A Robust Cardiovascular Disease Predictor Based on Genetic Feature Selection and Ensemble Learning Classification

Original Scientific Paper

Sadiyamole P. A.
Research Scholar, Department of Computer Science,
Karpagam Academy of Higher Education
Coimbatore 21, India
sadiya.pa@gmail.com

Dr. S. Manju Priya
Professor, Department of CS, Karpagam Academy of Higher Education
Coimbatore 21, India
smanjupr@gmail.com

Abstract *– Timely detection of heart diseases is crucial for treating cardiac patients prior to the occurrence of any fatality. Automated early detection of these diseases is a necessity in areas where specialized doctors are limited. Deep learning methods provided with a decent set of heart disease data can be used to achieve this. This article proposes a robust heart disease prediction strategy using genetic algorithms and ensemble deep learning techniques. The efficiency of genetic algorithms is utilized to select more significant features from a high-dimensional dataset, combined with deep learning techniques such as Adaptive Neuro-Fuzzy Inference System (ANFIS), Multi-Layer Perceptron (MLP), and Radial Basis Function (RBF), to achieve the goal. The boosting algorithm, Logit Boost, is made use of as a meta-learning classifier for predicting heart disease. The Cleveland heart disease dataset found in the UCI repository yields an overall accuracy of 99.66%, which is higher than many of the most efficient approaches now in existence.*

Keywords*: Cardiovascular disease prediction; Deep learning techniques; Genetic algorithms; Adaptive Neuro-Fuzzy Inference System; Multi-Layer Perceptron; Radial Basis Function; Logit Boost*

1. INTRODUCTION

The World Health Organization (WHO) considers coronary artery problems as the significant reason for people's death worldwide. Their estimation shows about 32% of the world's population is deceased yearly due to CVDs [1] Moreover, according to the Centers for Disease Control and Prevention (CDC) in the United States of America, one civilian dies every second succumbing to CVD [2]. In addition, the country spent nearly $407.3 billion from 2018 to 2019 on heart diseases. The American Heart Association's statistics for heart disease and stroke in 2023 estimate that the yearly direct costs of CVD increased from $103.5 billion in 1996–1997 to $251.4 billion in 2018–2019 [3]. So, CVDs have an impact not only on human fitness but also on the financial side and the expense of countries. A crucial way to decrease this record is to detect CVDs early. But, the identification of CVDs with traditional medical examinations is quite challenging, takes up a great deal of time, and is expensive. Thus, the early determination of heart disease in developing countries is risky due to the shortage of conventional examination tools and medical practitioners [4,5].

The growth in the artificial intelligence field, mainly in the areas of data mining and machine learning procedures, can be well utilized in clinical research. These techniques can be used to effectively diagnose heart diseases in a benignly and cost-effectively manner. In the literature, there are numerous cardiovascular disease diagnostic algorithms proposed by different researchers over the years. To extract needed knowledge in an organized manner from high-dimensional unorganized data, various data mining techniques have been proposed [6, 7]. However, the obtained data may still contain redundant and irrelevant features that can confuse for classification algorithms, resulting in high time complexity and poor performance. Incorporating a good feature selection process is necessary for any heart disease predictive system. Furthermore, among the available classification algorithms, deep learning methodologies are currently considered one of the researcher's favorite disciplines. Therefore, this work proposes a robust CVD prediction

algorithm using the evolutionary power of genetic algorithms along with an ensemble of three established deep learning neural networks such as Multilayer Perceptron (MLP), Radial Basis Function (RBF), and Adaptive Neuro-Fuzzy Inference System (ANFIS). The findings of the experiment demonstrate that, when compared to current approaches, this suggested ensemble learning system has a strong competence in predicting cardiovascular illness.

The rest of this work is put together as follows: An overview of the literature on alternative heart disease prediction techniques is presented in Section 2. A comprehensive explanation of the suggested methodology can be found in Section 3. The experimental findings are discussed in Section 4 along with a comparison to existing methods. The conclusion is presented in Section 5.

2. RELATED WORKS

Over the years of research on the automated diagnosis of cardiovascular diseases, many researchers have suggested different data mining and classification algorithms. Many researchers have tried to incorporate Genetic Algorithms (GA) into heart disease predictions to select the most relevant attributes. The significance of the GA is to categorize the relevant features from the raw data set that could represent the whole more accurately. The fitness function used in this GA identifies the best attributes from the whole dataset.

Gokulnath and Shantharajah [8] proposed GA-based Support Vector Machines (SVM) to diagnose the presence of heart diseases. Their results claimed that the SVM classifier attained only 83.70% accuracy without GA, while the same classifier improved its performance to 88.34% with GA aid. [9] presented a Genetic-based Crow Search Algorithm (GCSA) for selecting relevant features from the dataset and deep convolutional neural networks (DCNN) for the classification. They claimed that the proposed GCSA could increase the classification accuracy by achieving more than 94% in comparison with the other feature selection approaches. Also, Kanwal et al. [10] proposed a heart disease classification algorithm that utilized GA for feature selection. Their classification included different machine learning algorithms like Deep Learning (DL), Naive Bayes (NB), Neural Network (NN), Support Vector Machine (SVM), and Logistic Regression (LR). Their method achieved 92% accuracy.

A similar type of procedure was presented in [11]. They also used GA for feature selection. The performance was obtained using three different classifiers, namely, Nave Bayes (NB), clustering, and Decision Tree (DT). Their experiments show that NB performs consistently with and without the attribute selection procedure, while DT shows better performance after feature selection using GA. They obtained 99.2% accuracy with features selected using GA and DT for classification. Another author, Durga Devi [12], also used genetic algorithms for feature reduction. Furthermore, they used Radial Basis Function (RBF) for diagnosing heart diseases and achieved an accuracy of 85.48%.

