Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 9005278, 11 pages
https://doi.org/10.1155/2022/9005278

Research Article

AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease

T. R. Mahesh,[1] V. Dhilip Kumar,[2] V. Vinoth Kumar,[1] Junaid Asghar,[3] Oana Geman,[4] G. Arulkumaran,[5] and N. Arun[1]

[1] *Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bangalore, India*
[2] *Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India*
[3] *Faculty of Pharmacy, Gomal University, Dera Ismail Khan 29050, Khyber Pakhtunkhwa, Pakistan*
[4] *Stefan Cel Mare University of Suceava, Suceava, Romania*
[5] *Department of Electrical and Computer Engineering, Bule Hora University, Bule Hora, Ethiopia*

Correspondence should be addressed to G. Arulkumaran; erarulkumaran@gmail.com

Received 24 February 2022; Revised 24 March 2022; Accepted 30 March 2022; Published 18 April 2022

Academic Editor: Muhammad Ahmad

As a result of technology improvements, various features have been collected for heart disease diagnosis. Large data sets have several drawbacks, including limited storage capacity and long access and processing times. For medical therapy, early diagnosis of heart problems is crucial. Disease of heart is a devastating human disease that is quickly increasing in developed and also developing countries, resulting in death. In this type of disease, the heart normally fails to provide enough blood to different body parts in order to allow them to perform their regular functions. Early, as well as, proper diagnosis of this condition is very critical for averting further damage and also to save patients' lives. In this work, machine learning (ML) is utilized to find out whether a person has cardiac disease or not. Both the types of ensemble classifiers, namely, homogeneous as well as heterogeneous classifiers (formed by combining two separate classifiers), have been implemented in this work. The data mining preprocessing using Synthetic Minority Oversampling Technique (SMOTE) has been employed to cope with the imbalance problem of the class as well as noise. The proposed work has two steps. SMOTE is used in the initial phase to reduce the impact of data imbalance and the second phase is classifying data using Naive Bayes (NB), decision tree (DT) algorithms, and their ensembles. The experimental results demonstrate that the AdaBoost-Random Forest classifier provides 95.47% accuracy in the early detection of heart disease.

1. Introduction

Heart disease is mainly observed as the world's most dangerous and life-threatening chronic disease. During heart illness, the heart generally fails to deliver enough blood to different body regions so as to allow them to operate normally. The narrowing and occlusion of coronary arteries can cause heart failure. Heart disease is one among the leading reasons for death nowadays across the globe [1]. This leads to crucial requirement of monitoring the functioning organs in the human body and a critical aspect in monitoring health records of cardiovascular system. The coronary arteries control the entire circulation of blood to the heart. According to the latest survey, United States is one of the severely affected countries with relatively high ratio of heart disease observed in patients. The symptoms like breathing problem, physical body weakness, exhaustion, and swollen feet among various other symptoms are the most typical markers of heart disease [2]. Most of the cardiovascular diseases affecting people across the world are usually fatal. So, to overcome this problem, development of new technique may aid in detection of heart diseases in early stages as there is huge growth in the technology. Also, before causing substantial damage to avoid advantageous problems in terms of time, cost, and saving human lives machine learning techniques are used to focus on monitoring the heart diseases. Machine learning involves emerging techniques in manipulating and extracting features or relevant data information in possible way [3]. Machine learning is one of the complex fields and also has huge scope in various applications which is expanding all the time. Machine learning techniques consist of supervised learning, unsupervised learning, and also ensemble learning classifiers, which are mainly used to forecast the heart diseases in early stages with increase in accuracy results [4].

In the past years, academicians and researchers attempted to create and implement many intelligent programs by applying predefined procedures, which are similar to regular existing program works [2]. But, still there is a lag in monitoring many observations and instances in timely manner to overcome many societal challenges. Nowadays, very challenging tasks include photo tagging, identification of web-based ranking, identification of spam, or no spam Emails. To overcome these tasks or objectives, one of the options includes development of a program generating relevant rules to evaluate the data samples. It is also called training set, and one of the common emerging fields used for this is machine learning methods. Since 2010–2015, many intelligence software-based machine learning methods are applied including recognition systems on patient images to improve the accuracy results from 72% to 95% [5].

Most of the machine learning applications are evolving in present days and affecting every aspect in our daily lives. Machine learning is applicable in many emerging areas like healthcare monitoring systems, pattern recognition and feature extraction, text and speech recognition, education systems, military and defense applications, fraud detection, etc. Artificial intelligence takes the main lead in the development of ML technology systems. ML technology also simulates human learning systems from the input dataset or information. Many machine learning algorithms from firms such as Facebook, Amazon, or Flipkart are boosting the business trends in developing various brands [6]. With the help of past data or information, machine learning tries to discover new patterns in applying algorithms to achieve feasible outcome results. Also adding value to the business trends or organizations mainly focuses on monitoring future situations and outcome [7].

2. Related Work

A lot of research work is carried out using machine learning methods in achieving more accurate results and predicting outputs based on input dataset [8]. Machine learning plays a very important role in view of new trends and new techniques based on customer behavior or various input patterns, in the development of new products and new brands [9]. Enterprises can understand the customer needs at deeper level to overcome their needs using machine learning algorithms depending, for various applications, on their outcomes [10]. Machine learning also increases the importance in business operations and artificial intelligence is becoming practically high using today's ML models.

One of the new strategies for detecting cardiac diseases, mainly based on Co-Active Neuro-Fuzzy Interference Systems (CANFIS), is applied in one of the research work [11]. Most of the research study is based on regularity in detection of heart diseases based on their strategies as well as on their difficulties. Classifier strategies for the detection of heart diseases are demonstrated using machine learning algorithm, Naive Bayes classifier model. Most of the survey is carried out on various applications, in many research papers, by using data mining algorithms for prediction of heart diseases [12]. But traditional invasive-based approach is carried out using machine learning algorithms. The classifier models for diagnosing heart diseases are based on medical history of patients, patient test results, or scan results so that researchers or doctors can research on connected symptoms [13].

