IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 23, 2020, accepted January 19, 2021, date of publication January 22, 2021, date of current version February 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3053759

Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques

PRONAB GHOSH[1], SAMI AZAM[2], MIRJAM JONKMAN[2], (Member, IEEE), ASIF KARIM[2], F. M. JAVED MEHEDI SHAMRAT[3], EVA IGNATIOUS[2], SHAHANA SULTANA[1], ABHIJITH REDDY BEERAVOLU[2], AND FRISO DE BOER[2]

[1]Department of Computer Science and Engineering, Daffodil International University, Dhaka 1225, Bangladesh
[2]College of Engineering, IT, and Environment, Charles Darwin University, Casuarina, NT 0810, Australia
[3]Researcher and Developer, Information and Communication Technology Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh, Dhaka 1000, Bangladesh

Corresponding author: Sami Azam (sami.azam@cdu.edu.au)

ABSTRACT Cardiovascular diseases (CVD) are among the most common serious illnesses affecting human health. CVDs may be prevented or mitigated by early diagnosis, and this may reduce mortality rates. Identifying risk factors using machine learning models is a promising approach. We would like to propose a model that incorporates different methods to achieve effective prediction of heart disease. For our proposed model to be successful, we have used efficient Data Collection, Data Pre-processing and Data Transformation methods to create accurate information for the training model. We have used a combined dataset (Cleveland, Long Beach VA, Switzerland, Hungarian and Stat log). Suitable features are selected by using the Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) techniques. New hybrid classifiers like Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM) are developed by integrating the traditional classifiers with bagging and boosting methods, which are used in the training process. We have also instrumented some machine learning algorithms to calculate the Accuracy (ACC), Sensitivity (SEN), Error Rate, Precision (PRE) and F1 Score (F1) of our model, along with the Negative Predictive Value (NPR), False Positive Rate (FPR), and False Negative Rate (FNR). The results are shown separately to provide comparisons. Based on the result analysis, we can conclude that our proposed model produced the highest accuracy while using RFBM and Relief feature selection methods (99.05%).

INDEX TERMS Heart disease, machine learning, CVD, relief feature selection, LASSO feature selection, decision tree, random forest, K-nearest neighbors, AdaBoost, and gradient boosting.

I. INTRODUCTION

Cardiovascular disease has been regarded as the most severe and lethal disease in humans. The increased rate of cardiovascular diseases with a high mortality rate is causing significant risk and burden to the healthcare systems worldwide. Cardiovascular diseases are more seen in men than in women particularly in middle or old age [1], [2], although there are also children with similar health issues [3], [99]. According to data provided by the WHO, one-third of the deaths globally are caused by the heart disease. CVDs cause the death of approximately 17.9 million people every year worldwide and have a higher prevalence in Asia [4], [5]. The European Cardiology Society (ESC) reported that 26 million adults worldwide have been diagnosed with heart disease, and 3.6 million are identified each year. Roughly half of all patients diagnosed with Heart Disease die within just 1-2 years and about 3% of the total budget for health care is deployed on treating heart disease [6]. To predict heart disease multiple tests are required. Lack of expertise of medical

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Cusano.

staff may results in false predictions [7]. Early diagnosis can be difficult [8]. Surgical treatment of heart disease is challenging, particularly in developing countries which lack trained medical staff as well as testing equipment and other resources required for proper diagnosis and care of patients with heart problems [9]. An accurate evaluation of the risk of cardiac failure would help to prevent severe heart attacks and improve the safety of patients [10]. Machine learning algorithms can be effective in identifying the diseases, when trained on proper data [11]. Heart disease datasets are publicly available for the comparison of prediction models. The introduction of machine learning and artificial intelligence helps the researchers to design the best prediction model using the large databases which are available. Recent studies which focus on the heart-related issues in adults and children emphasized the need of reducing mortality related to CVDs. Since the available clinical datasets are inconsistent and redundant, proper preprocessing is a crucial step [12]. Selecting the significant features that can be used as the risk factors in prediction models is essential. Care should be taken to select the right combination of the features and the appropriate machine learning algorithms to develop accurate prediction models [13]. It is important to evaluate the effect of risk factors which meet the three criteria like the high prevalence in most populations; a significant impact on heart diseases independently; and they can be controlled or treated to reduce the risks [14]. Different researchers have included different risk factors or features while modelling the predictors for CVD. Features used in the development of CVD prediction models in different research works include age, sex, chest pain (cp), fasting blood sugar (FBS) – elevated FBS is linked to Diabetes [72], resting electrocardiographic results (Restecg), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope, number of major vessels coloured by fluoroscopy (ca), heart status (thal), maximum heart rate achieved (thalach), poor diet, family history, cholesterol (chol), high blood pressure, obesity, physical inactivity and alcohol intake [12], [15]–[19]. Recent studies reveal a need for a minimum of 14 attributes for making the prediction accurate and reliable [20]. Current researchers are finding it difficult to combine these features with the appropriate machine learning techniques to make an accurate prediction of heart disease [21]. Machine learning algorithms are most effective when they are trained on suitable datasets [22]–[25]. Since the algorithms rely on the consistency of the training and test data, the use of feature selection techniques such as data mining, Relief selection, and LASSO can help to prepare the data in order to provide a more accurate prediction. Once the relevant features are selected, classifiers and hybrid models can be applied to predict the chances of disease occurrence. Researcher have applied different techniques to develop classifiers and hybrid models [12], [20]. There are still a number of issues which may prevent accurate prediction of heart disease, like limited medical datasets, feature selection, ML algorithm applications, and a lack of in depth analysis. Our research aims to address some of these research gaps to develop a better model for CVD prediction. In this research, five datasets are combined, increasing the total size of the dataset. Two selection techniques, Relief and LASSO are utilized to extract the most relevant features based on the rank values in medical references. This also helps to deal with overfitting and underfitting problems of machine learning.

In this study, various supervised models such as AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB), K-Nearest Neighbors (KNN), and Random Forest (RF) together with hybrid classifiers are applied. Results are compared with existing studies.

The flow of the paper is as follows: Section II describes the aim and scope of this research. Section III provides an overview of related literature on the prediction of heart disease with various classifiers and hybrid approaches. Subsequently, section IV details out the proposed system and various performance metrics. The process of the data preparation, preprocessing and hybrid algorithms, Bagging and Boosting methods, are explained in section V. Section VI describes the implementation of the system and the results. Discussion on the statistical significance of the results, runtime and computational complexity and hyper-parameter tuning have been covered between section VIII and X respectively. Some recommendations for future works and conclusion are in section XII with a brief discussion on the limitations of the proposition in section XI.

II. RESEARCH AIM AND SCOPE OF THE PAPER

The aim of this research is to develop an effective method to predict heart disease, in particular Coronary Artery Disease or Coronary Heart Disease, as accurately as possible. Required steps can be summarized as follows:

1) Five datasets are combined to develop a larger and more reliable dataset.
2) Two selection techniques, Relief and LASSO, are utilized to extract the most relevant features based on rank values in medical references. This also helps to deal with overfitting and underfitting problems of machine learning.
3) Additionally, various hybrid approaches, including Bagging and Boosting, are implemented to improve the testing rate and reduce the execution time.
4) The performance of the different models is evaluated based on the overall results with All, Relief, and LASSO selected features.

III. LITERATURE REVIEW

The application of artificial intelligence and machine learning algorithms has gained much popularity in recent years due to the improved accuracy and efficiency of making predictions [25]. The importance of research in this area lies in the possibility to develop and select models with the highest accuracy and efficiency [26]. Hybrid models which integrate different machine learning models with information systems (major factors) are a promising approach for disease

prediction [27]. Various available public data sets are applied. In the study of Latha and Jeeva [28] ensemble technique was applied for improved prediction accuracy. Using bagging and boosting techniques, the accuracy of weak classifiers was increased, and the performance for risk identification of heart disease was considered satisfactory. They used the majority voting of Naïve Bayes, Bayes Net, C 4.5, Multilayer Perceptron, PART and Random Forest (RF) classifiers in their study for the hybrid model development. An accuracy of 85.48% was achieved with the designed model. More recently [29] machine learning and conventional techniques like RF, Support Vector Machine (SVM), and learning models were tested on the UCI Heart Disease dataset. The accuracy was improved by the voting-based model, together with multiple classifiers. The study showed that for the anemic classifiers, an improvement of 2.1% was achieved. In the study of NK. Kumar and Sikamani [30], different machine learning classification techniques were used to predict chronic disease. In their study, the Hoeffding classifier achieved an accuracy of 88.56% of in CVD prediction.

Ashraf *et al.* [15] used both the individual learning algorithms and ensemble approaches like Bayes Net, J48, KNN, multilayer perceptron, Naïve Bayes, random tree, and random forest for prediction purposes. Of these, J48 had an accuracy of 70.77%. They subsequently employed new-fangled techniques of which KERAS obtained an 80% accuracy. A multitask (MT) recurrent neural network was proposed to predict the onset of Cardiovascular disease with the attention mechanism at work [16]. The proposed model benefits by an Area under Curve (AUC) increase between 2 and 6%.

In the study of Amin *et al*. [12] the critical risk factors identified, machine learning models were applied (k-NN, DT, NB, LR, SVM, Neural Network, and a hybrid of voting with NB and LR) and a comparative analysis was performed. The outcome of their study indicates that the hybrid model, together with the selected attributes achieved an accuracy of 87.41%. The mean Fisher score feature selection algorithm (MFSFSA) together with the SVM classification model was used in the technique proposed by Saqlain *et al.* [31]. By using a SVM they obtained the selected feature subset and they used a validation process for MCC calculation. The features were selected based on a higher than average Fisher score. The combination of MFSFSA and SVM resulted in 81.19% accuracy, a 72.92% sensitivity, and an 88.68% specificity.

In the research work of Mienye *et al.* [22] prediction model for heart disease was proposed which involves the mean based splitting method, classification, and regression tree were used for randomly partitioning the dataset into smaller subsets. Afterwards, using an accuracy based weighted classifier ensemble, a homogenous ensemble was generated with the classification accuracies of 93% and 91% on the Cleveland and Framingham test sets. Two-tier ensemble-based coronary disease (CHD) detection model [24] was proposed in the study of Tama *et al*. Three different ensemble learners: random forest, gradient boosting machine, and extreme gradient boosting machine were used. The proposed model provides accuracy, F1, and AUC values of 98.13%, 96.6%, and 98.7%, respectively which exceeded other existing CHD detection methods.

A novel prediction model was introduced in the paper of Mohan *et al*. [32] with different combinations of features and several known classification techniques. An ANN with backpropagation and 13 clinical features as the input was used in the proposed HRFLM. DT, NN, SVM, and KNN were considered while making use of the data mining methods. SVM was useful for enhanced accuracy in disease prediction. The novel method Vote, in conjunction with a hybrid approach using LR and NB was proposed. An accuracy of 88.7% was obtained with the HRFLM method.

An improved random survival forest (iRSF) with high accuracy was used for the development of a comprehensive risk model in predicting heart failure mortality [33]. iRSF could discriminate between survivors and non-survivors using the novel split rule and the stop criteria. Patient demographics, clinical, laboratory information and medications were included in the 32 risk factors for the development of predictors. A data mining approach to detect cardiovascular has also been applied [34]. The Decision Tree, Bayesian classifiers, neural networks, Association law, SVM, and KNN data mining algorithms were used to detect the heart diseases. SVM resulted in an accuracy of 99.3%.

In works related to the prediction of patient survival [35], several machine learning classifiers were utilized. Feature relating to the significant risk factors were ranked and a comparison was performed between the traditional biostatistics tests and the provided machine learning algorithms. The result was that serum creatinine and ejection fraction were demonstrated to be the two most relevant features for accurate predictions. A model for CVD detection was developed with the AL Algorithm [36]. The dataset preparation and investigation was done with four algorithms. The precision was 99.83% for Decision Tree, and Random Forest methods and 85.32% and 84.49% respectively for SVM and KNN. Congestive heart failure (CHF) was effectively predicted using the ensemble method in another study [37] by analyzing the Heart rate variability (HRV) and using deep neural networks to solve the gap in related fields. The accuracy of the proposed system was 99.85%.

Yadav and Pal [3] used the UCI repository for their study. This dataset contains 14 attributes. The classification was carried out by four tree-based classification algorithms: M5P, random Tree, and Reduced Error Pruning and the Random forest ensemble method. The Pearson Correlation, Recursive Features Elimination, and Lasso Regularization were the three feature-based algorithms used in this work. The methods were then compared for accuracy and precision. The last method achieved the best performance. In recent work [38], Gupta *et al.* utilized the factor analysis of mixed data (FAMD) and RF-based MLA for developing a machine intelligence framework. RF was used for the prediction of disease by finding the relevant features using the FAMD. The proposed

method achieved a 93.44% accuracy, an 89.28% sensitivity and a 96.96% specificity.

Rashmi *et al.* [40] experimented on 303, a dataset that was extracted from the Cleveland dataset. The proposed algorithm, Decision Tree obtained 75.55% accuracy. Dinesh *et al.* [41] examined 920 datasets (Cleveland, Long Beach VA, Switzerland, and Hungarian) which from the UCI machine learning repository. Random forest achieved 80.89% accuracy; on the other hand, Saqlain has received 68.6% accuracy over the AFIC dataset [49]. Sharma *et al.* [43] and Dwivedi *et al.* [50] have applied the K-Nearest Neighbors algorithm to the same dataset. The results were 90.16% and 80% respectively. An accuracy of 46% was recorded by Enriko [48] when using the Kita Hospital Jakarta (450) dataset. An improved result was obtained, for instance 56.13%, using AdaBoost on the Cleveland dataset by Kaur *et al.* [51]. Shetty *et al.* [45] achieve 89% accuracy using the 270 datasets from the Statlog dataset, and Chaurasia *et al.* [39] have been used the same with a Boosting hybrid approach resulting in an accuracy of 75.9%. The UCI laboratory dataset was also used to evaluate the performance of the Boosting ensemble technique. Cheng *et al.* and Chaurasia *et al.* obtained accuracy of 82.5% by ANN model [46] and 78.88% [39] accuracy using a hybrid model. Using the Gradient Boosting technique, Dinesh *et al.* [41] obtained 84.27% accuracy using a combination of 4 different datasets where Bhuvaneeswari *et al.* [53] achieved 95.19% accuracy using 583 records from the Cleveland and Statlog dataset. A survey result has been generated on Rajaie cardio vascular medical dataset [44] using the hybrid approach, resulting in a 79.54% accuracy. On the other hand, the Bagging approach of Decision Tree [52] achieved more than 85.03% accuracy. Three different datasets were converted into one to obtain a more accurate result. A hybrid approach, achieved an accuracy of 88.4% by Mohan *et al.* [42]. Latha *et al.* [39] used 303 datasets of Cleveland heart disease by Bagging approach and gained 80.53% accuracy. Tan *et al.* [47] experimented on 303 datasets which were collected from Cleveland Heart disease dataset by hybrid approach and obtained 84.07% accuracy, while Latha *et al.* [39] achieved 85.48%.