Jothi Prakash and Karthikeyan [13] proposed a combination of GA and linear discriminant analysis (LDA) for feature selection and combined the classification results of MLP, NN, DT, and SVM using the ensemble bagging technique. The highest accuracy achieved by their procedure is 93.65% for the Statlog dataset. A hybrid deep learning technique was designed for predicting CVD by Kishore and Jayanthi [14]. To predict CVD, the attribute weights for neural network initialization in Artificial Neural Network (ANN) are performed using Analytic Hierarchy Processing (AHP), and Multilayer Back Propagation Neural Network (MLBPNN) is designed using the Gradient Descent algorithm. Their system obtained an average accuracy of 94.15%. Sharma and Parmar [15] employed an optimized DNN using Talos that produced 90.78% accuracy without any feature extraction procedures.

Mohan et al. [16] presented a novel technique called the Less Error Classifier for the dimensionality reduction of CVD. They produced an accuracy of 88.7% with their HRFLM (Hybrid Random Forest with a Linear Model). Ali et al. [17] proposed a CVD monitoring and prediction system based on ensemble deep learning and feature fusion. The feature combination procedure consisted of missing data filtering, normalization, information gain-based feature selection, and conditional probability-based feature weighting. The system produced an accuracy of 98.5%. Javeed et al. [18] exploited a feature selection method that uses a floating window with adaptive size (FWAFE). Additionally, they used two types of prediction structures: ANN and DNN. They obtained the best results with the DNN system of 93.33%.

Khourdifi and Bahaj [19] utilized an FCBF (Fast Correlation-Based Feature Selection) algorithm to select relevant features from the CVD dataset. The system consisted of classification methodologies based on various classification algorithms such as KNN, SVM, NB, RF, and MLP approaches optimized by Particle Swarm Optimization (PSO) combined with the Ant Colony Optimization (ACO) method. They obtained an accuracy of 99.65% using the FCBF, PSO, and ACO optimized KNN classifiers. Baksh [20] proposed a cluster-based improved deep GA for solving the problem of CVD diagnosis. They developed stochastic gradient boosting with recursive feature elimination (SGB-RFE) for dimensionality reduction, followed by the adaptive Harris Hawk optimization algorithm. Finally, they exploited EDGA for the classification purpose, achieving 99.77% on the UCSF heart disease database. Rani et al. [21] exploited a combination of GA and a recursive feature elimination method for feature selection and a hybrid system consisting of SVM, NB, LR, RF, and Adaboost classifiers for prediction. The system performed best with the RF classifier, giving 86.6% overall accuracy.

Once the literature on CVD is analyzed, it could note that most of the accepted procedures for CVD predic-

tion include a dimensionality reduction technique and a learning mechanism for classification. The dimensionality of the medical dataset is huge so, there is a need for an appropriate parameter selection mechanism to restrict the number of features and to increase classification accurateness. Also, a machine learning procedure is necessary to progress the working performance of the system predictions. In addition, with the modern developments in classification practices such as ensemble learning, there could exist a means of further refinement in terms of accuracy, efficiency, and robustness of the predictions. Evolutionary or metaheuristic attitudes open a wide range of benefits over old-fashioned approaches as they encompass least domain-specific information. For attaining a reduced dimensional robust attribute space, GA is exploited in this work. Furthermore, when compared to a single classification algorithm, ensemble techniques usually produce more accurate results. Consequently, the suggested method employs the ensemble approach to categorization.

3. PROPOSED METHODOLOGY

The sole intention of this research project is to confer a robust prediction method for CVD prediction. So, here the model proposes a system consisting of three major stages; data preprocessing, feature selection, and classification. But the proposed system could be accepted only if it produces acceptable results, so we carried out the experiments on the Cleveland heart- dataset from the UCI repository [22], which is a benchmark in heart disease prediction research. The comprehensive structure of the proposed system is shown in Fig. 1.

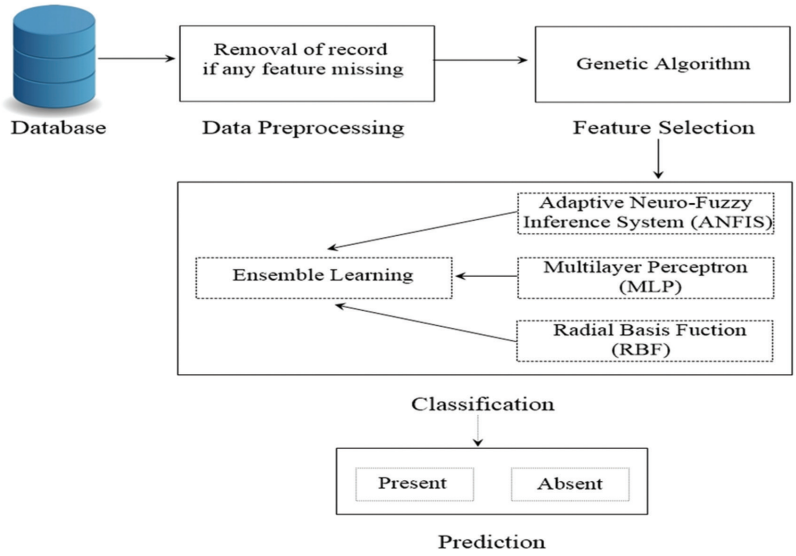

Fig 1. Structure of the Proposed CVD Prediction System

3.1. DATASET AND PRE-PROCESSING

The Cleveland Heart Disease Database formerly consisted of 303 instances with 75 parameters. But a subset of the database with 14 parameters is made available and utilized in most of the research, so this work is also using the same method in this experiment. The age and sex of the data providers along with other disease-related symptoms, such as type of chest pain, blood pressure level, cholesterol level, sugar level, etc., are some of the features included in the database. The 14 included features and their description are given in Table 1.

