Prediction of Heart disease using Calibration of Hyper Parameter Optimization (HPO)

Dr. Balaganesh. N, Anto Nidhish. M, Hariharan. C

ABSTRACT

Heart disease is a major global health concern and remains the leading cause of death worldwide. Early detection and accurate prediction of heart disease can help prevent disease progression and improve patient outcomes. This research study aims to improve the accuracy of heart disease prediction using a combination of various feature selection and hyperparameter tuning techniques. The UCI Cleveland heart disease dataset, which contains 14 features, is used in the study. The features include demographic, medical, and lifestyle factors that are known to influence the risk of heart disease. The study employs a range of feature selection techniques, including filter, wrapper, and embedded methods. Several classification algorithms, including K-Nearest Neighbors (KNN), XGBoost (XGB), and Random Forest, are used in the study. The performance of each model is optimized using a range of hyperparameter tuning techniques, including Grid Search, Hyperband, Randomized Search, TPOT, and Bayesian Optimizer. To combine the results of each model and further improve the accuracy of heart disease prediction, the study employs ensemble classifiers such as Voting and Stacking techniques. The results of the study show that the stacking technique with Decision Tree algorithm achieved the best performance, with an accuracy of 98.5% on the test set. The study highlights the importance of appropriate validation methods and statistical significance testing to ensure the robustness of the results.

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death globally. In 2019, approximately 17.9 million individuals passed away as a result of cardiovascular diseases (CVDs), which accounted for 32% of all deaths worldwide. Of these fatalities, 85% were attributed to heart attacks and strokes. More than 75% of CVD deaths occurred in low- and middle-income countries. Among the 17 million premature deaths (occurring before the age of 70) due to noncommunicable diseases in the same year, 38% were caused by CVDs.The number of cardiovascular disease-related deaths is predicted to reach 23.3 million by 2030 [1]. The heart's blood arteries carry oxygen, thus when they become clogged or constricted, it might result in cardiac condition or stroke [2]. WHO reports that each year, approximately 12 80% of persons with heart disease, which affects million people, dies from a cardiac condition [3]. High cholesterol, high blood pressure, stress, tension, alcohol use, sedentary behavior, obesity. The main factors that impact the heart are diabetes. These qualities aids in heart disease prognosis. because of elevated blood pressure, the arteries' walls harden, resulting in obstruction, which can the mortality rate will rise [4]. In real time, Angiogram can be used to diagnose the abnormal narrowing of heart vessels.

Numerous Supervised and Unsupervised Machine learning algorithms have been applied by a number of researchers in the medical field for diagnosis and prediction of the heart disease [5]. Utilizing machine learning techniques in healthcare can significantly enhance the quality of patient care. For instance, advanced deep learning algorithms could be leveraged to develop proactive monitoring systems that alert medical devices or electronic health records in case of any alterations in a patient's condition. This would ensure that patients receive appropriate care promptly. Currently, the applications of machine learning in healthcare have already shown promising results, but the full potential is yet to be harnessed. With the mounting clinical data sets, machine learning in healthcare will become increasingly crucial in the future. Machine learning techniques helps to derive useful knowledge to take decision from vast datasets [7]. These algorithms have been broadly used in the area of Health Care, Computer vision, Speech recognition, Social Science, Cosmology and in the Education field [8]. They provide a variety of algorithms to identify the different patterns in large dataset [9].

Massive amounts of data is collected by the healthcare industry and ML provides different models to train and analyse the data quickly [10]. These algorithms search through a large search space of solutions and finds an optimal solution by training the dataset. The performance of the models can be examined from the various performance metrics such as Accuracy, Sensitivity, Specificity, Precision and F1-Score.Machine learning models is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters. However, there is another kind of parameter, known as Hyperparameters, that cannot be directly learned from the regular training process.

Hyperparameters are typically established prior to the actual training of a model and serve to define key characteristics, such as the level of complexity or rate of learning.

2. RELATED WORKS

The performance of different datasets were analysed using Bayesian Optimization based on Gaussian process [11]. The performance of 6 different machine learning models were examined on the heart dataset and found logistic regression predicted the heart disease with the highest accuracy [12]. Backward feature elimination is used to eliminate least significant features. In order to predict heart disease, the genetic algorithms and frequent item sets were used to mine the association rules. [13]. Fitness function was used to remove the redundant rules and in the optimization of association rules were generated.

Fahd Saleh Alotaibi [19] improved the previous work using Rapid miner tool. The previous work used the dataset ―UCI, whereas the main idea of that research was to do the experiment in two different platform; i.e. Matlab and Weka, and then compare the results from both perspectives. Decision tree, logistic regression and SVM were the similar models used in both researches. The decision tree classifier has increased the accuracy more than 30% from the previous work. In the same way, logistic regression and SVM also outperform and were computed the better score than previous work. It illustrates that the performance of the Rapid miner has shown better accuracy and performance of the classifiers. And, this result indicated that the system can be useful and helpful for the doctors and heart surgeons for timely diagnoses the chances of heart attack in a patient

Three neural network model to were used to construct an ensemble model to diagnose the heart disease [14]. SAS enterprise miner 5.2 was used to evaluate the performance. The number of NN was increased but no improvement was observed in the performance.270 Patient records were trained and tested using Cascaded neural network, a Self-Organizing network and Support Vector machine with RBF function [15]. Naïve Bayes machine learning model was developed to predict the heart disease [16].

An enhanced SVM classifier was presented to classify the linear and nonlinear inputs. PSO was used in feature extraction and Fuzzy C-means Clustering was applied to improve the accuracy [17]. Bhatla and Jyoti [18] employed Weka Tool on different data mining techniques for heart disease prediction and observations showed that Neural network showed good results compared to other data mining techniques. Krishnaiah et al. [19] incorporated a fuzzy approach to remove the uncertainty in the data and applied a KNN Classifier to classify the heart patients. Amin et al. [20] identified the risk factors of 50 patients and implemented an integrated model of genetic algorithm and neural network to predict the presence of heart disease. Abdeldjouad et al. [21] used two different software's Weka and Keel tool to build two different models. First model was built by applying the PCA feature selection method to extract the significant features and 3 different classification algorithms Multi-Objective Evolutionary Fuzzy Classifier (MOEFC), Logistic Regression (LR), Adaptive Boosting (AdaBoostM1) using Weka tool for classification. The second model was built by applying the Wrapper method for feature extraction and Genetic Fuzzy System LogitBoost (GFS-LB), Fuzzy Unordered Rule Induction Algorithm (FURIA) and Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) for classification under Keel tool. The two models were completely trained and the performance were evaluated.

