Research Article

A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms

Abdul Saboor,[1] Muhammad Usman ⓘ,[1] Sikandar Ali ⓘ,[2] Ali Samad ⓘ,[3] Muhmmad Faisal Abrar ⓘ,[4] and Najeeb Ullah ⓘ[4]

[1]*Department of Computer Software Engineering, University of Engineering & Technology, Mardan, Khyber Pakhtunkhwa, Pakistan*
[2]*Department of Information Technology, The University of Haripur, Haripur 22620, Khyber Pakhtunkhwa, Pakistan*
[3]*Department of Computer Science, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*
[4]*Department of Computer Science, University of Engineering & Technology, Mardan, Khyber Pakhtunkhwa, Pakistan*

Correspondence should be addressed to Sikandar Ali; sikandar@uoh.edu.pk

A great diversity comes in the field of medical sciences because of computing capabilities and improvements in techniques, especially in the identification of human heart diseases. Nowadays, it is one of the world's most dangerous human heart diseases and has very serious effects the human life. Accurate and timely identification of human heart disease can be very helpful in preventing heart failure in its early stages and will improve the patient's survival. Manual approaches for the identification of heart disease are biased and prone to interexaminer variability. In this regard, machine learning algorithms are efficient and reliable sources to detect and categorize persons suffering from heart disease and those who are healthy. According to the recommended study, we identified and predicted human heart disease using a variety of machine learning algorithms and used the heart disease dataset to evaluate its performance using different metrics for evaluation, such as sensitivity, specificity, F-measure, and classification accuracy. For this purpose, we used nine classifiers of machine learning to the final dataset before and after the hyperparameter tuning of the machine learning classifiers, such as AB, LR, ET, MNB, CART, SVM, LDA, RF, and XGB. Furthermore, we check their accuracy on the standard heart disease dataset by performing certain preprocessing, standardization of dataset, and hyperparameter tuning. Additionally, to train and validate the machine learning algorithms, we deployed the standard K-fold cross-validation technique. Finally, the experimental result indicated that the accuracy of the prediction classifiers with hyperparameter tuning improved and achieved notable results with data standardization and the hyperparameter tuning of the machine learning classifiers.

1. Introduction

As per the World Health Organization (WHO) report, 17.9 million deaths occurred from cardiovascular diseases (CVDs) in 2019, representing 32% of all global deaths [1] and having an annual mortality rate of greater than 17.7 million [2]. In 2018, the Australian Institute of Health and Welfare (AIHW) reported cardiovascular disease (CVD) as the leading cause of death in Australia, representing 42% of all deaths [3]. Researchers are attempting to develop an effective technique for the timely identification of heart diseases as existing heart disease diagnosis methods are ineffective in early detection for various reasons, including accuracy and computational time [4]. When advanced technology and healthcare experts are unavailable, diagnosing and controlling heart disease is incredibly challenging [5]. Many people's lives can be saved with a good, solid diagnosis and treatment [6]. A physician's evaluation of the patient's medical history, physical examination report, and analysis of concerning symptoms are used to diagnose heart diseases. However, the findings of this method of diagnosis are insufficient in detecting heart disease patients. Furthermore, it

is both costly and computationally challenging to examine [7]. Thus, we build a noninvasive prediction system to handle these issues using machine learning classifiers. Heart diseases are efficiently diagnosed using an expert decision system relying on machine learning classifiers and artificial fuzzy logic. As a consequence, the death ratio declines [8, 9]. Numerous researchers used the Cleveland heart disease dataset. For training and testing, the predictive models of machine learning require appropriate data. When a refined/ standardized dataset is used for training and testing, the accuracy of machine learning classifiers can be improved. Furthermore, by incorporating relevant and related data features, the predictive model capabilities can be enhanced. Therefore, data standardization and feature selection are important for machine learning classifiers' accuracy. Numerous researchers have used different predictive techniques in the literature, however, these approaches do not predict heart diseases effectively. Data standardization is necessary to enhance the machine learning classifiers' accuracy. There are different standardization techniques, such as standard scalar (SS), min-max scalar, and others that are used to remove the missing feature value instances from the dataset.