Once the dataset is analyzed, we notice that out of 303 clinical records, 6 records were missing with some feature values. Those 6 records with missing values were removed in the pre-processing stage as we acknowledged that these records could reduce model performance.

3.2 FEATURE SELECTION

One of the widely used Genetic Algorithms (GA) is employed for eliminating irrelevant and redundant features. Darwin's natural selection is the basic idea behind GA; that is, it works on the principle that the fittest shall survive. It is one of the commonly used feature selection methods [23, 24]. The major characteristics of GA that make it a favorite of researchers; beginning with a set of arguments, GA has the potential to reach a closer optimal solution through a quick convergence, a simple fitness function is only utilized for evaluation, and it supports parallel programming.

3.1. DATASET AND PRE-PROCESSING

The Cleveland Heart Disease Database formerly consisted of 303 instances with 75 parameters. But a subset of the database with 14 parameters is made available and utilized in most of the research, so this work is also using the same method in this experiment. The age and sex of the data providers along with other disease-related symptoms, such as type of chest pain, blood pressure level, cholesterol level, sugar level, etc., are some of the features included in the database. The 14 included features and their description are given in Table 1.

Table 1. Overall structure of the database

Attribute Number	Feature	Explanation
i.	Age	The age completed on the last birthday
ii.	Gender details	Male = 1; Female = 0
iii.	Chest ache is categorized into four types.	Typical angina = 1, atypical angina = 2, non-angina pain = 3, asymptomatic = 4
iv.	Level of BP at resting mode.	Represented anywhere from 94 mm/Hg to 200 mm/Hg
v.	Serum cholesterol level in mg/dl	Represented anywhere from 126 mg/dl to 564 mg/dl
vi.	Sugar levels on fasting >120 mg/dl	True is 1 and false is 0.
vii.	Electrocardiographic results at resting mode.	Hypertrophy equals 2, ST-T wave abnormalities equals 1, and the normal equals 0.
viii.	The highest possible heart rate was noted.	Anywhere from 71 to 202
ix.	Angina brought by exercise	1 represents "yes", and 0 represents "no".
x.	Generated ST depression vs. at rest.	Up-sloping equals 1, Flat equals 2, and down-sloping equals 3.
xi.	Peak exercise slope	SPE ranging from 0 to 6.2
xii.	Major vessel count (0–3) colored by fluorescence	From 0 to 3.
xiii.	Thallium, the status of the heart.	Normal is 3, the fixed defect is 6, and the reversible defect is 7
xiv.	Target value	The presence of heart disease equals 1, and no heart disease 0

Once the dataset is analyzed, we notice that out of 303 clinical records, 6 records were missing with some feature values. Those 6 records with missing values were removed in the pre-processing stage as we acknowledged that these records could reduce model performance.

3. 2. FEATURE SELECTION

One of the widely used Genetic Algorithms (GA) is employed for eliminating irrelevant and redundant features. Darwin's natural selection is the basic idea behind GA; that is, it works on the principle that the fittest shall survive. It is one of the commonly used feature selection methods [23, 24]. The major characteristics of GA that make it a favorite of researchers; beginning with a set of arguments, GA has the potential to reach a closer optimal solution through a quick convergence, a simple fitness function is only utilized for evaluation, and it supports parallel programming.

Any GA encompasses three basic stages: selection, crossover, and mutation.

i. Initially, a set of random populations is created from the data.

ii. Iterated until a generation is formed that satisfies the given fitness threshold value.

a. A fitness function is designed to find the fittest population. We used kNN to estimate the accuracy, then the error was evaluated to find the fitness of each population.

b. In the selection phase, a selected set of features known as chromosomes is placed in a mating pool and generates the next generation. Features in the mating pools are combined to form better generations.

c. The mutation is used to make sure that the next generation is not like the present ones.

3. 3. CLASSIFICATION

In this part, the final strategy of the proposed technique, that is, classification is discussed. The model is aiming to prognosticate the existence and non-existence of heart disease automatically from some of its related parameters. So, here it is dealing with a binary classification algorithm. From the literature analyzed, ensemble models are more reliable and robust in comparison with basic individual methods [25, 26]. Fundamentally, the idea of ensemble learning is that the data is trained using more than a single learning model, and these models are used as the input to the ensemble learning strategy. We have incorporated learning models such as ANFIS, MLP, and RBF to model our ensemble learning procedure.

3.3.1. Adaptive Neuro-Fuzzy Inference System (ANFIS)

Initially, the Adaptive Neuro-Fuzzy Inference System (ANFIS) was included in the scheme. J.S. Roger Jang created the ANFIS in 1993, and is extensively considered as a universal estimator or Takagi-Sugeno fuzzy system [27, 28]. It works with the combined characteristics of ANN and fuzzy inference systems (FIS). Fig. 2 puts forward the basic architecture of the ANFIS classification methodology. From Fig. 3, we could understand that the ANFIS basic architecture consists of 5 layers; Fuzzification, Multiplication, Normalization, De-fuzzification, and Summation.

In the ANFIS model, each rule forms an output based on a linear combination of an input variable and a constant. So, the ultimate output is formed by taking the weighted average of all the outputs from each rule. The following definitions apply to the IF-THEN rules for a Takagi-Sugeno system with two inputs.

$$R1 = a1p + b1q + c1, \text{ if } p = X1 \text{ and } q = Y1 \quad (1)$$

$$R2 = a2p + b2q + c2, \text{ if } p = X2 \text{ and } q = Y2 \quad (2)$$

Here p and q are the inputs in the feature set; Xi and Yi are the linguistic labels; ai, bi, and ci are the consequent parameters; $R1$ and $R2$ are the output fuzzy membership functions.

