Healthcare Analytics 3 (2023) 100133

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease

Subasish Mohapatra [a], Sushree Maneesha [a], Subhadarshini Mohanty [a], Prashanta Kumar Patra [a], Sourav Kumar Bhoi [b], Kshira Sagar Sahoo [c,*], Amir H. Gandomi [d,e,*]

[a] Odisha University of Technology and Research (Govt.), Bhubaneswar 751029, India
[b] Department of Computer Science and Engineering, Parala Maharaja Engineering College (Govt.), Berhampur 761003, India
[c] Department of Computing Science, Umeå University, SE-901 87, Umeå, Sweden
[d] Faculty of Engineering & Information Technology, University of Technology Sydney, Ultimo NSW 2007, Australia
[e] University Research and Innovation Center (EKIK), Óbuda University, 1034 Budapest, Hungary

ARTICLE INFO

Keywords:
Machine Learning techniques
Heart disease prediction
Electronic Health Records
Predictive modeling
Stacking classifiers

ABSTRACT

Cardiovascular Diseases (CVDs), or heart diseases, are one of the top-ranking causes of death worldwide. About 1 in every 4 deaths is related to heart diseases, which are broadly classified as various types of abnormal heart conditions. However, diagnosis of CVDs is a time-consuming process in which data obtained from various clinical tests are manually analyzed. Therefore, new approaches for automating the detection of such irregularities in human heart conditions should be developed to provide medical practitioners with faster analysis by reducing the time of obtaining a diagnosis and enhancing results. Electronic Health Records(EHRs) are often utilized to discover useful data patterns that help improve the prediction of machine learning algorithms. Specifically, Machine Learning contributes significantly to solving issues like predictions in various domains, such as healthcare. Considering the abundance of available clinical data, there is a need to leverage such information for the betterment of humankind. Researchers have built various predictive models and systems over the years to help cardiologists and medical practitioners analyze data to attain meaningful insights. In this work, a predictive model is proposed for heart disease prediction based on the stacking of various classifiers in two levels(Base level and Meta level). Various heterogeneous learners are combined to produce strong model outcomes. The model obtained 92% accuracy in prediction with precision score of 92.6%, sensitivity of 92.6%, and specificity of 91%. The performance of the model was evaluated using various metrics, including accuracy, precision, recall, F1-scores, and area under the ROC curve values.

1. Introduction

Data analytics merged with the power of Machine Learning (ML) has attracted a lot of attention across various domains due to its problem-solving ability. ML has diverse applications throughout these domains, such as speech recognition, medicine, business, social media, etc. Many breakthroughs have been driven by machine learning's use of neural networks, referred to as deep learning, which is a set of algorithms that enables the discovery of patterns and insights in large datasets. These techniques and frameworks can be deployed for information extraction, predictions, representation learning, outcome predictions, and de-identification.

In the healthcare sector, there are numerous areas in which ML has proven to be very beneficial. Considering the exponential growth of digital real-time information generated by the healthcare sector (e.g., Electronic Health Records(EHRs), wearable devices, diagnostics reports, etc.) [1], it is pertinent to develop smart systems to process such medical data. Encouraging the healthcare sector to adapt to digitization would help to increase the storage of a lot of useful data. Such data can cater to a variety of research, such as population aging, the recent advancement of new treatment plans and their effectiveness, health habits across age groups, and medical expense reports. In the bigger picture, these analytics would provide statistics for institutions, private parties, and governments across the region to make better medical policies and also refine the existing ones in use [2].

According to recent findings, cardiovascular diseases are the top listed cause and disease responsible for deaths in individual worldwide [3]. The World Health Organization (WHO) predicts a steep increase in deaths due to poor heart health and identifies various factors that are harmful to the heart [4]. In the USA, heart diseases rank as the highest cause of adult deaths. Most of the time, individuals are not even aware of underlying issues that may lead to heart failure, strokes, or even blocked arteries. Changes in lifestyle and health-related factors, such as food habits, smoking, consumption of high saturated fat, lack

* Corresponding authors.
E-mail addresses: ksahoo@cs.umu.se (K.S. Sahoo), gandomi@uts.edu.au (A.H. Gandomi).

