Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.035817
Article

Tech Science Press

Classifying Big Medical Data through Bootstrap Decision Forest Using Penalizing Attributes

V. Gowri[1,*] and V. Vijaya Chamundeeswari[2]

[1]Department of CSE, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamilnadu, India
[2]Department of CSE, Saveetha Engineering College, Chennai, Tamilnadu, India
*Corresponding Author: V. Gowri. Email: gowrimurthy83@gmail.com
Received: 05 September 2022; Accepted: 14 November 2022

Abstract: Decision forest is a well-renowned machine learning technique to address the detection and prediction problems related to clinical data. But, the traditional decision forest (DF) algorithms have lower classification accuracy and cannot handle high-dimensional feature space effectively. In this work, we propose a bootstrap decision forest using penalizing attributes (BFPA) algorithm to predict heart disease with higher accuracy. This work integrates a significance-based attribute selection (SAS) algorithm with the BFPA classifier to improve the performance of the diagnostic system in identifying cardiac illness. The proposed SAS algorithm is used to determine the correlation among attributes and to select the optimum subset of feature space for learning and testing processes. BFPA selects the optimal number of learning and testing data points as well as the density of trees in the forest to realize higher prediction accuracy in classifying imbalanced datasets effectively. The effectiveness of the developed classifier is cautiously verified on the real-world database (i.e., Heart disease dataset from UCI repository) by relating its enactment with many advanced approaches with respect to the accuracy, sensitivity, specificity, precision, and intersection over-union (IoU). The empirical results demonstrate that the intended classification approach outdoes other approaches with superior enactment regarding the accuracy, precision, sensitivity, specificity, and IoU of 94.7%, 99.2%, 90.1%, 91.1%, and 90.4%, correspondingly. Additionally, we carry out Wilcoxon's rank-sum test to determine whether our proposed classifier with feature selection method enables a noteworthy enhancement related to other classifiers or not. From the experimental results, we can conclude that the integration of SAS and BFPA outperforms other classifiers recently reported in the literature.

Keywords: Data classification; decision forest; feature selection; healthcare data; heart disease prediction; penalizing attributes

1 Introduction

The explosion of heterogeneous clinical records is increasing exponentially due to the digitization of the medical industry. The massive and devastating volumes of healthcare data and unremitting flow of colossal

records make big data processing a more critical and imperative process. Medical data processing provides a real-time analytic capability for disease surveillance and accurate assessment to help healthcare specialists make the appropriate clinical decision during emergencies [1]. The provision of better quality medical facilities that are reasonably priced to patients is a serious issue in the healthcare industry. The provision of a better medical facility needs both accurate diagnosis and identification of appropriate therapy for victims while circumventing erroneous diagnoses.

Cardiovascular disease is a deadly illness that causes possibly life-threatening problems including cardiac arrests. The WHO (world health organization) lists heart attack as the primary cause of fatality worldwide and accounts for about 30% of total deaths (around 17.9 million individuals) per year [2]. This condition is critically more solemn in middle- and low-income countries [3]. In most countries, there is a deficiency of cardiac specialists and a substantial amount of inaccurately identified cases. This could be solved by implementing an effective and precise cardiac illness classification and decision-making system using electronic health records [4,5]. The timely identification of heart disease helps medical professionals create appropriate verdicts about patients' health to avert them from further harm. Moreover, timely recognition of the cardiac disease is an essential factor to reduce mortality and cost.

Of late, we have witnessed the utilization of non-invasive clinical methods including a smart computational system to develop a diagnosis system for timely identification of cardiac illness at a low. Different classification models employ machine learning (ML) algorithms to classify the clinical records into normal (negative) and diseased (positive) accurately according to the prevailing information [6]. Generally, ML classifiers are trained using a colossal volume of raw electronic health records. ML classification algorithms including Naïve Bayes (NB), K-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), decision tree (DT), random forest (RF), and adaptive boosting method (ABM) perform a vital role in classifying heart diseases [7]. Nevertheless, the classification of the cardiac problem is a very challenging endeavor. The growing size of health databases has made it a complex process for physicians to understand the correlations among attributes. According to data science, strong ML algorithms might be beneficial for mining patient information and conducting a rapid, lucrative study on huge clinical data. In the learning phase, ML algorithms need a large number of data samples to evade overfitting problems [8]. But, adding more attributes is not obligatory due to the curse of dimensionality [9].

To handle the high-dimensional datasets, only the attributes which are meticulously associated with the output class should be selected and given as inputs to ML techniques. Feature (attribute) selection is an important phase in healthcare informatics for finding an optimal subset overhead of attribute space and solving problems related to the imbalanced dataset to improve the effectiveness of the classification algorithms. By selecting the best attribute, we can select the most significant disease risk factors. Most significant attributes selection aids to eliminate redundant and irrelevant attributes, which enables fast and improved classification performance. In view of that, this study aims to select the most significant risk factors from a high-dimensional database that aids to predict cardiovascular problems with improved classification accuracy. The key contributions of this study are four-fold.

1. To realize an efficient and correct prediction of heart disease, we propose a diagnostic model that takes advantage of the attribute selection approach and a DF classifier.
2. To decrease the size of the attribute space, we propose a significance-based attribute selection algorithm to calculate the relationship among attribute pairs. The proposed SAS algorithm implements Enhanced RelieF (ERelief) algorithm to compute the quality (merit) of features. Then, we rank the attributes based on this merit and we use top-ranked attributes for learning and testing processes.
3. We propose a BFPA using penalizing attributes classifier to classify the given clinical record of a patient into normal or heart disease with higher accuracy.

4. The combined BFPA-SAS algorithm is implemented and the results are compared with other ML-based classification models in terms of performance measures on a workbench using a real-world database viz., the Heart disease dataset.

The following sections of this manuscript are organized as follows. In Section 2, we explore related cardiac disease classification methods. Section 3 discusses the intended classifier in detail. Sections 4 and 5 discuss the experimental arrangement and outcomes achieved from a heart disease dataset. We analyze the effectiveness of the proposed classifier by relating its performance with other germane classification approaches. Finally, we conclude this work in Section 6.