The fuzzification layer is in charge of mapping the given inputs in the range of 0–1 with the fuzzy sets set up on membership functions (MFs) described using the universal bell function. Every node 'i' in the input

layer is an adaptive node whose node function is represented as follows:

$$O_{1,i} = \mu_{Xi}(p), \quad \text{where } i = 1,2,\ldots \tag{3}$$

$$O_{1,j} = \mu_{Yi}(q), \quad \text{where } j = 1,2,\ldots \tag{4}$$

Where $\mu_{Xi}(p)$, and $\mu_{Yi}(p)$ are the membership functions of the linguistic labels Xi and Yj respectively.

The multiplication layer estimates the number of fuzzy rules to be involved in the system. The output of the multiplication layer defines the dismissal strength of IF-THEN rules which is stated as follows:

$$O_{2,i} = w_i = \mu_{Xi}(p), x\, \mu_{Xi}(q) \tag{5}$$

Where i=1, 2,…, and w_i denotes the firing strength of i-th rule.

The neurons in the normalizing layer are fixed and are normalized using the weights of all the neurons in the layer. The following equation is utilized to calculate the node output is calculated.:

$$O_{3,i} = \overline{w}_i = w_i / (\sum_i w_i), \quad \text{where } i = 1, 2,\ldots \tag{6}$$

In the de-fuzzification layer, every node is an adaptive node that embraces the resulting parameters of the model. The output node can be obtained as follows:

$$O_{4,i} = \overline{w}_i z_i = w_i (a_i p + b_i q + c_i), \quad \text{where } i = 1,2,... \tag{7}$$

The final summation layer appends all the outputs from the previous layer to draw the final output of the ANFIS architecture as follows:

$$O_{5,i} = \sum_i \overline{w}_i z_i = \sum_i w_i z_i / \sum_i w_i, \quad \text{where } i = 1, 2,... \tag{8}$$