Purusothaman and Krishnakumari [22] surveyed the different research findings based on single model approach and hybrid model and concluded hybrid model are better in prediction of disease compared to a single model. Khourdifi et al. [23] optimized KNN, RF, SVM, Naïve Bayes and ANN with the combination of Particle Swarm optimization and ant colony optimization. Kalaiselvi and Nasira [24] used PSO for extraction of data and ANFIS with AGKNN was used in classification.

Santhanam and Ephzibah [25] has taken genetic algorithm for feature extraction and fuzzy logic for prediction. Mohan et al. [26] developed a hybrid approach of random forest and Linear method for classification of heart disease. KaanUyar et al. [27] proposed a genetic algorithm based recurrent fuzzy neural networks (RFNN) to classify the data. None of the aforementioned studies have implemented hyperparameter optimization (HPO) techniques to boost the accuracy of the heart disease prediction system. Thus, in our proposed model we used HPO techniques to improve the accuracy of the model.

Alizadehsani et al. [28] employed the rule-based classifier and cost sensitive algorithm along with Sequential minimal optimization (SMO) to diagnose CAD. Alizadehsani et al. [29] handled data uncertainty. Different evolutionary algorithms were used for feature selection [33- 38,41]. Some of the existing studies presented the Real time predictions and the performance of detection of heart disease using hardware. Based on five characteristics Cardiac arrhythmias were differentiated using Multi-Level Support vector machine classifiers [42]. Patient Specific SCAD processor [43], Smart ECG processor [44], Wearable ECG Processor [45] was designed to discriminate the CA in real time. An ECG processor and STAC algorithm was presented to improve the accuracy of heart rate detection [46].

3. PROPOSED WORK

The proposed model uses several feature selection algorithms such as Filter (Mutual Information, ANOVA, Chi-square), Wrapper (Sequential Forward Selection, Sequential Backward Elimination, Boruta Feature Selection) and Embedded (LASSO technique, Decision Tree, Genetic Algorithm) methods to select the best features. The features that are selected from each feature selection algorithm is then provided to classifiers such as Random Forest Classifier, Xtreme Gradient Boosting (XGBoost) and K-Nearest Neighbors Classifier (KNN). The features that provide best accuracy for each model are then obtained. The model is then subjected to hyperparameter optimization using algorithms such as Grid Search, Random Search, TPOT, Bayesian Optimization and hyperband. The algorithm which provides best accuracy for each model is then obtained. The model is once again trained with the features and optimized hyperparameters. The three classifier models Random Forest Classifier, XGBoost and KNN are then ensembled using decision tree classifier to obtain better results. The proposed model's flowchart is depicted in Fig. 1.

3.1 Dataset description

The Cleveland Heart Disease (CHD) dataset is a heart disease prediction dataset available online in UCI repository [45]. The actual dataset contains 76 attributes but most of the published articles used only 14 attributes. The heart disease prediction dataset takencontains13 independent variables and one dependent target variable, a total of 14 columns. The target variable has two classes that represent the presence and absence of heart disease. It has 303 rows. The dataset is found to have no missing values. The dataset description is given in Table 1.

Table 1. Dataset Description

s.no	feature	description
1	age	age
2	sex	male, female
3	cp	chest pain type
4	trestbps	resting blood pressure
5	chol	serum cholesterol
6	fbs	fasting blood sugar
7	restecg	resting electrocardiographic results
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	slope of the peak exercise st segment
12	ca	number of major vessels
13	thal	type of defect
14	target	risk, no risk

3.2 Feature selection

Feature Selection selects a subset of features that are supposed to provide best performance for a model. Feature selection reduces the number of input variables while developing a model for prediction. Several methods of feature selection such as Filter methods, Wrapper methods and Embedded methods are used to obtain several subsets of features. Filter methods are statistical methods that ranks features based on their relevance to the target variable using several statistical tests. Wrapper methods obtain

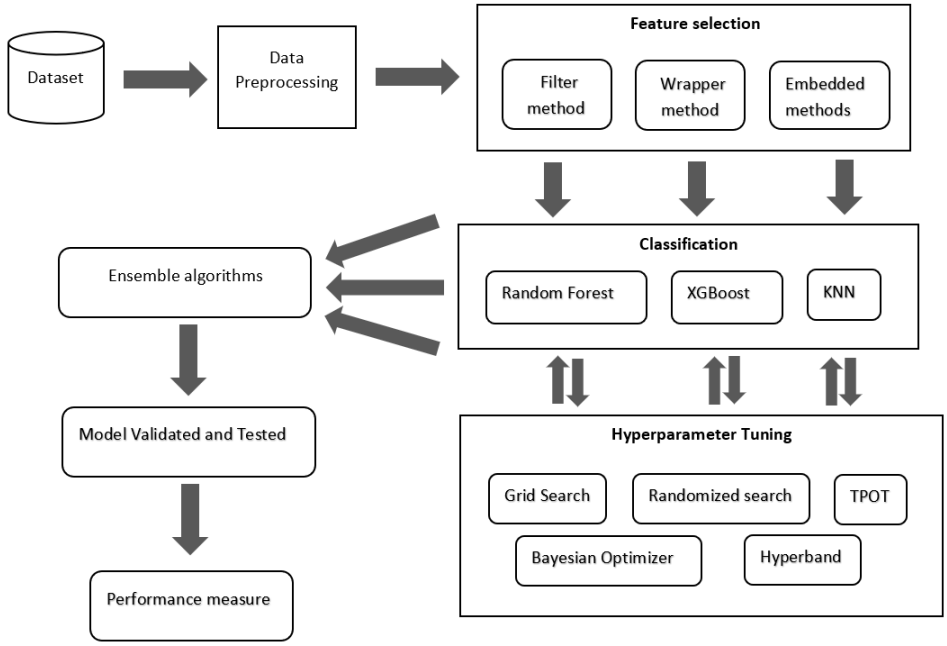

Fig. 1. The Proposed Model of the Heart Disease Prediction system.

optimal subset by building and evaluating the model on each subset. Embedded methods perform feature selection as a part of training process. They use regularization methods to penalize features that are irrelevant or has little or no impact to the model's performance.