https://doi.org/10.1016/j.health.2022.100133
Received 26 June 2022; Received in revised form 18 November 2022; Accepted 27 December 2022

of exercise, diabetes, and blood disorders, are some of the contributors. Patients who may be at risk of heart disease should be referred to cardiologists to determine the best-fit treatment to prevent any undesired events as soon as possible [5]. Traditionally, health practitioners have utilized various tests, such as blood work, electrocardiogram (ECG), echocardiogram, angiography, testing for diabetes, blood pressure, etc., and then screening the results. However, this screening can be a tedious process if a doctor has to go through hundreds of such reports. Also, the diagnosis and further treatment are costly and time-consuming. To reduce the time for manually checking the huge amount of data, we propose to incorporate ML algorithms to perform the data analytics, which would cut down the processing time significantly and allow cardiologists to spend more time preparing treatment plans.

ML algorithms are efficient in terms of detecting patients who might be at risk from an early stage, which will then help to reduce the overall costs of treatment [6]. Various machine learning classifiers are used for prediction as well as regression tasks. In the healthcare industry, they need to be reliable and show good performance with respect to medical data [7]. The main objective of these algorithms is the timely detection of cardiovascular disease before severe complications arise. Misclassifying a patient with heart disease as negative has much higher complications compared to misclassifying a healthy patient as having a disease [8]. Several studies report the use of machine learning-driven methods in achieving predictions and significantly reducing the cost of healthcare. With the amount of electronically available medical data, we aim to deliver better quality of healthcare service. As the world gradually moves towards complete digitization, computing devices are used to take notes and perform documentation, and the data availability issue is suspected to be gone in a few years. Therefore, it is pertinent to develop the proper tools to address the need for analytics with ML to improve the quality of human health.

Various studies have implemented ML techniques for the diagnosis of heart diseases. Traditional classifiers have been shown to perform well with proper model generation. The performance of such classifiers can be improved by implementing various techniques [9]. In the work described in this paper, the performance of various algorithms is improved by the implementation of the stacking technique. Specifically, in the concept of stacking, training is performed in level 1 with traditional ML classifiers, and then the output is fed to the next level, also known as a meta-level [10–12].

In subsequent sections of this paper, we discuss the proposed workflow, pre-processing of data, model generation, and the performance evaluation of the proposed workflow followed by a conclusive discussion and future work.

2. Literature review

Numerous studies have demonstrated the effective application of ML models in the detection of heart diseases. The UCI Heart Disease Dataset from UCI Machine Learning Repository is open to the public and is one of the most used datasets in this research area [13]. The Statlog dataset is also widely used [14]. In the clinical detection of diseases, such ML models aim to improve accuracy and reduce the total cost of the computation. For example, Verma et al. [15] proposed a hybrid model using particle swarm optimization (PSO) and two machine learning classifiers, namely K-nearest neighbor and multi-layer perceptron (MLP), for the prediction of heart disease, which achieved a 90.28% accuracy. Aakash Chauhan et al. (2018) introduced a model that extracts data from EHRs based on association rule generation and utilizes ML association mining for frequent pattern growth in a dataset. The model helps to achieve an overall outlook of a patient's data and underlying patterns in the dataset [16].

Saqlain et al. [17] developed a model using the Fisher score algorithm for feature selection and SVM classifier for the prediction model, which achieved an accuracy of 81.91%, sensitivity of 72.92%, and specificity of 88.68%. Latha and Jeeva [18] designed a hybrid model that implements four ML classification algorithms, namely NB, BN, RF, and MP, and incorporates various ensemble learning methods, obtaining an accuracy of 85.48%. Beunza et al. (2019) used a machine learning approach to how machine learning methods can be of great use for diagnosis with small datasets. They used R studio for the computations. Various methods such as decision tree, boosted decision trees (DTs), random forest, support vector machine (SVM), neural networks(NN), and logistic regression were tested. The highest accuracy obtained was 85% by boosted decision trees [19]. Subrat Kumar Nayak et al. (2020) emphasize the use of feature selection. The work includes 23 datasets, one of which is the heart disease dataset. Using filter methods, 13 feature subset was chosen followed by 10-fold cross-validation [20]. Liyuan Gao et al. (2020) proposed sampling and substitution methods for the Bayesian hyper-parameter optimization technique. Then compared various ML classifiers to detect irregularities. For breast cancer, 94% accuracy and for heart disease, 73.40% of accuracy was obtained [21]. Ivan Miguel Pires et al. (2021) experimented with multiple classifiers such as SVM KNN, DT, neural networks, combined nomenclature (CN2) rule inducer. All the selected classifiers underwent 5-fold cross-validation, 10-fold cross-validation, and 20-fold cross-validation. The best accuracy score of 87.69% was obtained by DT, SVM and SGD at 20,10, and 5 fold respectively [22].