Various techniques have been implemented on data of cardiovascular disease patients. Data are processed such that the K-Nearest Neighbors algorithm handles the missing data. The feature selection process is done following the Relief and LASSO. Various machine learning algorithms are implanted using the Bagging and Boosting approaches. One of the goals of the proposed approach is to analyze the accuracy and error rates of the algorithms in order to determine the best features.

IV. RESEARCH METHODOLOGY

An overall explanation is explained to build an intelligent machine learning system over the dataset of chronic heart disease.

A. OVERVIEW OF THE PROPOSED MODEL

Dataset is constructed by combining five different datasets (Cleveland, Hungary, Switzerland, and VA Long Beach and Statlog). This is included in the framework. Fig. 1 illustrates the workflow of recommended models. During data preprocessing, the combined dataset is analyzed to check for missing values which are then dealt with by the K-Nearest Neighbors imputation technique. To overcome overfitting issues and avoid long execution times, two different feature selection techniques are utilized: Relief and LASSO. This assists in extracting the best features. Performance of classifiers with the features selected by these techniques as well as with the original features is analyzed. After feature selection, the dataset is split into two parts: training and testing. Based on model learning rates, 80% of data is assigned for the training phase, and the remaining 20% d for the testing phase. All ensemble models with classifiers are implemented to make a comparison over the combined dataset; however, the generated outcome of our model is gained within a short period. Different training model has been given for testing the dataset so that we can pick the best model for our reliable dataset. The process resulted in RFBM being the most useful with 99.05% of accuracy. Furthermore, the most suitable features of a patient having affected by heart disease have been suggested in this diagnosis system.

B. PERFORMANCE MEASURE INDICES

The effectiveness and accuracy of the machine learning method can be evaluated using performance indicators. Positive classification occurs when a person is classified as having HD. When a person is not classified as having HD, he has a negative classification. The following formula from (1) to (7) has been applied to get all of this [54], [55].

TP = True Positive (when the model correctly Identified as having HD).

TN = True Negative (when the model correctly identified the opposite class, such as patients truly having no heart issues).

FP = False Positive (when the model incorrectly identified HD patients i.e., identifying non-HD patients as HD patients)

FN = False Negative (when the model incorrectly identified the opposite class, such as HD patients as normal patients).

$$\text{Accuracy (Acc)} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \quad (2)$$

$$\text{Recall or Sensitivity (Sen)} = \frac{(\text{TP})}{(\text{TP} + \text{FN})} \quad (3)$$

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

$$\text{False Positive Rate} = \frac{FP}{\text{FP} + \text{TN}} \quad (5)$$

$$\text{False Negative Rate} = \frac{\text{FN}}{(\text{TP} + \text{FN})} \quad (6)$$

$$\text{Negative predictive value} = \frac{\text{TN}}{(\text{TN} + \text{FN})} \quad (7)$$

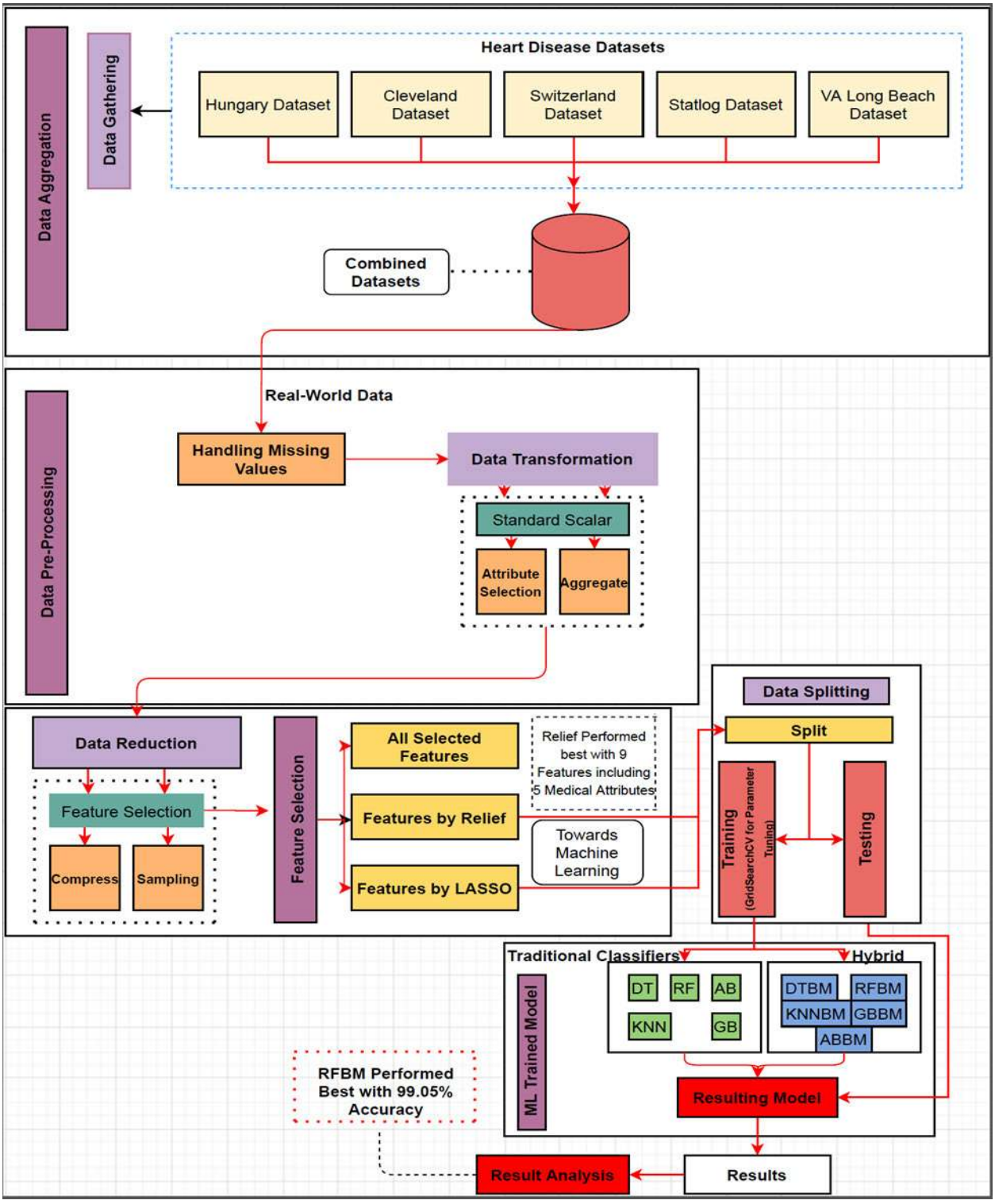

FIGURE 1. Working diagram of proposed model.

C. APPLICATION OF THE PROPOSED MODEL

Having a suitable application of the proposed model is key to the development of this unique system and will also help to deal with the real world challenges. The process has been illustrated in this section.

Fig. 2 picturises how a community health center can put the system to use, the following steps describes the procedures.

- Step 1: Reports are uploaded into the database.
- Step 2: Attributes are selected from the uploaded data to create input for the trained RFBM model.

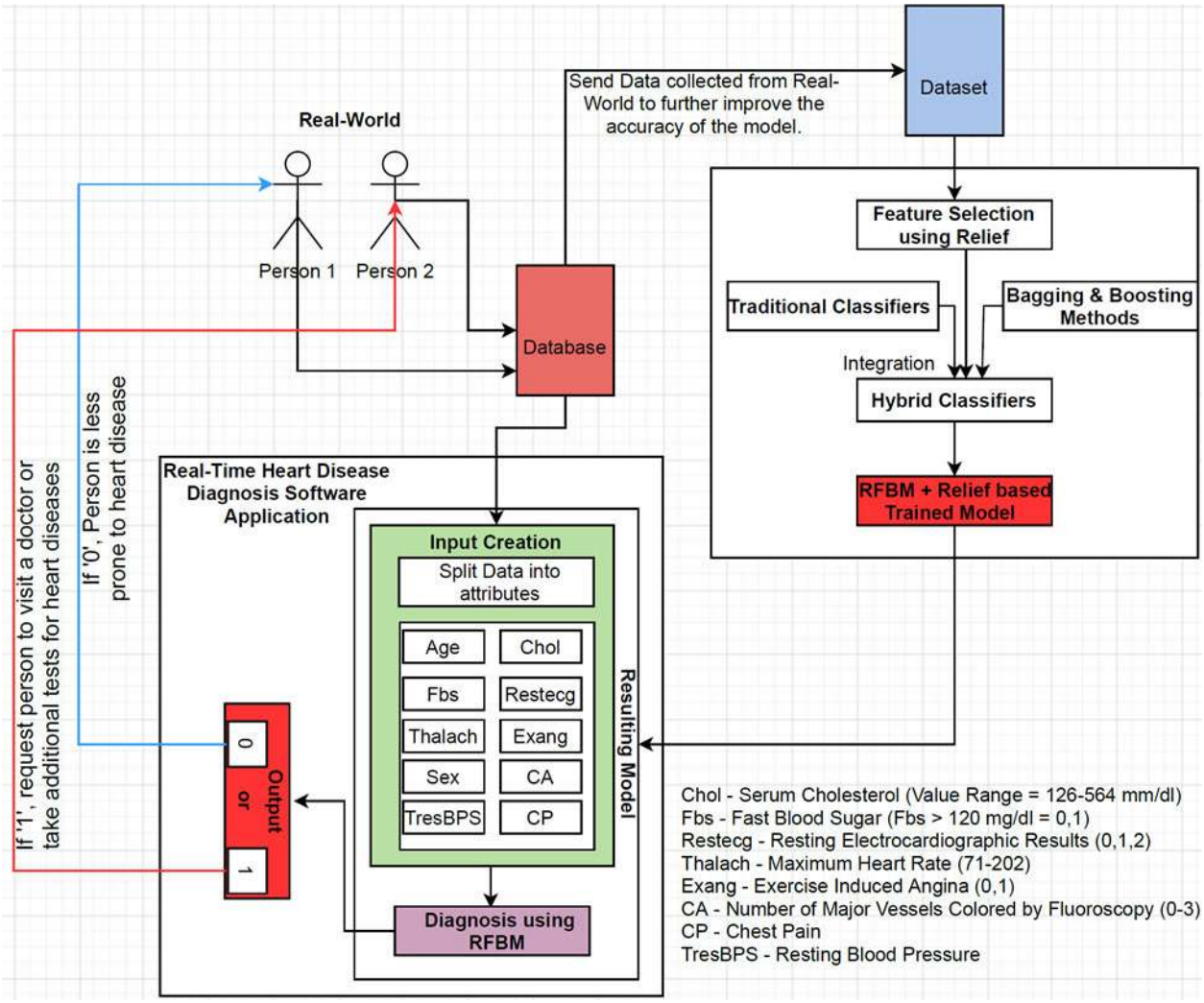

FIGURE 2. Suitable application of proposed model.

- Step 3: Selected attributes are processed in the trained model.
- Step 4: Output is generated in terms of 0 and 1.
  - 0 = A person is less prone to CVDs.
  - 1 = A person is prone to CVDs.
- Step 5: If '1', notify or request the person to consult a doctor or take additional tests.
- Step 6: Data uploaded to database is used to create trained model, to further improve the accuracy of hybrid classifiers and trained model.

D. JUSTIFICATION OF THE PROPOSED TECHNIQUE

This intelligent system has been developed based on the five classifiers. Subsequently, we used ensemble technique such as bagging and boosting to retain those algorithms as a base classifier. Numerous studies have already been conducted on different types of machine learning algorithms. Among them, we picked three most common techniques (DT, RF AND KNN) and two less common techniques (AB and GB). Some of the previous studies have actually shown that the predicted accuracy of DT [1], RF [1], [2] and KNN [3] algorithms were quite high compared to other existing techniques. Additionally, a limited number of studies also demonstrated AB [5], [6] as well as GB [53] can perform rather well with considerably high Accuracy. Our paper highlights some of the notable research attempts that deployed Bagging and Boosting ensemble techniques as well as proposed some hybrid frameworks, however, none of those research attempts closely resembled our introduced approaches as a base classifier except DT [8] and kNN [7]. As a consequence, in this work, all of those previous approaches have been further explored with the help of ensemble techniques to make the proposed model more efficient. Although from Literature Review it can be seen that propositions put forward in [1], [5], [24] and [27] yielded promising predictive accuracy, but was not high enough in comparison to our work.