3.2.1 Filter methods :

3.2.1.1 Mutual Information:

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. In particular, it measures how much information is communicated, on average, in one random variable about another. Mutual information is a measure of dependence or "*mutual dependence*" between two random variables. As such, the measure is symmetrical, meaning that ***I(X ; Y) = I(Y ; X)*.**

3.2.1.2 Chi-square test:

Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. It determines if the association between two categorical variables of the sample would reflect their real association in the population. Chi- square score is given by :

$$x^2 = \frac{(\text{Observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

Where**, Observed frequency** = No. of observations of class **Expected frequency** = No. of expected observations of class if there was no relationship between the feature and the target.

3.2.1.3 ANOVA test:

ANOVA is an acronym for "analysis of variance" and is a parametric statistical hypothesis test for determining whether the means from two or more samples of data (often three or more) come from the same distribution or not. Each of the features of the data will be ranked according to the F-statistic component, and the features with the higher scores can be selected as the optimal set of components from the data available.

3.2.2 Wrapper methods :

3.2.2.1 Sequential Forward Selection:

Sequential forward selection (SFS) is a greedy feature selection algorithm that starts with an empty set of features and iteratively adds the most promising feature to the set until a desired number of features is reached or no further improvement in performance can be achieved. The performance measure used to evaluate the feature subsets can vary depending on the problem and the learning algorithm being used. In classification problems, common performance measures include accuracy, precision, recall, and F1 score. One potential issue with SFS is that it can be computationally expensive, especially when the number of features is large. In addition, because SFS is a greedy algorithm, it may get stuck in a suboptimal solution if the optimal subset of features cannot be obtained by adding one feature at a time.

3.2.2.2 Sequential Backward Selection:

Sequential backward elimination (SBE) is a feature selection algorithm that starts with the full set of features and iteratively removes the least promising feature from the set until a desired number of features is reached or no further improvement in performance can be achieved. One potential issue with SBE is that, similar to SFS, it can be computationally expensive, especially when the number of features is large. In addition, because SBE is a greedy algorithm, it may get stuck in a suboptimal solution if the optimal subset of features cannot be obtained by removing one feature at a time.

3.2.2.3 Boruta Feature Selection:

Boruta is a feature selection algorithm that identifies the most important features in a dataset by comparing their importance to that of randomized shadow features. Boruta then runs a feature importance test on each feature in the original dataset, as well as on the shadow features. The feature importance test used can be any supervised learning algorithm that provides a feature importance score, such as Random Forest, XGBoost, or LightGBM. For each feature, the importance score of the original feature is compared to the distribution of importance scores of the shadow features. If the importance score of the original feature is higher than the maximum importance score of the shadow features, the feature is considered "confirmed", otherwise it is "rejected".

3.2.3 Embedded methods :

3.2.3.1 LASSO L1 regularization:

Least Absolute Shrinkage and Selection Operator, is a statistical formula whose main purpose is the feature selection and regularization of data models. Lasso regularization can zero out coefficients, making it useful for variable selection in machine learning. When the coefficients of some features are set to zero, those features can be safely removed from the data, leaving only the important features.

3.2.3.2 Decision Tree:

Decision tree for feature selection is a technique that uses decision trees to identify the most important features for a given task. Decision tree builds a tree in which the important features are

present in the root node. Feature importance for each node is calculated. Features are sorted in descending order by their feature scores. Then the desired number of features are selected based on high feature importance.

3.2.3.3 Genetic algorithm:

Genetic algorithms detect well-performing hyper-parameter combinations in each generation, and pass those parameters combinations to the next generation until the best-performing combination is identified.

3.3 Building Machine learning model

3.3.1 Random Forest (RF)

Random forest is a combination of tree predictors proposed by Breiman [35].He defines random forest as "A random forest is a classifier consisting of a collection of trees structured classifiers {h(x,Θk), k = 1, …} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x". A random forest is an ensemble classifier that fits a no.of decision tree classifiers on several sub-samples of a dataset and uses averaging to improve accuracy and to control overfitting. Many different trees are grown and depending upon how the trees are built and randomness introduced, a variety of random forest exists. Most commonly used hyperparameters to optimize the RF model aren_estimators, max_features, min_sample_leaf, max_features, max_depth and criterion. The parameters n_estimators, max_features and min_sample_leaf influences the accuracy prediction value.

3.3.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble technique [36], in which a set of weak learners are combined to improve the accuracy. Here, decision trees are created in a sequential form. Weights are assigned to all independent variables and are fed as inputs to decision tree which predicts the result. The weights of variables that are predicted wrong are increased and is fed to another decision tree. It uses the CART model for classification and regression problems. The model is trained by adding a tree and splitting the features in each iteration to grow the tree. At the end of the training, a score of each leaf node is obtained by multiplying the weight with the predicted value of the tree. The hyperparameters min_samples_split, min_samples_leaf and max_depth controls overfitting. These parameters influence each individual tree in the model. The parameters learning_rate, n_estimators, subsample enhances the boosting operation.

3.3.3 K Nearest Neighbor:

K-Nearest Neighbors (KNN) is a popular non-parametric algorithm used for classification and regression tasks in machine learning. The algorithm is based on the assumption that similar data points are likely to belong to the same class or have similar values for the target variable. KNN works by finding the k-nearest neighbors of a given data point based on a distance metric, and then using the majority vote of their classes (in the case of classification) or their average (in the case of regression) to make predictions.

