

A predictive analysis framework of heart disease using machine learning approaches

Shourav Molla¹, F. M. Javed Mehedi Shamrat², Raisul Islam Rafi¹, Umme Umaima³, Md. Ariful Islam Arif¹, Shahed Hossain¹, Imran Mahmud^{2,4}

¹Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

²Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

³Department of Aeronautical Engineering, Military Institute of Science & Technology, Dhaka, Bangladesh

⁴Graduate School of Business, Universiti Sains Malaysia, Malaysia

Article Info

Article history:

Received Apr 10, 2022

Revised May 16, 2022

Accepted Jun 29, 2022

Keywords:

Decision tree

Gradient boosting

Heart diseases

Random forest

Univariate feature selection

ABSTRACT

Heart disease is among the leading causes for death globally. Thus, early identification and treatment are indispensable to prevent the disease. In this work, we propose a framework based on machine learning algorithms to tackle such problems through the identification of risk variables associated to this disease. To ensure the success of our proposed model, influential data pre-processing and data transformation strategies are used to generate accurate data for the training model that utilizes the five most popular datasets (Hungarian, Stat log, Switzerland, Long Beach VA, and Cleveland) from UCI. The univariate feature selection technique is applied to identify essential features and during the training phase, classifiers, namely extreme gradient boosting (XGBoost), support vector machine (SVM), random forest (RF), gradient boosting (GB), and decision tree (DT), are deployed. Subsequently, various performance evaluations are measured to demonstrate accurate predictions using the introduced algorithms. The inclusion of Univariate results indicated that the DT classifier achieves a comparatively higher accuracy of around 97.75% than others. Thus, a machine learning approach is recognized, that can predict heart disease with high accuracy. Furthermore, the 10 attributes chosen are used to analyze the model's outcomes explainability, indicating which attributes are more significant in the model's outcome.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

F. M. Javed Mehedi Shamrat

Department of Software Engineering, Daffodil International University

Daffodil Road, Asulia, Dhaka 1341, Bangladesh

Email: javedmehedicom@gmail.com

1. INTRODUCTION

Heart disease encompasses various conditions causing cardiovascular problems, including abnormal heartbeats or blood artery problems. According to data from the World health organization, heart disease is one of the primary reasons for the high mortality rate worldwide, causing 17.9 million death every year [1]. Coronary artery disease, congenital heart disease, and arrhythmia are the most common forms of cardiovascular diseases. Machine learning is widely used in the medical sector as it is a unique and appropriate method for developing an algorithm to diagnose various types of diseases. If any disease is identified at the primary phases and treatment procedures are taken at the soonest possible, the death rate can be comprehensively controlled. Heart disease investigation employing a machine learning system has been applied in various research observations. Katarya and Meena [2] used the explained machine learning

algorithms to predict heart ailments and associated risk variables. Different machine learning techniques have been applied, such as logistic regression (LR), RF, K-Nearest Neighbor (KNN), support vector machine (SVM), and Naive Bayes. DT, LR, KNN, and Naive Bayes achieved an accuracy of 81.31%, 93.40%, 71.42%, and 90.10%, respectively. Random forest acquired the greatest outcomes, which was 96.50%. K and M [3] proposed a strategy for categorizing heart disease sufferers according to their risk factors. Five separate approaches were applied, among which the best accuracy was obtained from the SVM, which was 87.91%. The second-best model was both LR and KNN gaining an accuracy of 86.81%. Khair and Dasari [4] designed a classification technique using LR, SVM, KNN, and multi-layer perceptron (MLP), which are three different types of neural networks. The results of 10-fold cross-validation on the sample demonstrate that SVM marginally outperforms MLP neural network classifier, KNN, and Logistic Regression in terms of mean accuracy, which are 73.8%, 73.4%, 73.2%, and 72.7%, respectively. Kondababu *et al.* [5] proposed a technique that processes raw data by using machine learning methods for heart disease prevention. The HRFLM technique proposes the combination of appearances of the random forest (RF) and the linear method (LM). Using the HRFLM approach, the accuracy level was 88.7%. Rani *et al.* [6] designed a hybrid judgment support method using machine learning algorithms. They have found the highest 86.60% accuracy using RF classifiers for the proposed system.

The objective of this research is the prediction of heart disease using machine learning approaches. Better observation of heart patients can assist in preventing the death rate. Therefore, we present a comparative study to show whether a person is affected by heart disease or not. This paper used the UCI dataset to use the UCI dataset to compare five Machine learning classifiers: RF, DT, SVM, GB, and XGB, using the UCI dataset [7]. This dataset has been pre-processed to handle the accurate prediction. The classifiers are applied for selecting sets of essential attributes. A group of ten attributes is nominated through the Univariate feature selection technique, which helps resolve the overfitting and underfitting issues. A voting classifier has been applied and used hard voting to evaluate our method. This work also investigated the computational complexity of each algorithm proposed for the framework. Figure 1 depicted each step of the proposed methodology.