Alternatively, one more disadvantage is that the dye used is harmful as it affects kidneys, as it increases creatinine, including a high cost, a different kind of adverse effects, and a very good level of technological knowledge [14]. The traditional method is comparatively costly and also computationally intensive method for disease diagnosis which takes time to assess [15]. Researchers have tried to create various noninvasive smart healthcare systems which are based on predictive ML techniques, namely, SVM, K-NN, Naive Bayes (NB), and, also, decision tree (DT), among others, to overcome the challenges in conventional invasive-based methods for the identification of heart disease [16].

In the medical field, one of the most used classifiers is the decision tree. In this work [17] SEER medical datasets were used to predict the disease survivorship using classification and regression trees (CART).

In this work [18], use of neural networks was introduced to diagnose and forecast heart disease as well as blood pressure. A Deep Neural Network was built using the given disease attributes to generate an output that was accomplished by the output perceptron and almost included 120 hidden layers, which is the most basic and relevant method for ensuring an accurate result of having heart disease if the model is using the test dataset [19]. The use of a supervised algorithm for cardiac disease diagnostics is being recommended [17]. When the attributes of data are associated, the random forest approach has a tendency to favor the smaller group [20]. This is why, in order to alleviate the challenge of imbalanced data and limit the probability of bias against minorities in the dataset, the SMOTE method is being used. In this study [21], a combination of SMOTE and Artificial Neural Network (ANN) has been used to diagnose ovarian cancer using a publicly available dataset of ovarian cancer. The research demonstrates that, by using the preprocessing methodology of SMOTE to decrease the impact of data imbalance, we can improve the performance and efficiency of neural networks in cancer classification. On large datasets, most single classifier algorithms have the drawbacks of being computationally expensive and difficult. For large datasets, in particular, classification approaches do not give consistent and reliable results, making some individual classifier systems wasteful and unreliable [22]. For example, the DT approach is particularly good at managing intervariable interactions, but it struggles with linear relationships between variables [23].

In recent years, ensemble classifiers have become a popular strategy in machine learning and pattern recognition. In a nutshell, it is a method for combining the findings of many classifiers. The ensemble method's main goal is to improve classification efficiency by weighing multiple independent classifiers and thereby combining them into a single or an individual classifier that outperforms each one individually [22, 24, 25].

3. Exploratory Knowledge

One of the most well-known areas of medical research is the research for heart disease. Early identification and accurate projections of heart diseases have a significant impact on therapy and reduce patient mortality rates. The sections that follow provide brief descriptions of the algorithms used to detect heart disease in this study.

3.1. Decision Tree (DT) Classifier. A decision tree is a supervised ML algorithm that makes decisions based on a set of rules, very similar to how normally people do. A ML classification method is designed to make judgments, in one sense. Classification and regression problems can both be solved with this classifier [26].

There are different notions that define the model. They are given below.

(i) Entropy: Entropy is a measurement of a system's unpredictability or disorder. In the year 1850, a German physicist named Rudolf Clausius proposed this hypothesis. It is computed as shown in

$$\text{Entropy} = -\sum p(X)\log\, p(X), \tag{1}$$

where p(X) is a fraction of examples in a given class.

(ii) Gini Index: It is also called the Gini coefficient, which is a measure of income distribution in a population. Corrado Gini, an Italian statistician, created it in 1912. The Gini impurity is computed using

$$\text{Gini Inpurity} = 1 - \sum_{i=1}^{C} \left(p_i\right)^2. \tag{2}$$

(iii) Information Gain: The reduction in entropy achieved by changing a dataset is known as information gain, and it is frequently utilized in the training of decision trees. The entropy of a dataset before and after a transformation is used to calculate information gain. It is computed using

$$IG\left(D_p, f\right) = I\left(D_p\right) - \frac{N_{\text{left}}}{N} I\left(D_{\text{left}}\right) - \frac{N_{\text{right}}}{N} I\left(D_{\text{right}}\right), \tag{3}$$

where f is feature split on $D_p$which is parent dataset; D_{left} is left child node dataset;D_{right} is right child node dataset; I is impurity criterion; N is total number of samples; N_{left} is samples number of left child node; N_{right} is samples number of right child node.

3.2. The CART Algorithm. The CART algorithm was first introduced by Breiman et al. [27]. Hunt's algorithm is used to create the CART. To build a DT, it can process categorical as well as continuous attributes. It also accounts for missing data and constructs the DT by making use of Gini Index as an attribute selection criterion. CART divides the given datasets (training set) into binary segments and builds binary trees as a result. The Gini Index is not employed in the ID3 and C4.5 probabilistic assumptions. In order to increase accuracy of classification, CART algorithm increases the accuracy by making use of cost-complexity pruning for removing unpredictable branches from the DT.

3.3. Alternating Decision Tree (AltDTree). AltDTree is a classification ML method. It is related to boosting and generalizes decision trees. An AltDTree is made up of a series of decision nodes that indicates a predicate condition and prediction nodes that hold a single number [28]. Classic DTs, Voted DTs, and Voted Decision Stumps are all generalized into AltDTree. It allows any boosting implementation to extract the AltDTree model from the data as a learning method. In the context of the decision tree, AltDTree is an appealing extension of boosting. It enables the use of various boosting strategies to create an AltDTree model with unique properties that can handle a wide range of applications.

3.4. Random Forest (RF) Classifier. RF works by using the training data to create several decision trees. In the case of classification, every tree suggests output as a class; also the class with greatest number of outputs is selected as the final outcome [29]. In order to build, number of trees must be specified. RF is such a technique for aggregating or even bagging bootstrap data. This method is used to reduce an important parameter called variance in the outcomes.

3.5. Reduced Error Pruning Tree (RedEPTree). Top-down induction of decision trees has been observed to be hampered by the pruning phase's poor performance. It is known, for example, that the size of the resulting tree rises linearly with the sample size, despite the fact that the tree's accuracy does not improve. Errors are reduced. The RedEPTree technique is based on the notion of calculating information gain using entropy and backfitting to minimize variance-induced error [30].