3.4 Hyper parameter optimization (HPO)

The performance model is significantly impacted by choosing the best hyperparameters. To determine the ideal hyper parameter combination, experiments were conducted on various optimization approaches before being applied to the Random Forest , XG Boost and KNN. There are two different kinds of parameters in the machine learning model - Hyperparameters and model parameters. In order to train the machine learning model, the user needs to set the hyperparameters. These variables are utilised to regulate the learning process. For the same kind of machine learning model, different learning rates or weights are utilised to manage the learning process and find the patterns buried in the data. To reduce error and increase model accuracy, these hyperparameters are tweaked. The ideal hyperparameters strike a balance between over- and under-fitting. By adjusting the Hyper parameters, the RF and XG Boost model's performance can be enhanced. Grid and Random Search approaches are often used in hyper parameter optimization. Grid Search is a tuning technique that attempts find optimum values of hyperparameters for building a machine learning model. It is an exhaustive search that is performed in a specific parameter values of a model, also called as an estimator. Grid search brute-forces all possible combinations to find the best value of parameter. Grid Search uses learning rate and number of layers as hyper parameters. The grid search is not suitable for high dimensional space. Random search samples the search space from the equally distributed search space [37]. The researcher [38] stated that only few parameters had a meaningful impact in the optimization of model. Random Search is similar to Grid Search but it chooses random combination of hyperparameters rather than brute forcing. Here, we don't specify a set of possible values for every hyperparameter. A sampling distribution is defined for every hyperparameter. This technique allows us to control the number of attempted hyperparameter combinations. Random search provided better choice of hyper parameter combination compared to Grid Search. Random search works better under the assumption that all parameters are not equally important. These two approaches avoid the model falling in Local optima [39]. The disadvantages are these two approaches are time consuming and not suitable for data having high dimensional space.

The Tree-Based Pipeline Optimization Tool (TPOT) is a Genetic Programming (GP) based Auto ML system that optimizes the ML models automatically [40]. TPOT uses meta learning techniques to optimize the machine learning pipelines using GP primitives to solve a particular problem. The goal of TPOT is to automate the building of ML pipelines by combining a flexible expression tree representation of pipelines with stochastic search algorithms such as genetic programming. This AutoML considers multiple machine learning algorithms, multiple pre-processing steps and multiple ways to ensemble.

Bayesian methods differ from random or grid search in that they use past evaluation results to choose the next values to evaluate.The concept is limit expensive evaluations of the objective function by choosing the next input values based on those that have done well in the past. Bayesian Optimization is efficient with all types of Hyperparameters (continuous or categorical) and it also enable parallelization which converges to the best hyperparameter faster compared to other tuning algorithms.

Hyperband feature selection generate small-sized subsets and allocate budgets to each hyper-parameter combination based on its performance. Hyperband is essentially just a grid search over the optimal allocation strategy. So at each individual trial the set of hyper parameters is chosen randomly.

3.5 Ensemble techniques:

3.5.1 Voting classifier:

A voting classifier is a type of ensemble classifier that combines the predictions of multiple individual classifiers to make a final prediction. In a voting classifier, each individual classifier is trained on the same dataset using a different algorithm or set of hyperparameters, and their predictions are aggregated using a majority vote or a weighted vote.

3.5.2 Stacking classifier:

Stacking is a popular technique in ensemble learning that involves training a meta-model to combine the predictions of multiple base models. It is also known as stacked generalization or stacked regression.

3.6 Performance metrics

Confusion matrix was employed to evaluate the performance of the two models. True Positive (TP) is defined as the count of predicted values correctly identified the presence of disease. True negative is defined as the count of predicted values correctly identified the absence of disease. False Positive (FP) is defined as the count of predicted values incorrectly classified as positive (actually when it was negative). False Negative (FN) is defined as the count of predicted values incorrectly classified as negative (actually when it was positive). Once the model is trained, the risk of heart disease is predicted and evaluated with ten-fold cross validation. The analysis was done with the Performance metrics Accuracy, Specificity, Sensitivity, Precision and ROC-AUC values.

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

4. EXPERIMENTAL RESULTS

4.1 Experimental setup

The experiments were implemented in Python 3.8 on DELL, Intel (R) Core (TM) i5-10300H CPU @ 2.50GHz, RAM 8 GB with Windows11.

4.2 Results and analysis

In this study, the effect of feature selection and hyper parameter on the predictive performance of three different machine learning models Random Forest, K nearest neighbor and XGBoost were examined. Nine different feature selection techniques were used to pick the best subset of features for the ML models. Five different hyper parameter optimization methods: Grid Search, Randomized Search, Hyperband feature selection, Bayesian optimization and Genetic programming (TPOT Classifier) were used to optimize the ML models. A comparative analysis of the predictive performance of the 3 algorithms RF and XG Boost with 9 feature selection techniques and 5 hyper parameter techniques were carried out in the experiments. Each algorithm is analyzed by selecting different hyper parameters. We compared the different hyper parameter optimization results of Randomized Search, Grid Search, Genetic Programming algorithms with the results of existing techniques. 80% and 20% of data were taken for training and testing respectively.

4.3 Random forest

Selecting relevant features improves the accuracy of the proposed system. Table. 2, 3 and 4 shows the accuracy obtained and the no.of features selected with Random Forest applied with differnet feature selection methods. Table 5 shows the list of features selected by various feature selection tecnniques. Table 6 shows the Accuracy comparison with different hyperparameter tuning techniques.

Table 2. Random forest with filter based feature selection methods

Feature selection technique	Mutual information	ANOVA test	Chi - Square
Accuracy	91.8027	91.8027	**93.4426**
No.of features	11	13	**10**

Table 3. Random forest with Wrapper based feature selection methods

Feature selection technique	Sequential Frwd Selection	Sequential Bkwd Selection	Boruta Feature selection
Accuracy	**90.1639**	88.5246	88.5246
No.of features	**10**	9	8

Table 4. Random forest with embedded based feature selection methods

Feature selection technique	Lasso	Decision tree	Genetic algorithm
Accuracy	85.2459	83.6065	**88.5246**
No.of features	7	7	**9**

Table 5. Features selected by different Feature selection techniques

Method	Feature selection technique	No.of features	Features
Filter	Chi – square	10	**'thalach', 'oldpeak', 'ca', 'cp', 'exang', 'chol', 'age', 'trestbps', 'slope', 'sex'**
Wrapper	SFS	10	**'age', 'sex', 'cp', 'trestbps', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'**
Embedded	Genetic algorithm	9	**'age', 'sex', 'cp', 'trestbps', 'restecg', 'exang', 'slope', 'ca', 'thal'**

Table 6. Accuracy comparison with different hyperparameter tuning techniques

Hyperparameter tuning techniques	Chi-sq	SFS	Genetic algorithm
Grid search	**95.0819**	**91.8033**	90.1639
Randomized search	93.4426	**91.8033**	91.8033
TPOT	**95.0819**	90.1639	**93.4426**
Bayesian Optimization	93.4426	**91.8033**	90.1639
hyperband	90.1639	90.1639	90.1639

4.4 Extreme Gradient Boost:

Table. 7, 8 and 9 shows the accuracy obtained and the no.of features selected with XGBoost applied with differnet feature selection methods. Table 10. shows the list of features selected by various feature selection tecnniques. Table 11. shows the Accuracy comparison with different hyperparameter tuning techniques.

