








A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques

Azka Mir¹ , Attique Ur Rehman^{1,2} , Tahir Muhammad Ali² , Sabeen Javaid¹ ,
Maram Fahaad Almufareh³ , Mamoon Humayun^{3,4}  and Momina Shaheen^{4*} 

¹Department of Software Engineering, University of Sialkot, Sialkot, Pakistan; ²Department of Computer Science, Gulf University for Sciences and Technology, Hawally, Kuwait; ³Department of Information Systems College of Computer and Information Science, Jouf University, Sakaka, Saudi Arabia; and ⁴School of Arts Humanities and Social Sciences, University of Roehampton, London, UK

Abstract

Aims The objective of this research is to develop an effective cardiovascular disease prediction framework using machine learning techniques and to achieve high accuracy for the prediction of cardiovascular disease.

Methods In this paper, we have utilized machine learning algorithms to predict cardiovascular disease on the basis of symptoms such as chest pain, age and blood pressure. This study incorporated five distinct datasets: Heart UCI, Stroke, Heart Statlog, Framingham and Coronary Heart dataset obtained from online sources. For the implementation of the framework, RapidMiner tool was used. The three-step approach includes pre-processing of the dataset, applying feature selection method on pre-processed dataset and then applying classification methods for prediction of results. We addressed missing values by replacing them with mean, and class imbalance was handled using sample bootstrapping. Various machine learning classifiers were applied out of which random forest with AdaBoost dataset using 10-fold cross-validation provided the high accuracy.

Results The proposed model provides the highest accuracy of 99.48% on Heart Statlog, 93.90% on Heart UCI, 96.25% on Stroke dataset, 86% on Framingham dataset and 78.36% on Coronary heart disease dataset, respectively.

Conclusions In conclusion, the results of the study have shown remarkable potential of the proposed framework. By handling imbalance and missing values, a significantly accurate framework has been established that could effectively contribute to the prediction of cardiovascular disease at early stages.

Keywords cardiovascular disease prediction; data imbalance handling; multi-dataset approach; healthcare applications; CVD prediction using machine learning

Received: 19 December 2023; Revised: 19 May 2024; Accepted: 19 June 2024

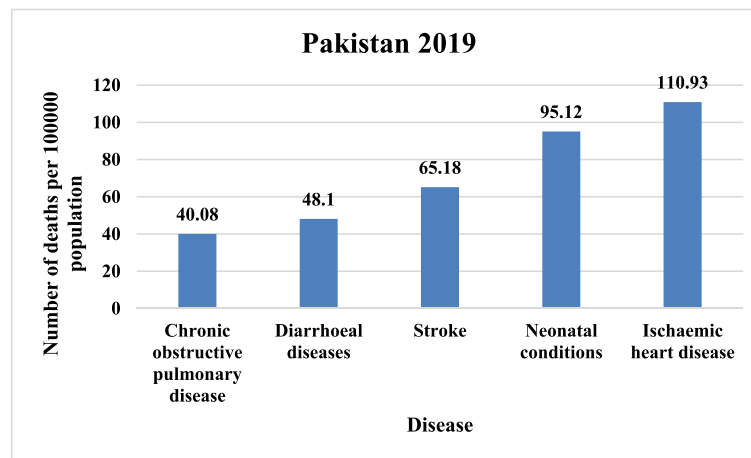
*Correspondence to: Momina Shaheen, School of Arts Humanities and Social Sciences, University of Roehampton, London SW15 5PJ, UK. Email: momina.shaheen@roehampton.ac.uk

Introduction

According to a World Health Organization (WHO) report, cardiovascular disease is one of the non-communicable diseases which are responsible for 32% mortality worldwide.¹ Cardiovascular diseases comprise diseases concerning blood vessels and heart such as stroke, heart attack, coronary artery disease, which is also known as coronary heart disease (Chd), and heart failure. In 2019, Pakistan reached 29.4% mortality, ranking among the top 20 most effected countries in the world.² The WHO forecasted the mortality due to

cardiovascular disease to reach its peak by 2030.³ In Pakistan, roughly 19% deaths occurred due to heart diseases, which have now increased to 29%.⁴ According to the WHO's 2019 global health estimation, the 10 major mortality causes in Pakistan are shown in Figure 1.

Ischaemic heart disorder and stroke are among the top three leading causes of death per 100 000 people. Disability-adjusted life year rate of ischaemic heart disorder and stroke is 3032.74 and 1755.79, respectively, indicating potential loss of healthy life by premature death or disability due to the disorder.⁵ The risk factor of developing heart

Figure 1 Death rate per 100 000 population caused by different diseases according to the global health estimation (2019).

failure is one out of five.⁶ With continuous increase in the prevalence of heart disorders, it is a serious concern for healthcare authorities to take measures in order to control it. There is lack of awareness on cardiovascular diseases exposing the population to its risks. Underdiagnosed or misdiagnosed heart disease can bring fatal results. Heart disorders such as heart attack and stroke are not identified at earlier stage, which may lead to disability and low survival rates.

Existing systems do not cover the wide range of cardiovascular diseases. In developing countries like Pakistan, diagnosis of heart disease at an early stage is complex due to lack of resources, which directly affects prediction of the disease.⁷ In Pakistan, health system defects are becoming a national crisis, and immediate steps need to be taken to resolve the issue. There is no accurate system available for early diagnosis,⁸ which causes 40% of the premature deaths due to insufficient care. Without efficient treatment recommendation, a disease can be left undertreated or over treated, which does more harm than good to patients. Clinical practices may go astray without accurate prediction of the disease which in turns increases treatment costs. Practitioners such as cardiologists, neurologists, radiologists, electro-physiologist and surgeons can utilize such diagnosis systems for the early prediction of disease and to recommend treatment accordingly.

Machine learning techniques are used to automate prediction models with its effective training algorithms to examine common behaviours and patterns in the data. Machine learning provides techniques to improve the ability of model to make decision based on the data. Machine learning has found its applications in different fields, including healthcare, where it is a leading way to understand today's huge amount of healthcare data. In view of the massive impact of cardiovascular diseases globally, the machine learning model becomes extremely useful for early prediction of the disorder.

In this research, we have tried to improve the performance of the model by conducting experiments with multiple machine learning classifiers to better utilize the data collected from different datasets. Health organizations gather data on various health issues, which can widely be used for research purposes. Because of the risks posed by cardiovascular disease, it is crucial to improve accuracy of the existing systems.⁹ The need of the hour is to develop an integrated framework that handles missing values, performs optimized feature selection and improves prediction accuracy in a systematic way.

Contributions

We have proposed a novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques to surmount the shortcomings of existing systems. In this paper, we have used five different datasets and developed a model using machine learning techniques. We used different feature selection and classification methods to improve the overall accuracy. The missing values have been handled by replacement method. After that, our model employs optimized selection technique for the selection of optimum set of features. Finally, a 10-fold cross-validation has been applied with AdaBoost ensemble using random forest (RF), which provides high accuracy rate for prediction. Overall, this paper contributes to handling missing values and imbalance data and using optimized selection of features for achieving high accuracy on 5 different datasets.

Organization of the paper

The paper is divided into following sections: Literature review section highlights literature review. Proposed solution section

presents proposed solution. Experimentation is presented in Experimental evaluation of proposed solution section, and comparative evaluation of the state-of-the-art with proposed model is presented in the Comparative analysis and discussion section. The conclusion and future work are discussed in the last section.

Literature review

This highlights the existing studies carried out on cardiovascular disease prediction systems. The detail of datasets previously used and methodology adopted is discussed in this section to identify the gap in the existing studies.