3.6. Naive Bayes (NB) Classifier. There are two steps of classified data in the Naive Bayesian approach [31]. The first stage involves evaluating the parameters of a probability distribution using the training input data. In the second stage, the test dataset is categorized based on the greatest posterior probability. The NB classifier's pseudocode is shown below.

3.7. AdaBoost. AdaBoost makes it possible to merge various "weak classifiers" into a single classifier which is called "strong classifier." Decision trees with one level, or decision trees with only one split, are the most popular algorithm used with AdaBoost. Decision Stump is another name for these trees [32]. This approach creates a model by assigning equal weights to all of the data points. It then gives points that are incorrectly categorized with a higher weight. In the next model, all points with greater weights are given more importance. It will continue to train models till a lower error is received [33].

The weight of the training set is used to start the AdaBoost algorithm. Let us consider training set $(x_1, y), \ldots (x_n, y_n)$, in which each x_i is in instance space X and each label y_i is in collection of labels Y, that is very much similar to the collection of $\{-1, +1\}$. Weight on training instance I on the round t is assigned as $D_{\mathrm{It}}(i)$. At the start, the same weight is used $(D_{\mathrm{It}}(i)) = 1/\mathrm{M}$, $i = 1, \ldots, \mathrm{M})$, where It is the iteration number. Then, weight of the misclassified case from the base learning algorithm is then increased in each round. The AdaBoost algorithm's pseudocode is shown below.

And

$$\alpha_{\mathrm{It}} = \frac{1}{2}\mathrm{In}\left[\frac{P_{+1} - P_{-1}}{P_{-1} + P_{-1}}\right]. \tag{4}$$

C_{It} is the normalization constant, α_{It} is used to allow the outcome to be generalized and to solve the problem of overfitting and noise sensitive situations [33]. The real value of $\alpha_{It} h_{It}\ (x)$ is built using a class probability estimate (P).

4. Proposed Methodology

The proposed approach contains two phases in this section. SMOTE is used in the initial phase to lessen the impact of data imbalance. Then, the second phase entails classification using Naive Bayes and DT methods (AltDTree, CART, RedEPTree, and RF) [33]. After that, AdaBoost Ensembles of the aforementioned algorithms are constructed and their performance is evaluated. Then, heterogeneous classifiers that are formed by combining two different individual classifiers are evaluated against different performance metrics to figure out the best model. Figure 1 depicts the flow of the suggested technique.

4.1. Dataset. The UCI repository provided the Heart Disease dataset. This dataset comprises 13 medical variables for 304 patients, which helps to determine whether the patient is in the danger of developing heart disease or not, as well as categorize patients who are at risk and those who are not. The pattern that leads to the discovery of patients at risk for heart disease is retrieved from this dataset. There are two aspects to these records: training and testing. Each row corresponds to a single record in this dataset, which has 303 rows and 14 columns. Table 1 lists all of the qualities and the heatmap is depicted in Figure 2.

4.2. Data Preprocessing. Most classification algorithms aim to gather pure samples to learn and make the borderline of each class as definitive as possible in order to perform better prediction. Synthetic instances that are far from the borderline are easier to categorize than those that are near to the borderline, which present a significant learning difficulty for the majority of classifiers. The authors in [32] describe an advanced strategy (A-SMOTE) for preprocessing imbalanced training sets based on these findings. It aims to clearly characterize the borderline and create pure synthetic samples from SMOTE generalization. This approach is divided into two parts, as follows.

Step 1. The SMOTE technique is used to create a synthetic instance using

$$N = 2 * (r - z) + z, \tag{5}$$

where r denotes majority class samples, z denotes minority class samples number, and N is the initial synthetic instance number (which is newly generated).

The synthetic instances generated by SMOTE can be approved or rejected based on two criteria, which correspond to the first stage: For example, consider $\widehat{x} = \{\widehat{x}_1, \widehat{x}_2, \widehat{x}_3, \ldots . \widehat{x}_N\}$ which is the collection of new synthetic instances, and $\widehat{x}_i{}^{(j)}$ is the jth attribute value of $\widehat{x}_i$, $j \in\in [1, M]$.Let$S_m = \{S_{m1}, S_{m2}, \ldots S_{mz}\}$ and $S_\alpha = \{S_{\alpha 1}, S_{\alpha 1}, S_{\alpha 1}, \ldots . S_{\alpha r}\}$ be the set of the minority samples as well as majority samples [32]. In order to make the rejection or acceptance decision, distance is computed between $\widehat{x}_i$ and S_{mk}, D $D_{\text{minority}}\ (\widehat{x}_i, S_{mk})$ and the distance between $\widehat{x}_i$ and $S_{\alpha l}$, D $D_{\text{majority}}\ (\widehat{x}_i, S_{\alpha l})$. For I from N steps, we calculate the distances as stated below, using equations (6) and (7).

$$DD_{\text{minority}}\left(\widehat{x}_i, S_{mk}\right) = \sum_{j=1}^{M} \sqrt{\left(\widehat{x}_i{}^{(j)} - \widehat{S}_{mk}{}^{(j)}\right)^2}, k \in [1, z], \tag{6}$$

$$DD_{\text{minority}}\left(\widehat{x}_i, S_{al}\right) = \sum_{j=1}^{M} \sqrt{\left(\widehat{x}_i{}^{(j)} - \widehat{S}_{al}{}^{(j)}\right)^2}, l \in [1, r]. \tag{7}$$

As per (6) and (7), we compute arrays A_{minority} and A_{majority} using (8) and (9).

$$A_{minority} = \left(DD_{\text{min}\,orit\,y}\left(\widehat{x}_i, S_{m1}\right), \ldots DD_{\text{min}\,orit\,y}\left(\widehat{x}_i, S_{mz}\right)\right), \tag{8}$$

$$A_{ma\,jorit\,y} = \left(DD_{ma\,jorit\,y}\left(\widehat{x}_i, S_{a1}\right), \ldots DD_{ma\,jorit\,y}\left(\widehat{x}_i, S_{ar}\right)\right). \tag{9}$$