Multiple tests are required for heart disease prediction. Timely identification is difficult. Cardiovascular disease prediction is complicated, especially in emerging nations, where there is a shortage of skilled medical personnel, testing equipment, and other resources needed for the identification and treatment of individuals with cardiac problems [10]. When trained using appropriate data, computational classifiers can be useful in diagnosing diseases [11]. Numerous machine learning-based methods have been proposed for predicting the risk of CSD. Most of these methods exploit the use of publicly available datasets for the purpose of model training and evaluation. The availability of these datasets has improved the performance of machine learning-based predictive models and opened up new research avenues for researchers to develop cutting-edge algorithms for predicting CVD risk. These datasets provide information about different risk factors and the patient's disease status (whether the patient has a disease). Preprocessing is required for designing predictive models for CVD because the clinical datasets available are inconsistent and duplicated [12]. Furthermore, information about different risk factors (features) is available, and the selection of an appropriate set of features is based on certain criteria, such as having a high prevalence in most populations, having a significant impact on heart disease on their own, and being able to be controlled or treated to lower the risks [13]. Various risk factors or features have been employed by different studies when modeling CVD predictors. When machine learning algorithms are trained on appropriate datasets, they are most effective [12, 14]. Limited medical datasets, feature selection, ML algorithm implementations, and a lack of in-depth analysis are all obstacles that may preclude the effective prediction of heart diseases. Our research intends to fill some of these knowledge gaps to construct a better CVD prediction model. Apart from that, the datasets used in existing studies also have some limitations. These datasets do not include sufficient risk factors or attributes from the detailed clinical data. This difference in clinical severity may affect the prediction accuracy. These limitations have not been sufficiently considered in previous studies. In the state-of-the-art research, dataset standardization and algorithm tuning were not performed.

For enhanced cardiac disease prediction, researchers have developed a variety of machine learning models, such as SVM, KNN, FR, DT, LR, NB, and so on. Heart disease prediction accuracy, on the other hand, remains a challenge. It is critical to develop a novel and cost-effective tool for predicting the risk of heart disease with high accuracy. The NB, BN, RF, and MLP total level of complexity has not been defined. The age element is the age risk factor, which is also excluded in NB, BN, RF, and MLP from dataset [15]. The system was studied using StatLog datasets. For the Cleveland dataset, important risk factors, such as age, RestECG, ST Depression (Slope), and so on are removed from the model [16]. For the standardization of the proposed approach, no significance tests are performed, and StatLog dataset [17] and Z-Alizadeh Sani dataset are used. The dataset has a smaller size. The obtained result was not compared to other datasets for standardization, and the Cleveland dataset was used [18].

In this research work, we proposed a machine learning classifier that includes random forest (RF), XGBoost (XGB), decision trees (CART), support vector machine (SVM), multinomial Naïve Bayes (MNB), logistic regression (LR), linear discriminant analysis (LDA), AdaBoost classifier (AB), and extra trees classifier (ET) for heart disease prediction. The standardization and hyperparameters are performed using the GridSearch CV method to select the best value for the hyperparameters for the best machine learning classifier. Apart from that, various performance evaluation parameters, such as accuracy, precision, sensitivity, recall, and F-measures, are used for the machine learning classifier's performance. The proposed method has been tested on the Cleveland HD dataset. Moreover, the proposed machine learning classifiers' accuracy has been compared to existing state-of-the-art methods in the literature, such as SVM, LR [19], and RF [20]. The proposed work has the following main contributions:

(1) Firstly, the authors attempt to address the issue of datasets and then refine and standardized the datasets. Then, the datasets are used to train and test classifiers and determine which classifiers provide the best accuracy results.

(2) Secondly, the authors, to identify the best values of hyperparameter, used the GridSearchCV method.

(3) Thirdly, apply the machine learning classifiers with the best hyperparameter values to achieve the highest accuracy performed using hyperparameter tuning.

(4) Finally, the proposed classifier (SVM) gives state-of-the-art accuracy.

The rest of the paper is organized as follows: in Section 2, a literature review of the existing machine learning techniques has been discussed. Section 3 describes research

goals, and Section 4 describes a proposed methodology to be followed during the study. Section 5 describes data collection, Section 6 discusses the experimental results, and Section 7 concludes the paper and gives future work.

2. Section II: Literature Review

The primary method used by the physicians was the auscultation method for distinguishing between normal and abnormal cardiac sounds [21]. Every heart disease was identified by the physicians listening to these sounds of the heart using stethoscopes [20]. The auscultation technique used by professional doctors to diagnose a heart disease has some drawbacks. The clarification and classification of distinct sounds in the heart are associated with the abilities and practices of the doctors, which are gained after lengthy examinations [22].