An ample amount of work has been done in this field. A hybrid approach is proposed in Mohan et al.¹⁰ to classify heart disease using Cleveland UCI repository dataset. The prediction model is a combination of hybrid RF algorithm and linear method. The result obtained by this model shows an accuracy of 88.7%. The proposed model has no restriction in feature selection; therefore, different feature selection methods can be used to increase accuracy using significant features. Reddy et al.¹¹ used 10 different machine learning classifiers to predict heart disease risk using Cleveland heart disease dataset. The performance was tested using 10-fold cross-validation. An accuracy of 86.468% was achieved with optimal attributes obtained from sequential minimal optimizer (SMO) classifier using χ^2 method. Jothi et al.¹² proposed a system to predict the risk of heart disease using *K*-nearest neighbours (KNN) and decision tree (DT) algorithms. The system uses 13 features of the dataset split into 80% for training and 20% for testing. The system applies data mining classifiers which displayed an accuracy of 81% with DT algorithm and 67% KNN algorithm. However, the accuracy rates can be increased using other machine learning techniques.

Additional to the classification of heart diseases, some studies have been done for accumulating the risk of heart diseases. A machine learning model is proposed in Motarwar et al.¹³ to predict the risk level of heart disease using five algorithms. The system used Cleveland dataset and split it into 80% training set and 20% testing set. The classifiers such as RF, support vector machine (SVM), Hoeffding DT, Naïve Bayes (NB) and logistic model tree (LMT) are applied on optimized features selected using feature selection. RF produced an accuracy rate of 95.08%. The result is obtained on the selected features only. To further explore and analyse the prediction rate of proposed model, the selected features should be varied to validate the accuracy of the proposed system. In Hossain et al.,¹⁴ prediction model is proposed for cardiovascular disease risk in T2D patients using Australia based dataset that utilizes network-based features and machine learning classifiers. The accuracy rate achieved by this proposed system is 79% to 88%. However, the dataset was col-

lected from a private health fund which is not sufficient for future analysis. Also, the proposed system did not include common types of cardiovascular diseases such as stroke and ischemic heart disease. Risk factors for the research purposes need to be expanded.

Dinh et al.¹⁵ developed weighted ensemble model for cardiovascular disease prediction using National Health and Nutrition Examination Survey (NHANES) dataset. The system used machine learning classifiers such as logistics regression (LR), RF, SVM and gradient boosting to measure the performance. The proposed model achieved 83.9% accuracy rate. The effectiveness of variables is yet to be explored and analysed for the evaluation of performance.

Sharma et al.¹⁶ used Heart Disease UCI dataset to develop a DNN model using Talos for the prediction of heart disease. The proposed system used 14 attributes of the dataset and applied Hyper-parameter optimization (Talos) to achieve the accuracy of 90.76%. Arunachalam et al.¹⁷ present a model for the prediction of presence of heart disease. The proposed system used six machine learning classification algorithms using 14 features of Cleveland dataset. The results obtained show an accuracy of 91.7% with both MLP and SVM algorithms. This system did not utilize ensemble methods which may improve the performance. Alotaibi et al.¹⁸ proposed heart failure prediction system using Rapid miner tool. The model used UCI heart disease dataset with 14 features. It performed 93.19% accuracy with DT algorithm using 10-fold cross-validation.

Ali et al.¹⁹ proposed a smart healthcare system for the prediction of heart disease using deep learning ensemble and feature fusion methods. Two different datasets, Cleveland and Hungarian, were used for this system. The proposed system used feature fusion and selection along with weighting techniques for evaluation using the ensemble deep learning model. This model obtained 98.5% accuracy rate. However, this system did not use refined data and handled missing values which can aid in achieving highly efficient result. Jindal et al.²⁰ proposed the prediction of risk of heart disease using machine learning algorithms like KNN and logistic regression. The dataset used for this model is obtained from UCI repository. The accuracy of KNN was obtained as the highest among the three algorithms used in the system which is 88.52%. The average accuracy of 87.5% was obtained on the proposed model with LR and KNN. Rani et al.²¹ proposed a hybrid system to predict heart disease. The dataset used for the proposed system is the Cleveland dataset obtained from the UCI repository. The system used multivariate imputation to manage missing values. A hybrid algorithm of the genetic algorithm (GA) and recursive feature elimination is used for feature selection. In addition, synthetic minority oversampling technique (SMOTE) is applied and machine learning algorithms have been used such as LR, SVM, RF, NB and AdaBoost classifiers. The proposed system achieved 86.6% accuracy with RF algorithm. However, different

methods of feature selection can be used to improve the accuracy rate and improve the performance of the system.

In existing works, researchers mostly focused on accuracy rate through traditional methods. To analyse the gap between researches over the year has been patching up the efforts to design an accurate predictive system. The issues with the data such as missing values and imbalanced data have not been dealt with in an effective manner. Handling missing values is highly important as it may influence accuracy by reducing the samples for the specific data. Therefore, if missing values are not handled in an efficient way, predictions can become ineffective. The current studies show deficiency of analysis on various datasets. Researchers presented the best results of their over-fitted models. In the light of existing studies, we may conclude that the previous systems still have not provided satisfactory accuracy rate. There is a lucid gap in existing studies for cardiovascular disease prediction systems. The main objective of this paper is to perform experiments on various datasets to get satisfactory accuracy rate for prediction using machine learning techniques.

Proposed solution

In this paper, a novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques has been proposed. The objective of this paper

is to propose a model to predict cardiovascular diseases with high accuracy rate. In this approach, following major steps are involved:

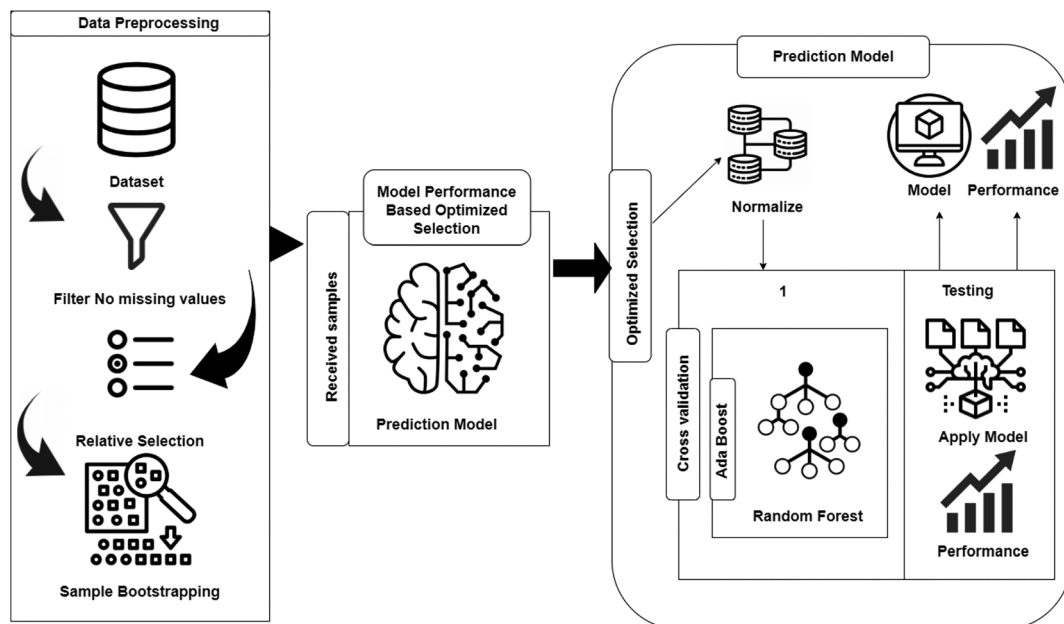
- 1 Missing values: to handle missing values in dataset,
- 2 Optimize Selection: to select optimized features, and
- 3 Classification using AdaBoost ensemble.