Then we choose the minimum value out of A_{minority}, $\min(A_{\text{minority}})$and the minimum value out of A_{majority}, $\min(A_{\text{majority}})$. If *min* (A_{minority}) is lesser than $\min(A_{\text{majority}})$, the new samples are accepted else, rejected.

$$\min(A_{\text{majority}}) < \min(A_{\text{majority}}) \text{ (Accepted)}.$$

```
Input required: TDS: Training Dataset TDS = $u_i\,(1, 2, 3, 4, \ldots\ldots n)$ Output expected: Class Labels YES and NO
Step_1: The Given Dataset TDS, consists of symptoms pertaining to different classes, say YES and NO
Step_2: Calculate prior-probability of "YES" class = No of attributes of class YES/Total no of attributes Compute prior-probability of
   "NO" class = No of attributes of class NO/Total no of attributes
Step_3: Compute $enum_i$, total no. of attributes that are frequent for each class
   $num_{\text{yes}}$ = Sum of frequent attributes of class YES
   $num_{\text{no}}$ = Sum of frequent attributes of class NO
Step_4: Compute the conditional probability
   $P_{(\text{attribute1/class YES})} = \text{attributeCount}/n_i\,(\text{YES})$
   $P_{(\text{attribute1/class NO})} = \text{attributeCount}/n_i\,(NO)$
   $P_{(\text{attribute2/class NO})} = \text{attributeCount}/n_i\,(\text{YES})$
   $P_{(\text{attribute2/class NO})} = \text{attributeCount}/n_i\,(\text{NO})$
     . . .
   $P_{(\text{attribute2/class NO})} = \text{attributeCount}/n_i\,(\text{NO})$
Step_5: Classify a new record of attributes of a patient based on the probability P (NEW/feature).
   Compute $P_{(\text{YES/attribute})} = P_{(\text{YES})} * P_{(\text{attribute1/class YES})} * P_{(\text{attribute1/class YES})} \cdots P_{(\text{attributen/class YES})}$
   Compute $P_{\text{NO/attribute}} = P_{(\text{NO})} * P_{(\text{attribute1/class NO})} * P_{(\text{attribute1/class NO})} \cdots P_{(\text{attributen/class NO})}$
Step_6: Assign the new record of patient to either class YES or NO which has higher probability.
```

Algorithm 1: Pseudocode of NB classifier.

```
    Input required: TDS: Training Dataset TDS = $x_i\,(1, 2, 3, 4, \ldots\ldots n)$ labels $y_i \in$ Y
    It: Iteration Number
    Steps a to h
(a)  Assign TDS sample $(x_1, y_1), \ldots (x_n, y_n)$; $x_i \in X$, $y_i \in \{-1, +1\}$
(b)  Initialize weights of $D_{\text{It}}(i) = 1/\text{M}$, $i = 1, \ldots,$ M
(c)  for It = 1, ... , T
(d)  Train the learner that is weak using distribution $D_{\text{It}}$
(e)  Get hypothesis of weak $h_{\text{It}}: X \longrightarrow \{-1, +1\}$ along with its error = $\varepsilon_{\text{It}} = \sum_{h_{\text{It}(x)} \neq y_i}^{n} D_{\text{It}}(i)$
(f)  Update distribution $D_{\text{It}}$: $D_{\text{It}+1}(i) = D_{\text{It}}(i) exp(-\alpha_{\text{It}} y_{\text{It}} T_{\text{It}}(x_{\text{It}}))/C_{It}$
(g)  Next It that, It + 1
(h)  Final hypothesis Outputs: $H_{(x)} = \text{sign}[\sum_{\text{It}=1}^{T} \alpha_{\text{It}} h_{\text{It}}]\ (x)]$
```

Algorithm 2: Pseudocode of AdaBoost classifier.

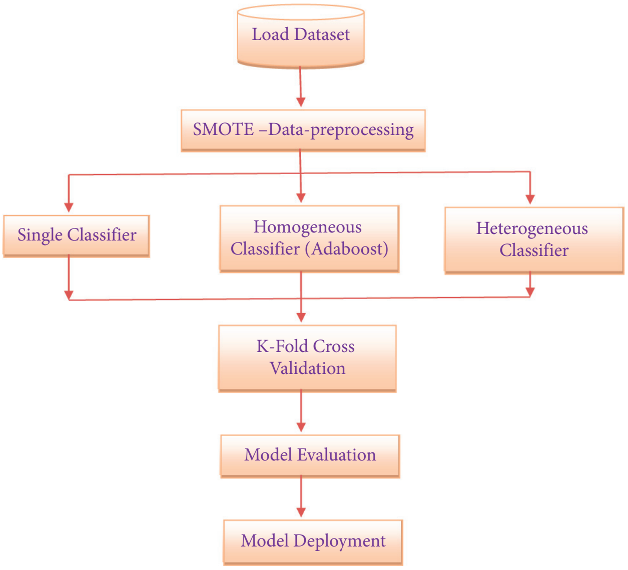

Figure 1: Proposed flow diagram.

Table 1: Attributes of the dataset.