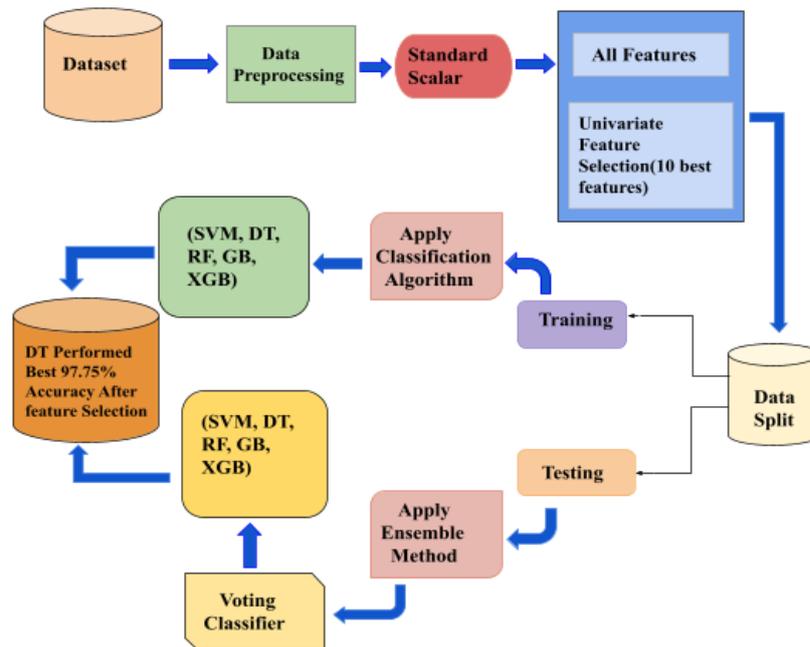


Figure 1. The proposed system for heart disease prediction

2. RESEARCH METHODOLOGY

2.1. Dataset collection

The dataset used comes from the 'UCI machine learning repository'. We gathered 1190 instances as a text file from the five merged datasets which are available in the UCI database, together with 14 distinctive characteristics. These combined datasets have 13 attributes used as diagnosis inputs, with the 'num' feature nominated as an output. All of them consist of numerical values. The 'num' attribute can have a value of 0, 1,

2, 3, or 4. The projected number '0' indicates that the patient is free of heart disease, while a score of 1 to 4 indicates different heart disease phases.

2.2. Overview of the proposed model

The dataset introduced had five different classes, namely 0, 1, 2, 3, and 4. Since this study is done based on whether a patient has heart disease or not. Therefore, we transform all scores in the series of 1 to 4 to a 1. This means that the attribute now has 0 and 1 values. Afterward, feature scaling is applied through a standard scaler to convert all the values to a similar scale. Here standard scaler is selected because it performs better than the min-max scaler. The min-max scalers can be used in convolutional neural networks. Because if the value is scaling, then it is easy to calculate weight because scaling has already been done in (0,1). The min-max scaler scales all data features between 0 and 1. It cannot manage outliers well, which is a fairly significant downside of min-max scaler.

The performance of classifiers is evaluated by some characteristics selected using the univariate feature selection technique and the original features. When the feature selection technique is applied, the dataset is divided into two portions: training and testing. 70% of the data is allotted to the training stage based on model learning rates, and the remaining 30% is allocated to the testing stage. All ensemble designs with approaches are developed for evaluation throughout the merged dataset [8].

2.3. Justification of proposed model

The correlations of features are presented to understand input and output features better. We have used 14 features in the framework. These are: 'age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num'. According to the importance of features, we have selected the 10 most important features including 'age', 'sex', 'cp', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal'. The following Figure 2 is illustrated depending on the 10 strongly associated features with the target attribute (num) chosen using the Univariate feature selection approach. The attribute values ranging from 0.3 to -0.4 are presented on the right side of the figure. Thus, it is clear that oldpeak, cp, and exang attributes have a powerful connection through an age where the score was approximately 0.3; elsewhere, the deepest correlation, roughly -0.4 remains seen in thalach. Similarly, cp has a strong relationship with exang. On the other hand, the correlations score between other attributes was not as strong, ranging from 0.15 to -0.3.