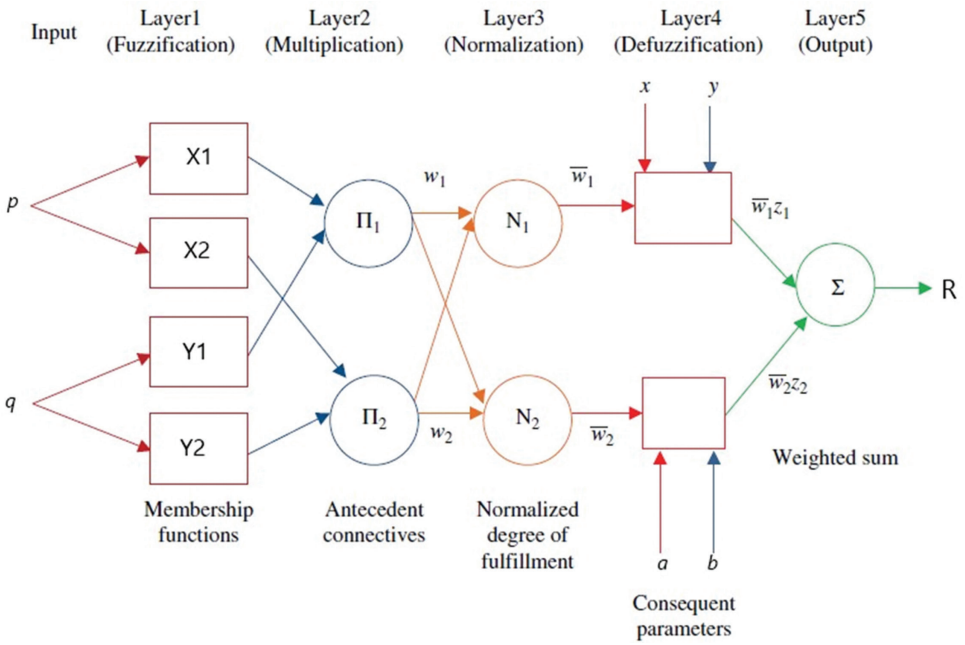

Fig 2. The basic structure of ANFIS

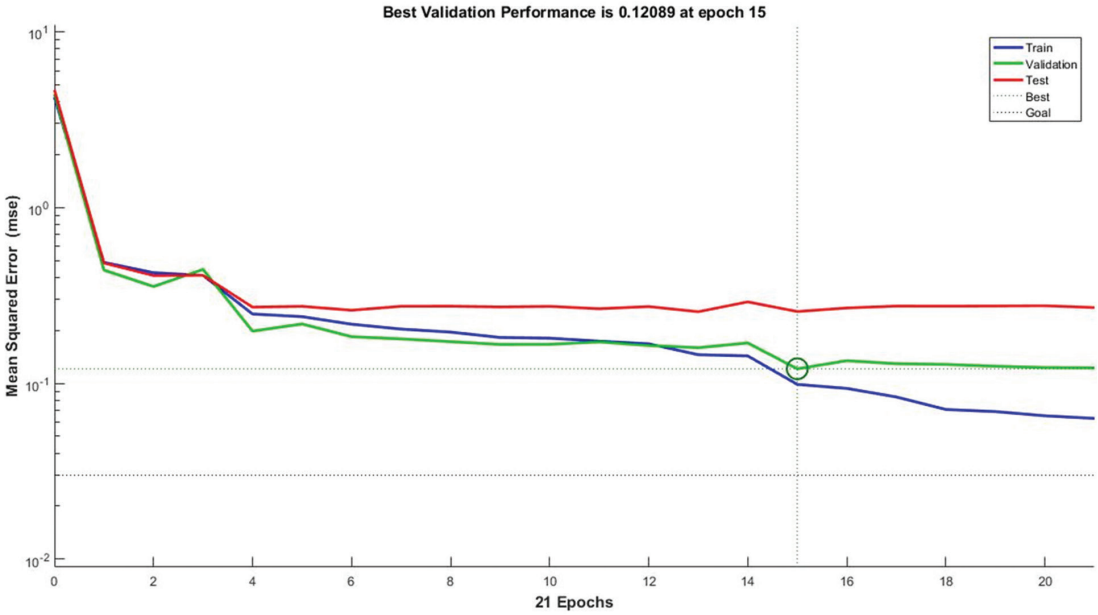

Fig 3. Performance graph of MLP model

3.3.2. Multi-Layer Perceptron (MLP)

The multilayer perceptron is a widely used feed-forward neural network due to its faster execution, easier implementation, and need for comparatively lesser training values [29, 30]. Generally, MLP encompasses three sequential layers, namely, the input layer, the hidden layer, and the output layer. The hidden layer is accountable for processing and executing output from the input layer to the output layer. The number of neurons should be analyzed according to the subject and data distribution to solve the issues of bad generalization and overfitting.

Here hidden layer with 7 neurons has been used for experimental purposes. The input signals xi that each neuron j in the hidden layer receives are multiplied by the corresponding connection weights wij before summing them together as follows:

$$Z_i = f(\sum w_{ji} x_i) \quad (9)$$

where in *f* is a function of activation using the weighted sums of the inputs. If the hyperbolic tangent function is taken as the activation function, which can be estimated as follows

$$f = tansig(x) = 2 / (1+ exp(-2*x)) -1 \quad (10)$$

The performance graph of MLP classification based on Mean Squared Error (MSE) at 21 epochs of the system is presented in Fig. 3.

3.3.3 Radial Basis Function (RBF)

Another widely used ANN method is the Radial Basis Function (RBF), which also consists of three layers similar to the MLP network [31, 32]. Each neuron in the hidden layer contains a radial basis function as the activation function, while each neuron in the input layer contains a certain predictor variable. A weighted sum of the hidden layer's outputs, which is a Gaussian function (Ψ), is included in the output layer and is shown below.

$$\Psi i = exp(-(X-Ci^2)/(2\sigma^2)) \quad (11)$$

here *ci* is the i-th neuron's centre vector and σ is a vector representing basis width. The output of the i-th RBF network can be computed using the following:

$$y_i = \sum_{q=1}^{p} W_{iq} \Psi_q \quad (12)$$

where Wiq is the weight of the *q-th* hidden neuron to the *i-th* output. The performance graph based on the MSE of the train set of our suggested system is presented in Figure 4.

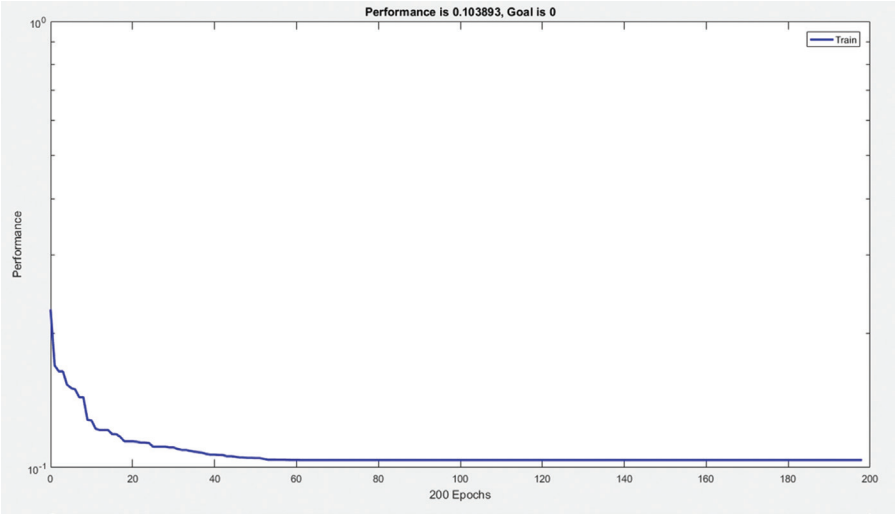

Fig. 4. Performance graph of the RBF model

3.3.4. Ensemble Learning

One of the most advanced strategies for addressing the concerns on classification and prediction problems are ensemble learning methods. It works on the principle that a combination of predictions could lead to better predictions [33]. Ensemble learning techniques provide benefits such as reduced overfitting, reduced finishing with local minimums, reduced impact of the curse of dimensionality, best-fit data, and improved class balances [34, 35].

There are different ways to form an ensemble learning mechanism and we have used the infamous boosting strategy to combine the results of previously mentioned learning methods.Boosting models carry out predictions by placing weak learners in a sequential manner and they reduce overfitting by re-weighting misclassified data of classifiers. We have used a boosting algorithm known as LogitBoost as a meta-learning classifier [13, 36, 37]. LogitBoost has proved better at handling noisy data by minimizing bias and variance. The main concept of the logitboost algorithm is described below:

i. Set the initial weight to $w_i = 1/N$, for $i = 1, 2, \ldots N$, and probability estimate $p(x_i)=1/2$.

ii. Iterate for $m = 1, 2, \ldots M$:

   a. Calculate the responses and weights below:
   $z_i = (yi^* - p(x_i))/(p(x_i)(1-p(x_i))$
   $w_i = p(x_i)(1-p(x_i))$

   b. Compute the function $f_m(x)$ by a weighted least square regression of zi to *xi* using weights *wi*.

c. Update $F(x) \leftarrow F(x)+1/2\, f_m(x)$ and $p(x) \leftarrow (e^{F(x)}) / (e^{F(x)} + e^{-F(x)})$.

iii. Output the classifier $sign[F(x)] = sign[\sum^{M}_{m=1} f_m(x)]$.

The flow of data in our ensemble learner is shown in Fig. 5.