Sl. No.	Features	Description	Values
1	Age	Age in years	Continuous
2	Sex	Gender of patient	Male/female
3	CP	Chest pain	Four types
4	Trestbps	Resting blood pressure	Continuous
5	Chol	Serum cholesterol	Continuous
6	FBS	Fasting blood sugar	<, or >120 mg/dl
7	Restecg	Resting electrocardiograph	Five values
8	Thalach	Maximum heart rate achieved	Continuous
9	Exang	Exercise induced angina	Yes/no
10	Oldpeak	ST depression when working out compared to the amount of rest taken	Continuous
11	Slope	Slope of peak exercise ST segment	Up/flat/down
12	Ca	Gives number of major vessels colored by fluoroscopy	0–3
13	Thal	Defect type	Reversible/fixed/normal
14	Num (disorder)	Heart disease	Not present ("NO")/present in the four major types ("YES")

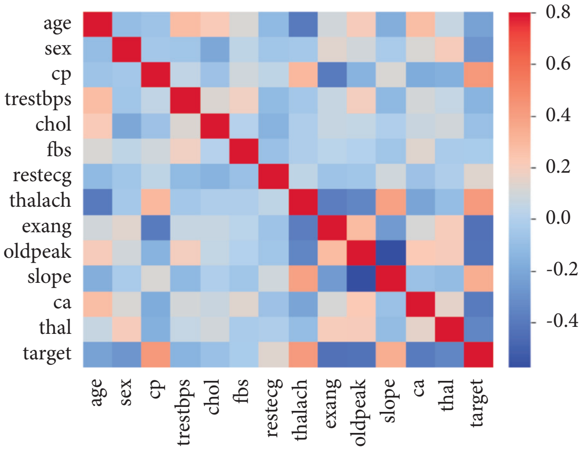

Figure 2: Heatmap depiction of the dataset.

$$\min(A_{\text{minority}}) \geq \min(A_{\text{majority}}) \text{ (Rejected).}$$

Step 2. Then, using the accepted synthetic instances, the following steps are taken to remove the noise.

Suppose $\widehat{S} = \{\widehat{S}_1, \widehat{S}_2, \widehat{S}_3, \ldots . \widehat{S}_n\}$ is a new synthetic minority received by Step 1. We then compute the distance between $\widehat{S}_i$ with each original minority $S_m, \mathrm{Min}_{\mathrm{Rap}}(\widehat{S}_i, \widehat{S}_m)$, defined using

$$S_m, \mathrm{Min}_{\mathrm{Rap}}\left(\widehat{S}_i\widehat{S}_m\right) = \sum_{k=1}^{z}\sum_{j=1}^{M}\sqrt{\left(\widehat{S}_i^{(j)} - S_{mk}^{(j)}\right)^2}, \qquad (10)$$

where

$S_m, \mathrm{Min}_{\mathrm{Rap}}(\widehat{S}_i.\widehat{S}_m)$ samples rapprochement including all minority and as per (10), L is obtained as follows:

$$L = \sum_{i=1}^{n}\left(\mathrm{Min}_{\mathrm{Rap}}\left(\widehat{S}_i, S_m\right)\right). \qquad (11)$$

Step 3. Compute the distance between $\widehat{S}_i$, and each original majority S_a, $Maj_{\mathrm{Rap}}(\widehat{S}_iS_a)$, described using

$$Maj_{\mathrm{Rap}}\left(\widehat{S}_iS_a\right) = \sum_{i=1}^{r}\sum_{j=1}^{M}\sqrt{\left(\widehat{S}_i^{(j)} - S_{al}^{(j)}\right)^2}. \qquad (12)$$

$Maj_{\mathrm{Rap}}(\widehat{S}_i, S_a) \longrightarrow$ samples rapprochement including all majority and as per equation (13) H is obtained as follows:

$$H = \sum_{i=1}^{n}\left(Maj_{\mathrm{Rap}}\left(\widehat{S}_i, S_a\right)\right). \qquad (13)$$

Then, we remove half of synthetic samples which have most likely less distance between $\widehat{S}_i$and S_a to obtain the data, that is, of high purity.

5. Performance Evaluations

The different ML algorithms, namely, Naive Bayes, AltD-Tree, RedEPTree, CART, and RF, are applied on the dataset as individual classifiers. Their performance is compared in terms of several metrics as described in the next section.

5.1. Performance Metrics. If the dataset is not balanced, accuracy may not be a good measure [34]. The number of accurately classified examples divided by total number of data instances is referred to as accuracy. The accuracy is computed using

$$\text{Accuracy} = \frac{TNs + TPs}{TNs + TPs + FPs + FNs}. \tag{14}$$

Precision is one of the performance metrics that is going to measure how many correct positive forecasts have been done. So, precision estimates the accuracy of the minority class; then, the ratio of accurately predicted positive instances divided by the total number of positive examples predicted, is used to compute it using

$$\text{Precision} = \frac{TPs}{TPs + FPs}. \tag{15}$$

A good classifier should have a precision of 100% (high); only when both numerator and denominator are identical, i.e., TP = TP + FP, can precision become 100% [33].

Recall is a metric that measures how many correct positive predictions were produced out of all possible positive predictions. Unlike precision, which only considers the right positive predictions out of all positive predictions, recall considers the positive predictions that were missed. In this approach, recall provides some indication of the positive class' coverage. The recall is computed using

$$\text{recall} = \frac{TPs}{TPs + FNs}. \tag{16}$$

We want both accuracy and recall to be of the value one in a good classifier, which also means FP and FN should be zero. As a result, we require a statistic that considers both precision and recall. The F1-score is a measure that takes precision and recall into account and is defined as follows:

$$F1\,\text{Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{17}$$

To compute error rates in forecasted value, let P^N denote a collection of data having the form (t_1, r_1) , (t_2, r_2),. . . (t_p, r_p)such that t_i denotes n-dimensional tuples of test with respective values of r_i for a given response r and denotes count of tuples in P^N.

In all test instances, the mean-absolute-error (MAE) is the mean of the difference among the projected and guanine value. It is the standard deviation of the prediction error calculated using

$$MAE = \sum_{i=1}^{p} \left| r_i - {r_i}^T \right|. \tag{18}$$

The root mean squared error (RMSE) is a well-known approach for calculating numeric prediction success. The mean of the squared discrepancies among every value is computed and its matching true value is used to calculate this value using

$$RMSE = \sqrt{\frac{\sum_{r=1}^{p} \left(r_i - {r^T}_i\right)^2}{p}}. \tag{19}$$

The total absolute mistake is made relative to what the error would have been if the prediction had just been the average of the actual numbers known as Relative Absolute Error (RAE). It is computed using

$$RAE = \frac{\sum_{r=1}^{p} \left(r_i - {r^T}_i\right)^2}{\sum_{r=1}^{p} \left(r_i - \overline{r}_i\right)^2}. \tag{20}$$

The total squared error made is compared to what the error would have been if the prediction had been the average of the absolute value, known as relative squared error (RRSE). It is computed using

$$RRSE = \sqrt{\frac{\sum_{r=1}^{p} \left(r_i - {r^T}_i\right)^2}{\sum_{r=1}^{p} \left(r_i - \overline{r}_i\right)^2}}. \tag{21}$$

Table 2 depicts that Random Forest is the best model as it takes only 2.27 seconds for model building (TTBM: Time to Build Model), while the AltDTree has taken 60.18 seconds for model building.