2 Literature Survey

Several researchers have suggested various diagnostic models using ML techniques to identify heart disease. Le et al. proposed an automated cardiac disease classification model by applying an attribute selection and data processing approach [10]. In this technique, features of heart problems are selected using their merit and weights are allotted through the infinite latent attribute selection technique. The authors used an SVM classifier to differentiate a set of the designated features into various types of heart disease. Ahmad et al. proposed an effective cardiac disease classification model using an ML algorithm [11]. The authors employed mixed ML approach for effective classification of heart disease by recognizing the categorical and numerical attributes.

Jackins et al. developed a smart cardiac problem prediction model using NB and RF classifiers to test whether a person is affected by a cardiac problem or not [12]. The authors analyzed the effectiveness of the intended system for NB and RF classification algorithms individually. Maheswari and Pitchai proposed a heart disease prediction system using DT [13]. Ayon et al. used different computational intelligence techniques including SVM, RF, LR, KNN, DT, NB, and deep neural networks on different real-world datasets such as Cleveland and Statlog heart disease datasets with numerous assessments methods [14]. Amarbayasgalan et al. developed an effective classification approach for coronary heart disease risk using two deep neural networks trained on efficient databases [15]. The authors used a two-step process to organize the learning database: (1) divide the initial learning database into two groups, commonly distributed and highly biased using principal component analysis, (2) improve the highly biased group by variational autoencoders. Then, two deep neural network classification algorithms are trained by learning data samples individually. In the classification phase, a neural network-based classification algorithm is employed to identify the cardiac illness by determining the relationship between attributes and the target class.

Mohan et al. proposed an efficient cardiac problem classification model using a hybrid ML technique called hybrid RF linear model (HRFLM) to identify cardiovascular disease on the UCI Cleveland dataset [16]. The HRFLM exploits ANN with backpropagation. Also, it calculates and applies DT entropy for selecting appropriate features. Absar et al. proposed four ML methods including DT, RF, KNN, and ABM to identify cardiac problems [17]. A comprehensive algorithm is used to examine the potential of the germane parameters that participate in identifying the cardiac problem. Also, the study utilized the cloud computing paradigm, Streamlit to design an automatic and intelligent model for disease classification. This study demonstrates that the ML techniques will perform a vital role in classifying heart disease in a more expedient way. Almustafa carried out a comparative study of different ML-based prediction methods including KNN, NB, DT, SVM, ABM, and stochastic gradient descent used to predict cardiac disease from the Heart Disease dataset with a minimum number of attributes [18]. Albeit the classification performance of these heart disease diagnostic systems is enhanced using the aforementioned ML algorithms, the accuracy, and other computational overheads are still deprived. This motivates us to develop a new classifier to classify medical data into normal and diseased accurately. The proposed

model exploits an attribute selection approach by calculating correlation among features and eliminates unnecessary features based on the rank of the features. A BFPA is proposed to classify the data sample into normal or diseased with higher classification accuracy.

3 Description of Proposed Method

This research develops and implements a BFPA to identify heart disease in online mode with an optimum number of decision trees. This work integrates a significance-based attribute selection approach with the BFPA algorithm to improve the performance of the diagnostic system. SAS algorithm is used to determine the significance of attributes by applying the concept of correlation between attributes and to select the optimum subset for learning and testing processes. BFPA selects an optimal number of samples in learning and testing datasets as well as the number of trees in the forest so as to realize higher prediction accuracy in classifying imbalanced and multi-class datasets effectively.

3.1 Significance-Based Attribute Selection

Feature engineering is the process of identifying optimal attributes as well as eliminating insignificant and redundant information from the dataset. Thus, this process reduces the size of the feature space and enables learners to run fast and proficiently. Sometimes, the performance of the classifier can be improved; in others, the feature selection process provides a more compact result and consequently, the interpretation of the target class is easily achieved. This study attempts to select optimal features by applying the concept of correlation between attributes (i.e., a degree of reliance or predictability of one parameter with another). The following sections describe the effectiveness of SAS in cardiac disease identification under various conditions and show that SAS can excerpt suitable as well as most appropriate features for BFPA classifiers employed in this work.

3.1.1 Estimation of Significance

The input to any ML method is a set of n training data samples. Each record in dataset $\mathcal{D}$ contains a set of features such as $\mathcal{D}_1 \times \mathcal{D}_2 \times \ldots \mathcal{D}_m$ where $\mathcal{D}_i$ is the domain of the i^{th} feature. Generally, the training dataset is defined as a tuple $\langle \mathcal{D}, \mathcal{L} \rangle$ where L is the output (class label). For a particular dataset, the term d_i represents the value of feature $\mathcal{D}_i$. The induction algorithm is used to create a DT for a particular dataset and it is possible to compute the target class appropriately. This study sets a probability ρ on the feature space $\mathcal{D}_1 \times \mathcal{D}_2 \times \ldots \mathcal{D}_m \times L$. In other words, a feature is useful if it is predictive of or related to the class; or else it is irrelevant. Kohavi and John formulate the following definitions to describe the characteristics of the attributes [19].

Definition 1: A feature $\mathcal{D}_i$ is considered significant to a target output L if $\mathcal{D}_i$ exists in every Boolean expression that signifies L and insignificant otherwise.

Definition 2: A feature $\mathcal{D}_i$ is significant if and only if there is some d_i and l for which $\rho(\mathcal{D}_i = d_i)$ is a non-negative value as given in Eq. (1).

$$\rho(L = l | \mathcal{D}_i = d_i) \neq \rho(L = l) \tag{1}$$

The feature $\mathcal{D}_i$ in Eq. (1) is significant if it impacts the value of class label L (i.e., L is dependent on $\mathcal{D}_i$). It is noteworthy that this statement fails to indicate the significance of attributes in the notion of parity. However, it may be improved as follows. Consider $\mathcal{E}_i$ is the set of all attributes excluding $\mathcal{D}_i$, i.e., $\mathcal{E}_i = \{\mathcal{D}_1, .., \mathcal{D}_{i-1}, \mathcal{D}_{i-2} \ \ldots..\mathcal{D}_m\}$. The term e_i signifies a value assigned to each feature in $\mathcal{E}_i$.