Basically, we felt the need to improve the current studies in this field and analyzed previous models to determine what

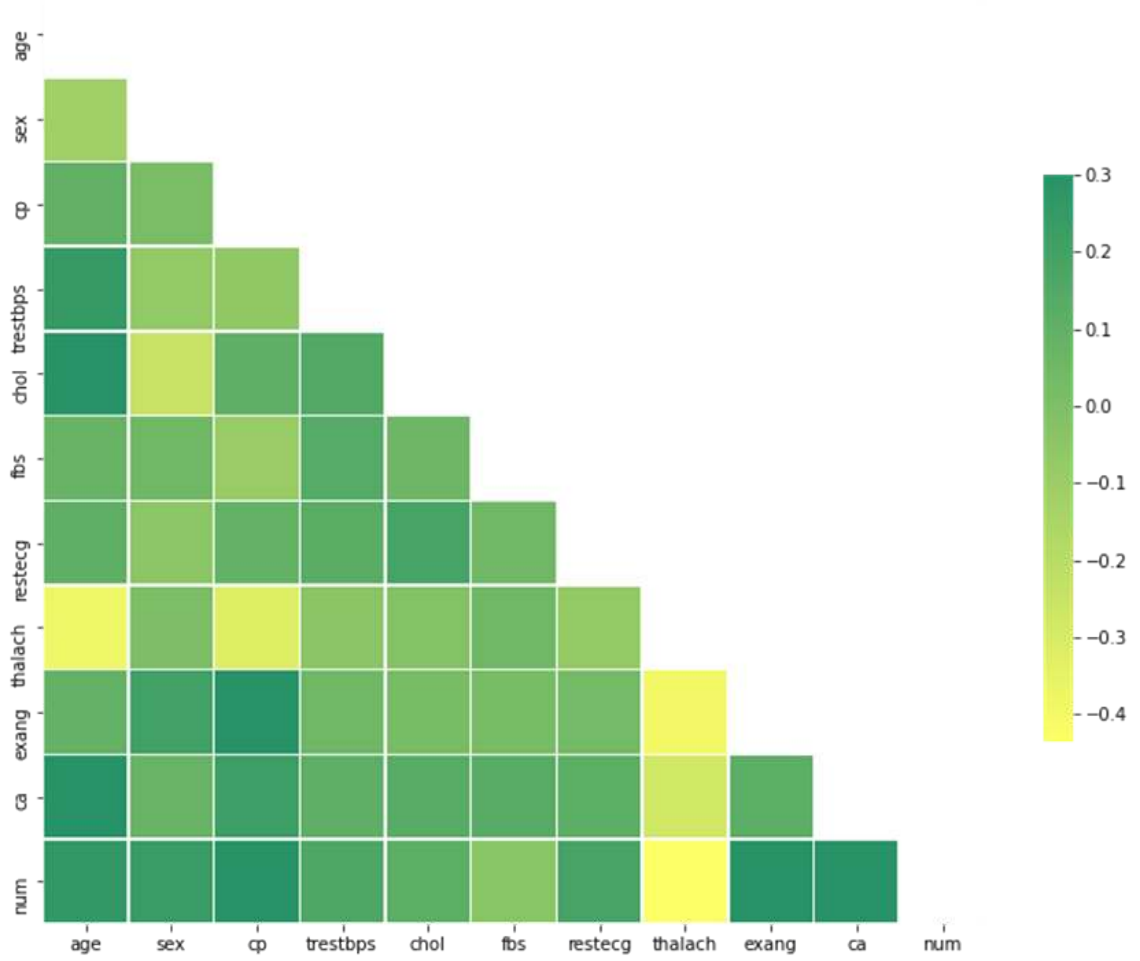

FIGURE 3. Highly correlated features of Relief approach.

might be lacking, after which we took the initiative to devise a solution that might reshape the current ideas and provide an acceptable level of results that makes the system suitable for practical implementation.

As has been discussed before, previous works that are somewhat related to this study and deal with the datasets used here are available, however, the performance of those systems were not as expected in most cases.

We believe one reason for the lack of performance of some systems is the inability of those systems to identify the most important and highly correlated features. We want to develop a method that will first identify the optimal group of features and then identify the algorithms that works best with those features.

In our understanding, algorithms that performed well benefitted from the tightly correlated feature-set, mainly derived from the use of Relief, whereas the algorithms that did not show strong performance, could not properly evaluate the correlative structure among the features used.

The following figure has been depicted based on the highly correlated 10 features with predicted attribute (num) which are selected by Relief feature selection technique. On the right side, the attribute values are shown (from 0.3 to −0.4). From Fig. 3, it is clearly seen that ca, chol and trestbps features have strong relationship with age where the value was approximately 0.3, on the other hand, the lowest correlation was observed for thalach that was about −0.4. Similarly, cp shows a significant correlation with exang. However, the correlated values among other features were not so high and fluctuated between 0.15 and −0.3.

V. IMPLEMENTATION

A. DIFFERENT MACHINE LEARNING LIBRARIES

The implemented model is written in Jupiter notebook's Python programming language using simple libraries like Panda [56], Pyplot [57] and Scikit-learn [58].

B. DATASET

Data is considered the first and most basic aspects of using machine learning techniques to get accurate results. The applied dataset is gathered from a well-known data repository, the 'UCI machine learning repository'. There are five different datasets: the Cleveland, Hungary, Switzerland, VA Long

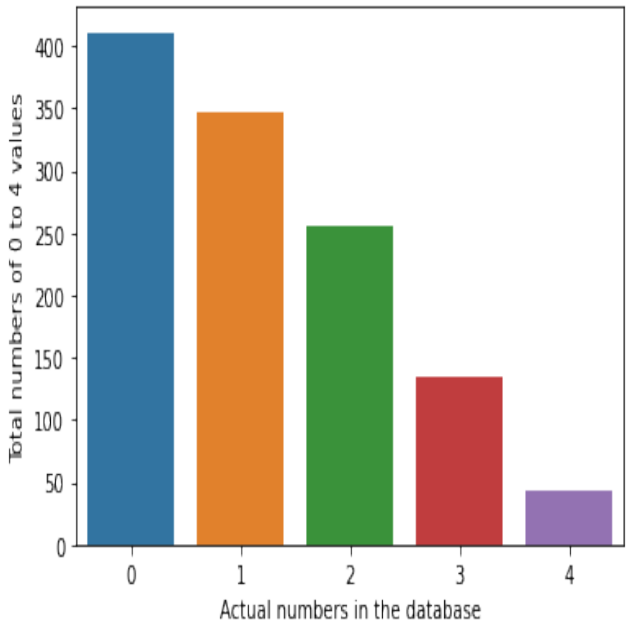

FIGURE 4. Actual data points in the datasets.

TABLE 1. Value range in dataset.

No.	Attributes	Data Types	Description	Value Range
1	age	Integer	Age in years	29 to 79
2	sex	Integer	Gender instance	0 and 1
3	cp	Integer	Chest pain type	1, 2, 3, and 4
4	trestbps	Integer	Resting blood pressure in mm Hg	94 to 200
5	chol	Integer	Serum cholesterol in mg/dl[	126 to 564
6	fbs	Integer	Fasting blood sugar > 120 mg/dl	0, 1
7	restecg	Integer	Resting ECG results	0, 1, and 2
8	thalach	Integer	Maximum heart rate achieved	71 to 202
9	exang	Integer	Exercise induced angina	0, 1
10	oldpeak	Real	ST depression induced by exercise relative to rest	1 to 3
11	slope	Integer	Slope of the peak exercise ST segment	1,2, 3
12	ca	Integer	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	Integer	Defect types	3,6,7
14	num	Integer	Diagnosis of heart disease	0, 1, 2, 3, and 4

Beach [59], and Statlog heart disease dataset [60]. We have combined all of them in this research to obtain more accurate outcomes. More than 1190 cases are collected as a text file along with 14 special features from their database. 13 attributes of these combined datasets are taken as diagnosis inputs, whereas the 'num' attribute is selected as output. Six features which are considered relevant in medical literature were present in all or most records: age in years (age), sex (sex), resting blood pressure (trestbps), fasting blood sugar (fbs), chest pain type (cp), and resting electrocardiographic results (restecg). Table 1 describes the different attributes and the range of values.

The value of the 'num' attribute can be 0, 1, 2, 3 or 4. The predicted value '0' represents that a patient does not have heart disease and the values from 1 to 4 reflect the various stages of chronic heart disease.

An overview of the total number of patients for each value of the num attribute in the combined dataset is shown in Fig. 4.

Since for the purpose of this research is to predict whether or not a patient is suffering from heart disease, we convert all values in the range of 1 to 4 to a 1. This means that the attribute now has the range of (0, 1).

C. AN OVERVIEW OF DATA PREPROCESSING AND CLEANING TECHNIQUES

There is a large amount of collected data in the modern world that can be gathered via the internet, surveys, and experiments, etc. Often the data to be used contain missing values, noise, and distortions, however. The combined dataset used for this research also contains missing or null values. There are some popular techniques, such as imputation and deletion that can be used to deal with missing values. In our dataset, this problem is resolved by using the K-Nearest Neighbors [62] imputation method. Before machine learning algorithms can be applied, data must also need to be normalized or standardized. Standardization converts the data to a mean of 0 ($\boldsymbol{\mu}$) and a standard deviation ($\sum$) of 1. The conversion formula of (8) is given below [63]:

$$Standardization, X = (X - \mu)/\sigma \tag{8}$$

D. FEATURE SELECTION TECHNIQUES

Feature selection techniques are important for the machine learning procedure as the best attributes for classification need to be extracted. This also helps to reduce the execution time. We have selected two algorithms: Relief feature selection and the Least Absolute Shrinkage and Selection Operator.

1) RELIEF FEATURE SELECTION TECHNIQUE

Relief is a selection attribute algorithm that gives a weight to all the features in the dataset. These weights can then be modified gradually [64]. The aim is to ensure that the important features have a large and that the remaining features have low weights. Relief uses the similar techniques as in KNN to determine feature weights. This well – known algorithm of feature selection approaches has been shown by Kira and Rendell [65]. $\boldsymbol{R_i}$ is for a randomly selected instance. Relief searches for its two nearest neighbours: one from the same class, called closest hit $\boldsymbol{H}$, and one from the opposite class, called closest miss $\boldsymbol{M}$. It adjusts the consistency calculation $\boldsymbol{W\,[A]}$ for feature $\boldsymbol{A}$ according to the $\boldsymbol{R_i}$, $\boldsymbol{M}$, and $\boldsymbol{H}$ values. If there is a large difference between $\boldsymbol{R_i}$ and $\boldsymbol{H}$ occur this is not desirable, so the performance value $\boldsymbol{W\,[A]}$ is reduced. On the

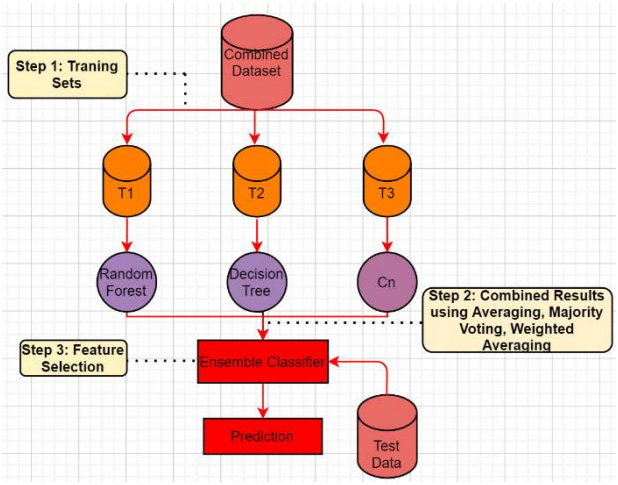

FIGURE 5. The working techniques of ensemble process.

other hand if there is a large difference between R_i and M for attribute A then A may be used to distinguish different classes, so the weight $W[A]$ is increased. This process will is continued form times where m is a parameter that can be adjusted.

2) LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR ALGORITHM (LASSO)

The minimum selection and shrinkage functionality of this operator depends on modifying the absolute value of the coefficient of functions. Some coefficient values of the features are zero, and features with negative coefficients can also be removed from the subset of features. The LASSO has a very good performance for feature values with small coefficients. Features which have large coefficient values will be available in the chosen subsets of features. Unnecessary features can be found with LASSO [66]. Moreover, the reliability of this feature can be enhanced by repeating the above procedure many times eventually taking the most frequently found features in as the most important ones. This is called the randomized LASSO feature, which was introduced by Meinshausen and Buhlmann, in 2010 and Wang in 2011 [67]. It should be implemented on a powerful computer as it uses parallel programming. It also demonstrates its realization for the present application, where q^{-i} represents the vector of the related i^{th} sub-region keys.

E. ENSEMBLE METHODS OF MACHINE LEARNING

Ensemble techniques mix multiple classifiers of a Decision Tree to achieve better classification results than only one Decision Tree classifier. The core idea behind the ensemble method is that a combination of weak learners can work together to f a strong learner, thus improving the model's accuracy and precision [39]. Fig. 5 depicts the ensemble process [39]. When we seek to identify the target feature using any machine learning method, key reasons for in the difference real and identified outcomes are noise, uncertainty, and bias. Ensemble techniques assist in dealing some of these

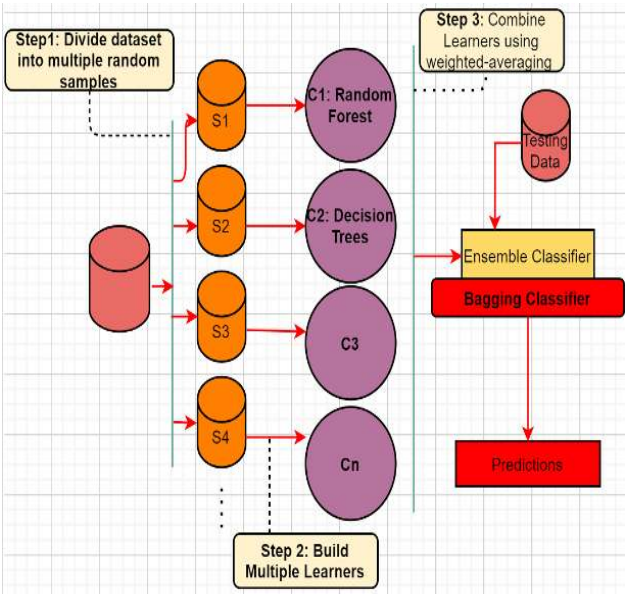

FIGURE 6. Bagging method.

variables, particularly uncertainty and bias. In this study, we apply two ensemble techniques: Bagging and Boosting to obtain more accurate results. These techniques are explained below.

1) BAGGING TECHNIQUE

Bagging is used when the goal is to reduce the variance of Decision Tree classifiers. The objective is to create several subsets of data from the training samples. [68] Randomly chosen collections of subset data are used to train their Decision Tree. As a result, we get an ensemble of different models. The average of all predictions from different trees is then used. This is more robust than a single Decision Tree classifier. It helps not only to reduce the overfitting problem but also to handle higher dimensionality data properly. It resolves missing data issues and maintains accuracy. The process of the Bagging method is described in Pseudocode 1 and Fig. 6.

With the help of the Bagging technique, three ensemble hybrid models, based on DT, RF, and KNN, are constructed. The three hybrid models: DTBM, RFBM, and KNNBM are applied in both the training and the testing phase.

2) BOOSTING TECHNIQUE

Boosting is a repetitive process which depends on the last prediction and changes the weight. Fig. 7 are added to better understand the workflow.

If an instance is incorrectly classified its weight is increased. Usually, Boosting constructs good predictive models [69]. It generates different loss functions and works by combining the weak models to boost their performance. For this research, we have applied the Boosting technique on two classification algorithms: AB and GB to construct our hybrid models. The resulting ABBM and GBBM are applied in both the training and testing phases.