Model will be trained and prediction will be evaluated. The goal of this paper is to develop a model which provides high accuracy rate on different datasets. The detailed model is given below in Figure 2. Each step of proposed framework has been described in detail in the subsequent section.

Experimental evaluation of proposed solution

In this section, we have provided the details of our model presenting the experimental processes. The Experimental setup section gives details on experiments performed. Pre-processing experimentation for missing values and data sampling outlines the details of data pre-processing, and Experimentation for feature selection contains the optimize selection method. In Experimentation with various classifiers section classification method using different algorithms on four of the selected datasets is discussed. In Experimenta-

Figure 2 A step-by-step framework of proposed system. Data pre-processing stage involves dataset collection, filtering method to get no missing values, relative selection and sample bootstrapping. On received samples generated by sample bootstrapping method, model performance based on optimized features is obtained and normalized. Prediction model includes training and testing sets. Model uses cross-validation, AdaBoost and random forest classifier; applied model and performance operators are used to obtain performance metrics of the model.



tion-classification via ensemble section, performance of algorithms is analysed on the selected datasets. From Performance measures to Comparison of proposed methodology sections. classification using ensemble method on the five selected datasets is discussed and the results of different machine learning algorithms are presented for comparison.

Experimental setup

In the experimental setup, we have described datasets as well as tools and techniques used in the experiment in sub-sections. The setup used to develop this model is explained in the sections below.

Dataset description

We have used five different datasets to expand the scope of our framework. *Four of the datasets focus mainly on coronary heart disease while one is specifically used to predict the risk of stroke.* All the datasets were collected from online resources. The first dataset used is Statlog Heart disease²² dataset which has been collected from Kaggle. It contains classification attributes for the presence of cardiovascular disease. The description of the dataset *in the order of attribute name (type): value* is shown below.

Statlog Heart Disease dataset description

1. Patient age (in years) (numerical):	29 to 77
2. Gender (binary):	0 = female 1 = male
3. Chest pain type (nominal):	1 = typical angina, 2 = atypical angina 3 = nonanginal, 4 = asymptomatic
4. Resting blood pressure (numerical):	94 to 200
5. Serum cholesterol (numerical):	126 to 564
6. Fasting blood sugar (binary):	(>120 mg/dL) 0 = false 1 = True
7. Resting electrocardiographic result (nominal):	0 = normal, 1 = ST-T wave abnormality 2 = left ventricular hypertrophy
8. Maximum heart rate (numerical):	71 to 200
9. Exercise induced angina (binary):	0 = no 1 = yes
10. Old peak (numerical):	Continuous (0 to 6.2)
11. Slope of peak exercise ST segment (nominal):	1 = upsloping 2 = flat 3 = downsloping
12. number of major vessels (nominal):	0 to 3
13. Defect type (nominal):	3 = normal, 6 = fixed defect, 7 = reversible defect
14. Class (binary):	0 = absence, 1 = presence

This dataset has 14 features and 270 values. This framework uses class attribute to predict absence (1) or presence (2) of cardiovascular disease. Chd²³ dataset is also used in this experiment. It has 10 features related to the patient's symptoms and 462 values. Target variable is Chd, which predicts

presence or absence of Chd as 0 or 1. The details of Coronary heart disease dataset are given below.

Coronary heart disease dataset description

1. Sbp (integer):	Systolic blood pressure
2. Tobacco (real):	Yearly tobacco use (in kg)
3. Ldl (real):	Low density lipoprotein
4. Adiposity (real):	Adiposity
5. Famhist (binominal):	Family history (0 or 1)
6. Typea (integer):	Type A personality score
7. Obesity (real):	Body mass index
8. Alcohol (real):	Alcohol use
9. Age (integer):	Patient's age
10. Chd (nominal):	Diagnosis of Chd (0 or 1)

Framingham²⁴ dataset is also used for the purpose of this research. It contains 16 attributes and 4238 values. The description of Framingham dataset is given below.

Framingham dataset description

1. Sex (nominal):	Male = 1 or female = 0
2. Age (continuous):	Age of patient in the whole number
3. Education (continuous):	Values = 1–4, some high school = 1, high school or GED = 2, some college or vocational school = 3, college = 4
4. Current smoker (nominal):	Yes = 1 or no = 0
5. Cigarettes per day (continuous):	Number of cigarettes smoked per day
6. BP meds (nominal):	Yes = 1 or No = 0 was BP patient or not
7. Prevalent stroke (nominal):	Yes = 1 or No = 0 was Stroke patient or not
8. Prevalent Hyp (nominal):	Yes = 1 or No = 0, whether the patient was hypertensive
9. Diabetes (nominal):	Yes = 1 or No = 0 was he a diabetes patient
10. Tot Chol (continuous):	Total cholesterol level
11. Sys BP (continuous):	Systolic Blood pressure
12. Dia BP (continuous):	Diastolic blood pressure
13. BMI (continuous):	Body mass index
14. Heart rate (continuous):	Heart rate or pulse rate
15. Glucose (continuous):	Glucose level
16. Ten-year Chd (nominal):	Yes = 1; no = 2, the 10 year risk of coronary heart disease

Heart UCI dataset²⁵ is used in this paper. It consists of different datasets such as Cleveland dataset, Hungarian dataset, Switzerland and Long Beach V datasets. It has 76 features out of which 14 features are utilized as a subset of the attributes from the above-mentioned datasets with size of 1025. The 'target' attribute is used for the prediction of disease referring to 0 = absence and 1 = presence of the disease. The description of the dataset is shown below.

Heart UCI dataset description

1. Age (integer):	Age of patient
2. Sex (integer):	Male = 1 or female = 0
3. Cp (integer):	Chest pain type (4 values) Resting blood pressure

(Continues)

4. Trestbp (integer):	
5. Chol (integer):	Serum cholesterol in mg/dL
6. Fbs (integer):	Fasting blood sugar > 120 mg/dL
7. Restecg (integer):	Resting electrocardiographic results (values 0,1,2)
8. Thalach (integer):	Maximum heart rate achieved
9. Exang (integer):	Exercise induced angina
10. Oldpeak (real):	Oldpeak = ST depression induced by exercise relative to rest
11. Slope (integer):	Slope of the peak exercise ST segment
12. Ca (integer):	Number of major vessels (0–3) coloured by fluoroscopy
13. Thal (integer):	Thal: 0 = normal; 1 = fixed defect; 2 = reversible defect
14. Target (integer):	Diagnosis of target disease (0 or 1)

Stroke dataset²⁶ is also used in this paper to predict the likelihood of getting stroke depending on given symptoms. This dataset contains 5110 observations and 12 attributes as described below.