Apart from the manual method, various machine learning methods have been proposed for CVD detection. Research was conducted by Amin et al. [19] to classify the most relevant attributes of heart disease prediction. Seven classification algorithms are used, which consist of NB, KNN, LR, DT, NN, SVM, and Vote. The Cleveland datasets were obtained from the UCI repository of machine learning, which consists of 303 records and 76 attributes. The 10-fold cross-validation method is used for model training and testing. We used 10-fold cross-validation because, in the dataset, we have fewer training examples, and using data split, such as train-test split, will give us an underestimate of the model predictive performance because we will have fewer number of examples in the training set. However, using 10-fold validation, the model will have 90% of the data to learn from. The Vote Classifier achieved a higher accuracy of 87.4%. A study carried out by Ketut Agung Enriko et al. [23] used a KNN classifier with minimal parameters for heart disease prediction and had an accuracy rate of 81.85%. When using KNN, the performance drops as the number of parameters increases, and it uses 90% of the input for training, which is computationally expensive. Subhadra et al. [24] conducted the study. The used training algorithm is a multilayer perceptron neural network (MLP-NN) with backpropagation for heart disease prediction. To evaluate the system's performance, recall, accuracy, precision, and F-measure are employed, and model training and testing are carried out using the UCI repository of machine learning Cleveland dataset, which consists of the records of 303 instances and has 76 attributes. Through preprocessing, missing values were removed from the data, which consisted of six records, and the 14 most relevant attributes of the heart disease were used. The results generated during the experiment showed that MLN-NN obtained a higher accuracy of 93.39%, with a running time of 3.86 seconds. Another study conducted by Khan et al. [25] used a comprehensive prediction of heart disease based on an analysis using some of the most popular machine learning classifiers. For training and testing, only 14 features are employed from the Cleveland (UCI) datasets, which consist of 303 records. There was a data preprocessing activity carried out, resulting in a dataset consisting of 296 records. The results of SVM

classifiers achieved a higher accuracy of 90.00%. Tarawneh et al. [26] have conducted a study using the hybrid approaches of data mining classifiers to predict heart disease. The datasets were obtained from the UCI repository of machine learning, which consists of 303 records and has 76 attributes. Model training and testing were performed on 14 attributes. The data was preprocessed to minimize the features from 14 to 12. KNN, NN, SVM, GA, J48, RF, and NB are the classification algorithms used to assess the precision, recall, and accuracy of cardiac disease prediction. The accuracy obtained by SVM and NB was 89.2%, and they made better predictions of heart disease. Anitha et al. [27] have conducted a study using learning vector quantization algorithms for the prediction of cardiac disease. The accuracy achieved by this algorithm is 85.55%. The datasets were taken from the University of California, Irvine's (UCI), machine learning library, which consists of 303 records and has 76 attributes. The data were preprocessed because of missing values, resulting in a sample of 302 records, with only 14 features used for heart disease. The dataset is categorized into two sections: 70% for model training and 30% for model testing. Another study developed by Jagtap et al. [28] developed a web-based application for heart disease prediction using machine learning techniques. For the classification algorithms, LR, NB, and SVM are used for model training and testing. Using the UCI machine learning repository, the Cleveland datasets were divided into 75 percent and 25 percent for training and testing, respectively. The data were preprocessed to eliminate discrepancies and missing values, and SVM achieved a higher accuracy of 64.4%. The study's limitation was its inability to detect the risk factors of human heart disease patients at an early stage. Another study developed by Dulhare et al. [29] combined the common feature selection algorithms of particle swarm optimization (PSO) and Naïve Bayesian algorithms for an efficient prediction of heart disease. The model training and testing processes were conducted using the UCI repository of the machine learning dataset of VA Long Beach, which consists of 270 records and 14 attributes, however, only 7 attributes out of the 14 attributes of heart disease were used to predict it. When combined with PSO and NB, the performance accuracy of NB increases to 87.91%. It has been shown that accuracy improves by 8.79% as compared to NB accuracy. Another study was developed by Kim et al. [30] using machine learning algorithms to predict heart disease. The datasets were collected from the repository of machine learning at the University of California, Irvine (UCI), which consists of 303 records and uses 14 attributes. For training and testing, the 10-fold cross-validation approach is utilized. The DT algorithm performs with a better accuracy of 93.19% prediction of heart disease. Siontis et al. [31] describe the present and future condition of AI-enhanced electrocardiogram (ECG) in the diagnosis of heart disease in at-risk communities, summarize its consequences for healthcare decisions in patients with cardiovascular disease, and assess its potential drawbacks. Linda et al. [32] proposed a unique health information system for prescribing exercise to heart disease patients. According to their early findings, clinicians are confused about how to establish an exercise prescription