Table 7. XGBoost with filter based feature selection methods

Feature selection technique	Mutual information	ANOVA test	Chi - Square
Accuracy	**90.1639**	88.5246	88.8246
No.of features	**9**	13	10

Table 8. XGBoost with Wrapper based feature selection methods

Feature selection technique	Sequential Frwd Selection	Sequential Bkwd Selection	Boruta Feature selection
Accuracy	91.803	88.5246	**91.8033**
No.of features	5	9	**6**

Table 9. XGBoost with filter based feature selection methods

Feature selection technique	Lasso	Decision tree	Genetic algorithm
Accuracy	78.6885	**80.3279**	80.327
No.of features	7	**3**	5

Table 10. Features selected by different Feature selection techniques with Xtream Gradient Boost Classifier

Method	Feature selection technique	No.of features	Features
Filter	Mutual Information	9	**'cp', 'thal', 'ca', 'slope', 'oldpeak', 'exang', thalach', 'chol', 'sex'**
Wrapper	Boruta FS	6	**'cp', 'thal', 'ca', 'slope', 'exang', 'sex'**
Embedded	Decision tree	3	**'cp', 'ca', 'thal'**

Table 11. Accuracy comparison with different hyperparameter tuning techniques

Hyperparameter tuning techniques	Mutual Information	Boruta FS	Model based
Grid search	**93.4426**	**93.4426**	**93.4426**
Randomized search	88.5245	**93.4426**	**93.4426**
TPOT	90.1639	91.8033	90.8033
Bayesian Optimization	88.5246	**93.4426**	**93.4426**
hyperband	90.1639	90.1639	88.6344

4.5 K – Nearest Neighbor:

Table. 12, 13 and 14 shows the accuracy obtained and the no.of features selected with KNN applied with differnet feature selection methods. Table 15. shows the list of features selected by various feature selection tecnniques. Table 16. shows the Accuracy comparison with different hyperparameter tuning techniques.

Table 12. KNN with filter based feature selection methods

Feature selection technique	Mutual information	ANOVA test	Chi - Square
Accuracy	**91.8023**	85.2459	85.2459
No.of features	**5**	8	5

Table 13. KNN with Wrapper feature selection methods

Feature selection technique	Sequential Frwd Selection	Sequential Bkwd Selection	Boruta Feature selection
Accuracy	**75.4098**	67.2131	Not applicable
No.of features	**8**	8	-

Table 14. KNN with Embedded feature selection methods

Feature selection technique	Lasso	Decision tree	Genetic algorithm
Accuracy	91.803	83.6065	**91.8033**
No.of features	6	7	**7**

Table 15. Features selected by different Feature selection techniques with K Nearest Neighbor

Method	Feature selection technique	No.of features	Features
Filter	Mutual Information	5	**'thal', 'cp', 'ca', 'oldpeak', 'slope'**
Wrapper	SFS	8	**'age', 'sex', 'cp', 'fbs', 'exang', 'slope', ca', 'thal'**
Embedded	Genetic algorithm	7	**'sex', 'cp', 'fbs', 'oldpeak', 'slope', 'ca', 'thal'**

Table 16. Accuracy comparison with different hyperparameter tuning techniques

Hyperparameter tuning techniques	Mutual Information	SFS	Genetic algorithm
Grid search	**91.8033**	80.3277	**91.8233**
Randomized search	**91.8033**	**80.3279**	90.1639
TPOT	90.1639	**80.3279**	90.1639
Bayesian Optimization	90.1639	78.6885	90.1639
hyperband	83.6344	78.6885	83.6344

Table 17. Overall comparison of results achieved from the combinations of classifier, feature selection algorithm and hyperparameter optimization techniques

Classifier	Feature selection technique	Hyperparameter tuning algorithm	Accuracy
Random forest	Chi-square	Grid search, Bayesian optimization	95.0819
XGBoost	Boruta feature selection	Grid, random, TPOT	93.4426
KNN	Genetic algorithm	Grid search	91.8233

Random Forest achieved the highest accuracy of 95.0819 using the Chi-square feature selection technique and hyperparameter tuning using Grid Search and Bayesian Optimization. XGBoost achieved an accuracy of 93.4426 using Boruta feature selection technique and hyperparameter tuning using Grid Search, Random Search, and TPOT. KNN achieved an accuracy of 91.8233 using Genetic Algorithm for feature selection and Grid Search for hyperparameter tuning.

Table 18. Overall comparison of results achieved by combining Random forest, XGBoost and KNN

Classifier	Accuracy	Recall	Precision
Voting classifier	93.44%	96.88%	91.18%
Stacked Classifier with Random forest	91.80%	87.55%	96.55%
Stacked Classifier with Decision Tree	98.36%	96.88%	100%

Voting classifier that combines the predictions of KNN, Random Forest, and XGBoost models. This classifier achieved an accuracy of 93.44%, which is quite good. Additionally, it has a high recall of 96.88%, meaning it correctly identifies a high percentage of positive cases, and a precision of 91.18%, indicating that it has a relatively low false positive rate.

Stacked classifier with a Random Forest model at the top level achieved an accuracy of 91.80%, which is slightly lower than the voting classifier. It has a lower recall of 87.55%, indicating that it misses more positive cases, but a higher precision of 96.55%, meaning that it has a lower false positive rate.