Stroke dataset description

1. Id (integer):	Unique identifier
2. Gender (binominal):	'Male', 'female' or 'other'
3. Age (integer):	Patient's age
4. Hypertension (integer):	0 if the patient does not have hypertension, 1 if the patient has hypertension
5. Heart_disease (integer):	0 if the patient does not have any heart diseases, 1 if the patient has a heart disease
6. Ever_married (nominal):	'No' or 'Yes'
7. Work_type (nominal):	'Children', 'Govt_jov', 'Never_worked', 'private' or 'self-employed'
8. Residence_type (nominal):	'Rural' or 'urban'
9. Avg_glucose_level (real):	Average glucose level in blood
10. BMI (nominal):	Body mass index
11. Smoking_status (nominal):	'Formerly smoked', 'never smoked', 'smokes' or 'unknown'
12. Stroke (binominal):	1 if the patient had a stroke or 0 if not

It is important to choose the datasets carefully to examine their characteristics for the research. The data can be analysed using statistical visualization of the datasets. The features of the dataset form relationships with the other features of that dataset, which is beneficial for prediction analysis. A heat map provides a visual representation of variables of the dataset in correlation in the form of a matrix.²⁷ To visualize the correlation between the features of each dataset, we have built heat maps below. In Figure 3, the heat map of the datasets illustrates the relation between variables. The value range at the right side shows the variation between 1.0 and −0.4. Heat map of Heart UCI dataset illustrates that target value has high relational value with chest pain,

exercise angina and ST slope. Low density lipoprotein, tobacco consumption and adiposity show 0.3 value range with Chd. For Framingham dataset, prevalent hypertension, sys Bp and age appears in the 0.2 range with respect to 10 year Chd variable. The representation of correlational variables shows the relation range of 0.0 to 1.0 between variables of stroke dataset.

Tools and technologies

All the experiments are performed using Rapid Miner Studio. It is an open-source tool having a wide range of pre-programmed machine learning operators. We have used rapid miner for data visualization and graph plotting. The processes of feature selection and classification algorithms have been carried out by using rapid miner operators.

Pre-processing experimentation for missing values and data sampling

In our proposed framework, we have used sample bootstrapping method in pre-processing step. This section provides detail of pre-processing experimentation for handling missing values and class imbalance. In cardiovascular disease datasets, numerous features are encountered which may have redundant or irrelevant values or features. It can increase the difficulty in the classification of disease. For missing values and class imbalance, we have used replacement method and resampling method respectively. The detailed procedure is given below.

Missing values

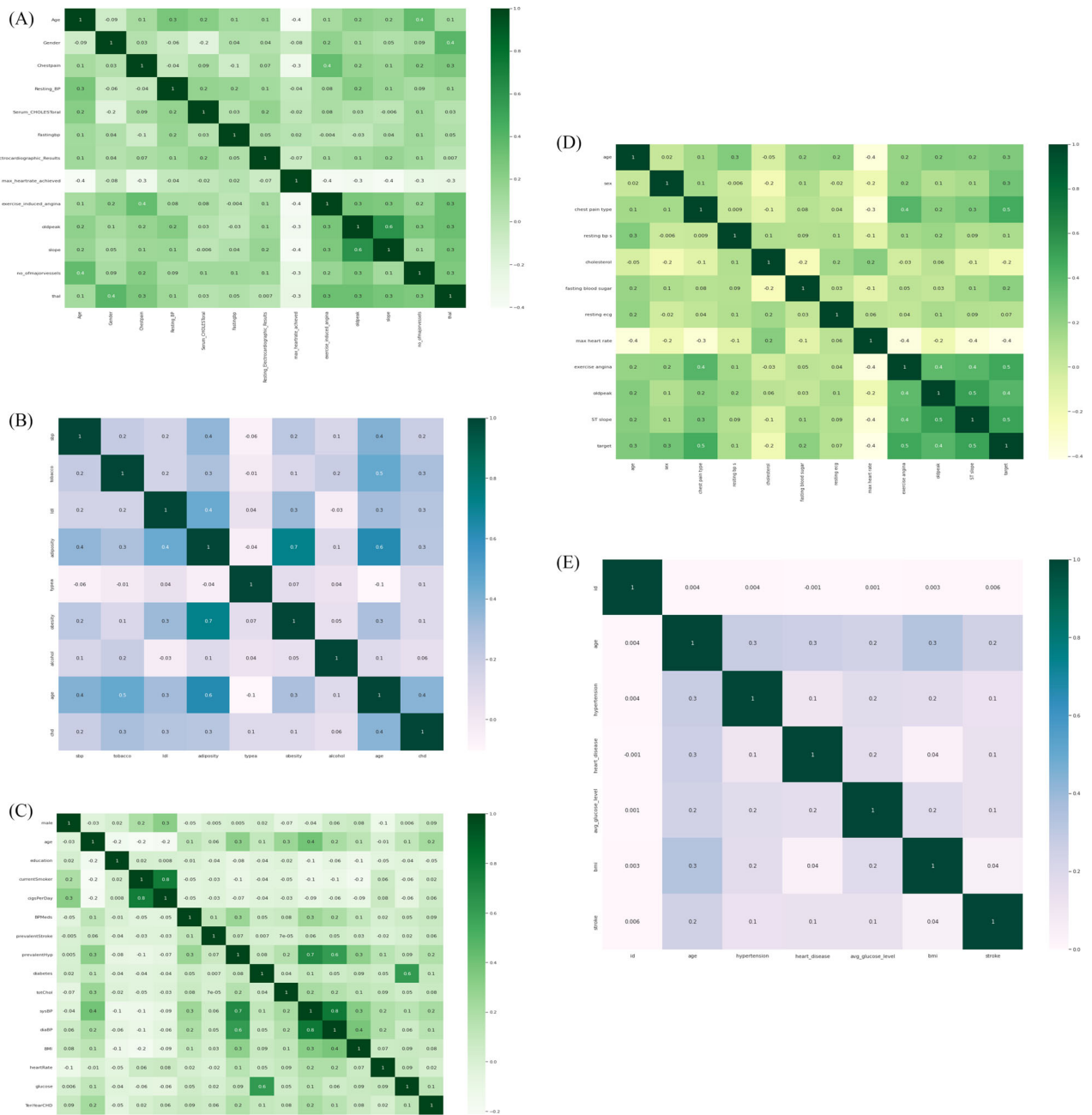
Missing values exist due to the absence of data value for a particular feature in a dataset. Statistically, it can affect the results because the dataset under study contains samples with unrepresentative data. It is important to handle the missing values; otherwise, biased results will be produced. One out of five of the datasets contained 30 missing values, which are handled in the pre-processing stage. We have replaced missing values by mean of the relative attributes in our proposed system.

Sample bootstrapping

In classification problems, data imbalance is a familiar term that is used to define unequal distribution of classes. As a result of imbalance, majority group is overclassified and accuracy is affected due to biasness. Figure 4 illustrates the class imbalance in the datasets used in this paper.

In Heart Statlog dataset, the number of observations in absent class is 150 and present class has 120 number of observations. In the coronary heart disease dataset, Chd class 1 has 160 values and Chd Class 0 has 302 values. The target value of Framingham dataset, that is, TenYearChd has class 0 with 3594 instances whereas class 1 has 644. Heart dataset shows a target with value 0 containing 138 observations and value 1

Figure 3 (A) Heat map of Heart Statlog dataset; (B) heat map of Coronary heart disease dataset; (C) heat map of Framingham dataset; (D) heat map of Heart UCI dataset; (E) heat map of Stroke dataset.

