Arora, S. (2016). A novel approach to notarize multiple datasets for medical services, Imperial Journal of Interdisciplinary Research, 2(7), 325-328.

C.-z. Gao, Q. Cheng, P. He, W. Susilo, & J. Li. (2018). Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack, Information Sciences, 444. 72-88.

G. Parthiban & S. K. Srivatsa, (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems. 3(7). 2249-0868.

H. Yan, Q. Ye, T.a. Zhang, D.-J. Yu, X. Yuan, Y. Xu, & L. Fu, (2015). Least squares twin bounded support vector machines based on L1-norm distance metric for classification. Pattern Recognition. 434-447.

J. N. Khiarak, M.F.Derakhshi, K. Behrouzi, S. Mazaheri, Y. Zamani-Harghalani, & R. M.Tayebi, (2019). New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. Health and Technology.1-12.

K. C. Tan, E. J. Teoh, Q. Yu, & K. C. Goh, (2009). A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications,36(4).8616-8630.

K. Vembandasamy, R. Sasipriya, & E. Deepa, (2015). Heart Diseases Detection Using Naive Bayes Algorithm. IJISET-International Journal of Innovative Science, Engineering & Technology.2.441-444.

Latha, C. Beulah Christalin, & S. Carolin Jeeva, (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 16.1-7.

M. Fatima & M. Pasha, (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications. Journal of Intelligent Learning Systems and Applications, 9(1).

M. Jaworski, P. Duda, L. Rutkowski, (2018).New Splitting Criteria for Decision Trees in Stationary Data Streams, IEEE Transactions on Neural Networks and Learning Systems, 29.2516-2529.

N. Pereira, (2019).Using machine learning classification methods to detect the presence of heart disease.

N. Kundu, G. Rani & V. S. Dhaka, (2020).Machine Learning and IoT based Disease Predictor and Alert Generator System, Fourth International Conference on Computing Methodologies and Communication (ICCMC),764-769.doi:10.1109/ICCMC48092.2020.ICCMC-000142.

Nitesh Pradhan, Geeta Rani, Vijaypal Singh Dhaka, & Ramesh Chandra Poonia,(2020) Diabetes prediction using artificial neural network, Editor(s): Basant Agarwal, Valentina Emilia Balas, Lakhmi

C. Jain, Ramesh Chandra Poonia, Manisha, Deep Learning Techniques for Biomedical and Health Informatics, Academic Press.327-339, ISBN 9780128190616, https://doi.org/10.1016/B978-0-12-819061-6.00014-8.

Pradhan, N., Dhaka, V.S., Rani, G. & et al., (2020). Transforming view of medical images using deep learning. Neural Computation & Application. 32.15043–15054. https://doi.org/10.1007/s00521-020-04857-z

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

S. h., M. Agarwal & R. Kawatra, (2020). Prediction of Educationist's Performance using Regression Model. 7th International Conference on Computing for Sustainable Global Development (INDIACom).88-93.doi: 10.23919/INDIACom49435.2020.9083708

S. Zhang, D. Cheng, Z. Deng, M. Zong, & X. Deng, (2018). A novel kNN algorithm with data-driven k parameter computation, Pattern Recognition Letters, 109. 44-54.

S. H. Wijaya, G.T. Pamungkas, & M. Burhanis Sulthan. (2018). Improving classifier performance using particle swarm optimization on heart disease detection. International Seminar on Application for Technology of Information and Communication.603-608. IEEE.

Set, Z.-A.S.D. (2017).UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems. Available online: https://archive.ics.uci.edu/ml/machine-learning-databases/00412/ (accessed on 8 April 2019).

X. Liu, X. Wang, Q.Su, M. Zhang, Y. Zhu, Q.Wang, & Q.Wang, (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method.Computational and mathematical methods in medicine.

V. Chaurasia & S. Pal, (2014). Data mining approach to detect heart diseases, International Journal of Advanced Computer Science and Information Technology, 2(4).56-66.

Zipes, D.P., Libby, P., Bonow, R.O., Mann, D.L., Tomaselli, & G.F. Braunwal's, (2018). Heart Disease E-Book: A Textbook of Cardiovascular Medicine; Elsevier Health Sciences Wiley: San Francisco, CA, USA.

Authors Biographies

Ms Savita is pursuing her Ph.D degree from GD Goenka University in the area of Artificial Intelligence. She has completed MCA from KIIT College of Engineering Gurugram in 2013. Experience of teaching as an adjunct faculty in KR Mangalam University and Gurugram University (Gurugram). She is currently working as an Assistant professor at DPG Degree College Gurugram. Her key research areas are machine learning, Deep Learning, and Artificial Intelligence.

Geeta rani has 12 years' of 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 and eighteen copyrights are registered in her name for the software works. She ha expertise in writing and filing IPR. She has Qualified test of women scientist in IPR. She has published more than 30 papers in SCI and SCOPUS indexed Journals. She is Chief Editor of the book "Disease Prediction using Machine Learn-ing". She also has published many book chapters in SCOPUS indexed book series. She has delivered talks as dis-tinguished speakers at several workshops, FDPs and conferences.

Apeksha Mittal received her 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. Cur-rently she is working as an Assistant Professor at GD Goenka University, Gurugram. Her key research areas are Artificial Neural Networks, Deep Learning, Machine Learning and Artificial Intelligence.