3. Materials

Fig. 1 presents the workflow of the model, where data are initially acquired from the source, converted into a dataset, and then pre-processed. The model generation subsequently occurs, followed by analysis of the results. Each step of the model is discussed in detail in subsequent sections.

3.1. Dataset description

In this research, the UCI Heart Disease Dataset from the UCI Machine Learning Repository was selected as the open dataset taken, named which is available online. It is a combination of 4 datasets collected from Cleveland Clinic Foundation, Medical Centre Long Beach, Hungarian Institute of Cardiology and University Hospital Switzerland.

The dataset is comprised of 303 instances of records, out of a total of 76 attributes. In this study, only 13 attributes and one target attribute were taken into consideration. Table 1 describes the attributes of the UCI dataset, specifically 8 categorical and 6 numeric attributes. The dataset s a combination of different clinical test result data, such as serum cholesterol, fasting blood sugar, vessel count, and thalassemia detected from blood work. ST depression and slope of ST-segment were obtained from the electrocardiogram.

4. Methods

4.1. Pre-processing of the dataset

In this study, for the initial step of data pre-processing, we performed outlier detection. To improve the model's performance, Z-score outlier detection was used. Based on the empirical rule, the data point is considered to be an outlier in a distribution where the z-score is greater than 3. A z-score, or standard score, is the value that represents how far a data point is from the mean value, indicating the variability of an attribute's value in a dataset.

$$Z_{score} = \frac{x - \mu}{\sigma} \tag{1}$$

$$x_{normalized} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

For categorical attributes, such as sex, chest pain (cp), resting electrocardiograph(restecg), and slope of the st segment(slope), one

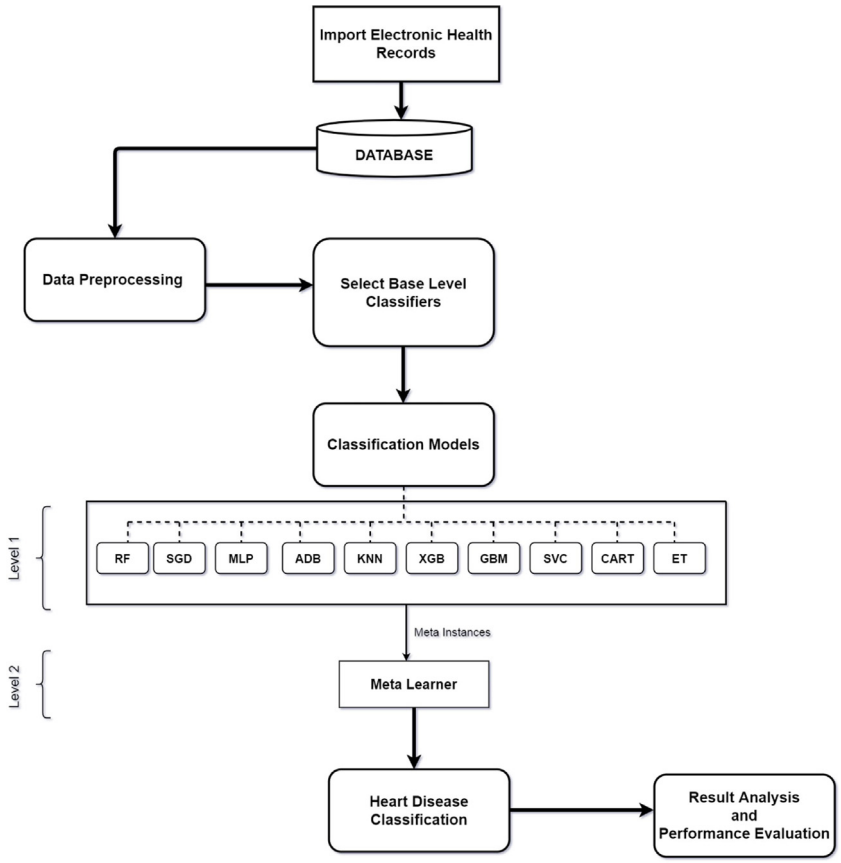

Fig. 1. Workflow of the model.