Pseudocode 1 Pseudocode for Bagging Method

BEGIN

1. Let D = {$d_1, d_2, d_3, \ldots d_n$} be the given dataset
2. E = { }, the set of ensemble classifiers
3. C = {$c_1, c_2, c_3, \ldots c_n$}, the set of classifiers
4. X = the training set, X D
5. Y = the test set, Y D
6. L = n(D)
7. for i = 1 to L do
8. S(i) = {Bootstrap sample I with replacement} I X
9. $M_{(i)}$ = Model trained using $C_{(i)}$ on $S_{(i)}$
10. E = E $C_{(i)}$
11. next I
12. for i = 1 to L
13. $R_{(i)}$ = Y classified by $E_{(i)}$
14. next i
15. Result = max($R_{(i)}$: i = 1, 2, ... , n)

END

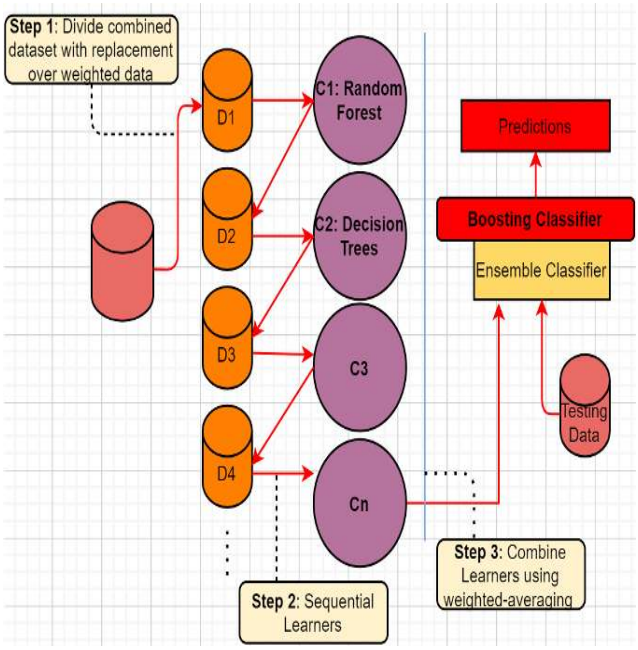

FIGURE 7. Boosting method.

F. PROPOSED APPROACH FOR THE CLASSIFICATION MODEL

This section discusses the machine learning approaches that are used in this research to generate an intelligent prediction system for heart disease.

1) DECISION TREE

The Decision Tree algorithm, which has only 2 numClasses, is one of the most powerful and well-known predictive instruments [70]. Every interior node in the structure of a Decision Tree refers to testing a property, every branch corresponds to a test outcome, and each leaf node is a separate class [71], [87].

'Learning' based on Decision Tree (DT) often applies an upside-down tree based progression technique. The algorithm is capable of resolving both classification and regression problems. The tree grows from the root node by determine a 'Best Feature' or 'Best Attribute' from the set of attributes available at hand, 'splitting' is then applied. Selection of the 'Best Attribute' is often carried out through the calculation of two other metric, 'Entropy' as shown in (9), and Information Gain, shown in (10). The 'best attribute' is the one that provides the most useful information. Entropy indicates how homogeneous the dataset is and Information Gain is the rate of increase or decrease in Entropy of attributes [100].

$$\begin{aligned} \text{E (D)} = &-\text{P (positive) log2 P (positive)} \\ &-\text{P (negative) log2 P (negative)} \end{aligned} \quad (9)$$

Equation (9) calculates the Entropy E, of a dataset D, which holds the positive and negative 'Decision Attributes'.

$$\begin{aligned} \textit{Gain (Attribute X)} = &\textit{Entropy (Decision Attribute Y)} \\ &-\textit{Entropy}(X, Y) \end{aligned} \quad (10)$$

Non-parametrically supervised learning methods, such as C4.5 are used for classification and regression. This aim of the method is to develop a model that predicts the value of the dependent variable by studying basic rules for decision making.

Baihaqi *et al.* [73] applied the C4.5 classifier to diagnose CAD using and obtained 78.95% accuracy. However, the classifier C4.5 usually does not allow small datasets. The RF classifier (describer below) may perform better [74], for heart disease detection or alternatively the combining strategy using bagged decision trees [75].

2) RANDOM FOREST

The Random Forest (RF) classifier is an ensemble algorithm [76]. This implies that it consists of more than one algorithm. Usually In this case, it consists of several DT algorithms [77]. RF build up an entire forest from several uncorrelated and random Decision Trees during training segment [101]. Ensemble learning methods employ multiple learning algorithms to generate an optimal predictive model, which can provide better results than any of the individual model's prediction [101]. Computational complexity may increase as RF uses more features than a standalone DT, but it generally has a higher accuracy when dealing with unseen datasets. The result of the Random Forest algorithm is the mean result of the total number of Decision Tree algorithms. Illustration. Fig. 8 gives and graphical description of Random Forest [87].

The Random Forest ensemble classifier builds and integrates multiple decision trees to get the best result. It primarily refers to tree learning through aggregating bootstraps. Let the provided data be $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots\ldots, \boldsymbol{x_n}\}$ with responses $\boldsymbol{Y} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots\ldots, \boldsymbol{x_n}\}$ with a lower limit of $\boldsymbol{b} = \boldsymbol{1}$ and an upper limit of $\boldsymbol{B}$: The prediction for sample $\boldsymbol{x'}$ is made by averaging the predictions $\sum_{b=1}^{B} \boldsymbol{f_b}(\boldsymbol{x'})$ from every

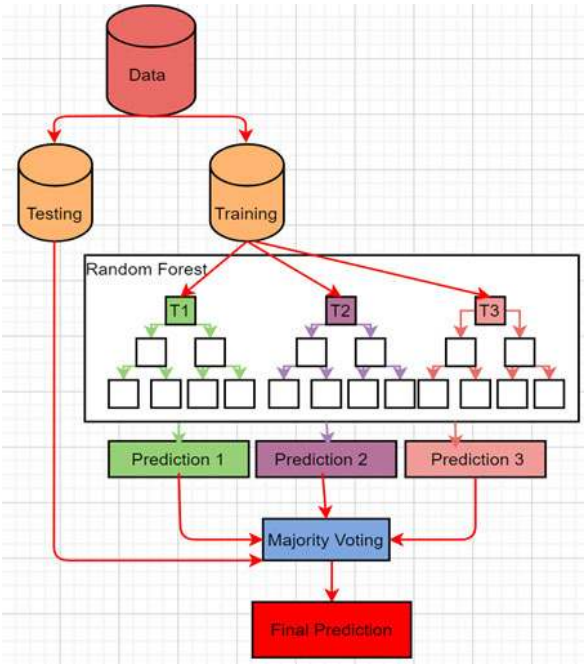

FIGURE 8. Random Forest algorithm.

individual trees for x' that is shown using (11).

$$j = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{11}$$

The Random forest (RF) classifier, a combination of many different tree predictors, is often used for the analysis of big data. It is a learning method for grouping, regression, and other functions in an ensemble.

Banerjee *et al*. [79] used successfully applied the RF classifier using time-frequency characteristics from PCG signals to identify heart disease.

3) K-NEAREST NEIGHBORS

K-Nearest Neighbors (n_neighbors = 5) is amongst the most common classification technique in the field of machine learning. It has previously been used for coronary artery disease. KNN is considered nonparametric since the method does not use data distribution assumptions. KNN considers the equivalence of the new data and the existing data and places the new data in the class, which is nearest to the existing classes. KNN is used for regression problems as well as for recognition problems. It is also known as the lazy learner algorithm [80] as it does not immediately learn from a collection of training data. KNN calculates the Euclidean distance between new $\boldsymbol{A}\ (\boldsymbol{x_1}, \boldsymbol{y_1})$ data and previously accessible $\boldsymbol{B}(\boldsymbol{x_2}, \boldsymbol{y_2})$ data, using the equation (12) [81].

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{12}$$

The Euclidean formula may be used to evaluate the distance between two data points $(\boldsymbol{x_2}, \boldsymbol{x_1})$ and $(\boldsymbol{y_2}, \boldsymbol{y_1})$ in two-dimensional space. KNN puts the new data into the class which has the least Euclidean distance to the new data.

Previous research [82] has used KNN as an automated classification technique for coronary artery disease. When conducting linear discriminant analysis KNN had a better accuracy than SVM and NN [85]. Rajkumar and Reena obtained an accuracy of just 45.67% [83] using KNN to diagnose CAD. However, Gilani *et al.* [84] subsequently compared the F1 score with many classification models and found that the KNN classifier performed best among the seven classifiers. A limitation of the method is that due to the high computational complexity, KNN is not appropriate for implementation in a low power or a real-time environment.

On a different note, in place of using Euclidean Distance, Suryawanshi and Sharma [102] have shown 'Spearman Correlation' [103] can also be employed as the distance measure for KNN based classification as shown in (13). P and Q are training and testing tuple respectively while n is the number of total observations. The values of f_{ij} usually lies between 1 and −1.

$$f_{ij} = 1 - \frac{6\sum_{i=1}^{n}\left(\text{rank}\,(P_i) - \text{rank}\left(Q_j\right)\right)^2}{\text{n}\left(n^2 - 1\right)} \tag{13}$$

The changes have demonstrated some enhancements over regular KNN model with nearly 50% improvement in accuracy (97.44% in 80%-20% Train and Test ratio).

4) ADABOOST

AdaBoost or Adaptive Boosting is a Boosting algorithm that is used for binary classification and combines a number of weak classifiers to make a more robust classifier [86]. This algorithm produces the predicted accuracy based on 1000 samples. The training dataset instances are weighted with a starting weight [87] as shown in (14).

$$\text{Weight (xi)} = 1/\text{N} \tag{14}$$

where $\boldsymbol{N}$ is the frequency of training instances, and $\boldsymbol{x_i}$ is $\boldsymbol{i^{th}}$ training instance. The decision stump gives an output for each input variable. The misclassification rate is then calculated using equation (15).

$$\text{Error} = (\text{correct}-\text{N})/\text{N} \tag{15}$$

where $\boldsymbol{N}$ is the frequency of training instances. Boosting simply means combining several simple trainers to achieve a more accurate prediction. AdaBoost (Adaptive Boosting) fixes the weights which vary for both samples and classifiers [88]. This causes the classifiers to focus on results that are relatively difficult to identify accurately. The final classification formula is shown in equation (16).

$$H_k\,(p) = +/-\,(\sum_{k=1}^{k} a_k h_k\,(p)) \tag{16}$$

Equation (15) is a linear combination of all the weak classifiers (simple learners), where $\boldsymbol{K}$ is the total number of weak classifiers $\boldsymbol{h_k}\ \boldsymbol{(p)}$ is the output of weak classifier $\boldsymbol{t}$ (this can be either −1 or 1). $\boldsymbol{a_k}$ is the weight of classifier $\boldsymbol{k}$.

5) GRADIENT BOOSTING

Gradient Boosting is a Boosting algorithm that required only 100 samples, used for classification and regression problems [89]. Gradient Boosting consists primarily of three factors [90]: An enhanced loss function, a weak learner to make predictions, and an additive model to combine weak learners to minimize the loss function [91]. Gradient Boosting is a technique that can increase the algorithm's efficiency by eliminating overfitting.

The application of gradient tree Boosting to the Tobit model, called as the 'Grabit' model, helps to improve the accuracy when there is an imbalance between the numbers in each class. Boosting rather basis methods also known as regression tree learners, to obtain higher predictive precision on a large variety of datasets, e.g. [92], but it utilizes familiarity in a specific area. The distinction between Boosting process and traditional machine learning is that function space excludes optimization. The optimal function $\boldsymbol{F(X)}$ is obtained after iterations $\boldsymbol{m^{-th}}$ [93] that is derived as per (17):

$$\mathrm{F}(\mathrm{X}) = \sum_{i=0}^{m} f_i(\mathrm{x}) \tag{17}$$

where $\boldsymbol{f_i\ (x)\ (i = 1, 2\ldots., M)}$ indicates feature increments, the $\boldsymbol{f_i\ (x) = -\ \rho_i\ x\ gm(X).}$ The latest base-learner is the largest loss function correlated with negative gradients [94]. The negative gradient for the $\boldsymbol{m^{-th}}$ iteration is (18).

$$\mathrm{gm} = -[\frac{\partial L(\mathrm{y}, \mathrm{F}(\mathrm{X}))}{\partial \mathrm{F}(\mathrm{X})}]_{\mathrm{F(X)=Fm-1(X)}} \tag{18}$$

where $\boldsymbol{gm}$ is the path where the loss function decreases the most rapidly when $\boldsymbol{F(X) = F_m\ -1(X)}$ [93]. A new decision tree aims to correct the error made by its previous base learner. $\boldsymbol{T}$ model is then modified to (19).

$$F_m(\mathrm{X}) = F_m - 1(\mathrm{X}) + \rho_m \mathrm{x} h_m(\mathrm{X}, \alpha_m) \tag{19}$$

In this system, several classifiers with ensemble techniques including DTBM, RFBM, KNNBM, ABBM and GBBM have been applied to compare these algorithms. Using DT as a base class does not always help to get a higher accuracy. The highest accuracy using ensemble techniques was achieved by using RFBM in our prediction system.

VI. RESULTS AND DISCUSSION

A. OUTCOMES OF FEATURE SELECTION PROCESSES

Relief [95], a feature selection algorithm, selects main features based on the weight of the data. The most important seven input features selected by Relief are given in Table 2. The most important feature for predicting heart disease is serum cholesterol (chol) which rank score is 0.869 according to the findings.

The LASSO treats closely related features as true, and the rest as false. After applying the LASSO, chest pain (cp) had the highest rank score (0.0796), whereas maximum heart rate (thalach) had a very low score.

Table 3 shows the score of the eight most essential features selected by LASSO for diagnosing heart disease.

TABLE 2. Features selected by Relief algorithms and their rankings.

Feature name	Feature code	Score
Age in years	age	0.19
Serum cholesterol	chol	0.867
fasting blood sugar	fbs	0.0233
resting electrocardiographic results	restecg	0.582
maximum heart rate	thalach	0.543
exercise induced angina	exang	0.0089
number of major vessels (0-3) colored by fluoroscopy	ca	0.581

TABLE 3. Features selected by LASSO algorithms and their rankings.

Feature name	Feature code	Score
Age in years	age	0.0012
Chest pain type	cp	0.0796
Resting blood pressure	trestbps	0.0018
Serum cholesterol	chol	0.0000
Maximum heart rate	thalach	-0.0013
ST depression induced by exercise	oldpeak	0.0229
slope of the peak exercise ST segment	slope	0.0316
Thal	thal	0.0114

B. COMPARISON OF VARIOUS ALGORITHMS AND HYBRID APPROACHES ON THE DIFFERENT FEATURES

This section compares on the outcomes of the different classification models with the different input features. First, five machine learning classifiers and five hybrid techniques were applied to all features of heart disease dataset. Secondly, Least Absolute Shrinkage and Selection Operator Features Selection Algorithm (LASSO) was implemented to extract some relevant features and the same five machine learning classifiers and five hybrid techniques were applied again. Finally, the most important features selected by the Relief model were used as input to the classifiers and hybrid methods. Different performance metrics are also evaluated to evaluate the predicted outcomes.