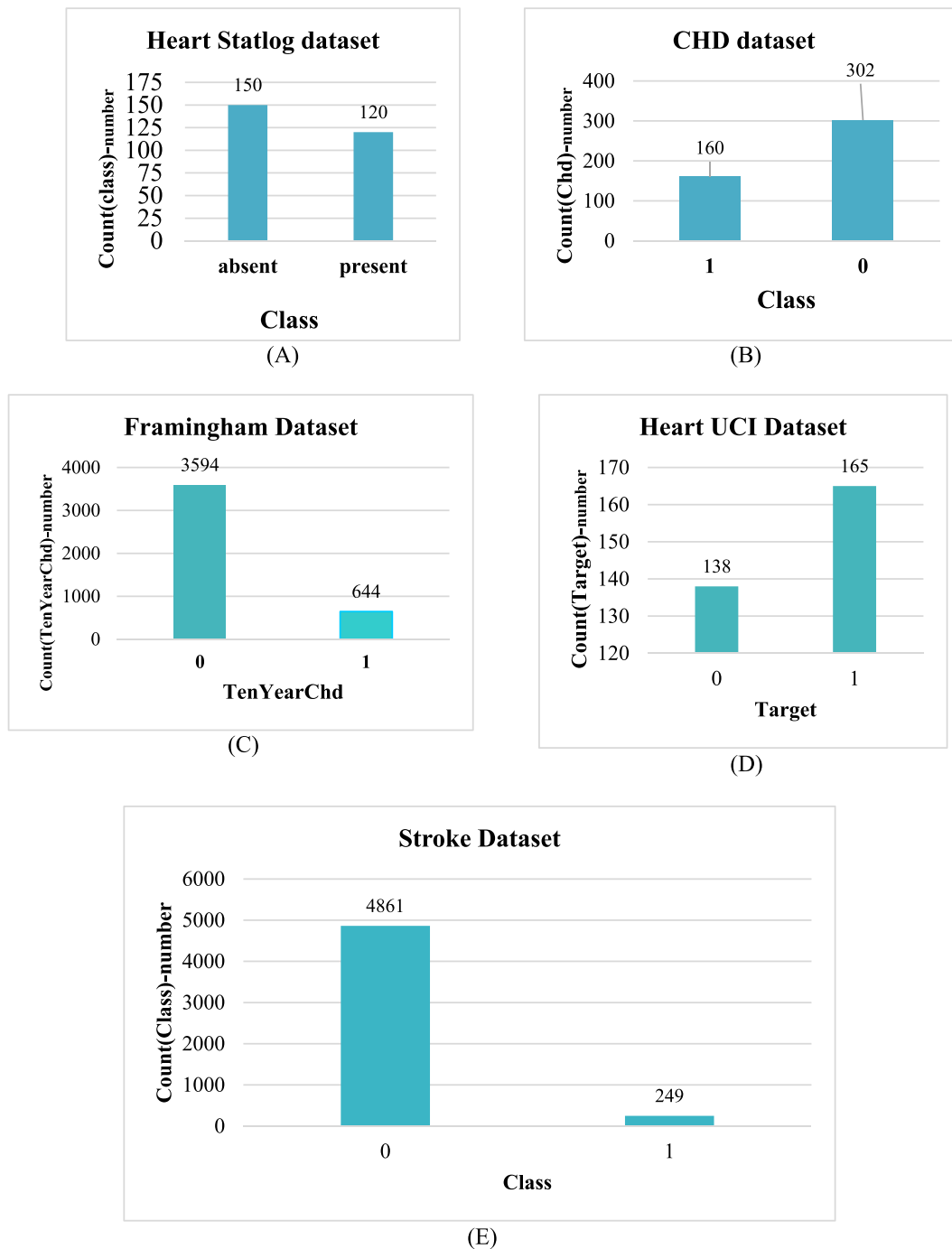


has 165 observations. In machine learning, different methods are used to handle data imbalance. One of the approaches used is data level method which aims at minimizing the class imbalance level using data sampling techniques. To handle class imbalance, sampling method is applied using randomly chosen samples so that results would have no biased effect. We have used bootstrap method as a resampling technique for our selected datasets. Sample bootstrapping method is a

resampling technique that generates bootstrapped sample from a dataset with replacement. It uses replacement method; therefore, the resultant samples may not contain unique values. This method involves two steps: size of sample and number of iterations.

- 1 Iterations: Iterations for resampling should be relatively sufficient to perform statistic calculations on the sample.

Figure 4 Class imbalance of (A) Heart Statlog dataset class absent and present; (B) Chd dataset class 1 (present), 0 (absent); (C) Framingham dataset TenYearChd class 1 (present), 0 (absent); (D) Heart UCI dataset target 1 (present), 0 (absent); (E) Stroke dataset class 1 (present), 0 (absent).



2 Sample size: Sample size can be determined on the basis of absolute or relative. Depending on our different datasets, we set the sample parameter to relative. Relative-based samples are generated as a ratio of total in-

stances of the dataset. The ratio can be provided according to the dataset. The sample ratio at which it provides the best accuracy rate for chosen datasets is 4.3. Sample ratio can only be used if sample parameter is relative.

Experimentation for feature selection

Feature selection plays a significant role in improving efficiency and reliability of the model. In machine learning, feature selection is used to select relevant features to reduce the computational cost and improving performance. It reduces the number of input features and optimizes the efficiency of the system. Therefore, we have used optimize selection operator for the selection of relevant features from the datasets with forward selection. It creates n number of features from our sample data and gives weightage to each attribute. In Statlog dataset, age, gender, chest pain, resting electrocardiographic results, slope, resting bp, number of major vessels, defect type (thal) and serum cholesterol obtained high weightage. In Chd dataset, tobacco, family history and age obtained high weightage. In Heart UCI, age, gender, chest pain, resting ecg, maximum heart rate (thalach), exercise induced angina (exang), old peak, slope, number of major vessels (ca) and thal obtained high weightage. In stroke, gender, age, hypertension, work type and smoking status obtained high weightage. In Framingham dataset, bp meds, diabetes and systolic bp obtained high weightage. After optimization, the framework follows classification through ensemble method.

Experimentation with various classifiers

We have used various machine learning algorithms for our proposed model using selected datasets. The derivation of our framework remained the same while using other algorithms such as SVM, DT and NB. We have highlighted the experimentation with different algorithms on the selected datasets using proposed framework. A comparative analysis has also been included, which contains different performance measures of the results on the datasets. Detailed results of individual algorithms have been provided as supporting information.

Logistic regression

Linear regression is a supervised classifier used to form a linear relation between input variable and target variable by adjusting the linear equation to data under observation.²⁸ The equation is given by

$$Y = a + bX + e \quad (1)$$

Linear regression classifier obtained 91.36% accuracy on the Statlog dataset, 77.7% accuracy on the Chd dataset and 85.29% accuracy on the Framingham dataset.

K-nearest neighbours

KNN algorithm predicts the output based on the similarity of k-nearest input examples in the data.²⁹ We used KNN algorithm on the selected datasets by choosing the distance for-

mula. Then we selected the value of k neighbours to be performed on the datasets. KNN algorithm is well suited for the datasets of bigger size so that it can generate more samples. It obtained 91.36% accuracy on the Statlog dataset, 77.82% accuracy on the Heart UCI dataset, 76.26% accuracy on the Chd dataset and 85.29% accuracy on the Framingham dataset.

Decision tree

DT algorithm is used for non-parametric classification predictions. It generates a tree of data and predicts the output on the basis of decision principles according to the features of the data.³⁰ The DT algorithm is well suited for the datasets where root node can be defined easily to split the data for accurate decision making. The DT performed on the datasets provided 87.65% accuracy on the Statlog dataset, 78.18% accuracy on the Heart UCI dataset, 74.1% accuracy on the Chd dataset and 85.29% accuracy on the Framingham dataset.