Table 1
Attribute description of the UCI heart disease dataset.

Sr no.	Attribute	Description
1	Age(age)	Age of the patient (in years)
2	Sex(sex)	Gender (0 = Female and 1 = Male)
3	Chest pain(cp)	1 = Typical angina, 2 = Atypical angina, 3:Non-anginal pain, 4:Asymptomatic pain
4	Resting blood pressure(trestbps)	Resting blood pressure (in mm Hg)
5	Serum cholesterol(Chol)	Serum cholesterol level (in mg/dl)
6	Fasting blood sugar(fbs)	Fasting blood sugar (>120 mg/dl 0 = False, 1-True)
7	Rest electrocardiograph(restecg)	Resting ECG (0 = Normal, 1 = ST-T wave abnormality, 2 = LV Hypertrophy)
8	Maximum heart rate(thalach)	Maximum heart rate achieved
9	Exercise-induced angina(exang)	Exercise-induced angina (0 = No, 1 = Yes)
10	ST depression(oldpeak)	ST depression induced by exercise relative to rest
11	Slope of ST segment(slope)	Slope of peak exercise ST segment (1 = up sloping, 2 = flat, 3 = down sloping)
12	Vessel count(ca)	Number of major vessels colored by fluoroscopy (range 0–3)
13	Thalassemia(thal)	Thalassemia type (normal, fixed defect, reversible defect)
14	Heart disease(target)	0 = negative of disease, 1 = positive for heart disease

hot encoding was applied [23]. In this method, the attribute is converted into a numerical interpretable form for better adaptability and performance with machine learning algorithms.

For feature scaling, two datasets were pre-processed for preliminary analysis. For the first dataset, attribute values were standardized using the standard scaler. For the second dataset, values were normalized using the min–max scaler. In the min–max scaler, the values are in the range of 0 to 1, where 0 is the minimum value found and 1 denotes the maximum. The rest of the data are decimals in the range of 0 to 1.

4.2. Correlation heatmap of the dataset

Visualization of the dataset is an important part of the preprocessing step. Various methods of visualization give an overall idea of how the dataset is in a broad picture. Graphs such as bar graphs, charts, histograms, density estimate plots, etc provide a visual representation of the data for analysis. The correlation heatmap Fig. 2(a), depicts how the attributes of the taken dataset correlate to the target attribute (the attribute that denotes if a person has heart disease or not). The matrix represents the correlation coefficient of all the pairs of attributes. Heatmap visualization is a 2-dimensional representation of

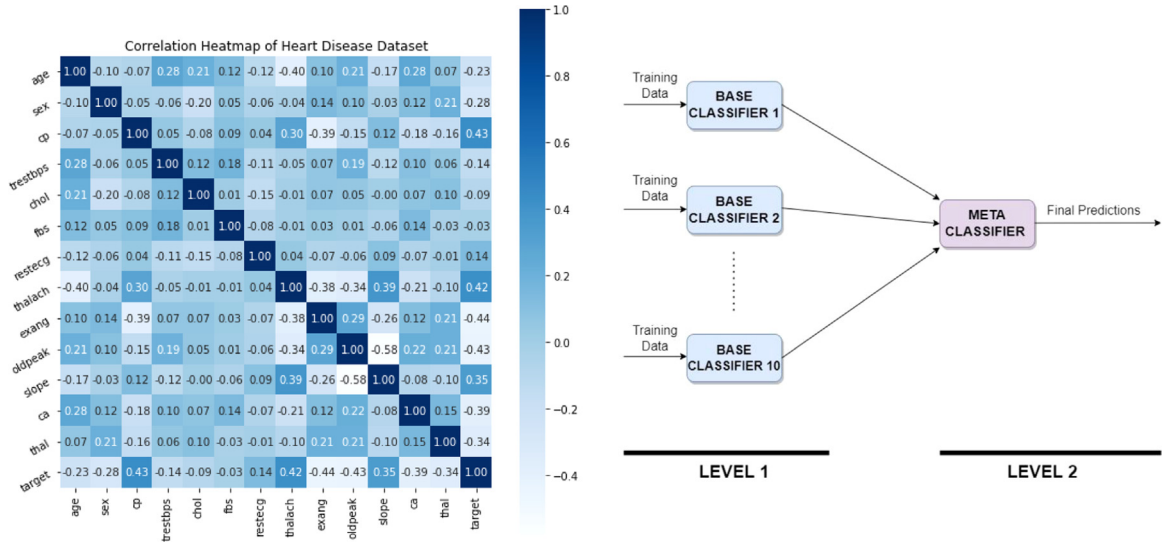

Fig. 2. (a) Heatmap of the dataset attributes; (b) Stacking concept diagram. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