for patients with numerous CVD risk factors. For patients, the supplied system is an easy-to-use, guided, and time-saving evidence-based method. Ali et al. [33] provided a three-phase PB-FARM approach for the assessment of disease-related risk factors. It was also used to analyze the factors that influence the incidence of this disease using the Z-Alizadeh Sani dataset. The findings revealed a clear link between the risk of coronary artery disease (CAD), elderly age, and normal chest pain. Rubini et al. [34] proposed a prediction model for heart disease prediction. Different classifiers, such as logistic regression, Nave Bayes, and SVM, were compared to the proposed algorithm. In the proposed article, random forest achieved the highest accuracy of 84.81%. Devansh Shah et al. [35] utilized a dataset of 303 examples and 76 attributes, 14 of which were used in supervised learning algorithms, such as decision tree, Nave Bayes, K-NN, and random forest. The results show that K-NN has attained the maximum level of accuracy. Archana Singh et al. [36] developed a heart disease prediction model using machine learning classifiers. The UCI, Cleveland dataset uses 14 attributes to train and test their models to achieve maximum accuracy. The results achieved by the classifiers were as follows: linear regression 78%, decision trees 79%, support vector machines 83%, and K-NN 87%. The results revealed that K-NN had the highest accuracy. In this article, Asif Khan et al. [37] use SVM, logistic regression, artificial neural networks, KNN, Nave Bayes, and decision tree as classification techniques. When compared to previous models, the new model achieved an accuracy of 92.37%. The fundamental goal of this article, according to Mohan et al. [38], was to uncover suitable features using machine learning techniques, such as decision trees, language models, SVM, random forests, Naive Bayes, neural networks, and KNN. The proposed hybrid HRFLM method was applied to merge the characteristics of random forests and linear techniques. This model's accuracy was 88.4%. Kumar et al.'s [39] various machine learning algorithms were utilized to predict cardiovascular disease. When compared to other classifier techniques, the proposed model revealed that random forests had the greatest accuracy of 85.71%.

3. Section III Research Goals and Objectives

The main goal of this research is to develop a heart disease prediction model with improved and enhanced accuracy. The specific objectives are to quickly identify new patients, reduce diagnostic time, reduce heart attacks, and save lives.

4. Section IV Methodology

In Section 4, we describe the proposed method and also explain that the method is defined by the subsequent steps, as shown in Figure 1.

(1) The first step is to select the dataset from the machine learning online repositories. There are many online repositories, such as the Cleveland heart disease dataset, Z-Alizadeh Sani dataset, StatLog Heart, Hungarian, Long Beach VA, and Kaggle Framingham dataset.

(2) In the second step, we refined and standardized the collected data sets. These datasets were not gathered in a controlled environment and had erroneous values. Hence, data preprocessing is an essential step for studying data and machine learning. Data normalization means when the risk factors of a dataset have different values. For example, Celsius and Fahrenheit are different measuring units of temperature. The standardization of data means scaling the risk factors and assigning the values that show the difference between standard deviations from the mean value. It rescales the risk factor value to improve the performance of machine learning classifiers with a standard deviation (σ) of 1 and a mean (μ) of 0. The mathematical form of standardization is given by (1).

$$\text{Standardization of } X = \frac{X - \text{Mean of } X}{\text{Standard Deviation of } X}. \quad (1)$$

(3) In this step, hyperparameter tuning is performed to select the best value for the hyper parameters and get high accuracy. For this purpose, we used the GridSearchCV method. Before applying machine learning classifiers, we adjust the hyper parameters values of machine learning classifiers to increase their performance. The Scikit-learn GridSearchCV class's fit approach provides a grid of tuning classification algorithms. It allows each machine learning algorithm to be trained and its corresponding hyper parameters to be adjusted in a single consistent environment. The entire training dataset is then used to achieve a precise model once the adequate values for hyperparameters have been achieved. The 10-fold CV is used to identify the optimum values for the adjustable hyperparameters based on the training dataset. During the CV process, the adjusted hyper parameter values are provided to achieve the overall best classification accuracy.

(4) The fourth step is to apply the machine algorithms (i.e., AdaBoost, logistic regression, extra tree, multinomial Naïve Bayes, support vector machine, linear discriminant analysis, classification and regression tree, random forest, and XGBoost) to the dataset obtained from step 2.

(5) In this step, the prediction model's performance is evaluated using different parameters, such as accuracy, precision, recall, and F-measure. The model that gives the highest prediction accuracy, precision, recall, and F-measures is selected. The accuracy metric assesses the precision or correctness of a machine learning or classifier model's predictions. Mathematically, it is given by equation (2).