Figure 3 shows the accuracy forecast for individual classifiers. Among all the aforementioned classifiers being used in the current research work, AltDTree provides the best accuracy of 93.56%. Random Forest provides 92.45% accuracy and NB classifier prediction is the lowest with 78.67% accuracy.

Figure 4 depicts the rates of errors obtained from the individual classifiers. AltDTree MAE rate is 0.28 and RMSE rate value is 0.41. This demonstrates that there is low error recorded during the prediction procedures. However, NB has a higher error rate, i.e., 0.60 MAE and 0.83 RMSE, respectively.

Table 3 demonstrates that AdaBoost-RF is the best model, as it has taken only 10.34 seconds to build the model. But the AdaBoost-CART is the worst model as it takes 295.45 seconds to build the model. Also, AdaBoost-RF has highest F1-value of 0.98 and AdaBoost-NB has the lowest F1-value of 0.81.

From Figure 5, AdaBoost-RF predictions are better than any other mentioned classification algorithm with an accuracy of 95.47%. However, AdaBoost-AltDTree provides 93.56% prediction accuracy and stands second. The AdaBoost-NB provides the least prediction rate of 80.6%.

Figure 6 depicts the different error rates that were recorded. AdaBoost Ensemble classifiers provide the lowest error rate of 0.14 for MAE and 0.38 for RMSE. However, AdaBoost-NB has a higher error rate, i.e., 0.54 and 0.76 for MAE and RMSE, respectively, whose values are almost the same as that of NB individual classifier.

Table 4 depicts the results of ensemble classifiers which are heterogeneous in nature. RF-CART and RF-RedEPTree take 7.34 seconds and 7.89 seconds for building the model,

Table 2: Single classifier evaluation comparison.

Performance metrics	Naive Bayes	AltDTree	RF	RedEPTree	CART
TTBM (sec)	4.56	60.18	2.11	10.25	52.24
Accuracy (%)	78.6	93.56	92.45	79.23	78.67
MAE	0.60	0.28	0.27	026	0.27
RMSE	0.83	0.41	0.42	0.42	0.56
RAE	120	67.71	77.87	79.12	68.91
RRSE	127.41	95.33	82.92	97.89	98.34
F1-score	0.3	0.85	0.84	0.83	0.81

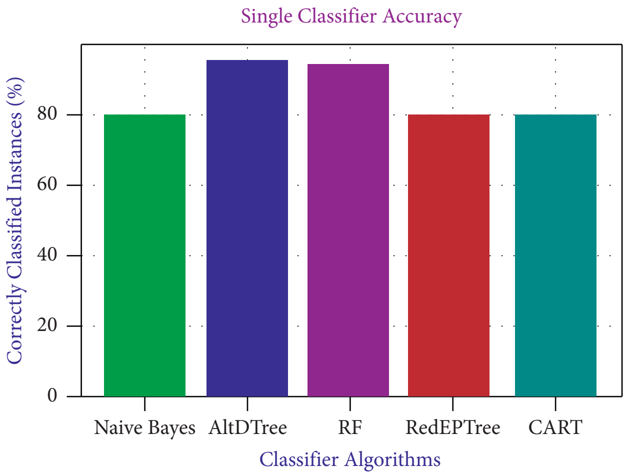

Figure 3: Accuracy prediction for single classifiers.

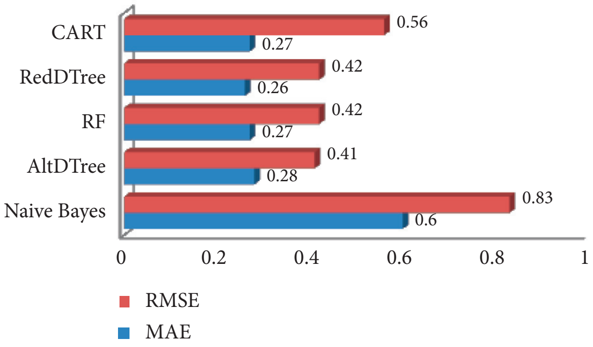

Figure 4: Error rates of individual classifier.

Table 3: AdaBoost classifier.

Performance metrics	AB-NB	AB-AltDTree	AB-RF	AB-RedEPTree	AB-CART
TTBM (sec)	18.32	30.01	10.34	64.35	295.45
Accuracy (%)	80.6	93.56	95.47	82.23	81.67
MAE	0.54	0.21	0.14	0.21	0.20
RMSE	0.76	0.43	0.38	0.41	0.41
RAE	129.79	57.78	35.87	45.19	41.61
RRSE	155.62	96.23	65.47	91.03	91.08
F1-score	0.81	0.94	0.98	0.83	0.87

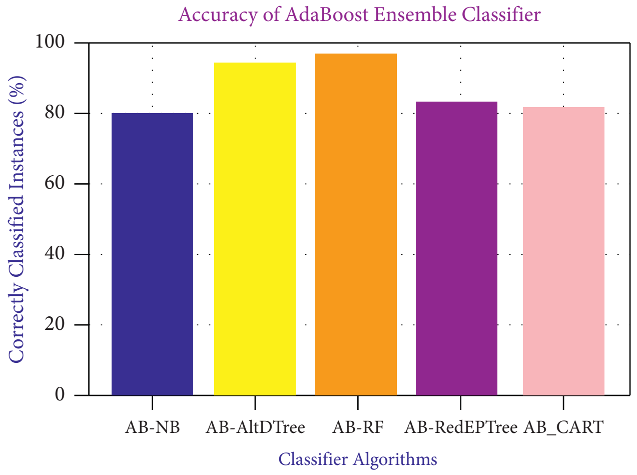

Figure 5: Accuracy of AdaBoost classifier.

