RESEARCH PAPER

Genetic Algorithms based optimization to enhance prediction accuracy of machine learning algorithms for cardiovascular heart disease.

Pratikkumar Parmar [1,3*], Nirali Shah [2]

[1]*Monark University, Gujarat, India*
[2]*Electronics & Communication Engg. Department, Hasmukh Goswami College of Engineering, Gujarat, India*
[3]*Electronics & Communication Engg. Department, Government Polytechnic College, Gandhinagar, India*
* Corresponding author's Email: pratik.parmar24192@gmail.com

Received: 21st July, 2026; Revised: 1st Aug, 2026; Accepted: 14th Aug, 2026; Available Online: 4th Sept, 2026

Abstract: Cardiovascular heart diseases rank among the leading global causes of mortality. However, accurate and early detection can considerably reduce fatality rates. In this context, to support early diagnosis and intervention in cardiovascular health, Machine Learning (ML) models have arisen as powerful tools. A robust predictive system for cardiovascular heart disease using a suite of ML-based algorithms is presented in this study. A comprehensive heart disease dataset from IEEE dataport is used to train the prediction models. The dataset is preprocessed by different processes, including data cleaning and standardization. This preprocessed dataset is bisected into training dataset and testing dataset. The machine learning algorithms, including Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF) are trained using the training dataset for the prediction process. Hyperparameters of these algorithms are tuned using Genetic Algorithm (GA) to optimize ML classifier algorithms and improve the efficiency of the prediction process. Standard performance parameters, accuracy, recall, precision, F1-score and ROC-AUC, are used for the evaluation of the performance of each optimized model and compared with state of the art model performances with raw data. Among all models, the GA optimized RF classifier demonstrated superior performance, achieving the highest accuracy of 95.10%, outperforming all other ML classifiers.

Keywords: CVD, ML algorithms, data cleaning, GA, Optimization, Accuracy
How to cite this article: Pratikkumar Parmar 13* Nirali Shah 2. Genetic Algorithms based optimization to enhance prediction accuracy of machine learning algorithms for cardiovascular heart disease.. Int J Drug Deliv Technol. 2026;16(80s):1310-1317. DOI: 10.25258/ijddt.16.80s.166.

1. Introduction

The heart plays a key role within the cardiovascular system, which also includes an extensive network of arteries, veins, and capillaries that facilitate blood transport. Cardiovascular diseases (CVDs) encompass a variety of heart-related conditions that typically arise from abnormalities or obstructions in normal blood flow. These disruptions can significantly impair the circulatory process. On a global level, heart-related illnesses remain among the most frequent and leading causes of mortality.

As per the World Health Organization(WHO), heart disease along with stroke, collectively lead to approximately 17.9 million deaths each year across the globe. Furthermore, heart attacks and strokes contribute to nearly 85% of all fatalities associated with cardiovascular diseases.

The diagnosis of heart disease typically involves evaluating the patient's reported symptoms along with a thorough physical assessment. Several factors are known to increase the risk of developing cardiovascular conditions, including usage of tobacco and alcohol, advancing age, a family history of heart-related illnesses, chest pain, high cholesterol levels, elevated blood pressure, lack of physical activity, hypertension, obesity, diabetes and psychological stress [1,2]. Some of these factors can be helpful for the prior diagnosis of cardiovascular disease. If a sufficient amount of accurate medical database is available about factors affecting the CVDs then by application of machine learning methodologies, it is possible to deal with heart-related disorders priorly, preventing the possibility of death by CVDs. Machine learning performs an essential role in this process by analyzing large and complex datasets from multiple perspectives and can uncover meaningful patterns from the historical medical database, based on which it can forecast the possible output of the fresh input data.[3]

With the increasing availability of patient data through modern infrastructure in healthcare systems, it has become possible to develop predictive models for detecting heart disease [4]. In recent scenarios, machine learning models like Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), and many more are widely used by researchers for CVDs detection. These algorithms provide a considerable level of accuracy in prediction. Besides the ML algorithms, the dataset also holds significant importance [5, 6, 7]. Before implementing machine learning algorithms, mandatory requirement is to visualize the data and clean and preprocess the data to make it ready to be utilized by ML algorithms. Inaccurate data mislead the model to give erroneous results in the prediction process. Many data repositories like UCI, Kaggle, IEEE Dataport provide authentic datasets to develop prediction models.

This work is focused towards the data preprocessing, ML classification algorithms and optimization of algorithms to elevate the accuracy of the prediction model. This work incorporates data preprocessing techniques,

including data cleaning, and data scaling, to make the dataset ready to train the ML algorithms for the prediction process. These algorithms are optimized by tuning the hyperparameters of the algorithms.