$$\text{Accuracy} = \frac{\text{true positive (TP)} + \text{true negative (TN)}}{\text{TP} + \text{TN} + \text{false negative (FN)} + \text{false positive (FP)}}. \quad (2)$$
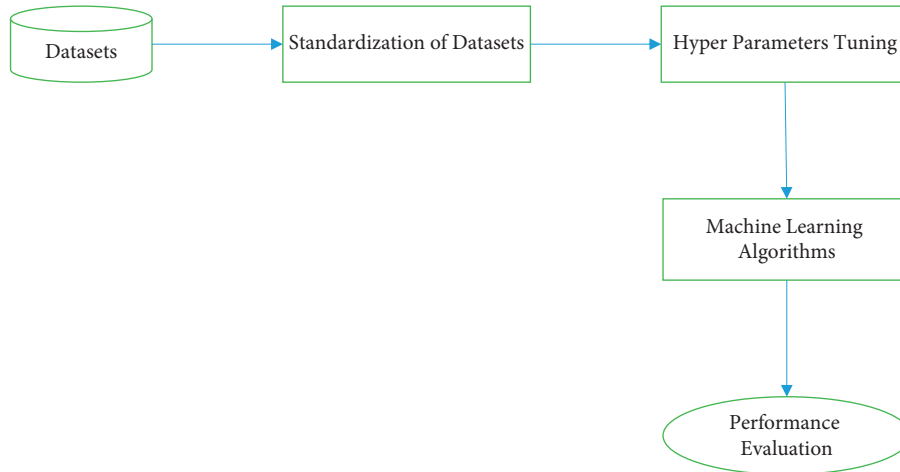
FIGURE 1: Block diagram of the proposed method.

Precision measures the predicted positive instances that are true/real positives. Mathematically, it is given by (3).

$$\text{precision} = \frac{\text{TP (true positives)}}{\text{TP (true positives)} + \text{FP (false positives)}}. \quad (3)$$

Recall evaluates the analysis of the total number of true/real positive instances as affected by the total number of false negative instances. Mathematically, it is given by (4).

$$\text{Recall} = \frac{\text{TP (true positives)}}{\text{TP (true positives)} + \text{FN (false negatives)}}. \quad (4)$$

An F-Measure is a harmonic mean of precision and recall. It takes the equilibrium between precision and recall, and mathematically, it is given by (5).

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5)$$

5. Section V: Data Collection

The Cleveland heart disease dataset, available from the University of California, Irvine (UCI) online repository for machine learning, is the most prominent dataset used by the researchers. There are 303 records, with 6 samples having missing values. The data has 76 features in its original form, however, all published work is likely to refer to 13 of them, while the other feature outlines the disease's effect. The Z-Alizadeh Sani dataset, which includes 303 patients' data with 55 input factors and a class label variable for each patient, is another popular dataset selected by researchers in the prediction process. The StatLog Heart, Hungarian, Long Beach VA, and Kaggle Framingham datasets are some of the additional datasets used by the researchers in the prediction process. The StatLog dataset has 270 records, each with 13 Cleveland-like attributes. The other two datasets, the Hungarian and Long Beach VA datasets, are collected from the UCI repository and consist of 274 records with 14 features each, similar to the Cleveland dataset. Researchers used publicly available datasets, such as Cleveland, Hungary, Switzerland, etc. There are different datasets available for this study. The first source for data is the Cleveland Clinic Foundation [40]. The second source for data is StatLog datasets that are accessible at [41]. The third source of data is the Z-Alizadeh Sani dataset, which is accessible at [42]. It contains 303 data samples and 55 attributes. To acknowledge the evaluation with literature, we used a publicly available resource, the Z-Alizadeh Sani dataset. Description and source of the datasets are given in Table 1.

6. Section VI: Experimental Results and Discussion

In Section 4, we discuss our experimental results. We collected the dataset from an online machine learning repository and refined and standardized it. After standardization, we performed hyperparameter tuning and applied machine learning classifiers. All the classifiers are trained and tested using 10-fold cross-validation. The accuracy of classifiers is also analyzed before and after standardized datasets. For evaluation purposes, the accuracy of the selected classifiers is plotted. Figure 2 shows the accuracy of classifiers before and after standardization data. From Figure 2, it is clear that most of the machine learning techniques (RF, CART, LDA, AB, LR, ET, and XGB) improved their accuracy, while MNB and SVM classifiers decreased their accuracy on the standardized dataset. Some classifiers, such as CART, ET, and AB, showed significant accuracy improvements on the standardized dataset. From Figure 1, it is evident that the ET and AB classifiers achieve the highest prediction accuracy of 90.16%. MNB shows the overall lowest performance and has the lowest accuracy of 59.01%. We also compare the accuracy before and after the standardization of the dataset. An accuracy of 90.16% is achieved by ET and AB classifiers, which shows the effect of the standardization of the dataset.

From the experimental results, it is clear that the accuracy of the classifier increased with hyperparameter tuning. We tune the selected classifiers by adjusting hyperparameter values to achieve the best accuracy. A set of accuracy with different hyperparameter combinations is achieved using 10-fold cross-validation. Since we have a small number of training examples, using test split is not a

TABLE 1: Risk factors of publicly available heart disease datasets.