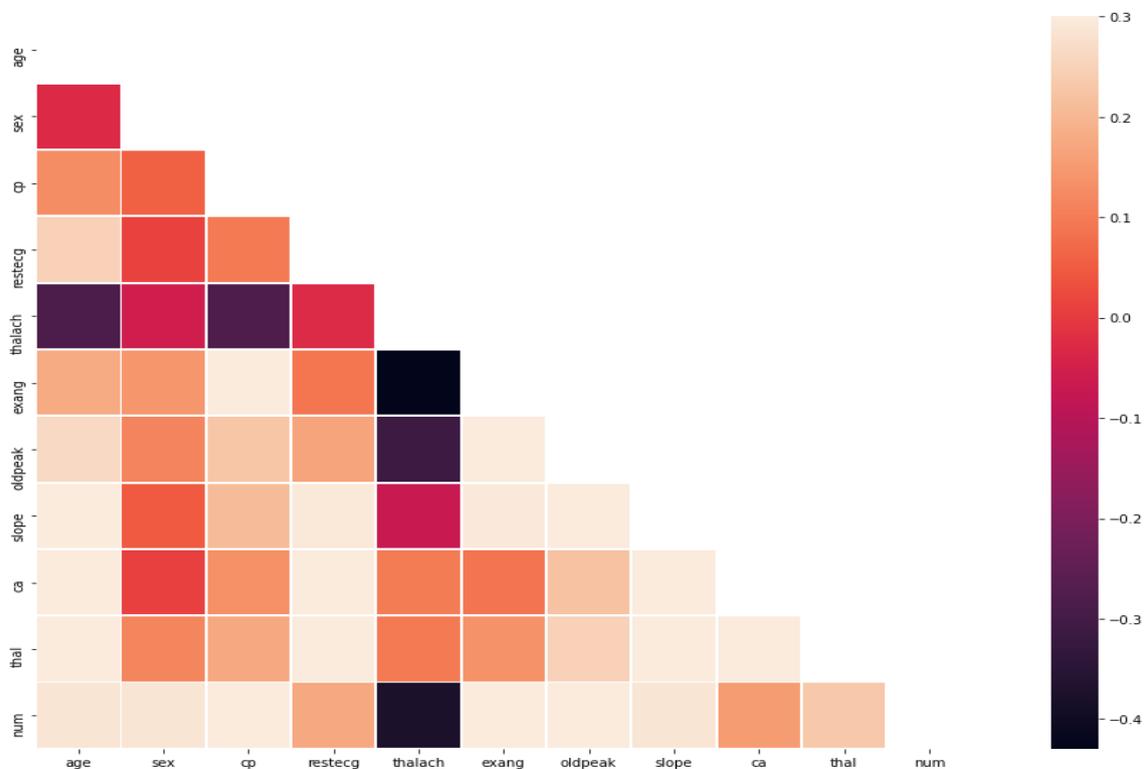


Figure 2. Correlations between different attributes (10 features heatmap)

2.4. Overview of the pre-processing

Using the standard scaler pre-processing technique, we have converted the attributes (num) on a similar scale. The input is represented as X , μ indicates the mean value and σ is denoted by the standard derivation. The procedure for converting (1) is provided by (1) [9]

$$\text{Standardization, } x = (x - \mu)/\sigma \quad (1)$$

2.5. Feature selection technique

Feature selection is a significant subject in machine learning approaches since it influences the model's performance. This also helps to decrease the processing time. Here, univariate feature selection has been used.

2.5.1. Univariate feature selection technique

A statistical test is used to pick a feature subset with the strongest association with a class label in univariate feature selection [10]. The univariate feature selection method includes sorting each component separately to accomplish the strength of the relationship between the component and the target variable. These strategies are straightforward to use and easy to perceive, and they are beneficial for acquiring a deeper interpretation of data. There have been a variety of univariate selection alternatives available. The benefits of feature selection are decreasing the number of features, minimizing overfitting issues, and increasing the models' accuracy. The select KBest and f_classif methods have been employed for this feature selection method.

2.6. Machine learning classifiers

2.6.1. Support vector machine

The SVM creates a decision boundary that separates the dataset into two classes, 0 and 1, on opposite sides of the hyperplane. There are two types of SVM: linear and non-linear. The data may be disjointed linearly using linear SVM, which denotes a single line. However, a single line cannot be utilized to split the data. In our model, linear SVM have used to get an improved accuracy. Let i =the i^{th} vector is represented by 1, 2, ..., n , $x_i \in R^n$ while the target item is represented by $y_i \in R^n$. The linear kernel function $f(x) = wt * x + b$, where w denotes a dimensional coefficient vector and b indicates an offset. This is accomplished by resolving the optimization problem by (2) and (3):

$$\text{Min}_{w,b,\xi_i} \frac{1}{2}w^2 + c \sum_{i=1}^n \xi_i \quad (2)$$

$$s.t. y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i \in \{1, 2, \dots, m\} \quad (3)$$

2.6.2. Random forest

The random forest (RF) classifier is an ensemble algorithm [11]-[13] which means that it is made up of multiple algorithms. It usually comprises numerous DT algorithms in this scenario [14]. During the training portion, RF constructs a full forest using multiple uncorrelated and random decision trees [15]. Multiple learning algorithms are used in ensemble learning approaches to create an ideal predictive model that can outperform any single model's prediction [16]. The total number of decision tree algorithms is the result of random forest. The random forest algorithm works in the following way. The ensemble technique constructs and combines many decision trees to achieve the optimal outcome. Allow the numbers given to be used $X = (x_1, x_2, x_3, \dots, x_n)$ through response $Y = \{x_1, x_2, x_3, \dots, x_n\}$ using a lowest bound of $b=1$ and a higher bound of B : averaging the guesses provides the estimate for the sample x' . $\sum_{b=1}^B f_b(x')$ is the equation from every individual tree for x' , that is shown using (4).

$$j = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (4)$$

2.6.3. Decision tree

This algorithm is applied in the solution of difficulties involving regression and classification. Every internal node in a decision tree's structure relates to the trial outcome, every leaf node corresponds to a distinct session, and each branch corresponds to a different test result. 'Learning' constructed on DT sometimes employs an upside-down tree-based progress approach. It is capable for classification and regression problems. The 'best attribute provides the most useful information. The entropy of a dataset

reveals how homogenous it is. The step at which the entropy elements upturn or reductions is known as information gain [17].

$$E(D) = -P(\text{positive}) \log_2 P(\text{positive}) - P(\text{negative}) \log_2 P(\text{negative}) \quad (5)$$

In (5) estimates the entropy E of dataset D, which holds the positive and negative ‘decision attributes’.

$$\text{Gain}(\text{Attributes } X) = \text{Entropy}(\text{Decision Attribute } Y) - \text{Entropy}(X, Y) \quad (6)$$

This approach aims to generate a framework that can evaluate the valuation of the reliant variable by reading up on essential standards for navigation.