some selected or all the features of a dataset, where the intensity level of color shows the magnitude of variation in comparison to the other features. The darkest hue of color represents the attribute having the highest correlation with the target attribute and the lightest hue is the least correlating attribute.

4.3. Model building

After the data are pre-processed, the next step is to generate the model. The proposed work primarily focuses on the ML method of stacking, as shown in Fig. 2(b), in which various machine learning classifiers are combined in two levels, which generate a higher predictive model performance. Level 1, also known as the "base level" or "base learners", contains the set of traditional ML algorithms. Following that, the second level, known as the "meta level" or "meta learners", takes the input from the former layer. The main advantage of stacking algorithms in two levels is the utilization of the heterogeneous nature of multiple classification algorithms. This heterogeneity is where weak learners prove to be essential because of their diverse nature. Every classifier comes with certain strengths and drawbacks. Stacking helps to combine the best scenarios from the chosen classifiers. At the base level, various classifiers fit the training data and give predictions. Then, the meta level figures out the best way to maximize the strengths of each classifier and produce the final optimal prediction results.

Algorithm 1 : Base Level (Level 1 model)

Let D = $\{d_1,d_2,d_3,...d_n\}$ given UCI dataset
X = training dataset
Y = test dataset, for the final validation
X ⊂ D
Y ⊂ D
Input : X_t, for training base models
X_v, for testing base models

S_{10}= {RF,MLP,KNN,ET,XGB,SVC,SGD,ADB,CART,GBM}
for every i in S_i

S_{10}={ RF,MLP,KNN,ET,XGB,SVC,SGD,ADB,CART,GBM}
for j= 1,2,3...10
do
train and test every fold for every classifier S_i
end **for**
//mean of test data results of baseline models, with standard deviation
end **for**
Output: Summarized model scores

For the base learners, classifiers include Random forest (RF), Multilayer perceptron(MLP), k-Nearest Neighbor(KNN), Extra Trees Classifier(ET), Extreme Gradient Boosting(XGB), Stochastic Gradient Descent(SGD), Support Vector Classifier(SVC), Adaptive Boosting(ADB), Decision Tree(CART), and Gradient Boosting (GBM) classifier. In addition, Logistic Regression (LR) and Naive Bayes (NB) were used in the preliminary experimental steps. As the number of classifiers used for the base level affects the overall performance, we selected 10 of the above-mentioned classifiers. Having a diverse set of base learners is essential as they produce results based on different assumptions.

After data were pre-processed, 10-fold cross-validation of the training data was performed to avoid model over-fitting. The total number of entries was divided into 10 sections, also known as folds, after reshuffling the data to avoid biased predictions. In every step of cross-validation, a particular fold was treated as test data and the rest as training data. This process follows a total of 10 iterations. Due to the limited number of instances available in the dataset, cross-validation was conducted for performance comparison. This ensures that the bias and variance of an algorithm are reduced to clearly show how well that particular algorithm performs with the taken dataset.

For both datasets, averaged-out scores of accuracies, precision, and recall were recorded with their standard deviation. The comparison is shown in Table 2. Specifically, 75% of the data was used for training and 25% for testing the models. The 10 ML classifiers selected for the base learners were RF, MLP, KNN, ET, XGB, SVC, SGD, ADB, CART, and GBM.

For the meta-learner level, various classifiers were tested, and their performance scores are provided in Table 3. Multi-Layer Perceptron (MLP) classifier was selected as the meta-learner. Due to its adaptive learning feature, MLP was chosen since it learns and can be trained in real-time and is best suited for non-linear data, which is a good fit for a classification problem with a predictor label. GBM and MLP performed the same in terms of accuracy, but GBM produced higher false-positive predictions, as shown in Fig. 3(a) and (b). Therefore, MLP is considered a better fit for the particular scenario where misclassification of a positive class is undesirable in clinical diagnosis.