2. Literature Review

Many researchers have worked on prior prediction of heart diseases based on publicly available datasets and developed various methods of prediction.

Ankur et al. [5] Factor Analysis of Mixed Data (FAMD) has been implemented for feature extraction with data imputation and standardization, which give the accuracy of 93.44% with RF classifier.

Ramesh et al. [6] have utilized UCI heart disease dataset on which data preprocessing methods applied included normalization and aggregation, along with the Isolation Forest method for anomaly removal. The study employed supervised machine learning classification algorithms. Among all comparisons, KNN with eight neighbors reached the highest accuracy level of 94.1%.

Arsalan et al. [7] have used data of 518 Pakistani patients for the Cardiovascular Disease (CVD) prediction process. After preprocessing data by data cleaning and data imputation, different classification algorithms were applied for the prediction process, from which RF gave the highest classification accuracy of 85.01%.

Niloy et al. [8] has used the UCI Cleveland dataset with 13 features. Data preprocessing involved imputation and applying Standard Scaling. Feature selection techniques created subsets of dataset and Supervised machine learning algorithms like LR, KNN, SVM, NB, DT and RF were applied. Out of them the highest value of accuracy was reached by Random Forest which is 94.51%.

Jafat et al. [9] with UCI Dataset, Relief, FCBF, and Genetic Algorithms (GA) Feature selection employed for data preprocessing. Work showed that Stacked-Genetic algorithm (ST-GA) with AdaBoost technique achieved the highest accuracy of 97.57%.

Rohit et al. [10] involved Isolation Forest as Data preprocessing for outlier detection, Lasso for feature selection, and data normalization to prevent overfitting. Along with that hyperparameter tuning and dropout layers were key optimization techniques. The Deep Learning model achieved the highest accuracy of 94.2%.

Santosh et al. [11] has used UCI repository dataset. Feature selection is done using GA and RF algorithm is used for classification. This GA-RF combination delivered an accuracy of 93.2%

Hosam et al. [12] have implemented the data balancing technique. After data preprocessing including handling inconsistencies, cleansing and normalization they have implemented Synthetic Minority Oversampling Technique (SMOTE) to balance the imbalanced classes. Using XGBoost classifier achieved the highest accuracy of 97.57%.

Zeinab et al. [13] considered Z-Alizadeh Sani dataset and selected the features by 'weight by SVM' method. GA was used as an optimization technique to enhance the initial weights of the neural network, thereby boosting its overall performance.

Ghulam et al. [14] has worked on merged Kaggle dataset and underwent processes like cleaning, encoding and scaling. After preprocessing on dataset, machine learning classifiers are used for prediction. Hyperparameters were tuned using RandomSearchCV method and classifiers were optimized for better performance. KNN exhibited the highest performance, achieving 91.8% accuracy with optimized classifier.

Chandra et al. [15] has implemented GA-SVM approach and selected 7 key features for classification and Z-score Normalization is applied. With this pre-processed data, classification is carried out in which SVM achieved higher accuracy of 88.34% compared to process on all features.

Mirza et al. [16] has investigated six machine learning algorithms with UCI datasets. It has implemented min max and standard scaler for scaling process and SMOTE-ENN technique for Data balancing as a part of data preprocessing. Hyperparameter were optimized via GridSearchCV and RandomSearchCV. RF with SMOTE-ENN and standard scaler has achieved highest accuracy of 90%.

Parmonangan et al. [17] has worked on combined (Cleveland and Hungarian) dataset containing 596 samples. Data preprocessing includes cleaning missing values and feature selection process. GA is used to optimize different hyperparameters of RF like max_features, min_sample_leaf, n_estimator, min_sample_split and max_depth. RF with optimized parameter with GA gave the highest accuracy of 85.83%.

However, state-of-the-art ML algorithms are lagging in the accuracy of prediction, which can be enhanced by supporting processes. Therefore, different data preprocessing techniques can be implemented to improve the quality of the dataset. Besides this ML algorithms can be optimized to improve the efficiency of the prediction process.

3. Methodology

This study follows a well-structured process to analyze data related to heart disease. It begins with using a combined dataset of five well-known datasets—Cleveland, Long Beach VA, Switzerland, Hungarian and Stalog—based on 11 shared features from IEEE Dataport. Then the workflow shifts towards cleaning the dataset to remove any errors or inconsistencies. To make the data suitable for ML models, data cleaning and scaling techniques are applied, ensuring it remains balanced and consistent. Using this preprocessed dataset, the training and testing process is carried out for several machine learning algorithms to evaluate their ability to predict heart disease accurately. Hyperparameters of algorithms are fine-tuned to enhance the performance of each model. Key performance parameters, including accuracy, recall, precision, F1-score, and ROC-AUC, are used to compare the results. Ultimately, the objective is to identify the most

effective and reliable ML method to predict heart disease, based on how well each performs across these parameters.