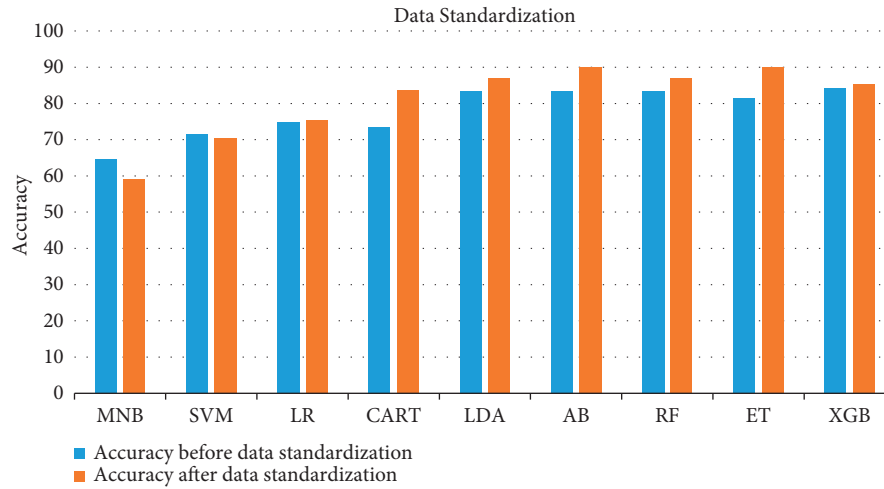| S.No.	Dataset	No. of samples	No. of risk factors	Source
1	UCI	303	13	[40]
2	StatLog	270	13	[41]
3	Z-Alizadeh Sani	303	55	[42]

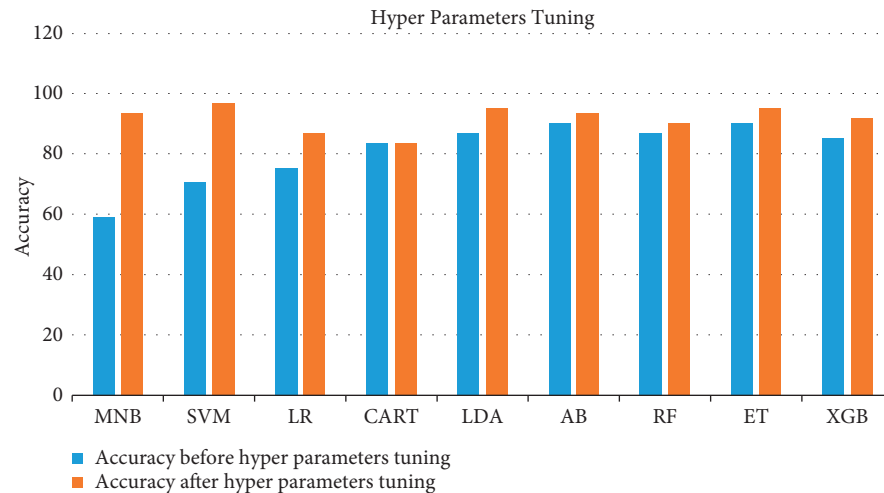FIGURE 2: Accuracy of classifiers before and after data standardization.

FIGURE 3: Accuracy of classifiers before and after hyper parameter tuning.

good option since we have fewer examples to train the model. Hence, we are using 10-fold cross-validation. The accuracy of the classifiers before and after hyperparameter optimization is presented in Figure 3. Most of the classifiers (MNB, RF, LR, LDA, AB, SVM, ET, and XGB) improved their accuracy on hyperparameter tuning, while the accuracy of CART alone was not changed. Table 2 shows the best combinations of hyperparameters for some algorithms to improve their accuracy.

Table 3 presents recall, precision, F-measure, and accuracy for the classifiers. A maximum precision of 98% is achieved by SVM for positive classes, although MNB and XGB classifiers have a maximum recall of 100%. However, SVM shows generally the best performance in recall, precision, F-measure,

and accuracy of 98%, 98%, 98%, and 96.76%, respectively. Precision is above 80% for all classifiers, whereas recall is above 90% for all classifiers. A maximum precision of 100% is achieved by XGB and MNB for the negative class. LR presents a small precision of 78%, and CART shows the lowest recall, F-measure, and accuracy of 61%, 69%, and 83.66%, respectively, where a maximum recall is achieved by SVM and LDA of 94.00%. Therefore, the negative class presented a comparatively poor recall of 61% and an F-measure of 69%, respectively. SVM shows comparatively good performance for negative classes, with a recall, precision, and F-measure of 94%, and an accuracy of 96.72%.