2.6.4. Gradient boosting

Gradient boost is a boosting technique for classification and regression issues requiring only 100 data [18]. Gradient boosting is made up of three main components [19]. We have utilized an upgraded destruction function to reduce the destruction function, a sluggish performer to produce predictions, and an additive prototype to merge weak performers [20]. This approach can reduce algorithm overfitting and increase algorithm efficiency. When there has been a disproportion between the numbers in each group, Gradient Boosting can help improve accuracy. The optimal function F(X) after several iterations [21] is generated according to (7).

$$F(x) = \sum_{i=0}^m f_i(x) \quad (7)$$

Where $f_i(x)$ ($i=1, 2, \dots, M$) denotes feature increases, the $f_i(x) = -\rho_i \text{ gm}(X)$. The most recent base-learner is the most significant destruction function connected through negative ascents [22]. For the m^{th} iteration, the negative gradient is:

$$gm = - \left[\frac{\partial L(y, F(x))}{\partial F(x)} \right] F(x) = F_m - 1(x) \quad (8)$$

Here, gm denotes the pathway along which the destruction function degrades fastest. $F(X) = F_m - 1(X)$ [21]. A fresh DT helps resolve the previous base learner's inaccuracy. The T prototype is then changed to:

$$F_m(X) = F_m - 1(X) + \rho_m x(X, \alpha_m) \quad (9)$$

2.6.5. Extreme gradient boosting

Extreme gradient boosting (XGBoost) is a machine learning technique. In gradient boosting machines (GBM), XGB is one of the most common boosting tree algorithms. It has already been widely used in industrial applications because of its great problem-solving performance and low feature engineering requirements. [23]. It is used for regression, classification and other works. $F = (f_1, f_2, f_3, f_4 \dots f_m)$ are the set of base learners. Therefore, the final prediction is given in (10).

$$\hat{y}_1 = \sum_{t=1}^m f_t(x_i) \quad (10)$$

2.7. Xplainability metrics

Assuming that the baseline estimation of ensemble tree classifiers, decision tree, is such a totally interpretable model, measures to quantify the effectiveness of such former black-box modeling techniques of readability are needed [24]. I (model) is the proportion of veiled characteristics that do not provide information to the ultimate classification result divided by the number of attributes. Furthermore, the model's fidelity, F (model), is frequently measured as the ratio of such completely explainable corresponding figure's accuracy and reliability to the model's accuracies. Ultimately, the fidelity-to-interpretability factor (FIR) of the model revealed how much of the model's readability is surrendered in exchange for performance. The well-adjusted fraction of 0.5 would be perfect, calculated as $FIR = F / (F + I)$.

3. RESULTS AND DISCUSSION

3.1. Outcomes of the features selection process

Using univariate feature selection, 10 features have been selected according to the importance of each feature. For example, we get the old-peak score (295.352) and exang score (294.59) which provides the feature scores. Thus, we find the importance of all features and select the 10 best features. Table 1 shows the univariate feature selection and their ranking.

Table 1. Univariate feature selection and their rankings

Name	Code	Score
ST depression induced by exercise.	oldpeak	295.352576
exercise induced angina	exang	294.598871
Chest pain type	cp	284.282977
Maximum heart rate	thalach	197.914520
slope of the peak workout ST phases	slope	106.835719
Sex	Sex	106.835719
Age in Years	age	105.845603
Thal	thal	69.080696
resting electrocardiographic results	restecg	36.169881
number of major vessels (0-3) colored though fluoroscopy	ca	28.644046

3.2. Ensemble technique (Voting)

In the sector of machine learning words, major voting means a meta-classifier for putting together the same kind of or conceptually various machine learning techniques for classification via majority voting. We have used 'hard' voting here. The anticipated output class in hard voting is the class with the largest majority of votes and the highest likelihood of prediction from each classifier (see Table 2).

Table 2. Result of ensemble technique (Voting)

Classifiers	All features (%)	10 features (%)
RF, SVM, DT, GB, XGB	94	94
GB, RF, SVM, DT	94	95
SVM, XGB, DT	92	95
SVM, DT, GB	95	96

3.3. Comparison among methods constructed on accuracy, recall, precision and F1-scores

Figure 3 illustrate that for the 14 features, SVM classifier got 95.23%, RF 93.27%, DT 92.99%, GB 91.03%, and XGB 89.91%. But for the 10 selected features, SVM obtained 82.07%, RF 92.99%, GB 91.59%, XGB 93.27%, and DT got the highest accuracy of 97.75%. SVM obtained the recall value of 95%, and XGB got a poor recall score of 90%. RF, DT, and GB obtained the recall score 94%, 93%, 91%, respectively, based on 14 input features. SVM gets an abysmal recall score of 83%, and decision tree gets a higher recall score of 98% compared to other classifiers. Gradually RF, GB, and XGB obtained recall 93%, 92%, 94% for the univariate feature. When 14 features were applied, SVM obtained 95% precision whereas XGB got 89%, DT and RF obtained 95%, and GB achieved 91%. When the 10-univariate feature was applied, DT got the highest precision which is 97%, and SVM acquired the lowest- 82%. GB and RF got 91% and 92% precision scores. XGB reached a 93% precision score. In the F-1 score, for the 14 features, SVM scored 95%, XGB 90%, RF and DT got 93%, and GB 93%. For 10 selected features SVM scored 82%, DT 97%, RF and XGB got 93% and GB classifiers reached 91%.