3.1 Dataset

In this work, the Comprehensive Heart Disease Dataset from IEEE Dataport is used [18]. It is a comprehensive mixture of independently available five popular heart disease datasets, namely Cleveland, Switzerland, Statlog, Long Beach VA and Hungarian (Heart) Data Set. This dataset consists of 11 independent features, each having 1190 samples, including sex, age, cholesterol, type of chest pain, Maximum heart rate, RestingBP, resting ECG, Fasting BS, old peak, exercise angina, ST slope and 1 target feature [19]. Out of these 11 features, 6 are categorical features and 5 are continuous-valued features. It includes a total of 561 records having heart disease, which are marked as 1 in the target class.

3.2 Data Pre-Processing

Data pre-processing is required to make raw data ready to be used by ML model. Raw data often contains noise, missing values or irrelevant features that can negatively impact model accuracy. This process improves the quality of data, enhance model performance and reduce computational cost. Various steps of data preprocessing are as follow:

3.2.1 Data Cleaning

In the data cleansing process, samples with erroneous data are identified, which are having missing data, outliers, duplicate values or other incorrect data values [20]. Dataset used in this work is not having any missing data or duplicate values. However, Cholesterol, RestingBPS, and Old peak features have outliers. Only the farthest outliers are removed, but soft outliers are considered part of the dataset because those datapoints contribute to increasing the model's generalization. In Cholesterol feature, many samples have 0 value, which is not practically possible, besides this 0 value of restbps and negative value of oldpeak is not possible. So those samples are removed from dataset.

3.2.2 Data Scaling

Data scaling is a process to bring feature values in an equal range. Various techniques like standardization, Normalization, min-max scaling can be used to scale the data [23]. In this work Standard Scaler standardization technique is used to scale the data. It brings the feature values on same level having mean at 0 and standard deviation of 1, allowing all features to contribute equally in the process.

3.2.3 Data Splitting

The full dataset is not directly utilized to train the model. It is randomly divided into two parts, out of which one subset trains the model, while the other evaluates its performance. The dataset can be divided into two categories with different training-to-testing ratios. In this work, an 80:20 training-to-testing ratio is used, based on which 80% of the total dataset serves as training data, while the remaining 20% dataset serves as testing data.

3.3 Machine Learning Techniques

Machine learning focuses on building statistical models, driven by data analysis, that enable computational systems to identify patterns and perform predictive tasks autonomously, without the need for task-specific programming instructions. This work has used different machine learning algorithms for classification purposes.

3.3.1 Logistic Regression (LR)

LR is a broadly used technique in statistics as well as machine learning for classifying data into one of two possible categories. The algorithm estimates the likelihood of a binary result by passing a linear mix of input variables through a sigmoid function, which maps output values in a range of 0 to 1. It is valued for its simplicity and efficiency, especially with large, balanced datasets. However, logistic regression is not ideal for datasets with strong correlations between variables and may be influenced by extreme values in the data [20, 21].

3.3.2 Gaussian Naive Bayes (GNB)

Naive Bayes is a ML technique grounded in Bayesian probability, used to evaluate the probability of different outcomes based on observed data. One popular version, Gaussian Naive Bayes (GNB), considers that the input data follow a Gaussian distribution. As a probabilistic model, GNB applies Bayes' theorem to classify data efficiently. To enhance reliability, the ‘var_smoothing’ parameter is used to add a small value to the variance estimates, preventing them from becoming too small and thus reducing the risk of overfitting [20, 24, 25].

3.3.3 K Nearest Neighbors (KNN)

KNN is a widely used supervised ML algorithm for classification that retains the entire training dataset and makes decisions at the time of prediction. To classify a new data point, KNN locates the ‘K’ nearest neighbours within the feature space and gives the label of the majority class to the sample. Proximity is calculated using distance metrics such as Euclidean, Manhattan, Minkowski or Hamming, based on the characteristics of the data. The number of neighbours 'K', is a key hyperparameter that needs tuning for optimal results. KNN is especially effective for pattern recognition tasks where the decision boundaries are not linear and can adapt well when the data is distributed meaningfully in feature space [21, 24, 26].