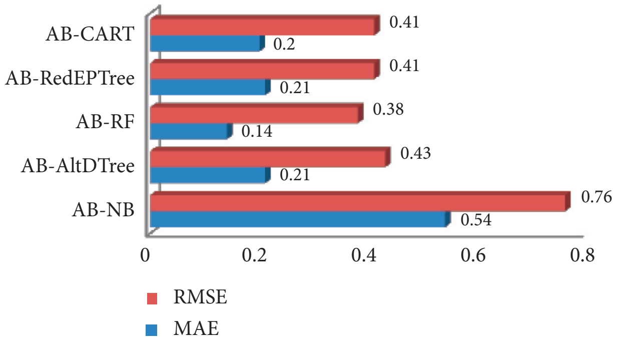

Figure 6: AdaBoost classifier error rate.

Table 4: Ensemble classifiers, heterogeneous.

Performance metrics	NB + AltDTree	NB + RF	AltDTree + RF	RF + RedEPTree	RF + CART	AltDTree + RedEPTree	AltDTree + CART
TTBM (sec)	30.03	32.05	398.12	7.89	7.34	357.77	598.02
Accuracy (%)	76.45	76.05	70.12	85.45	86.29	74.49	71.29
MAE	0.42	0.43	0.37	0.35	0.34	0.37	0.41
RMSE	0.42	0.39	0.49	0.36	0.36	0.37	0.42
RAE	99.23	92.23	80.12	71.01	70.89	73.23	89.23
RRSE	98.23	97.49	101.22	91.29	90.12	93.37	99.34
F1-score	0.74	0.75	0.68	0.84	0.85	0.73	0.69

respectively, which are very low. However, AltDTree-CART has taken 598.02 seconds being the worst time for building the model. So, it can be said that RF-CART has a higher F1-score of 0.85, and RF-RedEPTree is second with F1-score of 0.84. AltDTree-RF and AltDTree-CART have the worst F1-scores of 0.68 and 0.69, respectively.

From Figure 7, RF-CART provides the best accuracy of 86.29% in comparison to others, followed by RF-RedEPTree with 85.45% prediction accuracy. AltDTree-RF has the lowest accuracy value of 70.12%.

Figure 8 depicts error rates obtained by ensemble classifiers are heterogeneous in nature. RF-CART exhibits

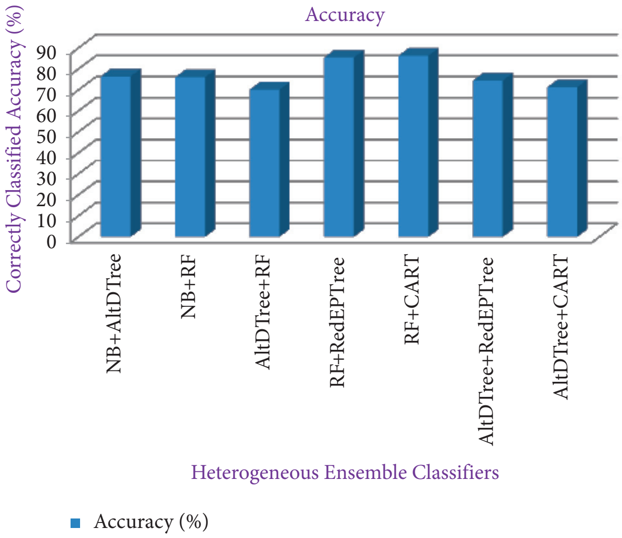

Figure 7: Accuracy of heterogeneous ensemble classifiers.

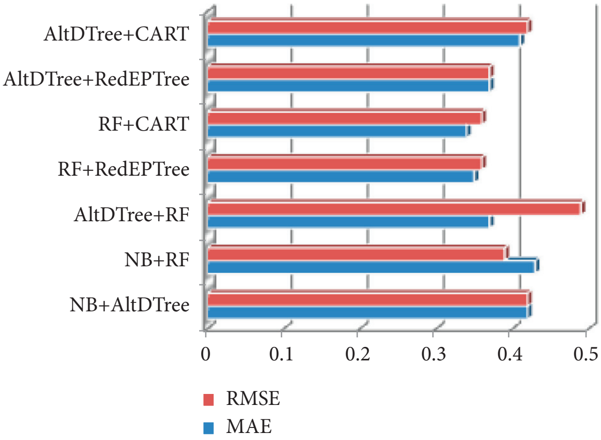

Figure 8: Error rates for heterogeneous ensemble classifiers.

the lowest error rate of 0.34 (MAE) and RMSE of 0.36. However, NB-RF has the highest MAE rate of 0.43 and AltDTree-RF has the highest RMSE rate of 0.49.

6. Conclusion

The AdaBoost Ensemble model for heart disease prediction has been proposed in this work, which is based on recognized feature patterns. In the diagnosis of cardiac disease, it can be compared with classic data mining methods. Ensemble classification approaches replace traditional methods of extracting meaningful information during the feature extraction step. The homogeneous classifiers and ensemble classifiers which are formed by combining multiple methods called heterogeneous classifiers were employed in this study. The data mining preprocessing technique using Synthetic Minority Oversampling Technique (SMOTE) is used to cope with the problem of class imbalance as well as noise present in the heart disease dataset. The best time to build the model for heterogeneous ensemble classifiers is 7.34 seconds for RF-CART and 7.89 seconds for RF-RedEPTree ensemble, according to the experimental results. NB-AltDTree has been observed to have taken the worst time of 598.02 seconds to build the model. With 86.29% prediction accuracy, RF-CART outperforms other classification algorithms, followed by RF-RedEPTree with 85.45% prediction accuracy. As per the results, AdaBoost-RF classifier exhibits 0.14 error rate for MAE which is the lowest and 0.38 for RMSE among the other AdaBoost Ensemble classifiers. In all the overall experiments, the performances of classifiers were compared, and the findings revealed that AdaBoost-RF is the best among other classifiers with 95.47% accuracy.