5. Results and discussion

The above-mentioned procedure was performed on a 64-bit machine with a 4th Gen Intel i5 CPU (8 GB DDR3+1 TB Hard drive+20 GB SSD). Python was chosen as the language for the machine learning tasks on the Jupiter notebook 3.7.2. Considering the nature of medical data, we

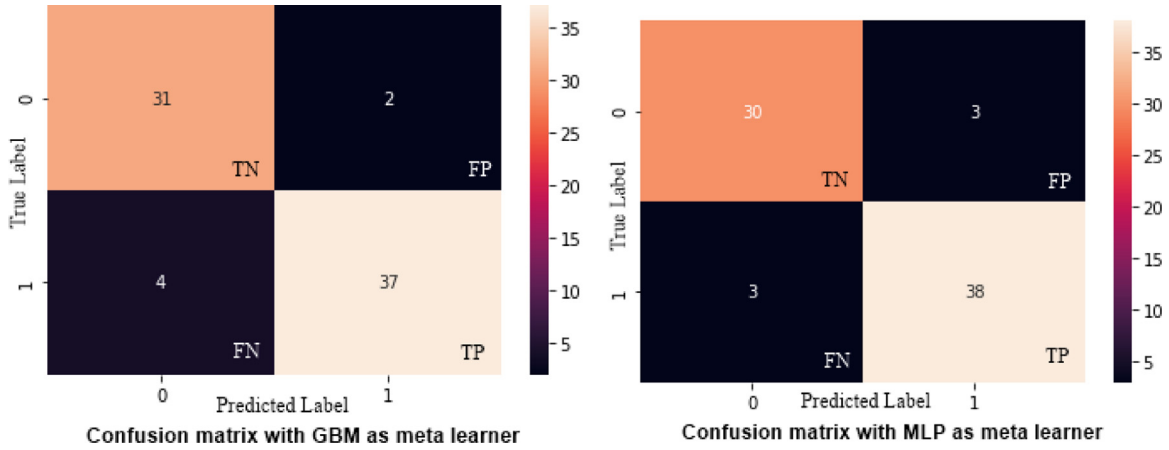

Fig. 3. Confusion matrix with (a) GBM as meta-learner; (b) MLP as meta-learner.

Table 2
Performance comparison of data.

Classifiers	Standardized data			Normalized data		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
RF	0.779 (±0.075)	0.807 (±0.087)	0.732 (±0.136)	0.805 (±0.068)	0.794 (±0.125)	0.773 (±0.126)
MLP	0.806 (±0.061)	0.791 (±0.102)	0.833 (±0.077)	0.815 (±0.098)	0.810 (±0.112)	0.833 (±0.111)
KNN	0.783 (±0.091)	0.782 (±0.127)	0.832 (±0.086)	0.769 (±0.092)	0.777 (±0.128)	0.831 (±0.102)
ET	0.787 (±0.087)	0.806 (±0.117)	0.828 (±0.111)	0.796 (±0.074)	0.805 (±0.114)	0.856 (±0.112)
XGB	0.783 (±0.061)	0.776 (±0.108)	0.851 (±0.079)	0.783 (±0.061)	0.776 (±0.108)	0.851 (±0.079)
SVC	0.734 (±0.122)	0.743 (±0.147)	0.782 (±0.177)	0.756 (±0.109)	0.750 (±0.114)	0.825 (±0.150)
LR	0.793 (±0.087)	0.789 (±0.116)	0.858 (±0.113)	0.783 (±0.102)	0.791 (±0.125)	0.825 (±0.121)
SGD	0.703 (±0.122)	0.765 (±0.126)	0.833 (±0.161)	0.719 (±0.114)	0.837 (±0.151)	0.693 (±0.180)
ADB	0.743 (±0.075)	0.749 (±0.113)	0.779 (±0.096)	0.743 (±0.075)	0.749 (±0.113)	0.779 (±0.096)
CART	0.721 (±0.077)	0.738 (±0.126)	0.773 (±0.126)	0.711 (±0.067)	0.750 (±0.107)	0.740 (±0.132)
GBM	0.775 (±0.062)	0.784 (±0.110)	0.833 (±0.070)	0.792 (±0.078)	0.806 (±0.112)	0.824 (±0.078)
NB	0.680 (±0.102)	0.632 (±0.112)	0.782 (±0.177)	0.680 (±0.102)	0.632 (±0.112)	0.693 (±0.180)

Table 3
Accuracy performance of various classifiers as meta-learner.