3.3.4 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that separates data into distinct classes within a multi-dimensional feature space. It accomplishes this by determining the best possible hyperplane to distinguish the data points into distinct categories. The algorithm focuses on finding the boundary that enlarges the gap between the closest samples from each class, known as support vectors. These data points are crucial in setting the location and alignment of the hyperplane. By maximizing this margin, SVM enhances classification accuracy and robustness, especially in multi-dimensional spaces. This makes it highly suitable for complex, non-linear data distribution tasks [24, 27, 28].

3.3.5 Decision Tree (DT)

DT is a supervised learning model that organizes data in a hierarchical structure. Within this framework, internal nodes are decision splits, determined by particular features, branches represent the output of these decisions, and leaf nodes are the final classifications. The decision tree building process starts with selecting the feature that

most effectively splits the data. The model recursively splits the dataset into more specific subsets. This partitioning continues until certain stopping conditions are met, such as reaching pure nodes or a maximum depth. Decision Trees are intuitive and powerful tools for classification tasks, especially when interpretability is important [26, 29].

3.3.6 Random Forest (RF)

RF algorithm is a popularly used supervised learning method suitable for classification as well as regression tasks. It operates using an ensemble of decision trees, combining their outputs to produce more accurate and reliable predictions. This approach is particularly effective even when some data entries are missing, offering robust performance under imperfect conditions. RF works in two main stages: first, several decision trees are developed, each trained on different segments of the training dataset and features; second, the collective predictions are integrated, most often through majority voting, while deciding final classification result. By leveraging the diversity among trees, Random Forest reduces overfitting and enhances model generalization [24, 25, 30].

3.4 Evaluation Parameters

All ML algorithms mentioned in the previous section provide the result in various parameters like Accuracy, Sensitivity, Specificity, F-1 Score. All these parameters can be derived from the values of confusion matrix as per following equations [31]:

$$Accuracy = \frac{TPs+TNg}{TPs+FPs+FNg+TNg} \quad (1)$$

$$Precision = \frac{TPs}{TPs+FPs} \quad (2)$$

$$Recall = \frac{TPs}{TPs+FNg} \quad (3)$$

$$F1\text{-}Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Where, classified instances are divided into four categories, namely True Positive (TPs), True Negative (TNg), False Positive (FPs) and False Negative (FNg).

3.5 Hyperparameter Optimization

Hyperparameters are the parameters that decide the nature of the working process of the ML algorithm. By properly tuning the parameters, the prediction performance of the algorithms can be improved [31]. In this work, hyperparameters are optimized using a GA approach. A genetic algorithm is an adaptive search strategy that is inspired by the biological evolutionary ideas of natural genetic selection [13]. It works with a population having potential solutions, evolving them through iterative processes that mimic natural evolutionary behaviours, including selection, crossover and mutation.

Here, each parameter is considered as a gene in a chromosome. Each chromosome in the population is evaluated by a fitness function that determines its suitability within the problem space. Over successive generations, solutions with higher fitness are more likely to be selected and recombined, gradually leading the population toward optimal or near-optimal solutions [17].

The stochastic nature of GA allows it to explore large and complex search spaces while avoiding local optima [33]. The key advantage of GA is to explore high-dimensional or nonlinear problems. In this work, averaging is used as a crossover process and the mutation rate is set to 0.5. The flow of the prediction process is shown in Fig. 1.