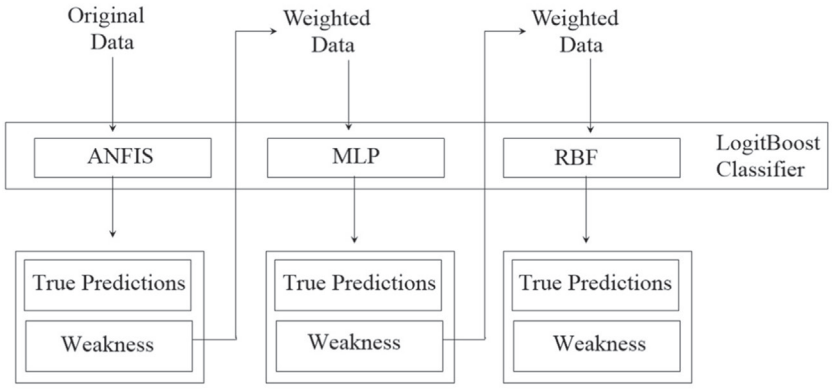

Fig. 5. The overall structure of proposed ensemble learning architecture

4. EXPERIMENTS AND RESULTS

Here, the experiments and tests that were conducted to establish the performance superiority of the suggested CVD diagnostic model will be discussed. This model has used MATLAB 2019 an environment to perform the system experiments with UCI-Cleveland heart disease datasets (discussed in Section 3.1). This technique has employed commonly used performance metrics like accuracy, specificity, sensitivity, precision, and f-measure to scale the results [21, 38]. The analysis of these metrics is performed using a confusion matrix, which produces a total outcome of four measures; true positive *(T_P)*, true negative *(T_N)*, false positive *(F_P)*, and false negative *(F_N)*. We could analyze the needed measures from these parameters with the following equations:

$$Accuracy = (T\_P+T\_N) / (T\_P+T\_N+F\_P+F\_N)$$

$$Specificity = T\_N / (T\_N+F\_P)$$

$$Sensitivity = T\_P / (T\_P+F\_N)$$

$$Precision = T\_P / (T\_P+F\_P)$$

$$F\text{-}measure = (2*T\_P) / (2*T\_P+F\_N+F\_P)$$

The results are validated by K-fold cross-validation with K = 10. Performance was assessed with the k=1 data for training and the balance one for testing, k times, with different sets for training and testing each time. Initially, it was experimented without using any feature selection strategy on the pre-processed data. It was noted that the ensemble model performed pretty well even without any feature reduction techniques. Ensemble learning outperformed the use of singular learning techniques as well. The performance of ANFIS is also acceptable in certain cases. Then it progressed to use GA for feature elimination and found a huge rise in performance metrics, so as inferred from the literature, the use of feature selection is highly useful in the case of CVD diagnosis databases.

Here the number of iterations is used as the factor of generation approval and set the number of iterations is 50. Additionally, the cross-over probability has set at 0.25 and the mutation probability at 0.01. The convergence graph for our fitness value vs.the number of iterations is presented in Fig. 6. We have utilized the K-NN algorithm to estimate the cost of the attributes selected in the genetic algorithm. The final set of features consisted of six attributes, numbered i, ix, x, xi, xiii, and xiv.

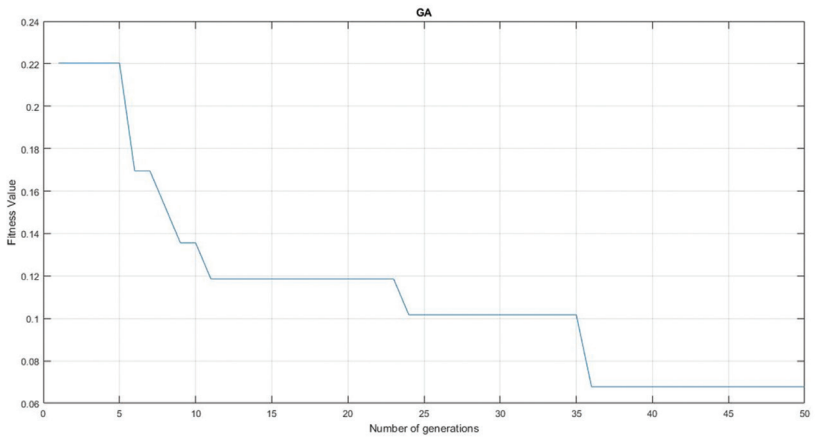

Fig. 6. Graph for Fitness Value VS. Number of Iterations in Genetic Algorithm

The model has obtained a rise of about 10% in most of the techniques with the introduction of GA. The overall accuracy of the ensemble learning was 92.31%, which has shifted to 99.66% with the incorporation of GA. If we take individual techniques of classifications, ANFIS saw a rise from 86.35% to 92.15%, MP hiked from 73.03% to 83.65%, and RBF accuracy changed from 70.79% to 83.17%. The complete experimental results are shown in Table 2. Additionally, for better understanding, we have presented a graphical representation of the comparative analysis of performances based on accuracy in Fig. 7.