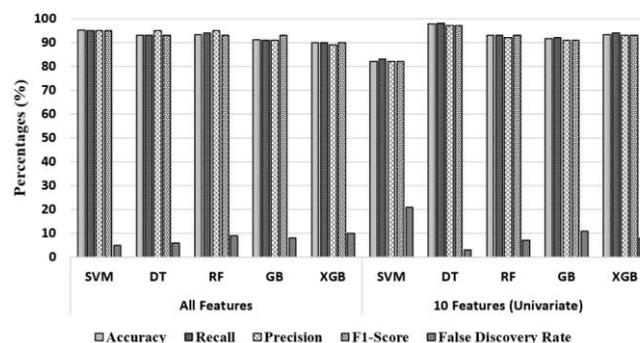


Figure 3. Classifiers results for all (14) features and 10 features

3.4. Comparisons based on specificity, negative predictive value, false-positive rate, false-negative value and AUC

The negative predictive score is the chance that issues through a negative showing test where the fact is that does not have the disease [25]. The NPV 96.47% was obtained from SVM and RF classifiers in 14

features. decision tree gets 92.25% NPV, whereas Gradient boost and Extreme Gradient Boost get 89.43% NPV. While evaluating based on the Univariate Feature selection technique, the maximum NPV is 98.59% achieved from DT and the lower NPV is 85.91%, achieved from SVM. 93.66% NPV was achieved by RF. GB and XGB received 95.07% and 95.77%, respectively. After applying univariate feature selection, the maximum false positive rate is 26.50% in the SVM model, and the lowest FPR is 4.76%, achieved by DT. Respectively, RF, GB, and XGB FPR are 10.73%, 14.55%, and 11.68%. Without univariate feature selection, the highest FPR in XGB, which is 14.18% and impoverished FPR present in SVM, the score is 8.05%. RF, DT and GB gradually achieve 12.17%, 10.88%, and 11.80% FPR. With univariate feature selection, the highest false negative rate is 10.47% in the SVM model and XGB got the lowest FNR which is 2.95%. The FNR of DT is 0.95% with this selective feature. RF and GB achieved 4.32% and 3.51%. Without the univariate feature selection method, the rate of GB is 7.04%. The false-negative rates of SVM, DT and RF are 2.40%, 5.23%, and 2.48%. Besides, the FNR of XGB is 7.17%. Figure 4, represents the specificity, negative predictive value, false-positive rate, false-negative value and AUC results of the all features (14 features) in Figure 4(a) and 10 features (univariate features) in Figure 4(b).

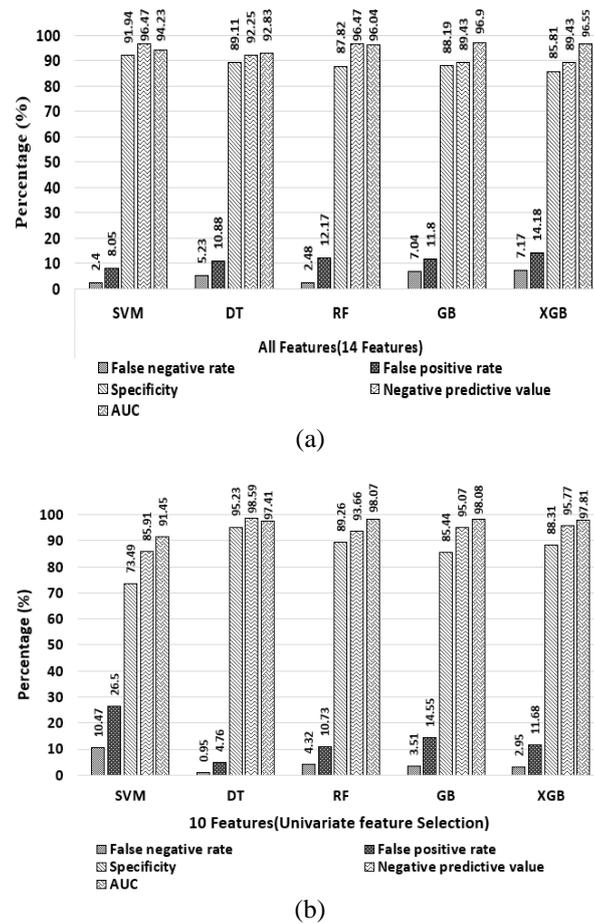


Figure 4. Specificity, negative predictive value, false-positive rate, false-negative value and AUC results for (a) all (14) features and (b) 10 features

3.5. Statistical analysis using means squared error, root mean squared error, log loss and kohen kappa scores

Here we analyze all features and the univariate feature selection. The highest RMSE is XGB (31.75%) for the all-input features, and the lowest is SVM (21.82%). The highest RMSE is SVM (42.34%) for univariate feature selection, and the lower RMSE is DT (15.87%). In all input features, RMSE of RF, DT and GB are (25.92%, 27.50%, and 29.93%). RMSE of RF, GB and XGB are (26.46%, 28.98%, and 25.92%) for the selected 10 features. log loss, which means logarithmic loss is a sorting loss function frequently used as an estimation metric in Kaggle competitions. In the machine learning sector, the work of log loss is to measure the amount of divergence of predicted probability with the actual label. For all input features, the

maximum LL is XGB (3.4829%) and the lowest LL is SVM (1.6447%). Gradually DT, RF, and GB LL are 2.6122%, 2.3219%, 3.0959%. For univariate feature selection, the highest LL is SVM (6.1919%) and the poor LL is DT (8.707%). 2.4187%, 2.9024%, 2.3219% LL are periodically for RF, GB and XGB. Figure 5 illustrates the result statistical analysis for all features in Figure 5(a) and 10 features in Figure 5(b).