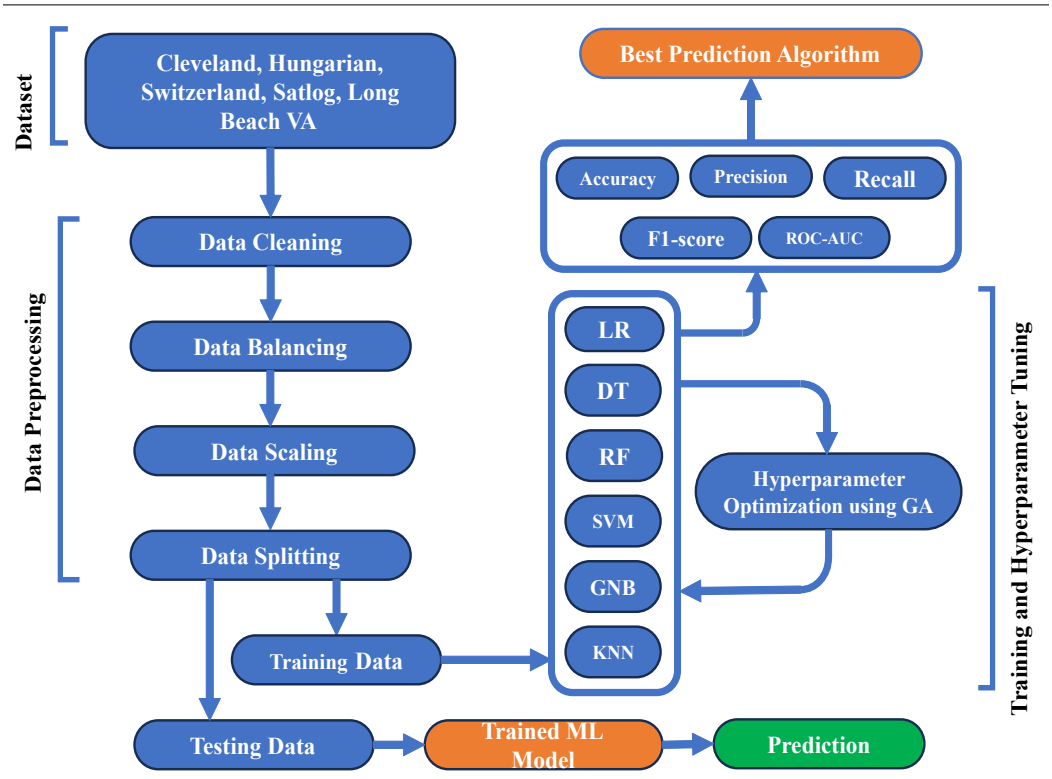

Figure. 1 Flow of the prediction process

3.6 Experimental Setup

For this work Python platform is used for evaluating different algorithms. Python is a strong programming language known for its flexibility and rich library support like Scikit-learn, NumPy and Pandas, which support a wide range of ML algorithms.

4. Results and Analysis

4.1 Performance Parameters

This work has used all the features of the Comprehensive Heart Disease dataset from IEEE Dataport. Results are recorded in two conditions, one with raw data and another with processed data and hyperparameter tuning. Six state-of-the-art ML algorithms are trained and tested using this dataset. Classified instances are divided into TPs, FPs, FNg and TNg. LR, DT, RF, SVM, GNB and KNN has classified 190, 205, 222, 201, 203 and 206 samples correctly out of 238 testing samples including TPs and TNg, which are 20% of the total raw data with 1190 samples. RF has correctly classified the highest 222 samples, whereas LR has classified only 190 samples correctly, which is the lowest among all other algorithms. DT and KNN performed almost equally, classifying 205 and 206 samples correctly, respectively, while SVM and GNB classified 201 and 203 samples correctly, respectively. RF leads all other algorithms with a classification accuracy of 93.28%, leaving behind KNN with 86.55%. DT, SVM and GNB obtained 84.03%, 84.45% and 85.27% accuracy for the test set, respectively. LR legs with an accuracy value of 79.83%. Performance comparison of all ML models is shown in Table 1.

Table 1. Result of ML models with raw data (%)

ML Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LR	79.83	81.40	81.40	81.40	79.7
DT	84.03	88.89	80.62	84.55	86.2
RF	93.28	93.13	94.57	93.85	93.2
SVM	84.45	83.82	88.37	86.04	84.1
GNB	85.29	87.30	85.27	86.27	85.3
KNN	86.55	86.47	89.15	87.79	86.3

Table 3. Result of ML models with optimization (%)

ML Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
GALR	81.86	79.8	82.29	81.03	81.89
GADT	91.18	90.62	90.62	90.62	91.15
GARF	**95.10**	94.79	94.79	94.79	95.08
GASVC_rbf	85.78	82.52	88.54	85.43	85.94
GNB	81.37	80.85	79.17	80.00	81.25
GAKNN	91.18	91.49	89.58	90.53	91.09

Table 2. Classification of samples by ML models after optimization

ML Algorithm	TPs	FPs	FNg	TNg
GALR	79	20	17	88
GADT	87	9	9	99
GARF	91	5	5	103
GASVC_rbf	85	18	11	90
GNB	76	18	20	90
GAKNN	86	8	10	100

After data cleaning by removal of extreme outliers, the data is scaled by the standard scaler method having total 1018 samples. Then, all classifier algorithms are optimized using GA by tuning the hyperparameters of the classifier. For LR 'max_iter', 'random_state', for DT 'max_depth', 'random_state', for RF 'max_depth', 'n_estimators' and 'random_state' are tuned. For SVC 'linear', 'poly', 'sigmoid' and 'rbf' kernels are used, among these, 'poly' has parameter 'degree' along with 'max_iter'. Considering higher result of SVC with 'rbf', optimization parameter 'max_iter' is tuned for 'rbf'. For KNN 'n_neighbors' parameter was tuned using GA. For GA, 50 population size, 20 generations and a 0.5 mutation rate are set. All Classifiers are trained with optimized parameters with GA for 20 different splitting of training and testing samples and mode value of accuracy is considered for the comparison.