Definition 3: A feature $\mathcal{D}_i$ is relevant if and only if there exists some d_i, l, and e_i for which $\rho(\mathcal{D}_i = d_i)$ is a non-negative number. That is, $\mathcal{D}_i$ is significant if the likelihood of the label (for each attribute) can fluctuate when the value of $\mathcal{D}_i$ is excluded as given in Eq. (2).

$$\rho(L = l|\mathcal{E}_i = e_i|\mathcal{D}_i = d_i) \neq \rho(L = l,\ \mathcal{E}_i = e_i) \tag{2}$$

Definition 4: A feature $\mathcal{D}_i$ is relevant if and only if there is some d_i, l, and s_i for $p(\mathcal{D}_i = d_i, \mathcal{E}_i = e_i)$ is a non-negative value as given in Eq. (3).

$$\rho(L = l|\mathcal{D}_i = d_i,\ \mathcal{E}_i = e_i) \neq \rho(L = l,\ \mathcal{E}_i = e_i) \tag{3}$$

The following instance exemplifies all these definitions. Let features $\mathcal{D}_1, \mathcal{D}_2, \ldots .\mathcal{D}_5$ be Boolean. The attribute space is in such a manner that $\mathcal{D}_2$ and $\mathcal{D}_3$ are unfavorably correlated to $\mathcal{D}_4$ and $\mathcal{D}_5$, correspondingly, i.e., $\mathcal{D}_4 = \overline{\mathcal{D}_2}$ and $\mathcal{D}_5 = \overline{\mathcal{D}_3}$. There are only eight possible combinations, and they are considered equiprobable. The class label (deterministic) is calculated as given in Eq. (4).

$$L = \mathcal{D}_1 \oplus \mathcal{D}_2 \tag{4}$$

where $\oplus$ symbol signifies XOR function. It is noteworthy that the class label has an equivalent Boolean function, viz., $L = \mathcal{D}_1 \oplus \overline{\mathcal{D}_4}$. The features $\mathcal{D}_3$ and $\mathcal{D}_5$ are irrelevant in the robust possible sense. $\mathcal{D}_1$ is significant, and one of $\mathcal{D}_2$, $\mathcal{D}_4$ can be excluded, but we must have one of them. The significant and trivial attributes of each definition are shown in Table 1. From Definition 1, the features $\mathcal{D}_3$ and $\mathcal{D}_5$ as well as $\mathcal{D}_2$ and $\mathcal{D}_4$ are trivial to each other since each one can be nullified by the other. From Definition 2, each attribute is trivial since, for any label l and feature significance d, two samples agree with the values. From Definition 3, each attribute is important, as calculating its value alters the likelihood of 4 out of the 8 probable combinations from $\frac{1}{8}$ to zero. From Definition 4, the features $\mathcal{D}_3$ and $\mathcal{D}_5$ as well as $\mathcal{D}_2$ and $\mathcal{D}_4$ are irrelevant as they do not provide any useful statistics to $\mathcal{E}_4$ and $\mathcal{E}_2$, correspondingly.

Table 1: Significant and trivial attributes

Significant attributes	Trivial attributes	Definition
$\mathcal{D}_1$	$\mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5$	1
–	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5$	2
$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5$	–	3
$\mathcal{D}_1$	$\mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5$	4

Albeit the weak confrontational correlations are implausible to occur, domain restrictions provide a similar result. When a trivial attribute is coded as input to a classification algorithm, it is common to exploit a local coding, where all the values are denoted by corresponding pointer variables. Definition 4 defines the most significant features. The most significant feature denotes that the feature is indispensable since it cannot be excluded without degrading the classification performance.

Definition 5: (Trivial or Least significant): A feature $\mathcal{D}_i$ is trivial if and only if it is not the most significant, and there is a subsection of attributes $\mathcal{E}'_i$ of $\mathcal{E}_i$ and there is some d_i, l, and e'_i with $p\left(\mathcal{D}_i = d_i, \mathcal{E}'_i = e'_i\right)$ is a non-negative number as given in Eq. (5).

$$p\left(L = l\middle|\mathcal{D}_i = d_i, \mathcal{E}'_i = e'_i\right) \neq p\left(L = l\middle|\mathcal{E}'_i = e'_i\right) \tag{5}$$

Trivial features specify that the feature can rarely participate in predicting the class of the test sample. Attributes are significant if they are either most or least significant, and are insignificant or else. Insignificant attributes can never participate in predicting the class label. In the above-mentioned example, feature $\mathcal{D}_1$ is the most significant; features $\mathcal{D}_2$ and $\mathcal{D}_4$ are trivial; and $\mathcal{D}_3$ and $\mathcal{D}_5$ are irrelevant. If the correlation between

the features in a data sample and the external parameter (k) is computed and the association between each pair of features is given, then the correlation between combined samples having the aggregated features and k can be calculated from Eq. (6).

$$A_{kc} = \frac{k\overline{A_{ki}}}{\sqrt{k + k(k-1)\overline{A_{ii}}}} \tag{6}$$

here i, k and c represent the component, external parameter, and composite variable, correspondingly. The term A_{kc} represents the correlation among the aggregated features and k is the number of attributes, and A_{ii} is the average value of correlation between features. Eq. (6) provides Pearson's relationship factor in which all parameters have been normalized. It divulges that the relationship between an external parameter and a composite is a function of k and the value of the correlation amongst them, along with the value of the relationships between the features. Eq. (6) is solved by utilizing two descriptive values for A_{ki} and permitting the values of k and A_{ii} to modify.

From this Eq. (6), we can draw the following conclusions: (i) the higher correlation between the features and k increases the correlation between k and the composite; (ii) the lower the correlation among the features reduces the value of the relationship between k and the composite; and (iii) as the number of features in the composite upturns (assuming the additional features are identical to the original features in terms of their mean value of correlation with the other features and with k), the correlation between k and the composite increases. In this study, Eq. (6) is used as a heuristic estimate of the quality of feature subsections in the data classification. In this case, k becomes L; the problem remaining is to determine suitable techniques for computing the feature-class correlation and feature-feature correlation.