Support vector machine

SVM algorithm uses support vectors to find hyperplane to classify and transform the data according to the optimal boundary. It utilizes kernel function that is a set of mathematical functions which takes input to transform the data into the required processing form. SVM algorithm is well suited for the datasets with higher number of dimensions as it does not require bigger data samples.³¹ It obtained 90.12% accuracy on Heart Statlog dataset.

Naïve Bayes

NB algorithm uses probabilistic techniques to predict the outcome of the observed data.³² It calculates the probability of input features and provides probability of the output by identifying the relevant symptoms to that of disease.³³ NB algorithm is well suited for datasets which are larger in size. It uses Bayes theorem with naïve approach which assumes that all attributes are independent. NB algorithm provided 88.39% accuracy on the Statlog dataset, 77.82% accuracy on the Heart UCI dataset, 75.54% accuracy on the Chd dataset and 85.52% accuracy on the Framingham dataset.

Random forest

RF algorithm uses DTs to form a forest and adds randomness to promote these forests. It takes samples of the data and makes DTs of the samples. RF algorithm seeks the optimal attributes from a random subset of discrete forest for prediction.³⁴ RF algorithm is well suited for large data and it can handle missing values as well. It is more flexible and diverse than the other algorithms. It generates high accuracy when used in ensemble with cross-validation. The performance of RF algorithm on the datasets is higher as compared with the other algorithms. The results obtained by RF algorithm are 87.65% accuracy on the Statlog dataset, 80.73% accuracy on the Heart UCI dataset, 74.82% accuracy on the Chd dataset and 85.44% accuracy on the Framingham dataset.

Experimentation-classification via ensemble

To improve the performance measures of machine learning model, several methods are used. Ensemble is an improvement technique that is used to combine multiple algorithms to achieve better results. Ensemble method usually provides more accuracy than a single model. Cross-validation uses sampling to increase data validation by learning from different training data samples, hence, it helps prevention from overfitting. In our proposed model, we have used AdaBoost method in cross-validation to improve the accuracy of the model. AdaBoost algorithm is a boosting algorithm which is used as an ensemble method. AdaBoost reassigns the weights of instances in the dataset and then assigns more weight to the misclassified instance so that it will appear with high probability in the training subset of classifier so that it will focus on difficult cases. Therefore, it provides a single strong classifier to enhance the results. The main process workspace of the model in rapid miner is shown in Figure 5. In the training sub-process of the cross-validation, we have used AdaBoost method. AdaBoost helps in reducing biasness and weighs the model on its performance. The cross-fold and number of iterations is set to 10. RF has been used to perform prediction on the datasets.

We used RF algorithm on the described datasets which formed DTs of the samples generated by sample bootstrapping. Each DT consists of complete set of attributes. The algorithm then used randomness to grow the DTs in to forest for samples generated. After the formation of discrete unrelated forests, it uses AdaBoost ensemble for prediction. AdaBoost is resistant towards overfitting. Therefore, boosting methods reduces biasness in general and RF utilizes diversity, the model provided the best accuracy of all algorithms.

The Apply Model operator in testing phase applies the RF classifier model generated by the AdaBoost. Training data are used to train a model for prediction. The trained model is then used to predict on testing data. Performance operator is used to measure the model performance.

Performance measures

The confusion matrix of the experimental model has been calculated to check the possibility of predicted class. Performance measures can be used to evaluate the effectiveness and efficiency of the applied classifiers. The calculations have been made using the following formulas of the mentioned performance metrics.

Accuracy is used to determine the number of times a classifier correctly predicted using the formula given below.

$$\text{Accuracy} = \frac{(TP + TN)}{\text{Total}} \quad (2)$$

where TP is the true positive rate and FP is the false positive rate. True positive rate or sensitivity describes the number of

times positive value of the class is correctly predicted. It is measured using the following formula. It is also known as recall.

$$\text{Recall} = \frac{TP}{\text{Actual yes}} \quad (3)$$

False positive rate describes the number of times positive class is not predicted correctly. We can measure it by using the formula below.

$$\text{FalsePositive} = \frac{FP}{\text{Actualnumber}} \quad (4)$$

True negative rate or specificity describes the number of times negative value of the class is correctly predicted. We can obtain true negative by using

$$\text{Truenegativerate} = \frac{TN}{\text{Actualnumber}} \quad (5)$$

Or

$$\text{Truenegativerate} = (1 - FP) \quad (6)$$

Precision is a measure of how often the positive prediction is actually correct. It can be measured as follows.

$$\text{Precision} = \frac{TP}{\text{Predicted yes}} \quad (7)$$

F measure or F1 score is weighted average of recall and precision rate.

$$F1 - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (8)$$

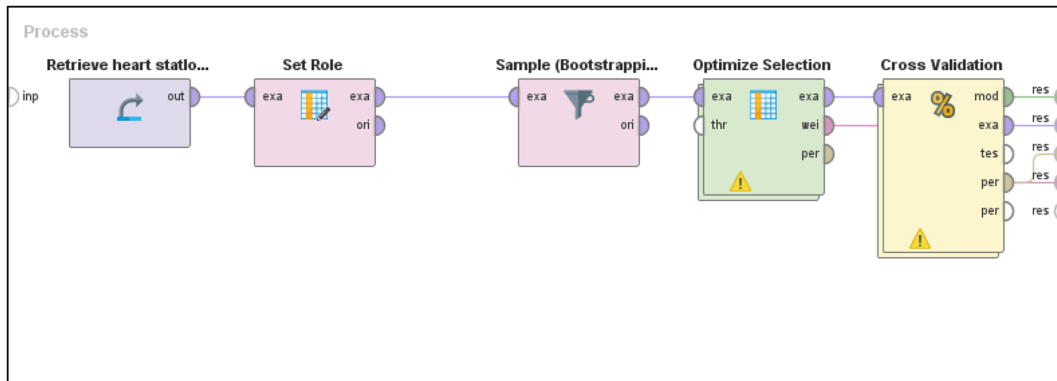
The confusion matrix of each dataset is given below. Table 1 shows the confusion matrix of all the datasets using the proposed model.

Results

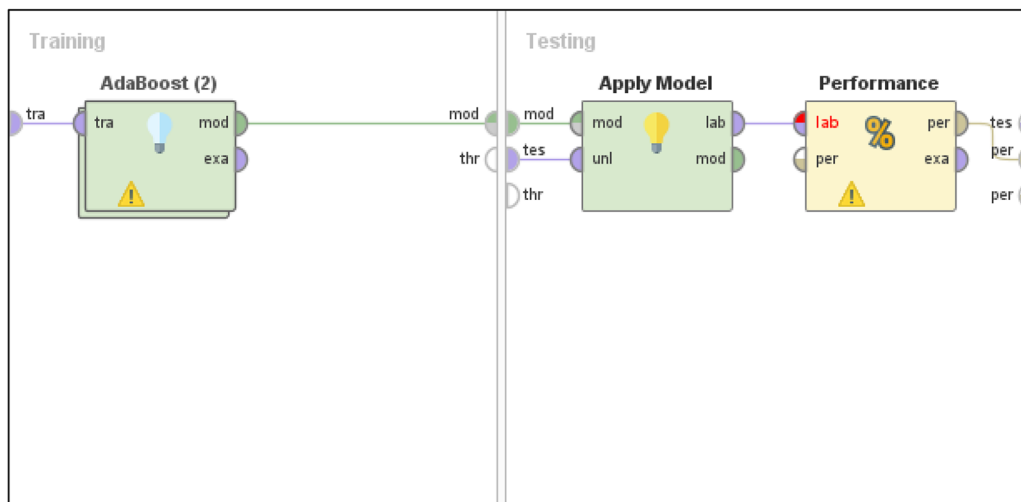
The results obtained from the experiments via ensemble are given in Table 2. This shows the performance measures of our proposed model.