From the analysis of the results, it is clear that the SVM classifier achieved the best accuracy during hyperparameter

TABLE 2: Best values for hyper parameter tuning to improve classifiers accuracy.

Algorithm	Best Hyperparameter Values
SVM	Kernel: sigmoid, C:0.5
CART	Max features = "auto," random state = 123, min samples split = 20, min samples leaf = 11
AB	N estimators = 50, learning rate = 0.05
LDA	Shrinkage = "auto," solver = "lsqr"
GBM	N estimators:250
RF	Criterion = "gini," n jobs = −1, min samples leaf = 2, min samples split = 5, n estimators = 15, random state = 123
ET	N jobs = −1, min samples leaf = 1, n estimators = 15, random state = 123, criterion = "gini," Min samples split = 6
XBoost	objective = "reg: linear," colsample bytree = 0.3, learning rate = 0.1, max depth = 15, alpha = 5, n estimators = 123

TABLE 3: Classifiers evaluation in terms of precision, recall, F-measure, and accuracy.

Class	Metrics	Classifiers								
		MNB	SVM	LR	CART	LDA	AB	RF	ET	XGB
Positive	Precision (%)	91	98	91	85	98	93	91	95	90
	Recall (%)	100	98	91	93	95	98	95	98	100
	F Measure (%)	96	98	91	89	96	95	93	97	95
Negative	Precision (%)	100	94	78	79	89	93	88	94	100
	Recall (%)	78	94	78	61	94	83	78	89	72
	F Measure (%)	88	94	78	69	92	88	82	91	84
	Accuracy (%)	93.44	96.72	86.88	83.66	95.08	93.44	90.16	95.08	91.80

tuning. By comparing results obtained before and after standardized datasets, it is determined that the standardization of datasets has a positive impact on the accuracy improvement of most of the classifiers, and some classifiers show an accuracy improvement of up to 8.78%, which is a huge performance improvement.

By comparing the classifiers' accuracy on the normal and standardized datasets, we observed an improvement in the accuracy of most of the classifiers. Therefore, the standardization of the dataset is a useful technique for accuracy improvement before applying machine learning classifiers. Similarly, we have observed a significant accuracy improvement after hyperparameter tuning of the classifiers. Therefore, algorithm tuning is also a useful technique for improving the accuracy of the algorithms. From the comparison of different classifiers, we conclude that XGB and ET classifiers show overall good accuracy. However, SVM shows the best accuracy in tuning the hyperparameters and achieved an accuracy of 96.72%.

7. Section VII: Conclusions

The drawback of the prior proposed systems is that their operation is considerably reduced if the size of the dataset is increased. The main problem with machine learning is that a dataset cannot be classified efficiently, although it can be enhanced if the [43] attributes of the dataset are efficiently extracted. Another flaw is that the classifier prediction accuracy improves with increasing dataset magnitude, however, after a certain point, increasing dataset magnitude has a negative impact on the classifier prediction accuracy. According to the proposed method, using machine learning techniques for heart disease prediction improves accuracy and minimizes the cost factor. We have used different classifiers of machine learning to classify the prediction of heart disease, including an accuracy of 96.72% achieved by SVM.

For future work, we plan to use XGBoost for heart disease prediction in children and compare if better accuracy can be achieved. If features are properly managed, then there will be significant performance in the classification of heart disease prediction. In future studies, the outcomes of our proposed methods will serve as the standard performance results on heart disease.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Cardiovascular Diseases (Cvds), "World health organization," https://www.who.int/news-room/fact%20sheets/detail/cardio vascular-diseases-(cvds).

[2] A. K. Dwivedi, S. A. Imtiaz, and E. R. Villegas, "Algorithms for automatic analysis and classification of heart sounds - a systematic review," *IEEE Access*, vol. 7, 2019.

[3] A Coronary, "Heart disease," Available from: https://www.aihw.gov.au/reports/australias-health/coronaryheart-disease, 2020.

[4] L. A. Allen, L. W. Stevenson, K. L. Grady et al., "Decision making in advanced heart failure: a scientific statement from the American heart association," *Circulation*, vol. 125, no. 15, pp. 1928–1952, 2012.

[5] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *Journal of Intelligent Learning Systems and Applications*, vol. 5, no. 3, Article ID 35396, 2013.

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

[7] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of The Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[8] S. I. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method," *International Journal of Recent Technology and Engineering*, vol. 7, no. 6S, pp. 1009–1015, 2019.

[9] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzy logic based decision support system for component security evaluation," *The International Arab Journal of Information Technology*, vol. 15, no. 2, pp. 224–231, 2018.

[10] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *Journal of Intelligent Learning Systems and Applications*, vol. 5, no. 3, pp. 176–183, 2013.