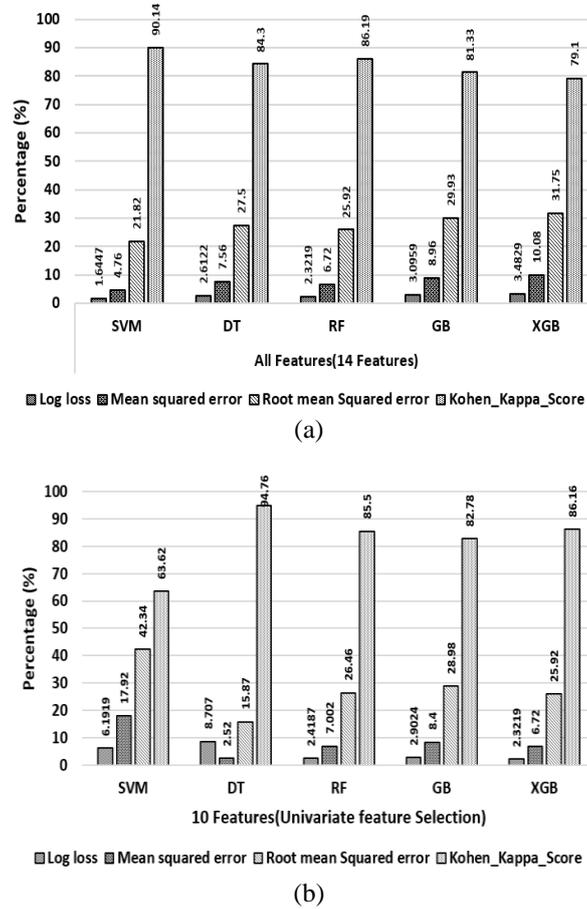


Figure 5. Statistical analysis results for (a) all (14) features and (b) 10 features

3.6. Computational complexity

The complexity of several models is depicted in Table 3. Computational complications comprise two sorts of complexities: training complexities and prediction complexities. The algorithms' algorithmic intricacies. The following approximations are obtained by representing n as the number of trees in the training section, and p denotes the number of features.

Table 3. Algorithmic complexities of the algorithms used

Models	Training	Training complexity calculation	Testing	Testing complexity calculation
SVM (Linear)	$O(n^2p + n^3)$	$O(1699320000)$	$O(n_{sv}p)$	$O(20)$
RF	$o(n^2pntress)$	$O(56644000)$	$O(pntress)$	$O(40)$
DT	$o(n^2p)$	$O(14161000)$	$O(p)$	$O(10)$
GB	$O(npntress)$	$O(4760000)$	$O(pntress)$	$O(4000)$

3.7. Hyperparameter tuning

GridSearchCV, which allocates hyperparameters, is a process of tuning which can determine the optimal value for a given model. In our proposed system, GridSearchCV has been used to obtain higher accuracy. The following parameters used on the examined algorithm are depicted in Table 4.

Table 4. Parameter used in classifiers

Algorithm	10 selected features (parameter)
SVM (Linear)	C=10, kernel='poly', gamma=100
RF	n_estimators=4, criterion='entropy', random_state=20
DT	Default
GB	n_estimators=400, max_features=4, random_state=42
XGB	"colsample bytree=0.3, learning rate=0.1, max_depth=5, alpha=10, n_estimators=400"

3.8. Compilation time

Based on all the features mentioned, we have found that the highest runtime is in SVM (0.0090) and the lowest is in DT (0.0059). In 10 selected features, the most excellent runtime is in SVM (0.0093) and the least runtime is in RF (0.0057). We also got that the maximum run time score is 0.0093 in SVM with 10 selected features using the univariate feature selection technique. Figure 6 portrays the compilation time of all the classifiers.

3.9. Explainability metrics

Explainability metrics necessitate a completely interpretable model as a form of benchmark against which the best model identified may be measured [26]-[28]. As a result, a decision tree is trained using the greatest model's variables and the set of features chosen to demonstrate the transparency of its classification choice. The classification performance of the decision tree generated is as follows: accuracy 97.75%, sensitivity (0.99), specificity (0.95), F1 score (0.97), and precision (0.97); the best model was utilized to assess the explainability, which received an interpretability of 23.07% a Fidelity of 100% and a FIR ratio of 81.25%. The interpretability measure uses the range of attributes throughout the main dataset as well as those deleted following feature selection, which is 14 and 10, respectively. The best model's accuracy of 97.75 was selected for the fidelity metric (F). It's worth noting that the FIR result is near the optimum score of 50%. Figure 7 illustrate the result of explainability metrics.