Comparison of proposed methodology

In this paper, we have proposed an effective framework for cardiovascular disease prediction with increased accuracy. We have achieved our primary goal of presenting a framework with enhance performance on various datasets. Therefore, the capability of our proposed framework can be illustrated by the comparison of the results achieved by individual classifiers with our model which provides distinct

Figure 5 Model using rapid miner operators (A) rapid miner main process; (B) cross-validation sub-process; (C) random forest classifier.

(A)



(B)



(C)

solution and is applicable on various datasets. A demonstration of the accuracy rates obtained on datasets by different algorithms in comparison with the proposed framework has been provided in the Supporting information section.

It illustrates the outperforming accuracy of the proposed model on all datasets as compared with the other algorithms. For Heart Statlog, it provides 99.48% accuracy. For coronary heart disease dataset, the comparison of obtained accuracy is 78.36% by the proposed model. On the Framingham dataset, the proposed model obtained an accuracy rate of

86.52%. Therefore, our proposed model outclassed the accuracy rate of other classifiers on the datasets mentioned above.

Comparative analysis and discussion

We have proposed a novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques with high accuracy rate. We have used five different datasets and processed the data using various machine learning techniques. After enlisting the previous studies and flaws in existing systems, we have incorporated the solution in our proposed model. We have distinctively achieved high accuracies on all the datasets. Using optimized selection and sample bootstrapping, mishandled data are utilized at early stages of machine learning processes. In order to demonstrate the capability of our framework, comparative analysis of our model with existing models is provided by Table 3.

Table 1 Confusion matrix of datasets,

Dataset	True positive	False negative	False positive	True negative
Heart Statlog	502	2	4	653
Heart UCI	2157	89	223	2648
Framingham	15 472	2450	6	295
Coronary heart disease	309	45	385	1248
Stroke	193	1	846	20 933

Table 2 Result of proposed framework.

Dataset	Feature selection	Accuracy	Recall	Classification error	F-measure	AUC optimistic	AUC pessimistic	Precision	Sensitivity	Specificity
Heart Statlog	Optimize selection	99.48	99.69	0.52	99.55	1	0.995	99.4	99.69	99.21
Heart UCI	Optimize selection	93.90	96.75	6.10	94.44	0.997	0.883	92.23	96.75	90.63
Stroke	Optimize selection	96.25	99.99	3.75	98.07	1	0.213	96.23	99.99	20.98
Framingham	Optimize selection	86.52	10.75	13.48	19.32	1	0.118	98.25	10.75	99.96
Coronary heart disease	Optimize selection	78.36	96.52	21.64	85.31	0.979	0.459	76.45	96.52	44.54

Table 3 Comparison of Framework on datasets.

Author	Year	Dataset	Data imbalance	Feature selection	Features	Classifier	Validation type	Accuracy achieved
P. Anuradha ³⁵	2021	Heart Statlog	—	—	3	Maj. vote ensemble	<i>k</i> -fold cross-validation	87.04
Faria Rahman ³⁶	2021	Heart Statlog	-	Recursive Feature Elimination	6	Bagging (Random Forest)	Cross-validation	85.18
Kondeth Fathima ³⁷	2021	Heart Statlog	—	—	—	Neural Networks	—	98.77
Karadeniz, T. ³⁸	2021	Heart Statlog	—	Feature Reduction	10	Shrunk Covariance classifier	Cross-validation	88.8
Walaa Adel Mahmoud ³⁹	2021	Framingham dataset	—	—	—	Random forest	10-fold cross-validation resampling	85.05
Hoda ⁴⁰	2017	Framingham	No	Feature importance	9	KNN	—	66.7
P. Anuradha ³⁵	2021	Heart Disease UCI cleveland	—	—	5	CatBoost	<i>k</i> -fold cross-validation	91.8
Kondeth Fathima ³⁷	2021	Heart Disease UCI cleveland	—	—	—	Neural Networks	—	96.7
Proposed model	2023	Heart Statlog dataset	—	Optimized selection	—	Random forest	Cross-validation	99.48%

We have discussed machine learning model exploring different techniques to overcome the shortcomings of the existing systems. We have evaluated missing data and data imbalance as a first step of pre-processing. We used replacement method and sampling method for handling missing values and imbalance class respectively. For the purpose of feature selection, we used optimization process to select the optimized features of the datasets. As a main component of our proposed framework, it plays a significant role in enhancing the performance of the model in early stages. In the experimental part, we have used ensemble method to elevate the accuracy on our datasets which are now past the pre-processing step. For classification, we have used different machine learning classifiers illustrated above. Random forest provided the high accuracy rates on all the datasets.

Conclusion and future work

In this paper, we have presented a novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. Health researches show different approaches for disease diagnosis; however, machine learning provides better accuracy rates. We have utilized machine learning techniques and classifiers to prediction heart disease. The system which provides high accuracy for the prediction of heart disease can play a significant role in saving a precious life. The proposed model consists of basic four steps. In the first step, missing values are replaced using mean replacement method. In the second step, sample bootstrapping is used to handle imbalance classes. Posterior step contains classification on the data. AdaBoost ensemble is applied on the features selected by optimized selection, and random forest classifier is used to predict the cardiovascular disease with improved accuracy. We used rapid miner as a tool for the implementation of this framework. The proposed framework provides the highest accuracy of 99.48% on heart Statlog, 93.90% on Heart UCI, 96.25% on stroke dataset, 86% on Framingham dataset and 78.36% on coronary heart disease dataset, respectively. A limitation of this study is that it has utilized five datasets, expanding which could potentially increase the accuracy.

In the future, we intend to work on more datasets to obtain a framework well-suited for all the data. The numbers of instances also influence the performance of the model. For the purpose of future experiments, more instances must be utilized. We will also deploy the web application of cardiovascular disease prediction system for its highly accurate assessment performance.

Conflict of interest statement

There is no conflict of interest.

Acknowledgements

The authors extend their appreciation to the Deputyship for Research Innovation, Ministry of Education in Saudi Arabia, for funding this research work through the project number 223202, and Joint Information Systems Committee UK, for funding the publication of this research study.

Conflict of interest

None declared.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Comparison of classifiers with proposed model (a) Heart Statlog Dataset (b) Heart UCI Dataset (c) Coronary Heart Disease Dataset (d) Framingham Dataset

Table S1. Linear Regression classifier results

Table S2. K-Nearest Neighbours classifier results

Table S3. Decision Tree classifier results

Table S4. Naive Bayes classifier results

Table S5. Random Forest classifier results

References

1. Cardiovascular Diseases (CVDs). <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>. Accessed 23 March 2022
2. Probability of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease. [https://www.who.int/data/gho/data/indicators/indicator-](https://www.who.int/data/gho/data/indicators/indicator-details/GHO/probability-(-)-of-dying-between-age-30-and-exact-age-70-from-any-of-cardiovascular-disease-cancer-diabetes-or-chronic-respiratory-disease)
3. Angell SY, McConnell M, Anderson CAM, Bibbins-Domingo K, Boyle DS, Capewell S, et al. The American Heart Association 2030 impact goal: a presidential advisory from the American Heart Association. *Circulation* 2020;**141**:E120-E138. doi:10.1161/CIR.0000000000000758
4. Heart attack cases in Pakistan|MMI. <https://mmi.edu.pk/blog/heart-attack-cases-in-pakistan/>. Accessed 23 March 2022
5. Global health estimates: leading causes of death. <https://www.who.int/data/>

- gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death. Accessed 23 March 2022
6. Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. *Nat Rev Cardiol* 2011;8:30-41. doi:10.1038/nrcardio.2010.165
 7. Haq AU, Li JP, Memon MH, Nazir S, Sun R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mob Inf Syst* 2018;2018:1-21. doi:10.1155/2018/3860146
 8. Bhattacharyya S, Berkowitz AL. Primary angitis of the central nervous system: avoiding misdiagnosis and missed diagnosis of a rare disease. *Pract Neurol* 2016;16:195-200. doi:10.1136/practneurol-2015-001332
 9. Butt MO, Rehman AU, Javaid S, Ali TM, Nawaz A. An application of artificial intelligence for an early and effective prediction of heart failure. In: *2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)*. IEEE; 2022: 1-6.
 10. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 2019;7:81542-81554. doi:10.1109/ACCESS.2019.2923707
 11. Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN, Pranavanand S. Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Appl Sci (Switzerland)* 2021;11:8352. doi:10.3390/app11188352
 12. Arul Jothi K, Subburam S, Umadevi V, Hemavathy K. Heart disease prediction system using machine learning. *Mater Today Proc* 2021; doi:10.1016/J.MATPR.2020.12.901
 13. Motarwar P, Duraphe A, Suganya G, Premalatha M. Cognitive approach for heart disease prediction using machine learning. In: *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE*. Vol. 2020; 2020. doi:10.1109/IC-ETITE.47903.2020.242
 14. Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Syst Appl* 2021;164:113918. doi:10.1016/j.eswa.2020.113918
 15. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak* 2019;19:211. doi:10.1186/s12911-019-0918-5
 16. (PDF) Heart diseases prediction using deep learning neural network model. https://www.researchgate.net/publication/341831889_Heart_Diseases_Prediction_using_Deep_Learning_Neural_Network_Model. Accessed 23 March 2022
 17. Arunachalam S. Cardiovascular disease prediction model using machine learning algorithms. *Int J Res Appl Sci Eng Technol* 2020;8:1006-1019. doi:10.22214/ijraset.2020.6164
 18. Saleh Alotaibi F. Implementation of machine learning model to predict heart failure disease. *IJACSA Int J Adv Comput Sci Applic* 2019;10: doi:10.14569/IJACSA.2019.0100637
 19. Ali F, el-Sappagh S, Islam SMR, Kwak D, Ali A, Imran M, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf Fusion* 2020; 63:208-222.
 20. Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. *IOP Conf Ser Mater Sci Eng* 2021; 1022:012072.
 21. Rani P, Kumar R, Ahmed NMOS, Jain A. A decision support system for heart disease prediction based upon machine learning. *J Reliab Intell Environ* 2021;7: 263-275.
 22. UCI Machine Learning Repository: Statlog (Heart) data set. [https://archive.ics.uci.edu/ml/datasets/statlog%2B\(heart\)](https://archive.ics.uci.edu/ml/datasets/statlog%2B(heart)). Accessed 25 March 2022
 23. Coronary Heart Disease. Kaggle. <https://www.kaggle.com/datasets/billbasener/coronary-heart-disease>. Accessed 25 March 2022
 24. Machine Learning—Heart Disease Framingham. Kaggle. <https://www.kaggle.com/code/lauriandwu/machine-learning-heart-disease-framingham/data>. Accessed 25 March 2022
 25. Heart Disease Dataset. Kaggle. <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>. Accessed 25 March 2022
 26. Stroke Prediction Dataset. Kaggle. <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>. Accessed 25 March 2022
 27. Voloshynskiy O, Vysotska V, Bublyk M. Cardiovascular disease prediction based on machine learning technology. In: *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*. Vol.1; 2021: 69-75. doi:10.1109/CSIT52700.2021.9648587
 28. Saboor A, Rehman AU, Ali TM, Javaid S, Nawaz A. An applied artificial intelligence technique for early prediction of diabetes disease. *Proceedings of 3rd International Conference on Latest Trends in Electrical Engineering and Computing Technologies, INTELLECT 2022* 2022; doi:10.1109/INTELLECT55495.2022.9969401
 29. Mir A, Rehman AU, Javaid S, Ali TM. An intelligent technique for the effective prediction of monkeypox outbreak. In: *3rd IEEE International Conference on Artificial Intelligence*. Vol.2023. ICAI; 2023:220-226. doi:10.1109/ICAIS8407.2023.10136662
 30. Islam S, Rehman AU, Javaid S, Ali TM, Nawaz A. An integrated machine learning framework for classification of cirrhosis, fibrosis, and hepatitis. *Proceedings of 3rd International Conference on Latest Trends in Electrical Engineering and Computing Technologies, INTELLECT 2022* 2022; doi:10.1109/INTELLECT55495.2022.9969404
 31. Waqar M, Rehman AU, Javaid S, Ali TM, Nawaz A. An applied artificial intelligence aided technique for effective classification of breast cancer. In: *2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC)*; 2023:1-6. doi:10.1109/ICEPECC57281.2023.10209518
 32. Mehreen F, Rehman AU, Ali TM, Javaid S, Nawaz A. A computer aided technique for classification of patients with diabetes. In: *Proceedings of 3rd International Conference on Latest Trends in Electrical Engineering and Computing Technologies, INTELLECT 2022*; 2022. doi:10.1109/INTELLECT55495.2022.9969392
 33. Dinesh KG, Arumugaraj K, Santhosh KD, Mareeswari V. Prediction of cardiovascular disease using machine learning algorithms. In: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*; 2018:1-7. doi:10.1109/ICCTCT.2018.8550857
 34. Aleem I, Ur Rehman A, Javaid S, Ali TM. An integrated machine learning framework for effective classification of water. In: *2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC)*; 2023:1-6. doi:10.1109/ICEPECC57281.2023.10209495
 35. Anuradha P, David VK. Feature selection and prediction of heart diseases using gradient boosting algorithms. In: *Proceedings—International Conference on Artificial Intelligence and Smart Systems, ICAIS*. Vol.2021; 2021:711-717. doi:10.1109/ICAIS50930.2021.9395819
 36. Rahman, F. & Mahmood, A. A dynamic approach to identify the most significant biomarkers for heart disease risk prediction utilizing machine learning techniques. Machine learning and Data Science View project Applied cryptography view project. 2022 10.1007/978-3-031-17181-9_2
 37. Fathima, K. & Vimina, E. R. Heart disease prediction using deep neural networks: a novel approach. Lecture notes in networks and systems 213, 725–736 (2022). 10.1007/978-981-16-2422-3_56
 38. Karadeniz T, Tokdemir G, Maras HH. Ensemble methods for heart disease prediction. *New Gener Comput* 2021;39: 569-581. doi:10.1007/s00354-021-00124-4
 39. Adel Mahmoud W, Aborizka M, Ahmed Elsayed Amer F. Heart disease prediction using machine learning and data mining techniques: application of

- Framingham dataset. *Turk J Comput Math Educ (TURCOMAT)* 2021;**12**: 4864-4870. doi:[10.17762/turcomat.v12i14.11445](https://doi.org/10.17762/turcomat.v12i14.11445)
40. Elsayed HAG, Syed L. An automatic early risk classification of hard coronary heart diseases using Framingham scoring model. In: *ACM International Conference Proceeding Series*; 2017. doi:[10.1145/3018896.3036384](https://doi.org/10.1145/3018896.3036384)