Stacked classifier with a Decision Tree model, which achieved the highest accuracy of 98.36%. This model has a very high recall of 96.88% and a perfect precision of 100%, meaning that it correctly identifies all positive cases without any false positives

4.6 Performance analysis with ROC_AUC for Cleave land Dataset:

Receiver operating characteristics is a visualization curve to compare the “true positive rate” and “false positive rate”. Area under the receiver operating characteristics is compared for both models for all the three hyper optimization techniques. The best model is found when the AUC value is closer or equal to 1. Fig. 2 shows the AUC-ROC curve for the experiments conducted with Random forest along with the features selected by chi-square and grid search hyperparameter tuning achieved an 95.08% AUC-ROC curve score. When experiments conducted with XGBoost along with the features selected by Boruta feature selection and grid search hyperparameter tuning achieved an 93.44% AUC-ROC curve which is illustrated in Fig. 3. When experiments were conducted with KNN classifier along with the features selected by Genetic algorithm and grid search hyperparameter tuning, it achieved an 91.82% AUC-ROC curve. Fig. 4 shows the AUC-ROC curve for the experiments conducted with KNN classifier. Among these classifers, Random forest showed the best result by achieving the accuracy of 95.08%.

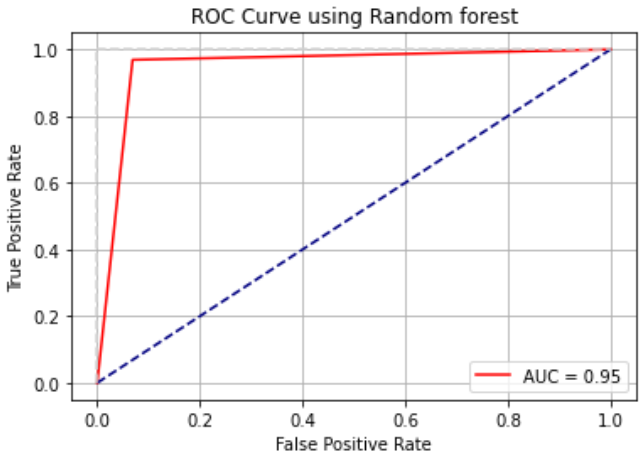

Fig. 2. ROC-AUC curve using Random forest with Chi-square feature selection and grid search.

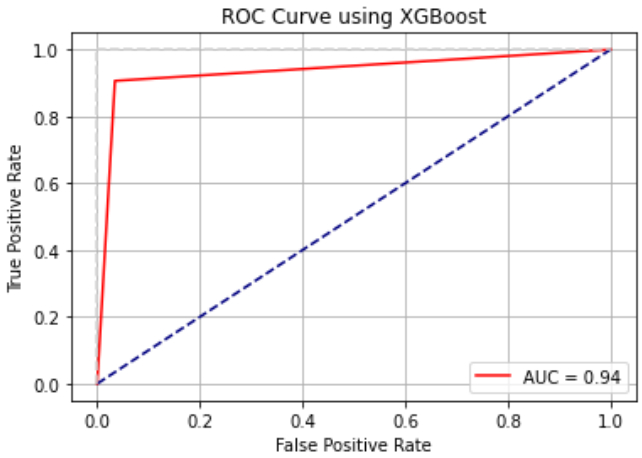

Fig. 3. ROC-AUC curve using XGBoost with Boruta feature selection grid search.

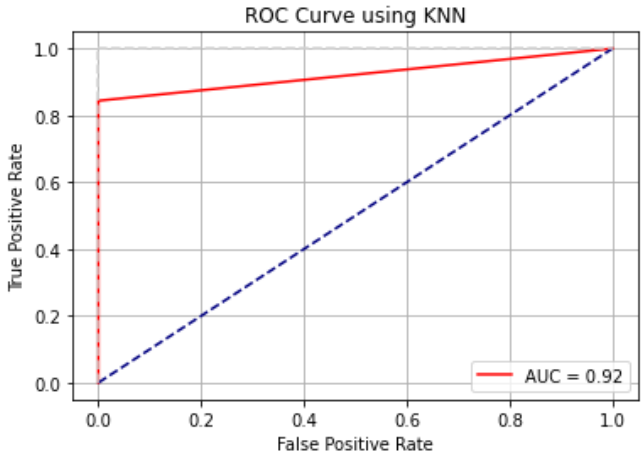

Fig. 4. ROC-AUC curve using KNN with Genetic algorithm and grid search.

Fig. 5 shows the AUC-ROC curve for the experiments conducted with Ensemble classifiers like voting classifier and stacking classifier. When experiments were conducted with Random forest, XGBoost and KNN, Voting classifier achieved an AUC of 93.44%,

Stacked classifier with Random forest model at the top achieved an AUC of 91.80% and Stacked classifier with Random forest model at the top achieved an AUC of 98.36% respectively.

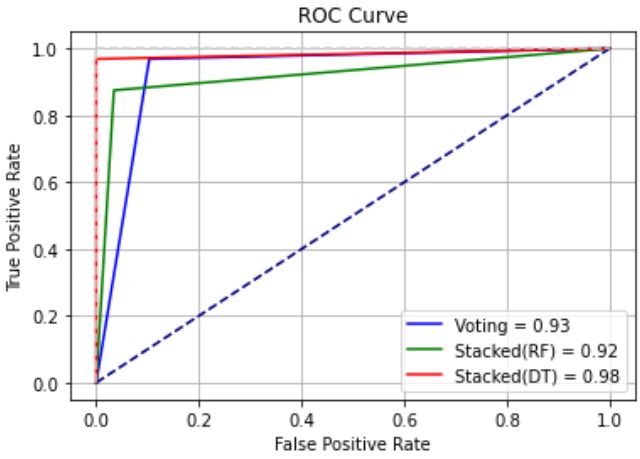

Fig. 5. ROC-AUC curve for Voting, and stacking classifiers

Table 19. Comparison of the Model from previous studies for Cleave land Dataset.