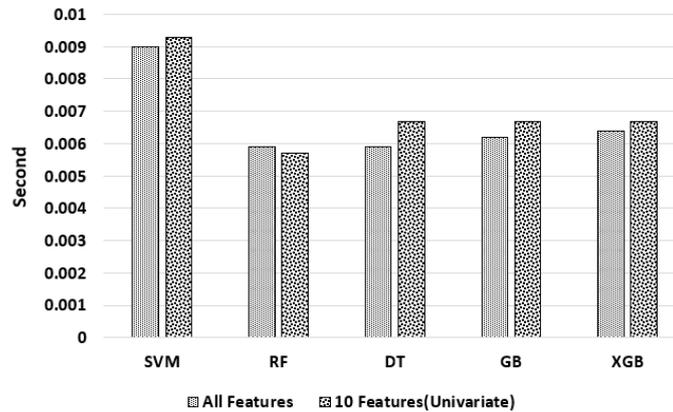


Figure 6. Compilation time of the classifiers

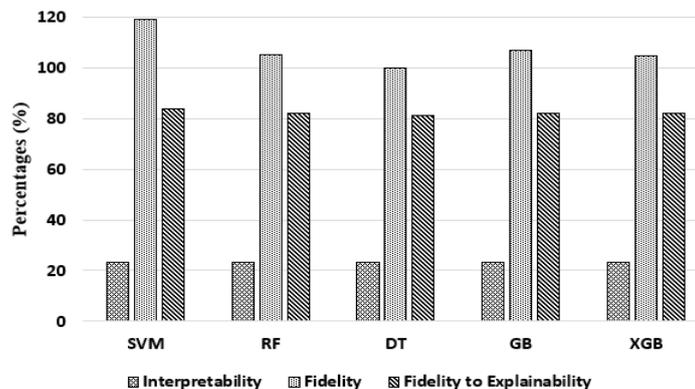


Figure 7. Explainability metrics result

4. CONCLUSION

This paper presents an effective machine learning-based diagnosis framework to determine heart disease based on the identified features. These machine learning techniques apply to processing raw data and developing a new framework and intelligence for diagnosing heart disease. Predicting heart disease is difficult but crucial. A small increase in prediction accuracy significantly influences the diagnosis of important cardiac characteristics, which helps to improve overall heart disease prediction accuracy. Machine learning approaches including SVM, RF, DT, GB, and XGB are used to design the system. A combination of the 5-most popular heart disease dataset has been implemented on the system. Our experimental results and evaluation show that DT achieved an accuracy of over 97% among ten top-selected features using the univariate feature selection technique. Moreover, a newer feature selection method could be developed to obtain a wider perception of the characteristics, which helps to increase overall heart disease estimate performance.

REFERENCES

- [1] "WHO: global cause of death due to heart disease," Accessed: March 31, 2022, [Online], Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.
- [2] R. Katarya and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Health Technology*, vol. 11, no. 1, pp. 87–97, 2021, doi: <https://doi.org/10.1007/s12553-020-00505-7>.
- [3] A. M, S. K and C. M, "Earlier Prediction on the heart disease based on supervised machine learning techniques," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 1696-1703, doi: 10.1109/ICICCS51141.2021.9432212.
- [4] H. Khadair and N. M. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021, doi: 10.14569/IJACSA.2021.0120505.
- [5] A. Kondababu, V. Siddhartha, B. H. K. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Materials Today: Proceedings*, 2021, doi: 10.1016/j.matpr.2021.01.475.
- [6] P. Rani, R. K. Gujral, N. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 6, 2021, pp. 263-275, doi: 10.1007/s40860-021-00133-6.
- [7] "Heart Disease Datasets From UCI Machine Learning Repository," Accessed: March 31, 2022, [Online], Available: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
- [8] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques," *International Conference on Smart Homes and Health Telematics*, vol. 12157, pp. 299–306, 2020, doi: 10.1007/978-3-030-51517-1_26.
- [9] A. Acharya, "Comparative study of machine learning algorithms for heart disease prediction," M.S. thesis, Helsinki Metropolia Univ. Appl. Sci., Helsinki, Finland, Apr. 2017.
- [10] R. Aggrawal, and S. Pal, "Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 4, pp. 2650-2665, doi: 10.17762/turcomat.v12i6.5765.
- [11] F. M. J. M. Shamrat *et al.*, "Analysing most efficient deep learning model to detect COVID-19 from computer tomography images," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 462-471, doi: 10.11591/ijeecs.v26.i1.pp462-471.
- [12] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, and M. Saifuzzaman, "Expert model of cancer disease using supervised algorithms with a LASSO feature selection approach," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2632-2640, 2020.
- [13] F. J. M. Shamrat, S. Chakraborty, M. M. Billah, M. Kabir, N. S. Shadin, and S. Sanjana, "Bangla numerical sign language recognition using convolutional neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 405-413, 2021, doi: 10.11591/ijeecs.v23.i1.pp405-413.
- [14] P. Ghosh *et al.*, "A comparative study of different deep learning model for recognition of handwriting digits," *International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020)*, pp. 857–866, January 19, 2021, doi: 10.2139/ssrn.3769231.
- [15] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: From early developments to recent advancements," *Systems Science & Control Engineering An Open Access Journal*, vol. 2, no. 1, pp. 602–609, Dec. 2014, doi: 10.1080/21642583.2014.956265.
- [16] M. O. Rahman *et al.*, "Internet of things based electrocardiogram monitoring system using machine learning algorithm," *International Journal of Electrical & Computer Engineering*, vol. 12, no. 4, pp. 2088-8708, 2022, doi: 10.11591/ijece.v12i4.pp3739-3751.
- [17] S. Hegelich, "Decision trees and random forests: Machine learning techniques to classify rare events," *European Policy Analysis*, vol. 2, no.1, pp. 98–120, 2016, doi: 10.18278/epa.2.1.7.
- [18] "An Overview of Gradient Boosting Algorithm," Accessed: March. 31, 2022. [Online]. Available: <https://machinelearningmastery.com/gentleintroduction-gradient-boosting-algorithm-machine-learning/>.
- [19] M. Almasoud and T.E.Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 10, no. 8, pp. 89–96, 2019, doi: 10.14569/IJACSA.2019.0100813.
- [20] K. C. Howlader *et al.*, "Machine learning models for classification and identification of significant attributes to detect type 2 diabetes," *Health information science and systems*, vol. 10, no. 1, pp. 1-13, 2022, doi: 10.1007/s13755-021-00168-2.
- [21] J. Cheng, G. Li and X. Chen, "Research on Travel Time Prediction Model of Freeway Based on Gradient Boosting Decision Tree," in *IEEE Access*, vol. 7, pp. 7466-7480, 2019, doi: 10.1109/ACCESS.2018.2886549.
- [22] Md. A. Talukder, M. Islam, Md. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine Learning-based Lung and Colon Cancer Detection using Deep Feature Extraction and Ensemble Learning," *arXiv e-prints, arXiv-2206*, 2022, doi: 10.48550/arXiv.2206.01088.