Our original dataset contains 14 individual attributes in which 13 input functions are used to generate the outcome of the disease. From these 13 features, 6 significant features of our dataset, which matched prominent medical books and guides, these are, age in years (age), gender (sex), resting blood pressure (trestbps), fasting blood sugar (fbs), chest pain type (cp), and resting electrocardiographic results (restecg) [61], [97]. Some features including age, and sex are

not modifiable, while risk factors associated with other features (fbs, restecg, cp and trestbps) are gradually modifiable.

After applying the Relief feature selection algorithm to the proposed dataset, 7 features: age in years (age), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate (thalach), exercise induced angina (exang), and number of major vessels (0–3) colored by fluoroscopy (ca) have been selected based on their ranking values. Some missing attributes, present in notable medical books, were added: sex, trestbps [61], and cp [97] as it was felt that is was important that these features were included.

Eight relevant features: age, cp, trestbps, chol, thalach, oldpeak, slope, and thal were selected according to their ranking by the LASSO feature selection algorithm. Chest pain was the feature with the highest score. Some missing attributes, present in all medical records, were added: sex, fbs and restecg, so that these features were part of all three feature sets.

Different machine learning techniques were applied to the selected features. The 2 × 2 confusion matrix was generated to produce the different performance metrics and provided a comparison of all mentioned algorithms. The performance metrics Accuracy, Error rates, Sensitivity, Precision, F1-Score, Negative Predictive Value, False Positive Rate, and False Negative Rates were used to evaluate the proposed models.

1) COMPARISON BETWEEN DIFFERENT METHODS BASED ON ACCURACY

Accuracy is usually considered to be the most important techniques to evaluate machine learning algorithms. As mentioned above, we use five classifiers and five hybrid classifiers. We applied the ten different methods on the original 13 input features then on the eleven input features selected by the LASSO approach, and on the 10 features selected with the Relief method. Fig. 9 shows the accuracy of the different types of classifiers, including the five hybrid classifiers.

Considering 13 features, the most accurate prediction [98] is 89.07% was obtained from the AB Classifier, whereas the accuracy of KNN is 83.61%. The accuracy of DT and GB are very similar to each other (86.97%). However, results are significantly better for some of the hybrid classifiers: the accuracy of RFBM is 92.65%. When only evaluating 11 selected features (LASSO), the RF Classifier generates the lowest accuracy (86.97%). We get 88.6%, 93%, 90.75%, 92.85% accuracy for DT, KNN, AB, and GB classifiers respectively with the 11 LASSO features. GBBM has an outstanding performance of 97.85% and the other four hybrid classifiers DTBM, RFBM, KNNBM, and ABBM also provide a good accuracy: 88.65%, 97.65%, 96.6% and 90.75% respectively.

Looking at the accuracy of these ten strategies with the Relief features, the Random Forest Bagging method (RFBM), which is a hybrid classifier, demonstrated an excellent accuracy of 99.05%. The results of the hybrid models of DT, AB, and GB were similar to the previous results.

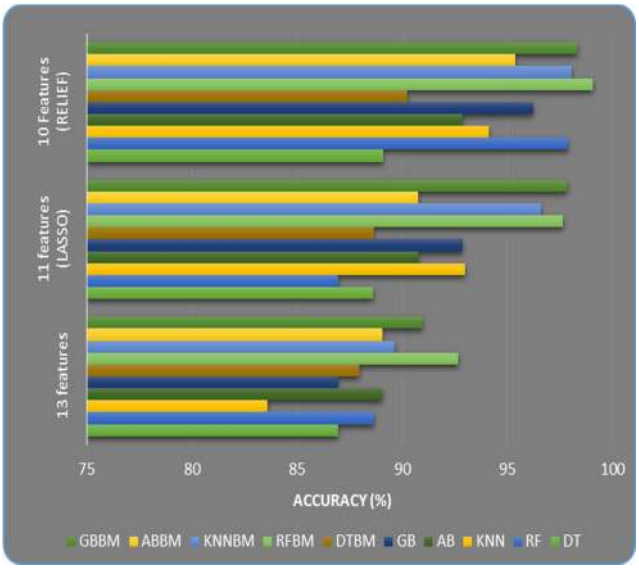

FIGURE 9. Accuracy.

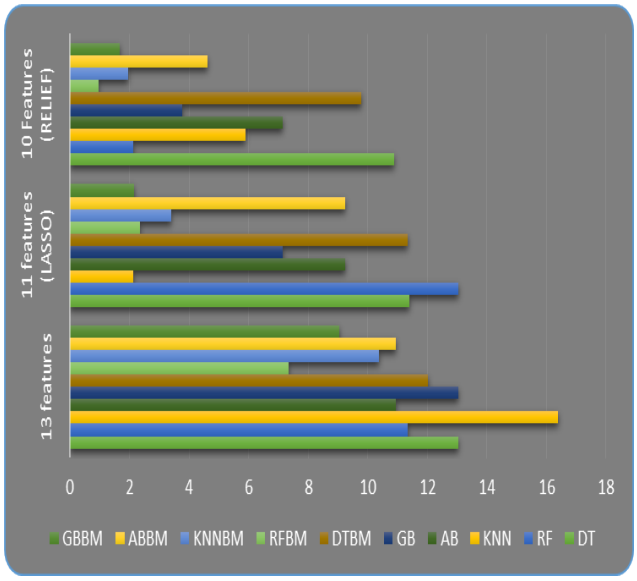

FIGURE 10. Error rates.

A dramatic improvement in accuracy with hybridization observed for the KNN model, from 94.11 % to more than 98 % accuracy.

2) COMPARISON BETWEEN DIFFERENT METHODS BASED ON ERROR RATES

Error rates also help to understand the model performance. The lowest error rate is generated by RFBM on the ten selected features by Relief, approximately 0.95%. However, for the eleven features selected by LASSO, the lowest error rate was obtain with KNN; just under 2.2%. Fig. 10 clearly shows that KNN had the highest error rate (16.39%) for 13 features, followed by RF for 11 features (13.03%) and DT for 10 feature (10.88%).

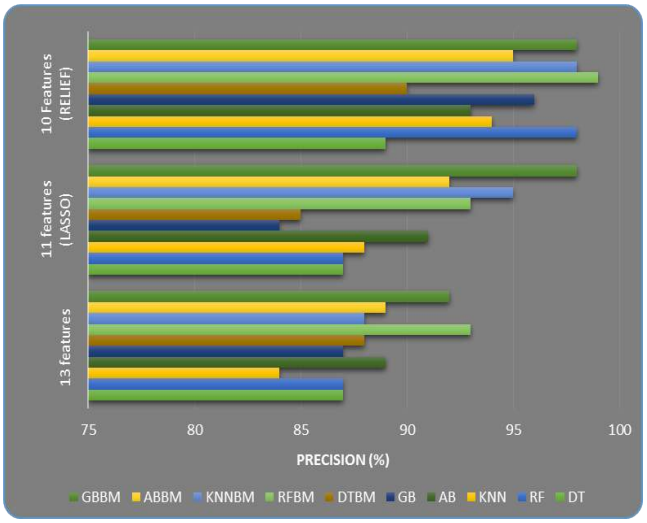

FIGURE 11. Precision.

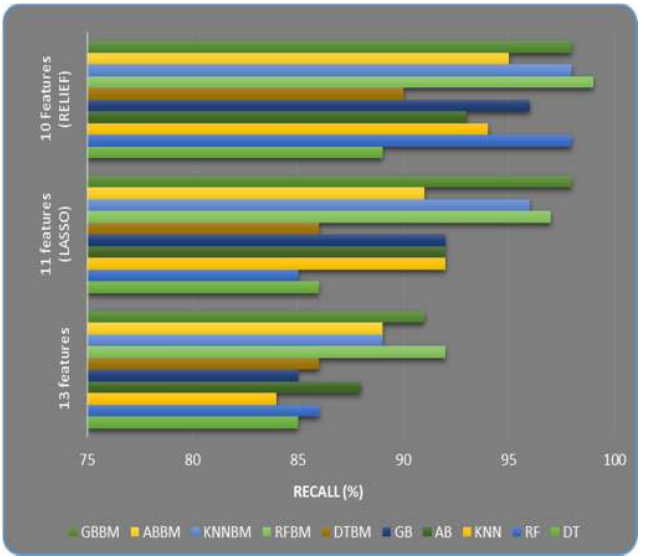

FIGURE 12. Obtained recall scores.

3) COMPARISON BETWEEN DIFFERENT METHODS BASED ON PRECISION

Other performance metrics such as precision have also been used to evaluate the performance of classifier and hybrid algorithms. Considering 13 input features, a noticeable result of over 93% was obtained for precision with the RFBM model KNN had the lowest precision score: 84%. Other models had precision scores between these values. When applied to the 11 LASSO features, the best precision was obtained with the GBBM (98%), and the lowest precision (84%) for the GB classifier. Both the Decision Tree (DT) and Random Forest (RF) classifiers achieved a precision score of approximately 87%. The best precision was obtained evaluating 10 Relief features by RFBM which was close to 99%. KNN also had a high precision score (94%). For the 10 Relief features, DT produced the lowest score but this was still 89%. The outcomes for precision are depicted in Fig. 11.

4) COMPARISON BETWEEN DIFFERENT METHODS BASED ON RECALL

Recall or sensitivity score is an important performance matrix as it is important that people with heart disease are accurately classified. Fig. 12 shows the recall scores for the different algorithms and feature sets. A very poor recall score (just over 84%) has been generated in was obtained with the KNN algorithm, while the RFBM achieved the highest recall score (92%) when applied to the original 13 features. ABBM, KNNBM, RF, and GBBM had recall scores of 89%, 89%, and 86%, and 91% respectively based on 13 features. For the 11 LASSO features, the RF algorithm had a low recall score (just over 85%) while RFBM and GBBM provided more satisfactory results over the 11 features. Similar recall scores of approximately 98% were obtained by the DTBM, RF, and KNNBM classifiers and hybrid models when using the 10 Relief features. The best recall score, however, was obtained with RFBM when applied to the 10 Relief features.

5) COMPARISON BETWEEN DIFFERENT METHODS BASED ON F1-SCORE

The F1-score is the harmonic mean of the precision and recall scores. For the 13 features, the highest F1-score (approximately 92%) is achieved with the RFBM which outperformed all other algorithms. KNN had the lowest F1 score for 13 features (84%), and the results for the DT and GB classifiers were similar: 87%, and 88% respectively.

After decreasing the number of features, the F1-score increased. For 11 features, GBBM had the highest score and most other classifiers also had better F1 scores than for 13 features. Result improved still further for the 10 Relief features with KNNBM and GBBM obtaining F1 scores of approximately 98% and DT, RF and AB of 90%, 98% and 93% respectively. The highest F1-score was obtained with the RFBM model that generates the highest outcome of f1-score (99%) and KNNBM provides the second highest score which is exactly 98%. The DTBM model had the lowest score for 10 features. F1-scores are shown in Fig. 13.

6) COMPARISON BETWEEN DIFFERENT METHODS BASED ON NEGATIVE PREDICTIVE VALUE

The negative predictive values (NPV) of the various algorithms have also been evaluated. The maximum NPV (98.59%) was obtained with RFBM, when applied to the Relief feature selection. The lowest NPV's were recorded for DT (86.47%) and DTBM (89.7%). For 13 features, the performance of the classifiers and hybrid model was not so good. The best NPV, for RFBM, was only 90.8%. NPV's for the features selected LASSO algorithm were less than for the features selected by Relief but still quite good compared with the 13 features values (93.6% for both RFBM and KNNBM). Overall, the lowest numClasses $= 2$ score was obtained by applying DT and KNN on the 13 features. The NPV outcomes are depicted in Fig. 14.

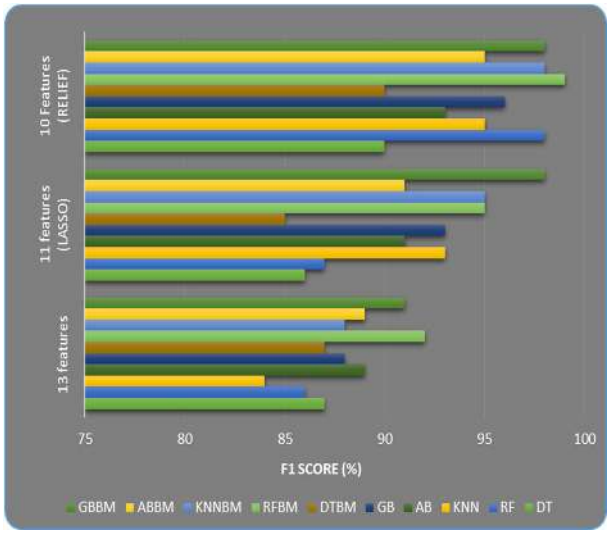

FIGURE 13. Obtained F1-scores.

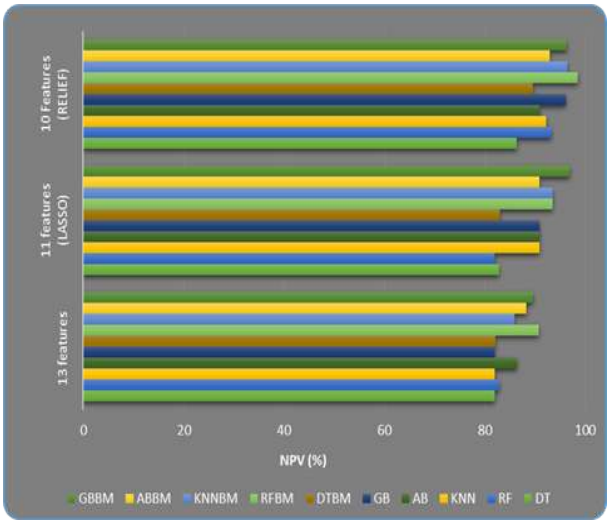

FIGURE 14. Outcomes of negative predictive values.

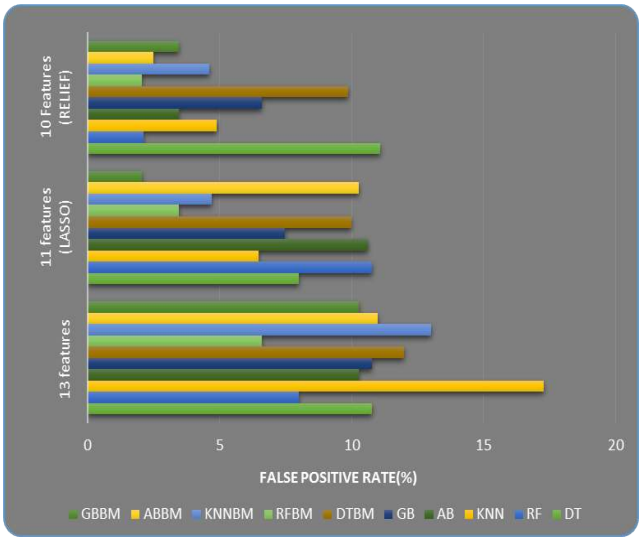

FIGURE 15. Metrics for false positive rates.