3.1.2 Attribute Discretization

Generally, ML techniques comprise different types of features such as ordinal, nominal, continuous, and binary. To have a common base for calculating the correlation, it is essential to have a common approach for dealing with different forms of features. Discretization is used as a preprocessing task to convert continuous features into nominal. The redundant attributes should be excluded to achieve effective classification. More precisely, if the classification performance of a particular feature is dependent on another feature then it can be eliminated. Several ML methods implement this feature elimination process to improve predictive performance. In data mining applications, where whole outputs are of the highest relevance, it is not always clear that redundant features should be eliminated. For example, a rule may give more sense to a user if an attribute is replaced by a most significant one. Our SAS approach accepts this condition by providing a report generation capacity. For a particular attribute in the final subset, SAS can list its close alternatives; either about the total quality of the final subset of the attribute was to be replaced by one of the alternatives or only correlated to the attribute.

3.1.3 Measuring the Quality Index of Attributes

When all features and the target labels are treated in the same way, the feature-class correlation and feature-feature correlation in Eq. (6) is calculated. Study on DF construction has given different methods for computing the merit of an attribute (i.e., how predictive one attribute is of another). Attribute quality measure define the uncertainty that presents in the set of data related to the values of a particular attribute. Hence, they are known as impure data samples [20]. A set of data is called pure if each data is the same with respect to the value of other attribute; the set of data is impure (to some degree) if it varies with the value of the other attribute. DF induction frequently defines in what way the predictive attributes are related to the target label. This is based on correlation between the feature and the class in Eq. (6). To estimate the quality of a feature, the feature-feature correlation is computed. Since DF executes a greedy simple-to-complex hill-climbing search, their complete inductive bias is used to construct compact trees

over bigger ones. One aspect that can impact both the dimension of the trees in the forest and in what way it generalizes to new inputs is the bias inherent in the attribute merit. This is used to select optimal attributes to check at the nodes of the tree.

Some quality metrics are selected to illegitimately favor attributes with higher values over those with smaller values. This can cause the generation of larger trees that may overfit the training dataset and poor generalization. Similarly, if such metrics are used as the relationship in Eq. (6), a feature set encompassing feature with more values may be designated—a situation that could cause poor performance by a DF model if it is regulated by such a subset. In a research, Kononenko analyzed the biases of measures for computing the merit of features [21]. For the correlation among two features, an estimate is essential that defines the analytical skill of one feature for another and vice versa. This works implements an algorithm known as Enhanced RelieF (ERelief) to calculate the merit of attributes. ERelief is a new version of the Relief algorithm.

3.1.4 Identification of Significant Attributes

The Relief algorithm computes an alternative value for each attribute that can be utilized to estimates feature merit (or quality or significance) to the class label (i.e., output) [22]. The conventional Relief computes the merit of attribute through a KNN algorithm that selects neighbors (samples) using the attribute value.

Algorithm 1: ERelief algorithm

Input: attributes ($\mathcal{D}_i$), number of top-ranked attributes (n) required,

Output: quality of each attribute

for i = 1 to n do

Execute Relief algorithm and compute feature merit

Sort features based on quality

Exclude features with the lower merits

end for

return last Relief merit of remaining attributes

The merit of each feature is computed based on whether the nearest neighbor (nearest hit, H) of a randomly selected sample from the same label and the nearest from the other label (nearest miss, M) have identical or diverse values. This procedure of adapting weights is repeated for m data points. But, the conventional Relief approach is restricted to handling binary classification applications and had no technique to manage missing values. Therefore, effective methods are required to solve this issue. This work implements an ERelief algorithm to compute the merit of features in the SAS process. The proposed ERelief uses a much weaker iterative method that can simply be used in any other basic relief-based algorithms. ERelief is a recursive feature exclusion technique. In each iteration, the attributes with lower merit are eliminated from further computation based on both feature weight updates and distance calculations.

3.2 Bootstrap Decision Forest Using Penalizing Attributes

After selecting the optimal features using the SAS algorithm, the DF is constructed, which is demonstrated to be more accurate and fast in producing prediction results. A new classifier called BFPA is presented in this study. It has the facility for penalizing attributes in the forest creation process. Contrasting some traditional classifiers found in the literature, BFPA uses a part of the non-class

attributes. This decision forest creates a set of trees with better accuracy by taking the advantage of the power of all non-class attributes present in a database. The BFPA employs the classification and regression trees (CART) as an effective decision-making algorithm. Analogous to the RF classifier, the proposed BFPA generates bootstrap samples b from the initial training dataset T. The trees are constructed upon these datasets in which the quality score (θ) of the attributes defines the optimum point of partition as defined in Eq. (7).

$$\theta = \delta_i \times \eta_i \tag{7}$$

here δ_i represents the classification ability and η_i indicates the weights of attributes. The primary tree is constructed with the default weight 1. Then, it is gradually improved in the tree creation procedure. The height of the tree defines the ultimate weight of the features. Therefore, weight distribution and weight increasing strategies are considered to improve prediction performance and strong diversity. BFPA will randomly select the weights for attributes that exist in the latest tree. The weight range (η_i) is computed using Eq. (8).

$$\eta^{\alpha} = \left\{ \begin{array}{ll} \left[0.00,\ e^{-\frac{1}{\alpha}}\right], & \alpha = 1 \\ \left[e^{-\frac{1}{\alpha-1}+\beta,}\ e^{-\frac{1}{\alpha}}\right], & \alpha > 1 \end{array} \right\} \tag{8}$$

In Eq. (8), the term α represents the feature level. The variable β is used to ensure that the range of weight for different hierarchies is non-overlying. We set $\alpha = 1$, if the attribute appears in the root node, we set $\alpha = 1$. We assign $\alpha = 2$ if the attribute presents at a child node. Similarly, to define the negative effect of assigned weights that do not present in the latest tree, BFPA has a method to gradually increase the feature weights. Assume a feature $\mathcal{D}_i$ is tested at hierarchy β of T_{k-1} th tree with height h and its weight is η_i. The addition of weight φ_i is calculated using Eq. (9).

$$\varphi_i = \frac{1 - \eta_i}{(h+1) - \alpha} \tag{9}$$

The method of adaptive weight allocation allows the DTs to process the unclassified data points. The decisions of the DTs are integrated and the ultimate decision is computed using a voting mechanism. In this study, we implement an average of probabilities (AoP) approach to aggregate the decisions of individual trees and provide a final result. In this study, the target class is designated according to the maximum mean likelihoods. Let n represents the number of trees in the decision forest $T = \{T_1,\ T_2, \ldots . T_n\}$ with L labels. The tree $T_i : R^n \rightarrow [0,1]^c$ accepts an input data $d_i \in R^n$ and provides the result as a vector $\rho_{ci}(\eta_1|d),\ \rho_{ci}(\eta_2|d), \ldots . \rho_{ci}(\eta_c|d)$, where $\rho_{ci}(\eta_k|d)$ denotes the likelihood assigned by T_i that input data sample d fits into η_k. Consider Δ_k is the average of probabilities assigned by the trees for each label. It can be calculated using Eq. (10).

$$\Delta_k = \frac{1}{n} \sum\nolimits_{i=1}^{n} \rho_{ci}(\eta_k|d) \tag{10}$$

Let $\Delta = [\Delta_1, \Delta_2, \ldots \Delta_c]$ is the collection of average probabilities for L label and d is assigned to the weight η_c if Δ_k represents the maximum value in Δ.