Authors	Method	Result
Verma et al [47]	CFS + PSO + K-means + MLP	90.28
Ayon et al [48]	RF	87.45
Valarmathi et al [49]	RF With Grid Search	91.32%
	RF with Randomized Search	95.04
	RF With TPOT Classifier	97.52%
	XGBoost with Grid Search	86.36%
	XGBoost with Randomized Search	92.14
	XGBoost With TPOT classifier	90.50
Proposed Model	Voting technique with RF, XGB and KNN	93.44
	Stacking technique with Random forest	91.80
	Stacking technique with Decision tree	98.36

5. Conclusion:

In this study, we proposed a heart disease prediction system that utilizes a wide range of feature selection and hyperparameter optimization techniques to develop an accurate and robust predictive model. The system incorporates various feature selection techniques such as Mutual Info, Chi sq test, Anova, SFS, SBE, Boruta FS, Genetic algo, Lasso L1, and decision tree to identify the most relevant features for the prediction task. Furthermore, the system employs three different classifiers, namely Random Forest, XGBoost, and K-Nearest Neighbors, and ensembles their outputs to improve the overall performance of the model. The results of our experiments demonstrate that the proposed system achieves high accuracy and robustness in predicting heart disease. The system outperforms several existing models in terms of accuracy, precision, recall, and F1 score. The comprehensive feature selection and hyperparameter optimization techniques utilized in our system ensure that the selected features are relevant and informative for the prediction task and that the classifiers are optimized for the specific problem at hand.

IV. REFERENCES

[1] C.D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030, Plosmed 3 (11) (2006).

[2] M. Akhil, Dr. Priti Chandra, Dr. B.L Deekshatulu, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical,Electronics and Communication TechnologiesICECIT, 2012.

[3] Ibrahim Umar Said, Abdullahi Haruna Adam, Dr. Ahmed BaitaGarko, "Association Rule Mining On Medical Data To Predict Heart Disease", International Journal of Science Technology and Management, August 2015, pp. 26-35.

[4] A. Chauhan, A. Jain, P. Sharma, V. Deep, Heart disease prediction using evolutionary rule learning, in: 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, 2018, pp. 1–4.

[5] Ilaria Castelli, Edmondo Trentin, Combination of supervised and unsupervised learning for training the activation functions of neural networks, Pattern Recogn. Lett. 37 (2014) 178–191.

[6] R. Safdari, T. Samad-Soltani, M. GhaziSaeedi, M. Zolnoori, Evaluation of classification algorithms vs knowledge-based methods for differential diagnosis of asthma in Iranian patients, Int. J. Inform. Syst. Serv. Sect. 10 (2) (2018) 22–26.

[7] F.S. Alotaibi, Implementation of machine learning model to predict heart failure disease, Int. J. Adv. Comput. Sci. Appl. 10 (6) (2019) 261–268.

[8] M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260, https://doi.org/10.1126/science.aaa8415.

[9] M. Motwani, D. Dey, D.S. Berman, G. Germano, S. Achenbach, M.H. Al-Mallah, D. Andreini, M.J. Budoff, F. Cademartiri, T.Q. Callister,"Machine learning for pre-diction of all-cause mortality in patients with suspected coronary artery disease": a 5-year multicentre prospective registry analysis, Eur. Heart J. 38(7), pp. 500–507,2016.

[10] T. Condie, P. Mineiro, N. Polyzotis, M. Weimer, Machine learning on big data: data engineering (ICDE), in: 2013 IEEE 29th International Conference on, IEEE, 2013, pp. 1242–1244.

[11] Wu. Jia, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, Si-Hao Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, J. Electron. Sci. Technol. 17 (1) (2019).

[12] Debabrata Swain, Preeti Ballal, Vishal Dolase, Banchhanidhi Dash, Jayasri Santhappan, An efficient heart disease prediction system using machine learning. Advances in Intelligent Systems and Computing, Springer Nature Singapore Pvt Ltd, 2020.

[13] M.A. Jabbar, B.L. Deekshatulu, P. Chandra, "An Evolutionary Algorithm for Heart Disease Prediction. In: K.R. Venugopal, L.M. Patnaik (eds.) Wireless Networks and Computational Intelligence". ICIP 2012. Communications in Computer and Information Science, vol 292. Springer, Berlin, Heidelberg. DOI:10.1007/978-3-642-31686-9_44.

[14] R. Das, I. Turkoglu, A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Syst. Appl. 36 (4) (2009) 7675–7680.

[15] R. Chitra, V. Seenivasagam, Heart disease prediction system using supervised learning classifier, Bonfring Int. J. Software Eng. Soft Comput. 3 (1) (2013) 1.

[16] K. Vembandasamy, R. Sasipriya, E. Deepa, Heart diseases detection using Naive Bayes Algorithm, Int. J. Innov. Sci. Eng. Technol. 2 (9) (2015).

[17] R. Kavitha, T. Christopher, "An effective classification of heart rate data using PSOFCM clustering and enhanced support vector machine," Indian Journal of Science and Technology, 8(30), 2015.

[18] N. Bhatla, K. Jyoti, An analysis of heart disease prediction using different data mining techniques, Int. J. Eng. Res. Technol. 1 (8) (2012) 1–4.

[19] V. Krishnaiah, G. Narsimha, N.S. Chandra, "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach. In: Satapathy S., Govardhan A., Raju K., Mandal J. (eds) Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1. Advances in Intelligent Systems and Computing, vol 337. Springer, Cham,2015. DOI:10.1007/978-3-319-13728-5_42.

[20] S. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in: 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, Tamil Nadu, India, 2013, pp. 1227-1231, doi: 10.1109/CICT.2013.6558288.

[21] F.Z. Abdeldjouad, M. Brahami, N. Matta. " A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques.", In: M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, S. Kallel (eds.) The Impact of Digital Technologies on Public Health in Developed and Developing Countries. ICOST 2020. Lecture Notes in Computer Science, vol 12157. Springer, Cham, 2020. DOI: 10.1007/978-3-030-51517-1_26.

[22] G. Purusothaman, P. Krishnakumari, "A survey of data mining techniques on risk prediction: heart disease", Indian J. Sci. Technol. 8(12), 2015.

[23] Youness Khourdifi, Mohamed Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, Int. J. Intell. Eng. Syst. 12 (1) (2018) 242–252.

[24] C. Kalaiselvi, G. Nasira, Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian J. Sci. Technol. 8(14), 2015.

[25] T. Santhanam, E. Ephzibah, Heart disease prediction using hybrid genetic fuzzy model", Indian, J. Sci. Technol. 8 (9) (2015) 797–803.