Classifiers	Accuracy scores	Precision	Recall
RF	0.8918	0.9230	0.8780
LR	0.9054	0.9250	0.9024
KNN	0.8918	0.8837	0.9268
NB	0.8918	0.9024	0.9024
MLP	0.9189	0.9268	0.926
ET	0.8783	0.9210	0.8536
XGB	0.8918	0.9230	0.8780
SVC	0.8783	0.9000	0.8780
SGD	0.8918	0.8837	0.9268
ADB	0.8783	0.8869	0.9024
CART	0.8918	0.9024	0.9024
GBM	0.9189	0.9487	0.9024

tested and recorded multiple performance metrics, including accuracy, precision, recall, and area under the ROC curve for evaluation [24–26].

As shown in Table 2, 12 different classifiers were tested on the dataset. Due to the limited number of records available in the dataset, cross-validation was performed as it gives the model multiple folds of data for training and testing in order to avoid over-fitting as well as reduce bias. Then, the results were recorded as mean values.

Algorithm 1 above shows a 10-fold cross-validation process. The 10 top-performing classifiers were selected for the next step. One dataset was standardized using the standard scaler function, and the other was normalized using the min–max scaler function. As discussed in the pre-processing section, outliers were removed prior to the standardization and normalization process. Normalized data performed better than standardized data in the majority of the classifier cases. Therefore, normalized data were used in subsequent experiments. Due to the heterogeneous nature of ML methods, as we can see some classifiers have good precision and recall scores but fail to achieve good accuracy. After the final predictions were made, the performance metrics were calculated based on the confusion matrix and the values of True Positive (TP), True Negatives(TN), False Positive(FP), and False Negative(FN) labels. In Fig. 3(b), the confusion matrix exhibits 3 false positives and 3 false negatives predictions.

Accuracy is the measure of a correctly classified class. Precision is the measure of the ratio between true positives to the total number

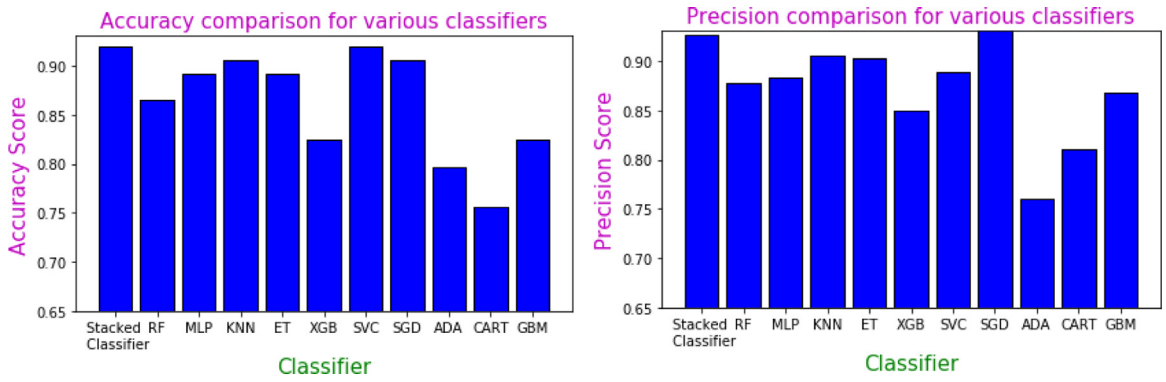

Fig. 4. (a) Comparison of accuracy scores; (b) Comparison of precision scores.

Table 4
Comparative result of model with classifiers.