4 Performance Evaluations

The effectiveness of the proposed BFPA-SAS model is evaluated by relating its performance with that of 6 similar classifiers such as SVM [10], RF [12], DT [23], LR [24], ABM [17], and HRFLM [16]. In our evaluation, the experimentation is carried out using an Intel Core i7-4790 CPU with 3.06 GHz, 16 GB

RAM, and Windows 10 operating system. In this study, Python version 3.8.5 is used for investigation, and Weka version 3.8.3 is used as a mining tool to carry out the experiments.

4.1 Dataset Preparation

This study uses the Heart disease dataset collected from the UCI repository to evaluate this BFPA-SAS model. The data samples in this database contain 13 attributes. Table 2 shows the statistical specifics of each attribute.

Table 2: Statistics of heart disease database

Label	Attribute	Description	Average value
F_1	cp	The type of chest pain (4 values)	3.16
F_2	ca	Number of major vessels (0–3) colored by fluoroscopy	Categorical
F_3	oldpeak	ST depression induced by exercise relative to rest	1.04
F_4	thal	normal = 1; fixed defect = 2; reversible defect = 3	Categorical
F_5	thalach	Maximum heart rate achieved	149.61
F_6	trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	131.69
F_7	age	Age in years	54.4
F_8	chol	Serum cholesterol in mg/dl	246.69
F_9	slope	The slope of the peak exercise ST segment	1.60
F_{10}	exang	Exercise induced angina (1 = yes; 0 = no)	0.33
F_{11}	fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)	0.15
F_{12}	restecg	Resting electrocardiographic results	0.99
F_{13}	sex	Female = 0; Male = 1	0.68

The database comprises 1025 medical data comprising 312 females and 713 males of various age groups where 48.68% (499) records are negative and 51.32% (526) records have cardiac problems. Among the cardiac patients, 42.97% (226) patients are female and 57.03% (300) patients are male. The database has been standardized in the range of [−1, +1] before applying our prediction algorithm. Actually, separate samples for testing do not exist in the dataset. Hence, we have intentionally fragmented the prevailing database into 70% for training and 30% for testing. To realize a more precise calculation of the performance of the classifier, the *k*-fold cross-validation (CV) is used. We set k = 10. This means that the whole database is split into 10 parts during training. For every iteration, one part is used to test the proposed model and the remaining parts are combined to build the learning samples. Then, the average value of standard deviation across all the autonomous trials is computed. It is important to note that a single repetition of the 10-fold CV cannot generate acceptable outcomes for assessment due to the uncertainty in data partitions. Hence, each output is calculated on a mean value of 10 runs to achieve accurate results.

4.2 Evaluation Measures

To demonstrate the effectiveness of the proposed diagnostic system using the BFPA-SAS approach, we used six imperative evaluation metrics such as prediction accuracy (ACC), specificity (SPE), sensitivity (SEN), precision (PRE), IoU, and *p*-value from Wilcoxon statistical test. The effectiveness is computed regarding the accuracy using Eq. (11).

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

In the above equation, TP represents true positive (i.e., the number of records correctly pigeonholed as cardiac patients); TN represents true negative (i.e., the number of records correctly categorized as a normal person); FN denotes false negative (i.e., the number of records incorrectly recognized as the normal person); and FP represents false positive (i.e., the number of normal persons incorrectly classified as the cardiac patient). The aptitude of a classification algorithm to classify only the appropriate instances is called precision. It describes the number of positive cases that belongs to the positive class. It is calculated using Eq. (12).

$$PRE = \frac{TP}{TP + FP} \tag{12}$$

Sensitivity and specificity represent how the classifier distinguishes positive and negative records. Sensitivity states the classification rate and specificity represents the false alarm rate as defined by Eqs. (13) and (14).

$$\text{SEN} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{SPE} = \frac{TN}{TN + FP} \tag{14}$$

The intersection over-union is an evaluation metric employed to compute the enactment of any ML classifier. For any dataset, the IoU gives the similarity between the standard reference and the classified data. The IoU metric can consider the class imbalance problem usually present in the dataset. IoU is calculated by Eq. (15).

$$\text{IoU} = \frac{TP}{TP + FN + FP} \tag{15}$$

The Wilcoxon statistical test is carried out to find out whether the intended BFPA-SAS model provides a significant enhancement related to other classifiers or not. This analysis is conducted by analyzing the impacts of the developed BFPA-SAS model and related to all other classifiers at the level of 5% significance. If the p-value <0.05, then the null hypothesis is excluded, i.e., there is a substantial difference at the level of 5% significance. The p-values >0.05 indicate that the difference among the related values is trivial.

5 Results and Discussion

Attribute significance and the corresponding merit score are calculated for all the employed classifiers and listed in Table 3 according to the participation in identifying cardiac disease. The same values are illustrated in Fig. 1 for an improved understanding of the calculation of attribute ranking and the significance of each classifier. Figs. 1a–1g show the attribute ranking using the significance-based merit scores for all the applied classifiers. The figure also tends to denote the highly accountable features for cardiac problem disease. The proposed BFPA-SAS classifier is trained using data samples collected from the Heart disease dataset and it is tested using testing samples. Table 5 displays the inclusive results realized by the proposed model. To prove the efficacy of the intended BFPA-SAS classifier, we compare its enactment with other state-of-the-art classifiers. The numerical outcomes obtained from different classifiers are listed in Table 6.