Table 2. Performance of the Proposed System with and without Genetic Algorithm for feature selection

Method		Accuracy	Specificity	Sensitivity	Precision	F-Measure
Without GA	ANFIS	86.35	82.53	86.80	88.91	89.63
	MLP	73.03	82.14	68.85	89.36	77.78
	RBF	70.79	7.14	100	70.11	82.43
	Ensemble	92.31	97.73	82.89	95.45	88.73
With GA	ANFIS	92.13	77.50	97.53	94.05	95.76
	MLP	83.65	90.91	71.05	81.82	76.06
	RBF	83.17	92.19	68.75	84.62	75.86
	Ensemble	99.66	100	99.27	100	99.63

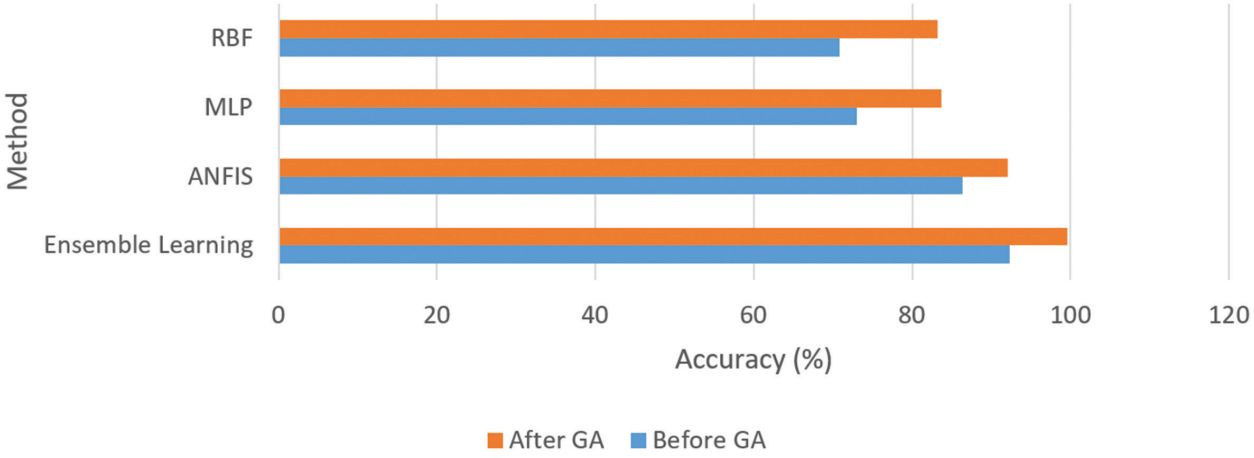

Fig 7. The Comparative Analysis of the Different Algorithms Before and After GA, Based on Accuracy

Furthermore, we have compared the performance of the suggested system against previous literature on the subject of interest. Table 3 presents the comparative analysis of the same. Here, it has included 10 recent works from the past few years in the literature on CVD prediction and diagnosis for comparison purposes. In addition, some comparisons have been made with only those works that provided results on the Cleveland heart disease database to ensure the worthiness of comparison. The results exhibit the superiority of the proposed system over the others based on the accuracy of the system's performance.

Table 3. The Comparison of the proposed system performance against the state-of-art techniques

Author	Feature Selection	Classification	Overall Accuracy
Gokulnath and Shantharajah [8]	GA-SVM	SVM	88.34%
Kanwal et al. [10]	GA	NN	89.00%
Durga Devi [12]	GA	RBF	85.48%
Kishore and Jayanthi [14]	MCDM	ANN	94.15%
Sharma and Parmar [15]	Talos	DNN	90.78%
Mohan et al. [16]	Less Error Classifier	HRFLM	88.70%
Ali et al. [17]	Conditional probabilistic approach	Ensemble deep learning model, and ontology-based recommendation	98.50%
Javeed et al. [18]	FWAFE	DNN	93.33%
Khourdifi and Bahaj [19]	FCBF, PSO and ACO	KNN	99.60%
Rani et al. [21]	GA and Recursive Feature Elimination	RF	86.60%
Proposed Method	GA	Ensemble Learning	99.63%

We could note that both [10] and [11] used a similar feature selection approach as ours, but the results projects differently, this can be due to the difference in the fitness functions used within the GA which has resulted in a different set of features selected in all the cases. Also, the choice of classifiers is another main reason for this disparity. The only system that comes close to the proposed model's performance is [19], but they used a lot of optimization strategies that could add complexities to the system.

5. CONCLUSION

Presently, the cause of the highest mortality rate in the world is cardiovascular disease. Early detection of these diseases could have a high impact on reducing these estimations. Therefore, in this research project, it is aimed to propose a robust automated heart failure prediction system. This method is not trying to replace doctors or heart specialists using the system, but to aid them with a quick diagnosis of the disease. Also, it is being sensed that these systems could be of great use in remote or rural areas where there is a lack of these specialists. In this research work, a GA-based ensemble learning technique that could predict the presence of heart diseases from commonly used clinical data is proposed. An overall accuracy of 99.6% was achieved with the proposed system. The proposed investigational out-turn reveals that the system surpasses some of the existing state-of-the-art works

6. REFERENCES

[1] "Cardiovascular diseases (CVDs)", https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed: 2022)

[2] Centers for Disease Control and Prevention, "heart disease facts", https://www.cdc.gov/heartdisease/facts.htm (accessed: 2022)

[3] C. W. Tsao et al. "Heart disease and stroke statistics-2023 update: A report from the American Heart Association", Circulation, Vol. 147, No. 8, pp. e93-e621, 2023.

[4] L. A. Allen, "Decision making in advanced heart failure: a scientific statement from the American Heart Association: A scientific statement from the American heart association", Circulation, Vol. 125, No. 15, 2012, pp. 1928-1952.

[5] H. Yang, J. M. Garibaldi, "Automatic detection of protected health information from clinic narratives", Journal of Biomedical Informatics, Vol. 58, 2015, pp. S30-S38.

[6] J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, N.-W. Chang, H.-J. Dai, "Coronary artery disease risk assessment from unstructured electronic health records using text mining", Journal of Biomedical Informatics, Vol. 58, 2015, pp. S203-S210.

[7] P. Melin, I. Miramontes, G. Prado-Arechiga, "A hybrid model based on modular neural networks and fuzzy systems for classification of blood pressure and hypertension risk diagnosis", Expert Systems with Applications, Vol. 107, 2018, pp. 146-164.

[8] C. B. Gokulnath, S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease", Cluster Computing, Vol. 22, No. S6, 2019, pp. 14777-14787.