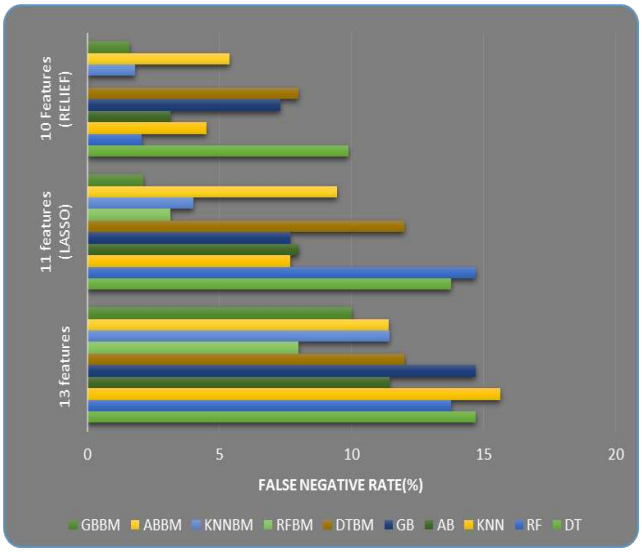

FIGURE 16. Obtained outcomes of false negative rates.

7) COMPARISON BETWEEN DIFFERENT METHODS BASED ON FALSE POSITIVE RATE

The false-positive rates of the various algorithms are illustrated before and after feature selection. After applying the Relief feature selection algorithm, the minimum false-positive rates was obtained with the RFBM, 2.05%, whereas the FPR was seen with DT. The outcomes of FPRs for RF, KNNBM, GBBM and others were just under 3.5%, a good result without applying the Relief or LASSO feature selection techniques, false-positive rates for classifiers and hybrid algorithms are considerably higher. The lowest FPR score for all 13 features was obtained by the RFBM, while the FPR for KNN was very high. FPRs are shown in Fig. 15.

8) COMPARISON BETWEEN DIFFERENT OUTCOMES BASED ON FALSE NEGATIVE RATES

The false-negative rates of various algorithms has been presented before and after applying the two feature selection techniques Relief and LASSO. With the features selected by LASSO, GBBM had the lowest FNR (2.1%).

RF had the highest FNR (14.7%). For the selective features by Relief, the FNR was approximately 0% for RBBM. For GBBM, the FNRs are low for both feature selection techniques. Without feature selection technique the false negative rates are higher. KNN had the highest FNR (15.62%). False Negative Rates are depicted in Fig. 16.

C. COMPARISON TABLE BETWEEN THE ACCURACY OF THE PROPOSED MODELS AND EXISTING TECHNIQUES

A combination of five different datasets has been employed for this study. Fig. 1 depicts the infrastructure of our proposed system and the outcomes based on the all features (13), The results of the features selected by LASSO (11) and Relief (10) were shown in Table 4 As a consequence, separate results have been reported based on these features.

TABLE 4. A Comparison of accuracy between the proposed system and some existing systems.

Our work				Other works			
Models	Accuracy of this model using ALL features (13)	Accuracy of this model using 11 Features (LASSO)	Accuracy of this model using 10 features (Relief)	Dataset	Existing Systems Accuracy	Dataset	Existing Systems Accuracy
DT	86.97%	88.6%	89.12%	Cleveland dataset (303)	75.55% [40]	Armed Forces Institute of Cardiology (AFIC) (500)	86.6% [49]
RF	88.65%	86.97%	97.89%	Cleveland, Long Beach VA, Switzerland, and Hungarian datasets (920)	80.89% [41]	Armed Forces Institute of Cardiology (AFIC) (500)	68.6 % [49]
KNN	83.61%	93%	94.11%	Cleveland dataset (303)	90.16% [43]	Cleveland dataset (303)	80% [50]
AB	89.07%	90.75%	92.85%	Kita Hospital (HKH) Jakarta (450)	46% [48]	Cleveland dataset (303)	54.13% [51]
GB	86.97%	92.85%	96.22%	Cleveland, Long Beach VA, Switzerland, Hungarian datasets (920)	84.27% [41]	Cleveland, Statlog (583)	95.19% [53]
DTBM	87.97%	88.65%	90.22%	Rajaie cardiovascular medical dataset (303)	79.54% [44]	Hungarian Institute of Cardiology (294)	85.03% [52]
RFBM	92.65%	97.65%	99.05%	Cleveland, Hungary, and Switzerland datasets	88.4 % [42]	Cleveland Heart Disease Data (303) by Bagging Approach	80.53% [39]
KNNBM	89.63%	96.6%	98.05%	hybrid method by Cleveland dataset (303)	84.07% [47]	Cleveland Heart Disease Data (303) by Hybrid Approach	85.48% [39]
ABBM	89.07%	90.75%	95.38%	Statlog Heart Disease Dataset (270) by Hybrid Approach	89% [45]	Cleveland Heart Disease Data (303) by Boosting Approach	75.9% [39]
GBBM	90.97%	97.85%	98.32%	Health insurance research database of Taiwan nation (317)	82.5% [46]	Cleveland Heart Disease Data (303) by ensemble Boosting Hybrid Approach	78.88% [39]

After changing the number of selected features by implementing selection algorithms, significant improvements have been noticeable. When an experiment has been gathered from all features, the best accuracy was achieved with the RFBM hybrid model (92.65%) and a low accuracy score was obtained with KNN (83.61%). Application of the LASSO selection algorithm leads to some dramatic changes. The highest accuracy was obtained with GBBM (97.85%), whereas the RF model performed the worst. The best results were obtained with the Relief feature selection technique. This achieves a 99.05% accuracy with RFBM. Our results have been compared to the existing models and datasets, see Table 4. Each row of the table deals with an algorithm that has been used in our studies, as well as two other related studies, and the results that have been reported. As an auxiliary information, we have also added the dataset that those studies have used. The table draws an overall picture of the performance of the algorithms in our study against other related works. The highest outcomes of previous results were just over 90.16% [43] and the performance of hybrid models was poor due to the limitations of the datasets [45]. The best result for hybrid models was only 89% (see Table 4). The highest accuracy achieved with previous research was 95.19% [53] and very poor performance of hybrid models [39]. Rashmi *et al.* [40] examined a 303-record dataset that had been extracted from the Cleveland dataset. That analysis showed that the Decision Tree achieved 75.55% accuracy. Dinesh *et al.* [41] worked on a 920-records datasets, combining the Cleveland, Long Beach VA, Switzerland and Hungarian datasets from the UCI repository and showed that RF could obtain an accuracy of 80.89%. Other authors in [49] applied the DT and RF to a dataset of 500 which was taken from the Armed Forces Institute of Cardiology (AFIC) and reported that DT achieved the best result (86.6 %). Hybrid classifiers were explored by several researcher [39], [52], obtaining an accuracy of 85.48% using the KNNBM approach. The performance of our proposed model is very good compared to previous research works as can be seen from Table 4.

FPR is used to show the percentage of wrongly detected heart disease whereas the FNR or miss rate measures the incorrect negative classifications. Fig. 17 shows FPR and

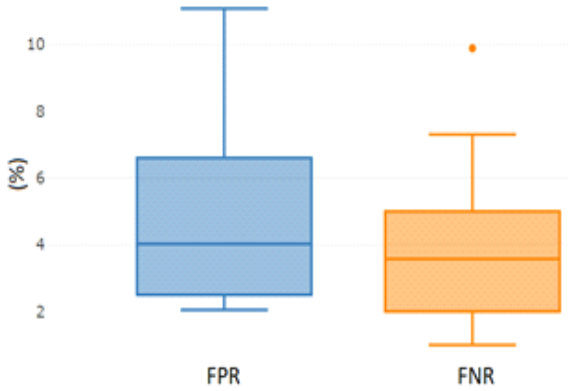

FIGURE 17. Comparison between FPR and FNR values.

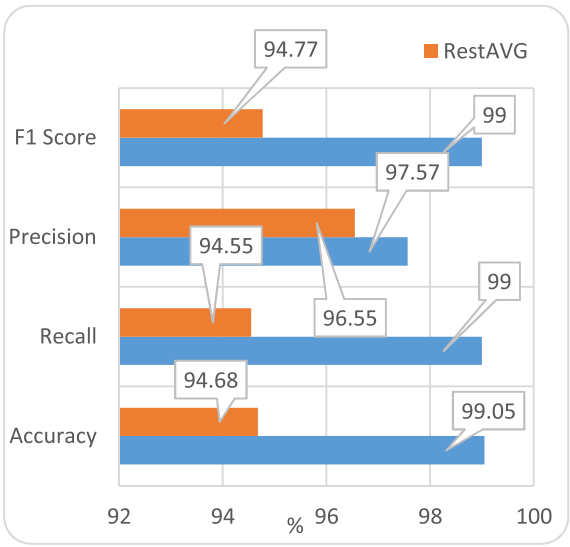

FIGURE 18. The experimental accuracy of between RFBM and RestAVG.

FNR values. The low FNRs represent a major outcome, based on the heart disease dataset. After evaluation, RFBM in combination with the Relief feature selection algorithm has been demonstrated to have the best performance.

The highest accuracy was obtained with the Relief feature selection algorithm and the Random Forest Bagging Method (99.05%). However, the outcomes of RestAVG scores were not bad for a diagnosis system. From Fig. 18, it can be observed that the values of the relevant performance indices were all about to 94% except precision values which was higher (96.55%).

Note that the remaining three features which were not used with Relief are Thal, oldpeak and slope.

VII. COMPARATIVE EVALUATION OF OUR PROPOSED MODEL

In our predicted model, ten features have been evaluated to make this comparison more unique. Our introduced algorithms were conducted based on the all features (13), LASSO selected features (11), and Relief selected features (10).

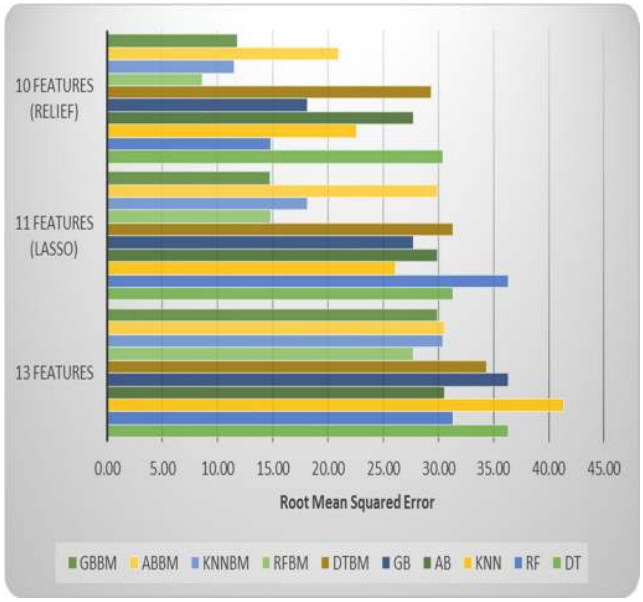

FIGURE 19. Root Mean Squared Errors of different algorithms on Relief selected features.

The obtained outcomes were compared to other works to show the percentage of improvement, while decrease in performance also noted in one occasion (KNN). The highest increment was noticed for AB approach as opposed to previous works which was about 46% [48] percentage improvement were calculated for 13 features (93.63%), 11 features (97.28%), and 10 features (101.85%) respectively. On the other hand, the lowest increment in percentage was seen for the ABBM model which was just under 2%, however, for the selected features of LASSO it was just over 4%. Significant higher values were witnessed in 10 features than 13 features percentage calculator for RF, RFBM, KNNBM, and GBBM. Table 5 has been given below.

VIII. STATISTICAL ANALYSIS

We applied the Root Mean Squared Error and Log Loss to the output of our algorithms. Results are described below.

A. ROOT MEAN SQUARED ERROR

Following Fig. 19 portraits the Root Mean Squared Error for the 10 models. Here we are analyzing the RMSE of each model for 13 features, 11 features (LASSO) and 10 features (Relief). It is clear that RFBM model has the lowest RMSE for 13 features and 10 features - 27.735 and 8.602 respectively. The GBBM model, produces the minimum RMSE for LASSO which is 14.732. Thus, we might say that RFBM model produces the best results for 13 features and 10 features (Relief), whereas the GBBM model produces the best result for 11 features (LASSO).

KNN, RF and DT have the highest values of the RMSE for 13 features (41.345), 11 features (36.313) and 10 features (30.38) respectively. Therefore, we might conclude that these three models are the most ineffective.

TABLE 5. A comparison of accuracy between proposed system and existed outcomes.

Models	Accuracy of all features (13)	Accuracy of 11 features (LASSO)	Accuracy of 10 features (Relief)	Existing Systems (Accuracy)	Percentage calculator for ALL features	Percentage calculator for 11 features	Percentage calculator for 10 features
DT	86.97%	88.6%	89.12%	75.55% [40]	15.11% increase	17.27% increase	17.97% increase
RF	88.65%	86.97%	97.89%	80.89% [41]	9.60% increase	7.51% increase	21.01% increase
KNN	83.61%	93%	94.11%	90.16% [43]	7.26% decrease	3.15% increase	4.38% increase
AB	89.07%	90.75%	92.85%	46% [48]	93.63% increase	97.28% increase	101.85% increase
GB	86.97%	92.85%	96.22%	84.27% [41]	3.20% increase	10.18% increase	14.18% increase
DTBM	87.97%	88.65%	90.22%	79.54% [44]	10.59% increase	11.45% increase	13.42% increase
RFBM	92.65%	97.65%	99.05%	88.4 % [42]	4.80% increase	10.46% increase	12.04% increase
KNNBM	89.63%	96.6%	98.05%	84.07% [47]	6.61% increase	14.90% increase	16.62% increase
ABBM	89.07%	90.75%	95.38%	89% [45]	0.0786% increase	1.97% increase	7.17% increase
GBBM	90.97%	97.85%	98.32%	82.5% [46]	10.26% increase	18.60% increase	19.17% increase

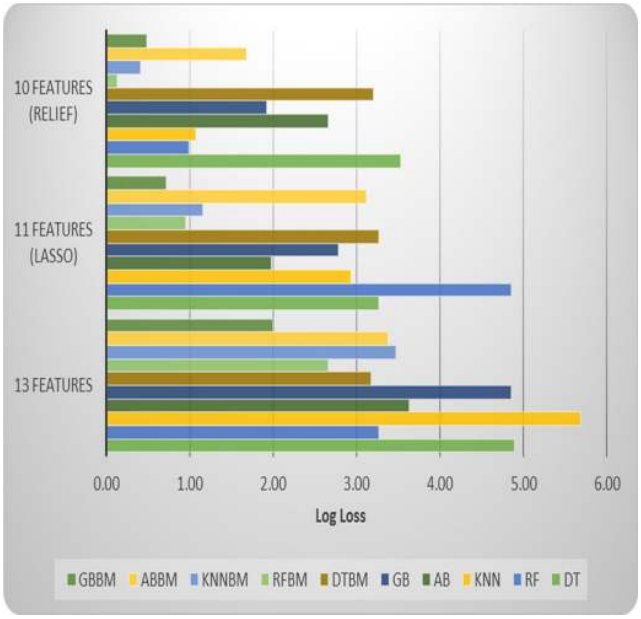

FIGURE 20. Log Loss of different algorithms on Relief selected features.