Table 3: Attribute significance in terms of merit score of different classifiers

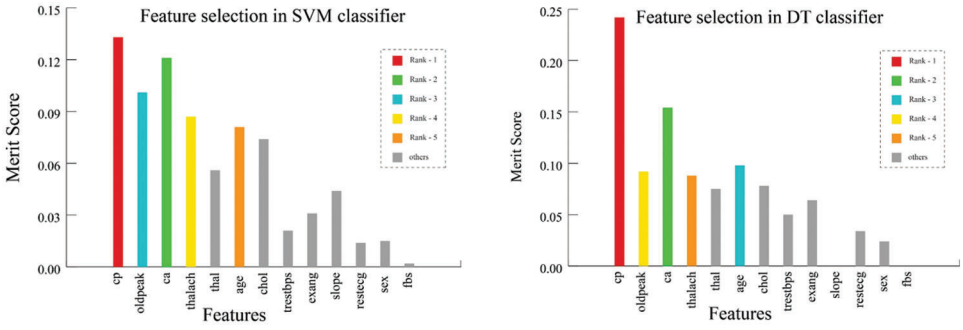
Feature	SVM	DT	RF	LR	ABM	HRFLM	BFPA-SAS
cp	0.133	0.242	0.135	0.883	0.041	0.145	0.345
oldpeak	0.101	0.092	0.121	−0.542	0.101	0.123	0.301
ca	0.121	0.154	0.115	−0.714	0.082	0.101	0.287
thalach	0.087	0.088	0.115	0.029	0.183	0.072	0.189
thal	0.056	0.075	0.112	−0.849	0.042	0.094	0.166
age	0.081	0.098	0.087	0.006	0.104	0.056	0.255
chol	0.074	0.078	0.075	−0.005	0.324	0.065	0.262
trestbps	0.021	0.050	0.071	−0.010	0.043	0.034	0.008
exang	0.031	0.064	0.056	−0.945	0.022	0.022	0.001
slope	0.044	0.000	0.049	0.733	0.022	0.011	0.001
restecg	0.014	0.034	0.033	−1.610	0.061	0.001	0.002
sex	0.015	0.024	0.019	0.399	0.001	0.052	0.084
fbs	0.002	0.000	0.009	−0.285	0.001	0.001	0.002

From these results, we observed that the SVM classifier has achieved reasonable prediction enactment with 76.4% accuracy, 97.6% sensitivity, 61.4% specificity, 64.2% precision, and 63.2% IoU. In the logistic regression model, the predicted parameters (trained weights) provide a better interpretation of the attribute significance. It realizes improved classification performance regarding prediction accuracy of 86.7%, sensitivity of 98.1%, specificity of 76.4%, precision of 78.8%, and IoU of 77.7%. The decision tree-based classifier achieves better results than LR with 87.7% accuracy, 98.2% sensitivity, 77.9% specificity, 80.4% precision, and 79.3% IoU.

Table 4 lists the 5 most important attributes based on their significance and association value. From the table, it is observed that chest pain (cp) is selected as the most important attribute by the majority of the classifiers considered in this study for identifying cardiac disease.

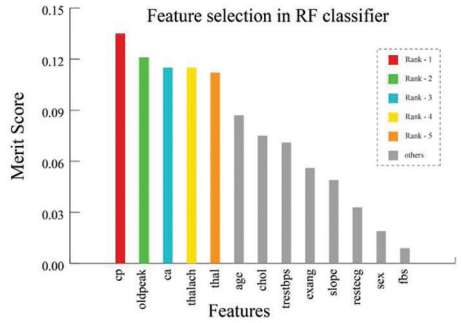
Our proposed SAS approach selects the five most important features in the order of merit score. It selects chest pain (cp) as the most vital attribute with a merit score of 0.345, ST depression generated by workout related to relaxation (oldpeak) with a merit score of 0.301, the number of major vessels stained by fluoroscopy (ca) with a merit score of 0.287, Serum cholesterol (chol) with a score of 0.262, and age with a score of 0.255. Then, we applied these selected features to six different ML classification algorithms, SVM, DT, RF, LR, ABM, and HRFLM to identify cardiac disease. The performance of each classifier is assessed based on the metrics such as accuracy, precision, specificity, sensitivity, IoU, and *p*-value.

In the RF classifier, a large number of DTs have been generated to construct the forest during the training process. All the DTs in the forest predict the output for each data sample at the testing phase. When an output is generated by every DT, then majority voting is employed to make the ultimate output for every testing sample. Hence, the RF achieves improved enactment with respect to accuracy (91.0%), sensitivity (93.4%), specificity (85.2%), precision (86.3%), and IoU (85.1%).

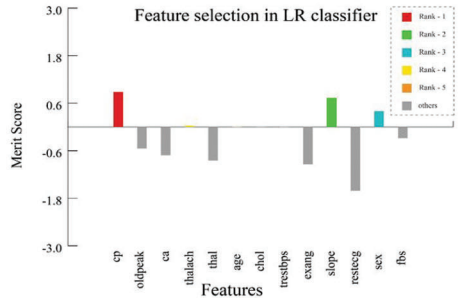


(a) Feature selection in SVM classifier

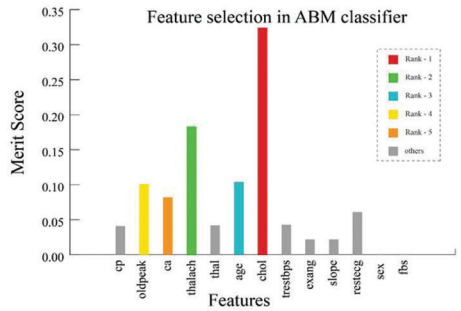
(b) Feature selection in DT classifier



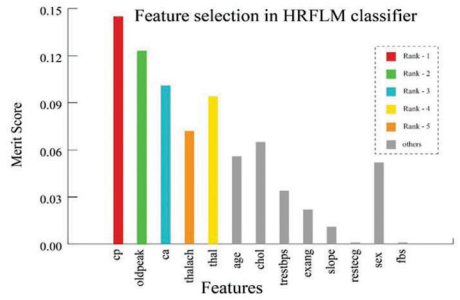
(c) Feature selection in RF classifier



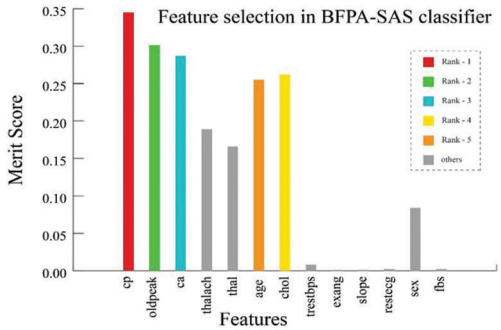
(d) Feature selection in LR classifier



(e) Feature selection in ABM classifier



(f) Feature selection in HRFLM



(g) Feature selection in BFPA-SAS classifier

Figure 1: (a) Feature selection in SVM classifier. (b) Feature selection in DT classifier. (c) Feature selection in RF classifier. (d) Feature selection in LR classifier. (e) Feature selection in ABM classifier. (f) Feature selection in HRFLM. (g) Feature selection in BFPA-SAS classifier