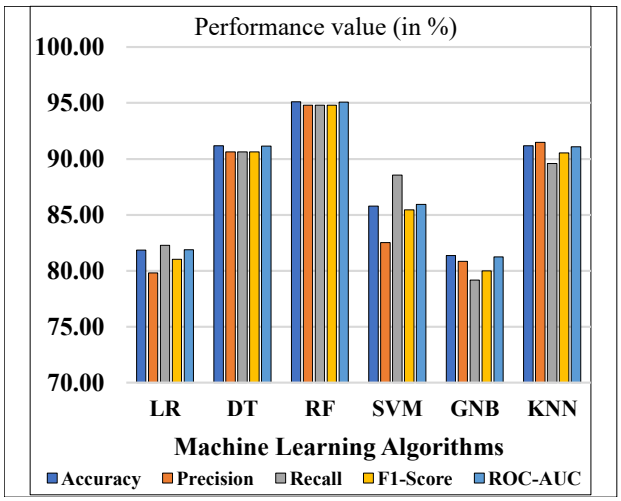

Figure. 2 Comparison analysis of models with optimization

After the hyperparameter tuning, it can been seen that the correctly classified samples are increased in number for classifiers. After optimization with GA, GALR, GADT, GARF, GASVC_rbf, GAGNB, GAKNN has classified 167, 186, 194, 175, 166 and 185 samples correctly out of 204 samples, which is 20% of pre-porecessed dataset of 1018 samples. Again, RF with GA optimization (GARF) has shown the highest number of truly classified samples than any other algorithms as shown in Table 2. GARF has

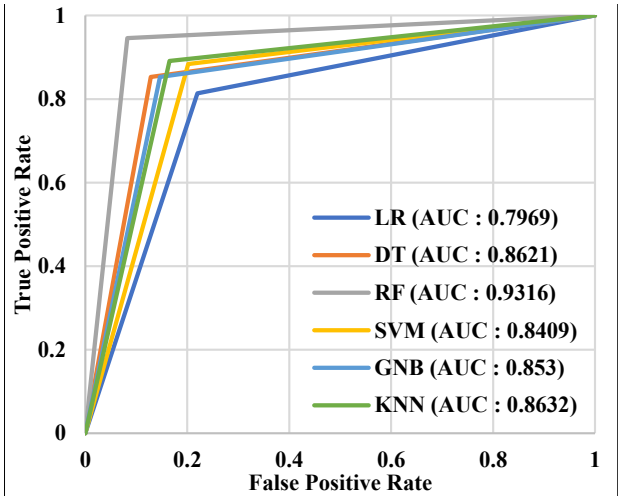

Figure. 3 ROC-AUC with raw data

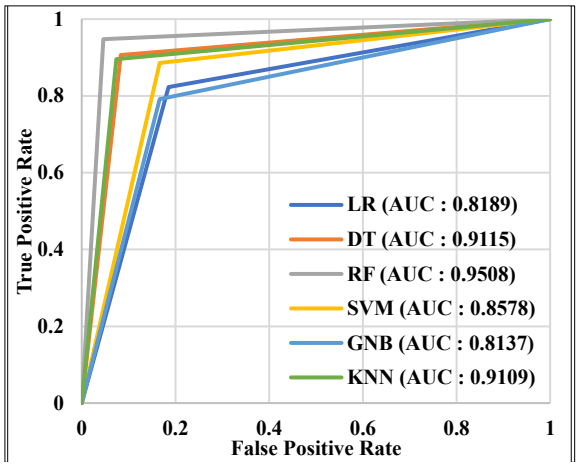

Figure. 4 ROC-AUC with optimization

an accuracy of 95.10%, precision 94.79% and recall 94.79%, which is the highest among all algorithms. GAKNN and GADT follows GARF, achieving high values of 91.18% accuracy by both the algorithm. GASVC_rbf (Support Vector Classifier with 'rbf' kernel) also perform well, especially in recall and F1-score, but fall short in precision. Performance of GALR and GNB algorithms are relatively lower in all parameters, achieving the lowest accuracy 81.86% and 81.37% respectively. Performance parameters, including accuracy, recall, precision, F1-Score, and ROC-AUC of all ML algorithms, are shown in Table 3. Comparison of all ML classifiers is shown in Fig. 2. It can be seen that GARF not only correctly classifies maximum instances but also maintains a strong balance between false positives and false negatives, making it highly trustworthy.