[9] S. M. Nagarajan, V. Muthukumaran, R. Murugesan, R. B. Joseph, M. Meram, A. Prathik, "Innovative feature selection and classification model for heart disease prediction", Journal of Reliable Intelligent Environments, Vol. 8, 2022, pp. 333-343.

[10] S. Kanwal, J. Rashid, M. W. Nisar, J. Kim, A. Hussain, "An effective classification algorithm for heart disease prediction with genetic algorithm for feature selection", Proceedings of the Mohammad Ali Jinnah University International Conference on Computing, 2021.

[11] M. Anbarasi, E. Anupriya, N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", International Journal of Engineering Science and Technology, Vol. 2, No. 10, 2010, pp. 5370-5376.

[12] A. Durga, D. M. Phil, "Enhanced prediction of heart disease by genetic Algorithm and RBF Network", International Journal of Advanced Information in Engineering Technology, Vol. 2, No. 2, 2015.

[13] V. Jothi Prakash, N. K. Karthikeyan, "Enhanced evolutionary feature selection and ensemble method for cardiovascular disease prediction", Interdisciplinary Sciences: Computational Life Sciences, Vol. 13, No. 3, 2021, pp. 389-412.

[14] A. H. N. Kishore, V. E. Jayanthi, "Neuro-fuzzy based medical decision support system for coronary artery disease diagnosis and risk level prediction", Journal of Computational and Theoretical Nanoscience, Vol. 15, No. 3, 2018, pp. 1027-1037.

[15] S. Sharma, M. Parmar, "Heart diseases prediction using deep learning neural network model", Interna-

tional Journal of Innovative Technology and Exploring Engineering, Vol. 9, No. 3, 2020, pp. 2244-2248.

[16] S. Mohan, C. Thirumalai, G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", IEEE Access, Vol. 7, 2019, pp. 81542-81554.

[17] F. Ali et al. "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion", Information Fusion, Vol. 63, 2020, pp. 208-222.

[18] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification", Mobile Information Systems, Vol. 2020, 2020, pp. 1-11.

[19] Y. Khourdifi, M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization", International Journal of Intelligent Systems, Vol. 12, No. 1, 2019, pp. 242-252.

[20] A. A. Bakhsh, "High-performance in classification of heart disease using advanced supercomputing technique with cluster-based enhanced deep genetic algorithm", Journal of Supercomputing, Vol. 77, No. 9, 2021, pp. 10540-10561.

[21] P. Rani, R. Kumar, N. M. O. S. Ahmed, A. Jain, "A decision support system for heart disease prediction based upon machine learning", Journal of Reliable Intelligent Environments, Vol. 7, No. 3, 2021, pp. 263-275.

[22] Heart disease dataset, https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ (accessed: 2022)

[23] S. Sayed, M. Nassef, A. Badr, I. Farag, "A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets", Expert Systems with Applications, Vol. 121, 2019, pp. 233-243.

[24] A. K. Das, S. K. Pati, A. Ghosh, "Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm", Knowledge and Information Systems, Vol. 62, No. 2, 2020, pp. 423-455.

[25] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, P. N. Suganthan, "Ensemble deep learning: A review", Engineering Applications of Artificial Intelligence, Vol. 115, No. 105151, 2022, p. 105151.

[26] Y. Cao, T. A. Geddes, J. Y. H. Yang, P. Yang, "Ensemble deep learning in bioinformatics", Nature Machine Intelligence, Vol. 2, No. 9, 2020, pp. 500-508.

[27] J.-S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 23, No. 3, 1993, pp. 665-685.

[28] S. Chopra, G. Dhiman, A. Sharma, M. Shabaz, P. Shukla, M. Arora, "Taxonomy of adaptive Neuro-Fuzzy Inference System in modern engineering sciences", Computational Intelligence and Neuroscience, Vol. 2021, 2021, p. 6455592.

[29] U. Orhan, M. Hekim, M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model", Expert Systems with Applications, Vol. 38, No. 10, 2011, pp. 13475-13481.

[30] H. Taud, J. F. Mas, "Multilayer Perceptron (MLP)", Geomatic Approaches for Modeling Land Change Scenarios, Springer International Publishing, 2018, pp. 451-455.

[31] E. Meng Joo, W. Shiqian, L. Juwei, T. Hock Lye, "Face recognition with radial basis function (RBF) neural networks", IEEE Transactions on Neural Networks and Learning Systems, Vol. 13, No. 3, 2002, pp. 697-710.

[32] E. Assareh, M. Biglari, "A novel approach to capture the maximum power from variable speed wind turbines using PI controller, RBF neural network and GSA evolutionary algorithm", Renewable and Sustainable Energy Reviews, Vol. 51, 2015, pp. 1023-1037.

[33] M. Bulmer, "Galton's law of ancestral heredity", Heredity, Vol. 81, 1998, pp. 579-585.

[34] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization", Machine Learning, Vol. 40, 2000, pp. 139-157.

[35] R. Polikar, "Ensemble based systems in decision making", IEEE Circuits and Systems Magazine, Vol. 6, No. 3, 2006, pp. 21-45.

[36] I. Palit, C. K. Reddy, "Scalable and parallel boosting with MapReduce", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 10, pp. 1904-1916, 2012.

[37] M. H. Kamarudin, C. Maple, T. Watson, N. S. Safa, "A LogitBoost-based algorithm for detecting known and unknown web attacks", IEEE Access, Vol. 5, 2017, pp. 26190-26200.

[38] P. Rani, R. Kumar, A. Jain, R. Lamba, "Taxonomy of machine learning algorithms and its applications", Journal of Computational and Theoretical Nanoscience, Vol. 17, No. 6, 2020, pp. 2508-2513.