Table 4: Selected attributes with top merits score for cardiac disease

Rank	SVM	DT	RF	LR	ABM	HRFLM	BFPA-SAS
1	cp	cp	cp	cp	chol	cp	cp
2	ca	ca	oldpeak	slope	thalach	oldpeak	oldpeak
3	oldpeak	age	ca	restecg	age	ca	ca
4	thalach	oldpeak	thalach	thalach	oldpeak	thal	chol
5	age	thalach	thal	age	ca	thalach	age

Table 5: The results obtained BFPA-SAS for the selected feature subset of the heart disease dataset

Fold	ACC	SEN	SPE	PRE	IoU	ρ-value
#1	0.930	0.995	0.866	0.900	0.890	0.015
#2	0.944	0.997	0.884	0.919	0.903	0.025
#3	0.948	0.981	0.913	0.883	0.914	0.035
#4	0.961	0.998	0.931	0.880	0.907	0.046
#5	0.963	0.998	0.915	0.888	0.911	0.041
#6	0.910	0.996	0.936	0.965	0.876	0.029
#7	0.962	0.986	0.886	0.896	0.916	0.008
#8	0.947	0.995	0.913	0.905	0.913	0.038
#9	0.948	0.986	0.891	0.941	0.924	0.023
#10	0.954	0.993	0.868	0.931	0.890	0.029
Mean	**0.947**	**0.992**	**0.901**	**0.911**	**0.904**	**0.029**
S.D	0.016	0.006	0.025	0.028	0.015	0.012

Table 6: The results gained from the heart disease database regarding evaluation metrics

Classifier	Criteria	ACC	SEN	SPE	PRE	IoU	ρ-value
SVM	Avg	0.764	0.976	0.614	0.642	0.632	0.045
	SD	0.037	0.010	0.054	0.042	0.039	0.020
LR	Avg	0.867	0.981	0.764	0.788	0.777	0.044
	SD	0.043	0.010	0.031	0.045	0.019	0.020
DT	Avg	0.877	0.982	0.779	0.804	0.793	0.051
	SD	0.060	0.009	0.025	0.042	0.039	0.020
RF	Avg	0.910	0.934	0.852	0.863	0.851	0.047
	SD	0.050	0.008	0.041	0.038	0.033	0.008
ABM	Avg	0.915	0.950	0.881	0.886	0.847	0.037
	SD	0.051	0.011	0.031	0.029	0.018	0.016

(Continued)

Table 6 (continued)

Classifier	Criteria	ACC	SEN	SPE	PRE	IoU	ρ-value
HRFLM	Avg	0.924	0.971	0.884	0.880	0.857	0.031
	SD	0.017	0.004	0.029	0.038	0.029	0.014
BFPA-SAS	Avg	0.947	0.992	0.900	0.911	0.904	0.029
	SD	0.016	0.006	0.025	0.028	0.015	0.012

The adaptive boosting method implements an adaptive improvement method and provides better prediction outputs by assimilating several weak classifiers into a sturdy classifier. It achieves improved performance in terms of accuracy (91.5%), sensitivity (95.0%), specificity (88.1%), precision (88.6%), and IoU (84.7%). The HRFLM employs various combinations of attributes and numerous recognized prediction methods with the voting mechanism. It exploits an artificial neural network with backpropagation. By applying these techniques HRFLM realizes 92.4% accuracy, 97.1% sensitivity, 88.4% specificity, 885 precision, and 85.7% IoU. Our proposed approach BFPA-SAS outperforms other approaches in terms of evaluation measures. It achieves superior classification performance with 94.7% accuracy, 99.2% sensitivity, 90% specificity, 91.1% precision, and 90.4% IoU. The benefits like significance-based feature selection, removal of irrelevant features, and integration of the power of all non-class attributes present in a database, BFPA-SAS enable an effective classification approach for the cardiac diagnosing system.

The average and SD values of evaluation metrics obtained from the Heart disease database by each classifier are demonstrated in Figs. 2 and 3. It can be observed that the BFPA-SAS outperforms all other methods regarding evaluation metrics. Figs. 2 and 3 also demonstrate that the p-values of BFPA-SAS related to other classifiers obtained by Wilcoxon's statistical test. This evaluation is carried out to determine whether the dissimilarity between the outputs obtained by BFPA-SAS and the outputs of other classifiers is noteworthy or not. More precisely, the p-value <0.05 denotes that the results achieved by BFPA-SAS have substantial deviations from those of the other classifiers utilized for performance assessment. If the p-value >5%, then there is no considerable difference between the results of BFPA-SAS and the other classifiers. In this study, it can be observed that the p-values <5% in most cases, which exhibits the improved results of BFPA-SAS. Furthermore, it is interesting to note that the standard deviation obtained by the BFPA-SAS classifier is lower than that of all the classifiers which demonstrate that the BFPA-SAS classifier can enable more accurate classification results.