4.2 ROC-AUC

The Receiver Operating Characteristic (ROC) curve illustrates how well a classifier performs with varying threshold values. It shows the relation between true positive rate and true negative rate, whereas the Area Under the Curve (AUC) summarizes the ROC in a scalar value. A higher AUC value indicates a higher classifier capacity to
differentiate between two classes [17]. Here, ROC curves and their respective AUC values of ML classifiers with raw data and pre-processed data with optimized classifiers are shown in Fig. 3 and Fig. 4, respectively. As per Fig. 3 the AUC of RF exhibited a favourable AUC score of 0.9316, while KNN and DT obtained an AUC score of 0.8632 and 0.8621. GNB and SVM achieved an AUC of 0.8530 and 0.8409, showing good performance. LR achieved an AUC of 0.7969. The effect of optimization can be observed after tuning the hyperparameters as per Fig. 4. The AUC of the LR improved slightly to 0.8189, surpassing GNB with 0.8125. GADT and GASVC_rbf achieved 0.9115 and 0.8594 AUC, respectively. The GAKNN showed improvement by reaching an AUC score of 0.9109 after tuning. Particularly, GARF achieved a significant rise in AUC, reaching 0.9508. Results show that GARF excels in the classification process of CVD prediction. Comparison of result with peers, having similar methodologies and algorithms, is shown in Table 4.

Table 4. Comparison of results with peers

Ref.	Year	Classification Model	Dataset	Performance
[5]	2019	LR, KNN, SVM, DT, RF	UCI Statlog (Heart) dataset	FAMD + RF method showed 93.44% accuracy
[6]	2022	NB, SVM, LR, DT, RF, KNN	Cleveland heart disease	KNN with eight neighbors gives 94.1% accuracy
[7]	2023	DT, NB, RF, SVM, LR	Collected data from two different largest teaching hospitals, the LRM and the KTH, in Khyber Pakhtunkhwa	RF model gave the highest accuracy of 85.01%
[8]	2023	LR, SVM, KNN, RF, NB, DT	UCI Cleveland dataset	RF achieves the highest accuracy 94.51%
[10]	2021	RF, KNN, LR, DT, SVM, XGBoost	Cleveland, Hungary, Switzerland, and Long Beach V	Highest accuracy with Random Forest is 88%
[14]	2023	KNN, RF, LR, SVM, GNB, DT	Kaggle encompasses	With KNN classifier 91.8% accuracy
[16]	2022	DT, GNB, RF, SVM, LR, KNN	UCI machine learning repository (299)	With RF 90% accuracy

5. Conclusion

The cardiovascular heart disease prediction framework presented in this research showcases the effectiveness of ML algorithms in anticipating the onset of cardiovascular diseases. The predictive model is constructed based on six different ML classifiers. Before training, the dataset underwent preprocessing, including cleaning, scaling and partitioning into training and testing subsets. Classifier training and evaluation are accompanied by performance optimization through hyperparameter tuning using Genetic Algorithm techniques. Among the six classifiers, the GA optimized RF algorithm delivered the most favourable results, attaining an accuracy score of 95.10%, recall score of 94.79%, F1-score of 94.79%, and a precision score of 94.79%. This shows that the pre-processed data produces better results compared to raw data, as the raw data can be erroneous. This work also focuses on the importance of optimization. A properly optimized algorithm with tuned hyperparameters can perform better in the classification process. Another important aspect addressed here is the reduction of false negatives (FNg). Since this is a medical prediction process, misclassifying an actual positive case as negative can lead to serious real-world consequences. This type of misclassification falls under false negative (FNs). The results demonstrate that GARF effectively minimizes false negatives, helping to mitigate potential real-life risks. This work highlights the potential use of ML-based systems in early prediction and risk stratification for cardiovascular diseases. Future scope

may involve leveraging larger and diverse datasets, integrating ensembled models for better performance.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Both authors have jointly conceptualized and designed the study. Data preprocessing, model development, optimization and experimentation were carried out by first author. The second author contributed to performance evaluation manuscript refinement. Both authors reviewed, edited, and approved the final version of the paper.

References

[1] WHO Newsroom anticlerical on a cardiovascular Diseases (CVDs) : https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] P. K. Mehta, S. Gaignard, A. Schwartz, and J. E. Manson, "Traditional and emerging sex-specific risk factors for cardiovascular disease in women," Reviews in Cardiovascular Medicine, vol. 23, Art. no. 8, 2022.

[3] B. Kolukısa, H. Hacılar, M. Kuş, Bakır-Ğ"unğ"orB., A. Aral, and Ğ"unğ"or, Vehbi Çağrı, "Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology," International Journal of Data Mining Science, vol. 1, Art. no. 1, 2019.