[26] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554, https://doi.org/10.1109/ACCESS.2019.2923707.

[27] Kaan Uyar, Ahmet ¨Ilhan, Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks, Procedia Comput. Sci. 120 (2017) 588–593.

[28] Alizadehsani Roohallah et al. "Exerting Cost-Sensitive and Feature Creation Algorithms for Coronary Artery Disease Diagnosis."IJKDB 3.1 (2012): 59-79. Web. 1 May. 2021. doi:10.4018/jkdb.2012010104.

[29] Roohallah Alizadehsani, Mohamad Roshanzamir, Moloud Abdar, Adham Beykikhoshk, Mohammad Hossein Zangooei, Abbas Khosravi, Saeid Nahavandi, Ru San Tan, U. Rajendra Acharya, Model uncertainty quantification for diagnosis of each main coronary artery stenosis, Soft Computing 24 (13) (2020) 10149–10160, https://doi.org/10.1007/s00500-019-04531-0.

[30] Roohallah Alizadehsani, Jafar Habibi, Mohammad Javad Hosseini, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, Zahra Alizadeh Sani, A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed. 111 (1) (2013) 52–61, https://doi.org/10.1016/j.cmpb.2013.03.004.

[31] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, M. Panahiazar, A. Koohestani, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, A database for using machine learning and data mining techniques for coronary artery disease diagnosis, Scientific Data 6 (1) (2019), https://doi.org/10.1038/s41597-019-0206-3.

[32] Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M. Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, U. Rajendra Acharya, Machine learning-based coronary artery disease diagnosis: a comprehensive review, Comput. Biol. Med. 111 (2019) 103346, https://doi.org/10.1016/j.compbiomed.2019.103346.

[33] Roohallah Alizadehsani, Mohammad Hossein Zangooei, Mohammad Javad Hosseini, Jafar Habibi, Abbas Khosravi, Mohamad Roshanzamir, Fahime Khozeimeh, Nizal Sarrafzadegan, Saeid Nahavandi, Coronary artery disease detection using computational intelligence methods, Knowledge-Based Syst. 109 (2016) 187–197, https://doi.org/10.1016/j.knosys.2016.07.004.

[34] R. Alizadehsani, M.J. Hosseini, Z.A. Sani, A. Ghandeharioun, R. Boghrati. Diagnosis of coronary artery disease using cost-sensitive algorithms. in: Paper presented at the 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium 2012.

[35] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32.

[36] T. Chen and C. Guestrin, ''XGBoost: A scalable tree boosting system,''in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785_794, doi:10.1145/2939672.2939785.

[37] Douglas C Montgomery, Design and Analysis of Experiments, John Wiley & Sons, 2017.

[38] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[39] R. Liessner, J. Schmitt, A. Dietermann, and B. Baker, ¨ "Hyperparameter Optimization for Deep Reinforcement Learning in Vehicle Energy Management", in: Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019), pages 134-144 ISBN: 978-989-758-350-6.

[40] R.S. Olson, J.H. Moore, TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning, in: F. Hutter, L. Kotthoff, J. Vanschoren (Eds.), Automated Machine Learning. The Springer Series on Challenges in Machine Learning, Springer, Cham, 2019, https://doi.org/10.1007/978-3-030-05318-5_8.

[41] Jalali, Seyed Mohammad Jafar, Khosravi, Abbas, Alizadehsani, Roohallah, Salaken, Syed Moshfeq, Kebria, Parham Mohsenzadeh, Puri, Rishi and Nahavandi, Saeid 2019, Parsimonious evolutionary-based model development for detecting artery disease, in: ICIT 2019 : Proceedings of the IEEE International Conference on Industrial Technology, IEEE, Piscataway, N.J., pp. 800-805, doi: 10.1109/ICIT.2019.8755107.

[42] M. A. Sohail, Z. Taufique, S. M. Abubakar, W. Saadeh and M. A. Bin Altaf, "An ECG Processor for the Detection of Eight Cardiac Arrhythmias with Minimum False Alarms," in: 2019 IEEE Biomedical

Circuits and Systems Conference (BioCAS), 2019, pp. 1-4, doi: 10.1109/BIOCAS.2019.8919053.

[43] S.M. Abubakar, M. Rizwan Khan, W. Saadeh, M. A. Bin Altaf, "A Wearable AutoPatient Adaptive ECG Processor for Shockable Cardiac Arrhythmia," in: 2018 IEEE Asian Solid-State Circuits Conference (A-SSCC) 2018 267 268 10.1109/ ASSCC.2018.8579263.

[44] Shihui Yin, Minkyu Kim, Deepak Kadetotad, Yang Liu, Chisung Bae, Sang Joon Kim, Yu Cao, Jae-Sun Seo, A 1.06- μ W Smart ECG Processor in 65-nm CMOS for Real-Time Biometric Authentication and Personal Cardiac Monitoring, IEEE J. Solid-State Circ. 54 (8) (2019) 2316–2326, https://doi.org/10.1109/ JSSC.2019.2912304.

[45] S. M. Abubakar, W. Saadeh and M. A. B. Altaf, "A wearable long-term single-lead ECG processor for early detection of cardiac arrhythmia," in: 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018, pp. 961-966, doi: 10.23919/DATE.2018.8342148.

[46] S. Izumi, et al., A 14 μA ECG processor with robust heart rate monitor for a wearable healthcare system, in: 2013 Proceedings of the ESSCIRC (ESSCIRC), 2013, pp. 145–148, https://doi.org/10.1109/ESSCIRC.2013.6649093.

[47] L. Verma, S. Srivastava, P.C. Negi, A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data, J. Med. Syst. 40 (7) (2016) 178, https://doi.org/10.1007/s10916-016-0536-z.

[48] Safial Islam Ayon, Md. Milon Islam, Md. Rahat Hossain, CoronaryArtery heart disease prediction: a comparative study of computational intelligence techniques, IETE J. Res. (2020), https://doi.org/10.1080/03772063.2020.1713916.

[49] Valarmathi R, Sheela T, Heart disease prediction using hyper parameter optimization (HPO) tuning, *IETE J. Res. (2021), https://doi.org/10.1080/03772063.2020.1713916.*