B. LOG LOSS

Following Fig. 20 depicts the log loss (LL) for 10 types of models. Here we are analyzing the changes in LL value of each model for 13 features, 11 features (LASSO) and 10 features (Relief). If we take a deeper look, we can observe that GBBM model has the lowest LL value for 13 features and 11 features which are 1.997 and 0.721 respectively. The RFBM model generates the least LL value for Relief which is 0.127. Therefore, that the GBBM model produces the best result for 13 features and 11 features (LASSO), whereas RFBM model produces the best result for 10 features (Relief). On the contrary, KNN, RF and DT give the highest value of LL for 13 features, 11 features and 10 features which are 5.683, 4.854 and 3.532 respectively. Thus, to recapitulate, all these three models are most inefficient for all three categories of features (13 features, LASSO and Relief).

IX. ANALYSIS ON RUNTIME AND COMPUTATIONAL COMPLEXITY

A. COMPUTATIONAL COMPLEXITY

The following Table 6 illustrates the Computational Complexity for seven different models. Two types of complexities are included in computational complexities: Training complexity and Prediction complexity. Only KNN and Boosting model have no training complexity associated with them. All other models have both training and prediction complexity which are given in the table.

Denoting $\boldsymbol{n}$ as the number of training sample, $\boldsymbol{p}$ as the number of features, ***ntrees*** as the number of trees (for methods based on various trees), and $\boldsymbol{k}$ as the number of neighbors, we have the following approximations: Bootstrap Aggregation or bagging is $O(n)$ (for n-sized trees) and random subspace is $\boldsymbol{O}(\boldsymbol{d'})$ (where $d' \ll d$). Complexity for $\boldsymbol{t}$ bagged trees of random subspaces is $\boldsymbol{O}$ (t∗d2∗n2∗log (n)) (taking $d' = d$ for big O notation).

B. RUNTIME

The Fig. 21 displays the run time (RT) for 10 models. Here we are trying to evaluate the RT value of each model for 13 features, 11 features (LASSO) and 10 features (Relief). We can clearly notice that the RFBM model has the lowest RT which are 0.0126 for 13 features, 0.0012 for LASSO and 0.0011 for Relief respectively. On the other hand, the GBBM model has the longest RT, 1.9973 for 13 features, 1.7021 for 11 features and 1.5547 for 10 features respectively.

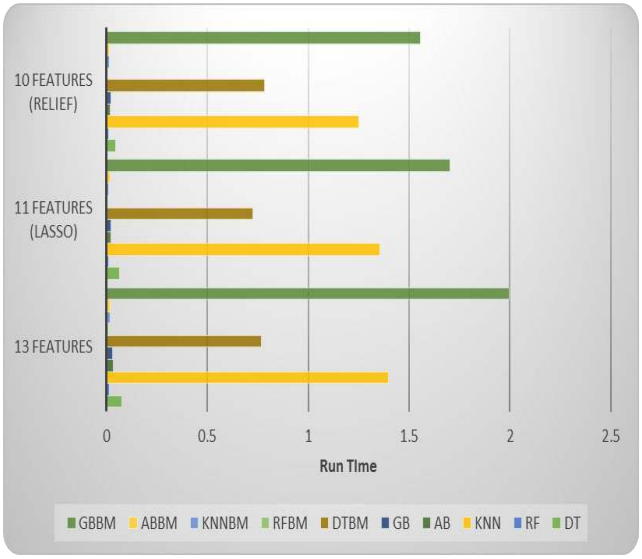

FIGURE 21. Run time performance of different algorithms on Relief selected features.

TABLE 6. Algorithmic complexities of the algorithms used.

Models	Training	Training Complexity Calculation	Prediction	Predicted Complexity Calculation
DT	O(n2p)	O(14161000)	O(p)	O(10)
RF	O(n2pntrees)	O(28322000)	O(pntrees)	O(20)
KNN	O (knp)	O(59500)	O(np)	O(11900)
AB	O(npntrees)	O(11900000)	O(pntrees)	O(10000)
GB	O(npntrees)	O(1190000)	O(pntrees)	O(1000)
DTBM	-	-	O(p*ntrees * n log(n))	O(365990.0884)
RFBM	-	-	O(ntress*p 2*n2*log(n))	O(4355282052)
ABBM	-	-	p * n2 * log(n) * ntrees	O(43552820520.3)
KNNBM	-	-	O(K*n*p* log(n))	O(182995.0442)
GBBM	-	-	p * n * log(n) * ntrees	O(3659900.88406)

X. HYPERPARAMETER TUNING

GridSearchCV, which allocates hyper parameters, is a process of tuning which can determine the optimal value for a given model. In our proposed system, GridSearchCV has been used in order to obtain a higher accuracy. The following parameters were used on the examined algorithms (see Table 7):

sklearn.model_selection.GridSearchCV (estimator, param_grid, scoring = None, n_jobs = None, iid = 'deprecated', refit = True, cv = None, verbose = 0, pre_dispatch = '2*n_jobs', error_score = nan, return_train_score = False)

For getting an accurate prediction, tuning is a fundamental part for all types of classifiers. As a result, we tuned our 5 classifiers including DT, RF, KNN, AB, and GB, however, the default parameter was used with base classifiers for ensemble technique.

TABLE 7. Parameters used.

Applied Algorithms	Parameters
DT	algo = "Classification"; numClasses = 2; maxDepth = 5; minInstancesPerNode = "auto"; minInfoGain = "auto"; maxBins = 32; maxMemoryInMB = 256 MB; subsamlingRate = "auto"; impurity = "gini"
RF	numClasses = 2; numTrees = "auto"; featureSubsetStrategy = "false"; subsamlingRate = "auto"; impurity = "gini"; seed = "false"
KNN	algorithm='auto', leaf_size=30, metric='minkowski', nmetric_params=None, n_jobs=1, n_neighbors=5, p=2, weights='uniform'
AB	n_samples = 1000, probability = True, kernel = 'linear', n_features = 4, n_informative = 2, n_redundant = 0, random_state = 0, shuffle=False
GB	Loss = "Log Loss"; numIterations and learningRate = "auto"; algo = "Classification"
DTBM	base_classifier = 'DT', parameter = 'default'
RFBM	base_classifier = 'RF', parameter = 'default'
KNNBM	base_classifier = 'KNN', parameter = 'default'
ABBM	base_classifier = 'AB', parameter = 'default'
GBBM	base_classifier = 'GB', parameter = 'default'

XI. LIMITATIONS OF OUR PROPOSED SYSTEM

The overall discussion has shown that the performance of different classifiers were good enough in comparison to previous studies, however, there are indeed few limitations, such as, the dependency on a specific Feature Selection technique, for instance more reliance on Relief in this case to produce highly accurate results. Additionally, high level of missing values in the dataset can have an adverse effect. We have demonstrated how to address the issue through the proper methods, and therefore other dataset when used with this model, must also take care of this issue if the missing value is quite significant. Furthermore, though our training dataset is reasonably extensive, larger dataset would make the model more precise.

XII. CONCLUSION

Identifying the risk of heart disease with reasonably high accuracy could potentially have a profound effect on the long-term mortality rate of humans, regardless of social and cultural background. Early diagnosis is a key step in achieving that goal. Several studies have already attempted to predict heart disease with the help of machine learning. This study takes similar route, but with an improved and novel method and with a larger dataset for training the model. This research demonstrates that the Relief feature selection algorithm can provide a tightly correlated feature set which then can be used with several machine learning algorithms. The study has also identified that RFBM works particularly well with the high impact features (obtained by feature selection algorithms or medical literature) and produces an accuracy, substantially higher than related work. RFBM achieved an

accuracy of 99.05% with 10 features. In the future we aim to generalize the model even further so that it can work with other feature selection algorithms and be robust against datasets where the level of missing data is high. The application of Deep Learning algorithms is another future approach. The primary aim of this research was to improve upon the existing work with an innovative and novel way of building the model, as well as to make the model useful and easily implementable to practical settings.

REFERENCES

[1] C. Trevisan, G. Sergi, S. J. B. Maggi, and H. Dynamics, "Gender differences in brain-heart connection," in *Brain and Heart Dynamics*. Cham, Switzerland: Springer, 2020, p. 937.

[2] M. S. Oh and M. H. Jeong, "Sex differences in cardiovascular disease risk factors among Korean adults," *Korean J. Med.*, vol. 95, no. 4, pp. 266–275, Aug. 2020.

[3] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *Int. J. Pharmaceutical Res.*, vol. 12, no. 4, 2020.

[4] World Health Organization and J. Dostupno, "Cardiovascular diseases: Key facts," vol. 13, no. 2016, p. 6, 2016. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[5] K. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Comput. Sci.*, vol. 120, pp. 588–593, Jan. 2017.

[6] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Inf. Syst.*, vol. 2018, pp. 1–21, Dec. 2018.

[7] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 204–207.

[8] J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, Dec. 2005.

[9] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 3, pp. 176–183, 2013.

[10] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," *Int. J. Comput. Sci.*, vol. 8, no. 2, pp. 150–154, 2011.

[11] F. M. J. M. Shamrat, M. A. Raihan, A. K. M. S. Rahman, I. Mahmud, and R. Akter, "An analysis on breast disease prediction using machine learning approaches," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 2450–2455, Feb. 2020.

[12] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics Informat.*, vol. 36, pp. 82–93, Mar. 2019.

[13] N. Kausar, S. Palaniappan, B. B. Samir, A. Abdullah, and N. Dey, "Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients," in *Applications of Intelligent Optimization in Biology and Medicine*. Cham, Switzerland: Springer, 2016, pp. 217–231.

[14] J. Mackay and G. A. Mensah, "The atlas of heart disease and stroke," World Health Org., Geneva, Switzerland, Tech. Rep., 2004.

[15] M. Ashraf, S. M. Ahmad, N. A. Ganai, R. A. Shah, M. Zaman, S. A. Khan, and A. A. Shah, *Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS*. Singapore: Springer, 2021, pp. 239–255.

[16] F. Andreotti, F. S. Heldt, B. Abu-Jamous, M. Li, A. Javer, O. Carr, S. Jovanovic, N. Lipunova, B. Irving, R. T. Khan, R. Dürichen, "Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units," 2020, *arXiv:2007.08491*. [Online]. Available: https://arxiv.org/abs/2007.08491

[17] W. Wiharto, H. Kusnanto, and H. Herianto, "Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 2, p. 1023, Apr. 2017.

[18] A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," in *Proc. 5th Int. Conf. Informat., Electron. Vis. (ICIEV)*, May 2016, pp. 145–150.

[19] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Comput. Math. Med.*, vol. 2017, pp. 1–11, Jan. 2017.

[20] D. Singh and J. S. Samagh, "A comprehensive review of heart disease prediction using machine learning," *J. Crit. Rev.*, vol. 7, no. 12, p. 2020, 2020.

[21] M. Shouman, T. Turner, and R. Stocker, "Integrating clustering with different data mining techniques in the diagnosis of heart disease," *J. Comput. Sci. Eng.*, vol. 20, no. 1, pp. 1–10, 2013.

[22] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100402.

[23] H. Wang, Z. Huang, D. Zhang, J. Arief, T. Lyu, and J. Tian, "Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in kawasaki disease," *IEEE Access*, vol. 8, pp. 97064–97071, 2020.

[24] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Res. Int.*, vol. 2020, Apr. 2020, Art. no. 9816142.

[25] J. Mishra and S. Tarar, *Chronic Disease Prediction Using Deep Learning*. Singapore: Springer, 2020, pp. 201–211.

[26] F. Z. Abdeldjouad, M. Brahami, and N. Matta, *A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques*. Cham, Switzerland: Springer, 2020, pp. 299–306.

[27] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," *Acta Sci. Nutritional Health*, vol. 3, no. 7, pp. 147–151, Jul. 2019.

[28] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203.

[29] I. Javid, A. Khalaf, and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, 2020.

[30] N. Kumar and K. Sikamani, "Prediction of chronic and infectious diseases using machine learning classifiers—A systematic approach," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 4, pp. 11–20, 2020.

[31] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, "Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, Jan. 2019.

[32] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[33] F. Miao, Y.-P. Cai, Y.-X. Zhang, X.-M. Fan, and Y. Li, "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest," *IEEE Access*, vol. 6, pp. 7244–7253, 2018.

[34] C. Raju, E. Philipsy, S. Chacko, L. P. Suresh, and S. D. Rajan, "A survey on predicting heart disease using data mining techniques," in *Proc. Conf. Emerg. Devices Smart Syst. (ICEDSS)*, 2018, pp. 253-255.

[35] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 16, Dec. 2020.

[36] E. Ahmad, A. Tiwari, and A. Kumar, "Cardiovascular Diseases (CVDs) Detection using Machine Learning Algorithms,"

[37] L. Wang, W. Zhou, Q. Chang, J. Chen, and X. Zhou, "Deep ensemble detection of congestive heart failure using short-term RR intervals," *IEEE Access*, vol. 7, pp. 69559–69574, 2019.

[38] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020.

[39] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, no. 2, 2019, Art. no. 100203.

[40] G. O. Rashmi and U. M. A. kumar, "Machine learning methods for heart disease prediction," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 5S, pp. 220–223, May 2019.