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

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

[13] J. Mackay and G. A. Mensah, "The atlas of heart disease and stroke," Techical Report, World Health Org., Geneva, Switzerland, 2004.

[14] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informat. Med. Unlocked*, vol. 20, Article ID 100402, 2020.

[15] C. B. C. Latha, S. C. Jeeva, and S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, Article ID 100203, 2019.

[16] M. Manur, A. Kumar Pani, and P. Kumar, "A prediction technique for heart disease based on long short term memory recurrent neural network," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 2, pp. 31–33, 2020.

[17] M. Abdar, W. Książek, U. R. Acharya, R.-S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 179, Article ID 104992, 2019.

[18] I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," in *Proceedings of the 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon)*, Bengaluru, India, August 2017.

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

[20] G. E. Guraksin, U. Ergun, and O. Deperlioglu, "Classification of the heart sounds via artificial neural network," *International Journal of Reasoning-Based Intelligent Systems*, vol. 2, no. 3-4, pp. 272–278, 2010.

[21] R. K. Sinha, Y. Aggarwal, and B. N. Das, "Backpropagation artificial neural network classifier to detect changes in heart sound due to mitral valve regurgitation," *Journal of Medical Systems*, vol. 31, no. 3, pp. 205–209, 2007.

[22] A. Kandaswamy, C. S. Kumar, R. P. Ramanathan, S. Jayaraman, and N. Malmurugan, "Neural classification of lung sounds using wavelet coefficients," *Computers in Biology and Medicine*, vol. 34, no. 6, pp. 523–537, 2004.

[23] I. Ketut Agung Enriko, M. Suryanegara, and D. Agnes Gunawan, *Heart Disease Prediction System Using K-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters*, Springer, Berlin, Germnay, 2016.

[24] K. Subhadra and B. Vikas, "Neural network based intelligent system for predicting heart disease," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 5, pp. 484–487, 2019.

[25] S. N. Khan, N. M. Nawi, A. Shahzad, A. Ullah, and M. F. Mushtaq, "Comparative analysis for heart disease prediction," *International Journal on Informatics Visualization*, vol. 1, no. 4-2, pp. 227–231, 2019.

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

[27] S. Anitha and N. Sridevi, "Heart disease prediction using data mining techniques," *Journal of Analysis and Computation*, vol. 8, no. 2, pp. 48–55, 2019.

[28] A. Jagtap, P. Malewadkar, O. Baswat, and H. Rambade, "Heart disease prediction using machine learning," *International Journal of Research in Engineering, Science and Management*, vol. 2, no. 2, pp. 352–355, 2019.

[29] U. N. Dulhare, "Prediction system for heart disease using naïve bayes and particle swarm optimization," *Biomedical Research*, vol. 29, no. 12, pp. 2646–2649, 2018.

[30] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *Journal of Healthcare Engineering*, Article ID 2780501, 2017.

[31] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and A. Paul, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Reviews Cardiology*, vol. 18, 2021.

[32] P. S. Linda, W. Yin, P. A. Gregory, Z. Amanda, and G. Margaux, "Development of a novel clinical decision support system for exercise prescription among patients with multiple cardiovascular disease risk factors," *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, vol. 5, no. 1, pp. 193–203, 2021.

[33] Y. Ali, R. Amir, and A.-M. Fardin, "Profile-based assessment of diseases affective factors using fuzzy association rule mining approach: a case study in heart diseases," *Journal of Biomedical Informatics*, vol. 116, Article ID 103695, 2021.

[34] P. E. Rubini, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. G. Kumar, and T. M. Nithya, "'A cardiovascular disease prediction using machine learning algorithms," *Annual Romanian Society Cell Biology*, vol. 25, no. 2, pp. 904–912, 2021.

[35] D. Shah, S. Patel, and S. Kumar Bharti, *Heart Disease Prediction Using Machine Learning Techniques*, Springer Nature Singapore Pte Ltd, Berlin, Germany, 2020.

[36] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," *International Conference on Electrical & Electronics Engineering*, pp. 452–457, 2020.

[37] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in E-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.

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

[39] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, and A. S. Sulthana, "Analysis and prediction of cardio vascular disease using machine learning classifiers," in *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 15–21, IEEE, Coimbatore, India, 2020 March.

[40] Cleveland Clinic Foundation, "Heart disease data set," Available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[41] Statlog, "Heart disease data set," Available at: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29.

[42] Z. Alizadeh Sani, "Heart disease data set," Available at: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani.

[43] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proceedings of the IEEE Symposium Computer Communication (ISCC)*, pp. 204–207, Heraklion, Greece, July 2017.