Data Availability

The [UCI repository] data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] World Health Organization, *Cardiovascular Diseases*, WHO, Geneva, Switzerland, 2020, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

[2] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou et al., “Forecasting the future of cardiovascular disease in the United States,” *Circulation*, vol. 123, no. 8, pp. 933–944, 2011 Mar 1.

[3] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Predictive data mining for medical diagnosis: an overview of heart disease prediction,” *International Journal of Computer Application*, vol. 17, no. 8, pp. 43–48, 2011.

[4] S. H. Jee, Y. Jang, D. J. Oh et al., “A coronary heart disease prediction model: the Korean Heart Study,” *BMJ Open*, vol. 4, no. 5, e005025 pages, 2014.

[5] The Economist, *From Not Working to Neural Networking*, http://www.economist.com/news/specialreport/21700756-artificialintelligence-boom-based-old-idea-modern-twist-not, 2016.

[6] S. Ben-David and S. Shalev-Shwartz, *Understanding Machine Learning," from Theory To Algorithms*, Cambridge University Press, Cambridge, UK, 2020.

[7] M. P. M., D. S. Bote and S. D. Deshmukh, "Heart disease prediction system using naive Bayes," *Int J. Enhanced Res. Sci. Technol. Eng*, vol. 2, no. 3, 2013.

[8] A. Ganna, P. K. E. Magnusson, N. L. Pedersen et al., "Multilocus genetic risk scores for coronary heart disease prediction," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 33, no. 9, pp. 2267–2272, 2013.

[9] American Heart Association, *Heart Failure*, American Heart Association, Chicago, IL, USA, 2020, https://www.heart.org/en/health-topics/heart-failure.

[10] A. Dandapath and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering & Technology*, vol. 7, no. 2, pp. 684–687, 2018.

[11] J. Soni, S. Soni, D. Sharma, and U. Ansari, "Intelligent and effective heart disease prediction system using weighted associative classifiers," *International Journal on Computer Science and Engineering*, vol. 3, no. 6, pp. 2385–2392, 2011.

[12] R. Subramanian and L. Parthiban, "Intelligent heart disease prediction system using CANFIS and genetic algorithm," *International Journal of Biological, Biomedical and Medical Sciences*, vol. 3, no. 3, 2008.

[13] O. W. Samuel, A. K. Asogbon, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_ AHP for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, pp. 163–172, 2017.

[14] Y. Kumaraswamy and S. B. Patil, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *European Journal of Scientific Research*, vol. 31, pp. 642–656, 2009.

[15] J. Singaraju and K. Vanisree, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks," *International Journal of Computer Application*, vol. 19, pp. 6–12, 2015.

[16] B. Edmonds, *Proceedings of AISB Symposium on Socially Inspired Computing*, pp. 1–12, Hatfield, 2005.

[17] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005 Jun.

[18] J. Y.. , S. Kiyasu, *Patent No. 4,338*, p. 396, U.S. Patent and Trademark Office, Washington, DC, 1982.

[19] M. Raihan, S. Mondal, A. More et al., "Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design," in *Proceedings of the 19th International Conference on Computer and Information Technology (ICCIT)*, pp. 299–303, IEEE, Dhaka, Bangladesh, December 2016.

[20] L. Tolo and T. Lengauer, "Classification with correlated features: un- reliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, p. 1986, 1994.

[21] A. Hambali Moshood and D. Gbolagade Morufat, "Ovarian cancer classification using hybrid synthetic minority oversampling technique and neural network," *Journal of Advances in Computer Research (JACR)*, vol. 7, no. 4, pp. 109–124, 2016.

[22] J. Vandar Kuzhali and S. Vengataasalam, "A novel ensemble classifier based classification on large datasets with hybrid feature selection approach," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 7, no. 17, pp. 3633–3642, 2014.

[23] D. Lavanya, R. K. Usha, Ensemble decision tree classifier for breast cancer data," *International Journal of Information Technology and Computer Science*, vol. 2, no. 1, pp. 17–24, 2012.

[24] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, no. 2, pp. 135–148, 2002.

[25] R. Lior, "Ensemble-based classifiers. Artificial intelligence," *Review*, vol. 33, pp. 1–39, 2010.

[26] J. Leskovec and A. Grover, "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 2016.

[27] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, C. A, 1984.

[28] K. Sharma, T. R. Mahesh, and J. Bhuvana, "Big data technology for developing learning resources," *Journal of Physics: Conference Series*, Conference Series IOP Publishing Ltd, vol. 1979, no. 1, p. 012019, May 2021.

[29] M. R. Sarveshvar, A. Gogoi, A. K. Chaubey, S. Rohit, and T. R. Mahesh, "Performance of different machine learning techniques for the prediction of heart diseases," in *Proceedings of the International Conference on Forensics, Analytics, Big Data, Security (FABS)*, pp. 1–4, Bengaluru, India, December 2021.

[30] I. H. Witten and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," *The United States of America, Morgan Kaufmann Series in Data Management Systems*, 2nd Edition, 2005.

[31] H. K. Shashikala, T. R. Mahesh, V. Vivek, M. G. Sindhu, C. Saravanan, and T. Z. Baig, "Early detection of spondylosis using point-based image processing techniques," in *Proceedings of the 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, pp. 655–659, Bangalore, India, August 2021.

[32] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: a new pre-processing approach for highly imbalanced datasets by improving SMOTE international journal of computational intelligence systems," vol. 12, no. 2, p. 1412, 2019.

[33] P. Chaitanya Reddy, R. M. S. Chandra, P. Vadiraj, M. Ayyappa Reddy, T. R. Mahesh, and G. Sindhu Madhuri, "Detection of plant leaf-based diseases using machine learning approach," in *Proceedings of the 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–4, Bangalore, India, December 2021.