[4] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: an effective heart disease prediction model for a clinical decision support system," Ieee Access, vol. 8, pp. 133034–133050, 2020.

[5] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," IEEE access, vol. 8, pp. 14659–14674, 2019.

[6] T. Ramesh, L. U. Kumar, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," 2022.

[7] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," Health & Social Care in the Community, vol. 2023, Art. no. 1, 2023.

[8] N. Biswas et al., "Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. BioMed Research International, 2023, 15 (2023). Article ID 6864343," 2023.

[9] J. Abdollahi and B. Nouri-Moghaddam, "A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation," Iran Journal of Computer Science, vol. 5, Art. no. 3, 2022.

[10] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," Computational intelligence and neuroscience, vol. 2021, Art. no. 1, 2021.

[11] S. Kumar and G. Sahoo, "A random forest classifier based on genetic algorithm for cardiovascular diseases diagnosis," International Journal of Engineering-Transactions B: Applications, vol. 30, Art. no. 11, 2017.

[12] El-Sofany, Hosam F, "Predicting heart diseases using machine learning and different data classification techniques," IEEE, 2024.

[13] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer methods and programs in biomedicine, vol. 141, pp. 19–26, 2017.

[14] G. Muhammad et al., "Enhancing prognosis accuracy for ischemic cardiovascular disease using K nearest neighbor algorithm: A robust approach," Ieee Access, vol. 11, pp. 97879–97895, 2023.

[15] C. B. Gokulnath and S. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," Cluster Computing, vol. 22, Art. no. Suppl 6, 2019.

[16] M. N. Mirza et al., "A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset," Scientific Programming, vol. 2022, Art. no. 1, 2022.

[17] Togatoropa, Parmonangan R, M. Sianturia, D. Simamoraa, and D. Silaena, "Optimizing random forest using genetic algorithm for heart disease classification," Machine Learning, vol. 2, Art. no. 3, 2022.

[18] A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, "A robust heart disease prediction system using hybrid deep neural networks," IEEE Access, vol. 11, pp. 121574–121591, 2023.

[19] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," Processes, vol. 11, Art. no. 4, 2023.

[20] K. Dissanayake and M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," Applied Computational Intelligence and Soft Computing, vol. 2021, Art. no. 1, 2021.

[21] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," IEEE access, vol. 9, pp. 106575–106588, 2021.

[22] A. Noor, N. Javaid, N. Alrajeh, B. Mansoor, A. Khaqan, and B. S. Hussain, "Heart disease prediction using stacking model with balancing

techniques and dimensionality reduction," IEEE Access, vol. 11, pp. 116026–116045, 2023.

[23] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," IEEE Access, vol. 8, pp. 107562–107582, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3001149.

[24] Chandra, S. S. Nee, L. Z. Min, and C. X. Ying, "Classification and feature selection approaches by machine learning techniques: Heart disease prediction," International Journal of Innovative Computing, vol. 9, Art. no. 1, 2019.

[25] A. Ishaq et al., "Improving the prediction of heart failure patients survival using SMOTE and effective data mining techniques," IEEE access, vol. 9, pp. 39707–39716, 2021.

[26] M. S. Amin, C. Y. Kia, and Varathan, Kasturi Dewi, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, pp. 82–93, 2019.

[27] G. N. Ahmad, H. Fatima, S. Ullah, and A. S. Saidi, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV," Ieee Access, vol. 10, pp. 80151–80173, 2022.

[28] H. Takci, "Improvement of heart attack prediction by the feature selection methods," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 26, Art. no. 1, 2018.

[29] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," Scientific reports, vol. 13, Art. no. 1, 2023.

[30] Ansyari, Muhammad Ridho, Mazdadi, Muhammad Itqan, F. Indriani, D. Kartini, and Saragih, Triando Hamonangan, "Implementation of random forest and extreme gradient boosting in the classification of heart disease using particle swarm optimization feature selection," Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 5, Art. no. 4, 2023.

[31] A. Jafar and M. Lee, "HypGB: High accuracy GB classifier for predicting heart disease with HyperOpt HPO framework and LASSO FS method," IEEE Access, vol. 11, pp. 138201–138214, 2023.

[32] A. K. Gárate-Escamila, El, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, p. 100330, 2020.

[33] L. Balliu, B. Zanaj, G. Basha, E. Zanaj, and Meçe, Elinda Kajo, "Enhancing heart disease prediction accuracy by comparing classification models employing varied feature selection techniques," Serbian Journal of Electrical Engineering, vol. 21, Art. no. 3, 2024.