- [23] T. Chen and C. Guestrin, "XGBOOST: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [24] T. Tagaris and A. Stafylopatis, "Hide-and-Seek: A Template for Explainable AI," *arXiv:2005.00130*, Apr. 2020, doi: 10.48550/arXiv.2005.00130.
- [25] Z. Tasnim *et al.*, "Classification of Breast Cancer Cell Images using Multiple Convolution Neural Network Architectures," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, pp. 308–315, doi: 10.14569/IJACSA.2021.0120934.
- [26] S. Das, M. S. Islam and I. Mahmud, "A Deep Learning Study On Understanding Banglish and Abbreviated Words Used in Social Media," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 1690–1695, doi: 10.1109/ICICCS51141.2021.9432339.
- [27] K. M. Hasib, M. A. Habib, N. A. Towhid and M. I. H. Showrov, "A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service," *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 450–455, doi: 10.1109/ICICT4SD50815.2021.9396879.
- [28] I. Mahmud, J. Akter, and S. Rawshon, "SMS based disaster alert system in developing countries: A usability analysis," *International Journal of Multidisciplinary Management Studies*, vol. 2, no. 4, 2012.

BIOGRAPHIES OF AUTHORS



Shourav Molla    studying at Daffodil International University in the Department of Computer Science and Engineering, under the B.Sc. program, since 2019. He has been involved in cooperative research activities with researchers from Bangladesh and researchers from Australia, especially in machine learning, deep learning, and image processing. His primary areas of interest in the study include deep learning, machine learning, image processing, web development, and artificial intelligence. He is interested in reading history books and comparing them with the current world. He can be contacted at email: shourav15-2438@diu.edu.bd.



F. M. Javed Mehedi Shamrat    graduated from Daffodil International University with a B.Sc. in Software Engineering in 2018. He was formerly employed with Daffodil International University. He is presently employed as a lecturer at the European University of Bangladesh in the Department of Computer Science and Engineering. He has been actively engaged in collaborative research with researchers from Bangladesh, the United States of America, Canada, China, Korea, and Australia. He has several research publications published in prestigious journals (Scopus) and conferences (Scopus). His primary areas of interest in the study include the internet of things, deep learning, data science, android and web apps, image processing, neural networks, artificial intelligence, robotics, bioinformatics, and machine learning. He can be contacted at email: javedmehedicom@gmail.com.



Raisul Islam Rafi    studying at Daffodil International University in the Department of Computer Science & Engineering. His area of interest is in the IT sector, machine learning, deep learning and android development. He likes to participate in different competitions, projects and events. He is good at programming languages such as C programming and java. He is a person who is positive about every aspect of life. His strengths are hard work and being a quick learner. His hobby is learning new things. He can be contacted at email: raisul15-2578@diu.edu.bd.



Umme Umaima    graduated from the Military Institute of Science & Technology with a B.Sc in Aeronautical Engineering in 2014. She was formerly employed with SAIC Institute of Science & Technology as a lecturer in Electrical Engineering Department. She has been actively engaged in collaborative research with researchers from Bangladesh and the USA on machine learning, deep learning, and image processing. She has several research publications published in prestigious journals and conferences. Her primary interest study includes physics & astronomy. She can be contacted at email: umaimasnigdha123@gmail.com.



Md. Ariful Islam Arif    studying at Daffodil International University in the Department of Computer Science & Engineering. He loves to explore and learn new things. He is a quick self-learner. Since his days at university, he has been teaching himself different skills. He worked hard in that sector, which he likes most. He wants to be a professional digital marketer and an entrepreneur. He also loves to find happiness in small things. His area of interest is machine learning, deep learning and image processing. He can be contacted at email: ariful15-2451@diu.edu.bd.



Shahed Hossain    student at Daffodil International University in Computer Science and Engineering. Among his research interests are computer vision, image processing, neural networks, network security, and deep learning. His several research papers were published in the prestigious conference (Scopus). He is the founder and CEO of Pre-Eminent4, which is a registered IT company in Bangladesh. He can be contacted at email: rkoshahed.cse@gmail.com.



Dr. Imran Mahmud    is an associate professor and the Department of Software Engineering at Daffodil International University. Additionally, he serves as an assistant director of research. Dr. Imran earned a doctoral degree in technology management at Universiti Sains Malaysia. His research interests are human-computer interface, usability testing, software engineering measurements/models, and management information systems. Dr. Imran has multiple publications in Sage and IEEE journals. He can be contacted at email: imranmahmud@daffodilvarsity.edu.bd.