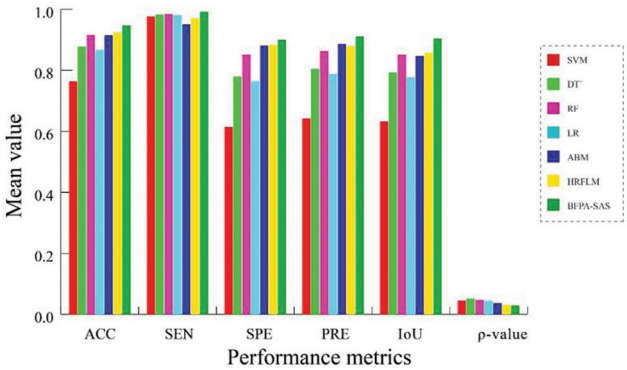


Figure 2: Results obtained by various classifiers in terms of average value of the performance measures

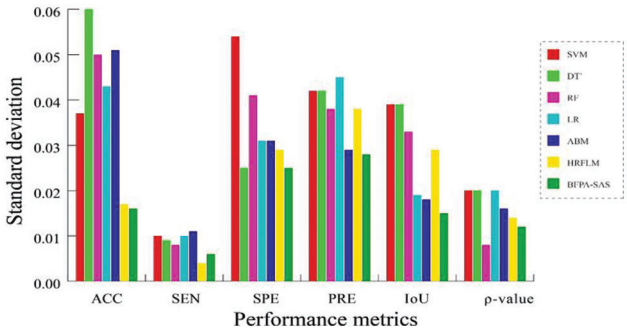


Figure 3: Results obtained by various classifiers in terms of SD value of the performance measures

6 Conclusions and Future Work

Cardiac problem is a deadly disease and it slays lots of individuals annually. The identification of the cardiac problems is a challenging endeavor. Using ML algorithms, the cardiac problem can be identified quickly and at a low cost. The growing dimension of health databases has made it a complex process for physicians to realize the intricate attribute relations and make disease classification complicated. This work intends a BFPA algorithm to identify the cardiac disease with higher accuracy. This work integrates a SAS algorithm with BFPA to improve the performance of the diagnostic system in identifying cardiac illness. The proposed SAS algorithm is used to determine the correlation among attributes and to select the optimum subset of feature space for learning and testing processes. BFPA-SAS selects an optimal number of learning and testing samples as well as the density of trees in the forest to realize higher prediction accuracy effectively. The efficiency of the proposed classifier is cautiously assessed on the Heart disease database by relating its enactment with many advanced classifiers in terms of ACC, PRE, SEN, SPE, and IoU. The empirical outcomes demonstrate that the developed approach outdoes other approaches with superior performance. Additionally, Wilcoxon's rank-sum test is carried out to determine whether our classification approach provides a significant enhancement related to other classifiers or not. Accordingly, we can conclude that the integration of SAS and BFPA is the better data stream classification model related to other state-of-the-art classifiers. However, the proposed method does not consider the false alarm rate including false postive and false negative rate as well as timing complexity of the proposed classofier. We intend to analyze the performance of this proposed model in terms of other evaluation metrics including false alarm rate and processing oevrheads in future.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. R. Panhalkar and D. D. Doye, "A novel approach to build accurate and diverse decision tree forest," *Evolutionary Intelligence*, vol. 15, no. 1, pp. 439–453, 2022.

[2] A. K. Gárate-Escamila, A. H. E. Hassani and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, no. 10, pp. 313–330, 2020.

[3] E. Maini, B. Venkateswarlu, B. Maini and D. Marwaha, "Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India," *Medical Journal Armed Forces India*, vol. 77, no. 3, pp. 302–311, 2021.

[4] J. R. Petrie, T. J. Guzik and R. M. Touyz, "Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms," *Canadian Journal of Cardiology*, vol. 34, no. 5, pp. 575–584, 2018.

[5] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn *et al.,* "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, no. 10, pp. 4672–4681, 2021.

[6] J. Yang, Y. Li, Q. Liu, L. Li, A. Feng *et al.,* "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 57–69, 2020.

[7] N. Absar, E. K. Das, S. N. Shoma, M. U. Khandaker, M. H. Miraz *et al.,* "The efficacy of machine-learningsupported smart system for heart disease prediction," *Healthcare*, vol. 10, no. 1137, pp. 1–17, 2022.

[8] S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, "Privacy risk in machine learning: analyzing the connection to overfitting," in *Proc. CSF*, Oxford, UK, pp. 268–282, 2018.

[9] O. O. Aremu, D. Hyland-Wood and P. R. McAree, "A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data," *Reliability Engineering and System Safety*, vol. 195, no. 1, pp. 706–716, 2020.

[10] H. M. Le, T. D. Tran and L. A. N. G. Van Tran, "Automatic heart disease prediction using feature selection and data mining technique," *Journal of Computer Science and Cybernetics*, vol. 34, no. 1, pp. 33–48, 2018.

[11] G. N. Ahmad, H. Fatima Shafiullah and M. Abbas, "Mixed machine learning approach for efficient prediction of human heart disease by identifying the numerical and categorical features," *Applied Science*, vol. 12, no. 15, pp. 1–33, 2022.

[12] V. Jackins, S. Vimal, M. Kaliappan and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021.

[13] S. Maheswari and R. Pitchai, "Heart disease prediction system using decision tree and naive bayes algorithm," *Current Medical Imaging Reviews*, vol. 15, no. 8, pp. 712–717, 2019.

[14] S. I. Ayon, M. Islam and R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," *IETE Journal of Research*, vol. 68, no. 4, pp. 1–20, 2020.

[15] T. Amarbayasgalan, V. H. Pham, T. U. Nipon, Y. Piao and K. H. Ryu, "An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well ordered training datasets," *IEEE Access*, vol. 9, pp. 135210–135223, 2021.

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

[17] N. Absar, E. K. Das, S. N. Shoma, M. U. Khandaker, M. H. Miraz *et al.,* "The efficacy of machine-learning-supported smart system for heart disease prediction," *Healthcare (Basel)*, vol. 10, no. 6, pp. 1137, 2022.

[18] K. M. Almustafa, "Prediction of heart disease and classifiers sensitivity analysis," *BMC Bioinformatics*, vol. 21, pp. 278, 2020.

[19] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.

[20] R. A. Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique," *Cybersecurity*, vol. 5, no. 1, pp. 25–36, 2022.

[21] J. Rezaei, "Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making," *Journal of Decision Systems*, vol. 30, no. 1, pp. 72–96, 2020.

[22] R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, no. 1, pp. 189–203, 2018.

[23] X. Luo, F. Lin, Y. Chen, S. Zhu, Z. Xu *et al.,* "Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[24] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–16, 2019.