[41] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in *Proc. Int. Conf. Current Trends Towards Converging Technol. (ICCTCT)*, Coimbatore, India, Mar. 2018, pp. 1–7.
[42] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
[43] S. Sharma and M. Parmar, "Heart diseases prediction using deep learning neural network model," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 3, pp. 1–5, Jan. 2020.
[44] R. Alizadehsani, J. Habibi, Z. A. Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features," *Res. Cardiovascular Med.*, vol. 2, no. 3, pp. 133–139, Aug. 2013.
[45] A. A. Shetty and C. Naik, "Different data mining approaches for predicting heart disease," *Int. J. Innov. Sci. Eng. Technol.*, vol. 5, pp. 277–281, May 2016.
[46] C. A. Cheng and H. W. Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 2566–2569.
[47] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "A hybrid evolutionary algorithm for attribute selection in data mining," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8616–8630, May 2009.
[48] I. K. A. Enriko, "Comparative study of heart disease diagnosis using top ten data mining classification algorithms," in *Proc. 5th Int. Conf. Frontiers Educ. Technol.*, 2019, pp. 159-164.
[49] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, "Identification of heart failure by using unstructured data of cardiac patients," in *Proc. 45th Int. Conf. Parallel Process. Workshops (ICPPW)*, Aug. 2016, pp. 426–431.
[50] A. K. Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation," *Neural Comput. Appl.*, vol. 29, pp. 685–693, Sep. 2016.
[51] A. Kaur, "A comprehensive approach to predict heart diseases using data mining," *Int. J. Innov. Eng. Technol.*, vol. 8, no. 2, pp. 1–5, Apr. 2017.
[52] V. Chaurasia and S. Pal, "Data mining approach to detect heart diseases," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, no. 4, pp. 56–66, 2014.
[53] R. Bhuvaneeswari, P. Sudhakar, and G. Prabakaran, "Heart disease prediction model based on gradient boosting tree (GBT) classification algorithm," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 41–51, Sep. 2019.
[54] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, A. Kazi, and S. Shultana, "Implementation of machine learning algorithms to detect the prognosis rate of kidney disease," in *Proc. IEEE Int. Conf. Innov. Technol.*, Nov. 2020, pp. 1–7.
[55] S. Shultana, M. S. Moharram, and N. Neehal, "Olympic sports events classification using convolutional neural networks," in *Proc. Int. Joint Conf. Comput. Intell. (IJCCI)*, Dhaka, Bangladesh, 2018, pp. 507–518.
[56] S. V. J. Jaikrishnan, O. Chantarakasemchit, and P. Meesad, "A breakup machine learning approach for breast cancer prediction," in *Proc. 11th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, Pattaya, Thailand, Oct. 2019, pp. 1–6.
[57] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Coimbatore, India, Mar. 2018, pp. 1275–1278.
[58] G. Singh, "Breast cancer prediction using machine learning," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 8, no. 4, pp. 278–284, Jul. 2020.
[59] *Heart Disease Datasets From UCI Machine Learning Repository*. Accessed: May 31, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[60] *Heart Disease Statlog Dataset of UCI Machine Learning Repository*. Accessed: May 31, 2020. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/statlog+(heart)
[61] S. Ralston, I. Penman, M. Strachan, and R. Hobson, *Davidson's Principles and Practice of Medicine*, 23rd ed. U.K.: Elsevier, Apr. 2018, pp. 219–225.
[62] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," in *Proc. Int. Conf. Intell. Comput. Control (IC)*, Jun. 2017, pp. 1–8.
[63] A. Acharya, "Comparative study of machine learning algorithms for heart disease prediction," M.S. thesis, Helsinki Metropolia Univ. Appl. Sci., Helsinki, Finland, Apr. 2017. [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/124622/Final%20Thesis.pdf?sequence=1&isAllowed=y
[64] A. M. D. Silva, *Feature Selection*, vol. 13. Berlin, Germany: Springer, 2015, pp. 1–13.
[65] S. Chikhi and S. Benhammada, "ReliefMSS: A variation on a feature ranking ReliefF algorithm," *Int. J. Bus. Intell. Data Mining*, vol. 4, pp. 375–390, Jan. 2009.
[66] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 73, no. 3, pp. 273–282, Jun. 2011.
[67] C. Zhou and A. Wieser, "Jaccard analysis and LASSO-based feature selection for location fingerprinting with limited computational complexity," in *Proc. 14th Int. Conf. Location Based Services (LBS)*, Dec. 2018, pp. 71–87.
[68] *Ensemble Techniques of Bagging*. Accessed: Jun. 31, 2020. [Online]. Available: https://quantdare.com/what-is-the-difference-between-Bagging-and-Boosting/
[69] *An Explanation of Ensemble Bagging Techniques*. Accessed: Jun. 31, 2020. [Online]. Available: https://towardsdatascience.com/ensemble-methods-Bagging-Boosting-and-stacking-c9214a10a205/
[70] P. Ghosh, M. Z. Hasan, and M. I. Jabiullah, "A comparative study of machine learning approaches on dataset to predicting cancer outcome," *Bangladesh Electron. Soc.*, vol. 18, nos. 1–3, pp. 1–5, 2018.
[71] F. M. Javed Mehedi Shamrat, Z. Tasnim, P. Ghosh, A. Majumder, and M. Z. Hasan, "Personalization of job circular announcement to applicants using decision tree classification algorithm," in *Proc. IEEE Int. Conf. Innov. Technol. (INOCON)*, Nov. 2020, pp. 1–5.
[72] M. M. Alam, S. Saha, P. Saha, F. N. Nur, N. N. Moon, A. Karim, and S. Azam, "D-CARE: A non-invasive glucose measuring technique for monitoring diabetes patients," in *Proc. Int. Joint Conf. Comput. Intell. Algorithms Intell. Syst.*, 2019, pp. 443–453.
[73] W. M. Baihaqi, N. A. Setiawan, and I. Ardiyanto, "Rule extraction for fuzzy expert system to diagnose coronary artery disease," in *Proc. 1st Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Yogyakarta, Indonesia, Aug. 2016, pp. 136–141.
[74] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Comput. Methods Programs Biomed.*, vol. 130, pp. 54–64, Jul. 2016.
[75] A. Mert, N. Kılıç, and A. Akan, "Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats," *Neural Comput. Appl.*, vol. 24, no. 2, pp. 317–326, Feb. 2014.
[76] 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," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, 2020.
[77] P. Ghosh, M. Z. Hasan, O. A. Dhore, A. A. Mohammad, and M. I. Jabiullah, "On the application of machine learning to predicting cancer outcome," in *Proc. Int. Conf. Electron. (ICT)*. Dhaka, Bangladesh: Bangladesh Electronics Society (BES), Nov. 2018, p. 60.
[78] *Responsible for Herat Disease Risk Factors*. Accessed:Jul. 15, 2020. [Online]. Available: https://www.texasheart.org/heart-health/heart-informationcenter/ topics/heart-disease-risk-factors/
[79] R. Banerjee, S. Biswas, S. Banerjee, A. D. Choudhury, T. Chattopadhyay, A. Pal, P. Deshpande, and K. M. Mandana, "Time-frequency analysis of phonocardiogram for classifying heart disease," in *Proc. Comput. Cardiol. Conf. (CinC)*, Vancouver, BC, Canada, Sep. 2016, pp. 573–576.
[80] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, and S. Shultana, "Implementation of machine learning algorithms to detect the prognosis rate of kidney disease," in *Proc. IEEE Int. Conf. Innov. Technol.*, Nov. 2020, pp. 1–7.
[81] *An Overview of K_Nearest Neighbors Algorithm*. Accessed: Jun. 31, 2020. [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor algorithm- for-machine-learning
[82] D. Giri, U. R. Acharya, R. J. Martis, S. V. Sree, T.-C. Lim, T. Ahamed, and J. S. Suri, "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform," *Knowl.-Based Syst.*, vol. 37, pp. 274–282, Jan. 2013.
[83] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using data mining algorithm," *Global J. Comput. Sci. Technol.*, vol. 10, pp. 38–43, Sep. 2010.
[84] M. Gilani, J. M. Eklund, and M. Makrehchi, "Automated detection of atrial fibrillation episode using novel heart rate variability features," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Lake Buena Vista, FL, USA, Aug. 2016, pp. 3461–3464.

[85] K. Padmavathi and K. S. Ramakrishna, "Classification of ECG signal during atrial fibrillation using autoregressive modeling," *Procedia Comput. Sci.*, vol. 46, pp. 53–59, Jan. 2015.

[86] S. H. Ripon, "Rule induction and prediction of chronic kidney disease using boosting classifiers, Ant-Miner and J48 Decision Tree," in *Proc. Int. Conf. Elect., Comput. Commun. Eng. (ECCE)*, Cox's Bazar, Bangladesh, 2019, pp. 1–6.

[87] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access*, vol. 7, pp. 168261–168295, 2019.

[88] P. Ghosh, F. M. J. M. Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optization of prediction method of chronic kidney disease with machine learning algorithms," in *Proc. 15th Int. Symp. Artif. Intell. Natural Lang. Process. (iSAI-NLP), Int. Conf. Artif. Intell. Internet Things (AIoT)*, 2020.

[89] *An Overview of Gradient Boosting Algorithm*. Accessed: Jun. 31, 2020. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-Boosting-algorithm-machine-learning/

[90] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 89–96, 2019.

[91] *Gradient Boosting Algorithm*. Accessed: Jun. 31, 2020. [Online]. Available: https://data-flair.training/blogs/gradient-Boosting-algorithm/

[92] T. Chen and C. Guestrin, "XGBOOST: A scalable tree boosting system," in *Proc. 22nd ACMSIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[93] J. Cheng, G. Li, and X. Chen, "Research on travel time prediction model of freeway based on gradient boosting decision tree," *IEEE Access*, vol. 7, pp. 7466–7480, 2019, doi: 10.1109/ACCESS.2018.2886549.

[94] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobotics*, vol. 7, no. 7, pp. 1–21, 2013.

[95] A. M. De Silva and P. H. W. Leong, *Grammar-Based Feature Generation for Time-Series Prediction*. Berlin, Germany: Springer, 2015.

[96] F. M. J. M. Shamrat, M. Asaduzzaman, P. Ghosh, M. D. Sultan, and Z. Tasnim, "A Web based application for agriculture: 'Smart farming system,'" *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2309–2320, Jun. 2020.

[97] *Responsible for Herat Disease Risk Factors*. Accessed: Jul. 15, 2020. [Online]. Available: https://www.texasheart.org/heart-health/heart-informationcenter/topics/heart-disease-risk-factors/

[98] F. M. J. M. Shamrat, P. Ghosh, I. Mahmud, N. I. Nobel, and M. D. Sultan, "An intelligent embedded AC automation model with temperature prediction and human detection," in *Proc. 2nd Int. Conf. Emerg. Technol. Data Mining Inf. Secur. (IEMIS)*, 2020.

[99] *Sex, Age, Cardiovascular Risk Factors, and Coronary Heart Disease*. Accessed: Dec. 29, 2020. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/01.cir.99.9.1165

[100] S. Hegelich, "Decision trees and random forests: Machine learning techniques to classify rare events," *Eur. Policy Anal.*, vol. 2, no. 1, pp. 98–120, 2016.

[101] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: From early developments to recent advancements," *Syst. Sci. Control Eng.*, vol. 2, no. 1, pp. 602–609, Dec. 2014.

[102] A. Sharma and A. Suryawanshi, "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure," *Int. J. Comput. Appl.*, vol. 136, no. 6, pp. 28–35, Feb. 2016.

[103] *Spearman's Rank-Order Correlation*. Accessed: Jul. 15, 2019. [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

PRONAB GHOSH received the B.Sc. degree from the Computer Science and Engineering Department, Daffodil International University, in 2019. He has been heavily involved in collaborative research activities with researchers in Bangladesh and researchers from Australia, especially in the fields of machine learning, deep learning, cloud computing, and the IoT.

SAMI AZAM is currently a leading Researcher and a Senior Lecturer with the College of Engineering and IT, Charles Darwin University, Casuarina, NT, Australia. He is also actively involved in the research fields relating to Computer Vision, Signal Processing, Artificial Intelligence, and Biomedical Engineering. He has number of publications in peer-reviewed journals and international conference proceedings.

MIRJAM JONKMAN (Member, IEEE) is currently a Lecturer and a Researcher with the College of Engineering, IT, and Environment. Her research interests include biomedical engineering, signal processing, and the application of computer science to real life problems.

ASIF KARIM is currently a Ph.D. Researcher with Charles Darwin University, Casuarina, NT, Australia, and lives in the port city of Darwin. His research interest includes machine intelligence and cryptographic communication. He is also working towards the development of a robust and advanced email filtering system primarily using Machine Learning algorithms. He has considerable industry experience in IT, primarily in the field of Software Engineering.

F. M. JAVED MEHEDI SHAMRAT received the B.Sc. degree in software engineering from Daffodil International University, in 2018. He used to work at Daffodil International University as a Research Associate. He is currently working in a Government Project under the ICT Division as a Researcher and Developer. He has published several research papers and articles in journals (Scopus) and international conferences. His research interests include the IoT, machine learning, data science, information security, android applications, image processing, neural network, cyber security, Artificial Intelligence, robotics, and deep learning.

EVA IGNATIOUS is currently a Ph.D. Researcher with Charles Darwin University, Casuarina, NT, Australia. Her research interests include biomedical signal processing (interesting features and abnormalities found in bio-signals), theoretical modelling and simulation (breast cancer tissues), applied electronics (thermistors), process control and instrumentation, and embedded/VLSI systems. She has considerable research experience with one U.S. patent and two Indian patents for the development of thermal sensor-based breast cancer detection at its early stages together with Centre for Materials for Electronics Technology (C-MET), an autonomous scientific society under Ministry of Electronics and Information Technology (MeitY), Government of India. She also has industrial experience as a Production Engineer and a Quality Controller, primarily in the Electronics and Instrumentation Engineering.

SHAHANA SHULTANA received the B.Sc. degree in computer science and engineering from Daffodil International University, where she is currently pursuing the M.Sc. degree in computer science and engineering. She is also working as a Lecturer with the Department of Computer Science and Engineering, Daffodil International University. Her research interests include computer vision, data mining, neural network, and Artificial Intelligence.

ABHIJITH REDDY BEERAVOLU received the M.S. degree in information systems and data science form Charles Darwin University. His goal is to live free and come up with ideas that can help the people and the societies near me and using those ideas to ship them into the world. He is a Computer Science Enthusiast who is interested in anything that is related to computers. Also, he is interested in reading books on history and making comparisons with the current world, to make sense of the reality and its progression. Mostly, he is interested in reading and analyzing information related to cognitive and behavioral psychology and trying to implement/integrate them into various technological ideas.

FRISO DE BOER is currently a Professor with the College of Engineering, IT, and Environment, Charles Darwin University, Casuarina, NT, Australia. His research interests include signal processing, biomedical engineering, and mechatronics.