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score	ROC	Log loss	Matthews correlation co-efficient
GBM	0.824	0.868	0.804	0.848	0.835	0.826	6.067	0.649
RF	0.864	0.878	0.878	0.848	0.878	0.863	4.667	0.726
MLP	0.891	0.883	0.926	0.848	0.904	0.887	3.733	0.781
KNN	0.905	0.904	0.926	0.878	0.915	0.902	3.267	0.808
ET	0.891	0.902	0.902	0.878	0.902	0.890	3.733	0.781
XGB	0.824	0.850	0.829	0.818	0.839	0.823	6.067	0.645
SVC	0.918	0.888	0.975	0.848	0.930	0.912	2.800	0.839
SGD	0.905	0.947	0.878	0.939	0.911	0.908	3.267	0.812
ADB	0.797	0.760	0.926	0.636	0.835	0.781	7.001	0.598
CART	0.756	0.810	0.731	0.787	0.769	0.759	8.401	0.516
Stacked classifier	**0.918**	**0.926**	**0.926**	**0.909**	**0.926**	**0.917**	**2.800**	**0.835**

Table 5
Parameters specified for various classifiers.

Classifiers	Parameters
KNN	k = 5
SVC	kernel = 'linear', gamma = 'auto', probability = 'true'
LR	penalty = l2
GBM	n_estimators = 100, max_features = 'sqrt'
RF	criterion = 'entropy', n_estimators = 10
MLP	Alpha = 0.0001
ET	n_estimators = 100
SGD	max_iter = 100

of cases classified as positives (true positives and false positives). For any medical predictive model, a good recall score (also known as sensitivity) and specificity should be maintained. Recall is a measure of how well the model correctly identifies true positive cases. Specificity is the number of correctly classified negatives by the model to actual negative cases.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$Sensitivity(recall) = \frac{TP}{TP + FN} \tag{5}$$

F1-score is a performance metric, which is the resulting score combination of precision and recall's harmonic mean. The Area Under the Curve (AUC) of the Receiver Characteristic Operator (ROC) is a graph of the true positive rate against the false positive rate, which is used for binary classification problems. A higher ROC score indicates that the model performs well.

Matthews Correlation Coefficient (MCC) is a performance metric that produces a high score when a model performs well across all four categories (TP, TN, FP, and FN).

$$Specificity = \frac{TN}{FP + TN} \tag{6}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

Table 4 displays the performance comparison of all classifiers based on the 8 performance metrics, including accuracy, precision, sensitivity or recall, specificity, F1-score, AUC-ROC, log loss value, and MCC.

The stacked classifier outperformed the other ML classifiers with an accuracy of 92%, precision of 92.6%, sensitivity of 92.6%, and specificity of 90.9%. SVC closely follows with an accuracy of 91%.

Table 5 lists the parameters defined for some of the tested classifiers, such as KNN, SVC, GBM, RF, ET, and SGD [30–49]. For the other classifiers, their default parameters were used. Values were pre-defined by the sci-kit learn library. Comparisons of all traditional classifiers with the stacked classifier based on accuracy, precision, and recall are illustrated as bar graphs in Figs. 4(a), (b), and 5(a) respectively. Table 6 presents the performance comparison of our proposed stacked model with existing approaches in the literature that used the same heart disease dataset.

The AUC-ROC curve was also used for performance evaluation, which depicts how well a classifier is able to distinguish between two classes [50]. In clinical scenarios, the ability to discriminate the data between positive and negative classes is of prime importance.

Fig. 5(b), compares the ROC score of the proposed stacked classifier with other classifiers, where a higher area under the curve indicates good performance. In the graph, we can see stacked classifier's curve has a greater area with an inclination towards the true positive rate axis.

6. Conclusion and future work

In this work, we propose an effective model that incorporates data pre-processing with outlier detection and the stacking of classifiers for predicting heart diseases. The data used was first normalized so as to ensure that the distribution of data is even and on a similar scale. This ensures the training stability of the model and gives better performance.

Table 6
Result comparison with existing approaches for heart disease predictions—UCI dataset.

Author	Methods	Accuracy	Precision	Recall
Christalin Latha, et al.	NB, BN, RF, MP with majority voting	85.48%	N/A	N/A
Mohan et al. [27]	HRFLM	88.4%	90.1%	92.8%
Haq et al. [28]	KNN with SBS feature selection	90%	N/A	N/A
Kavitha et al. [29]	Hybrid DT and RF	88%	N/A	N/A
Saqlain et al.	SVM with Fisher score + MCC feature subset selection	81.19%	N/A	N/A
Proposed model	**Stacked ensemble classifiers**	**91.8%**	**92.6**	**92.6%**

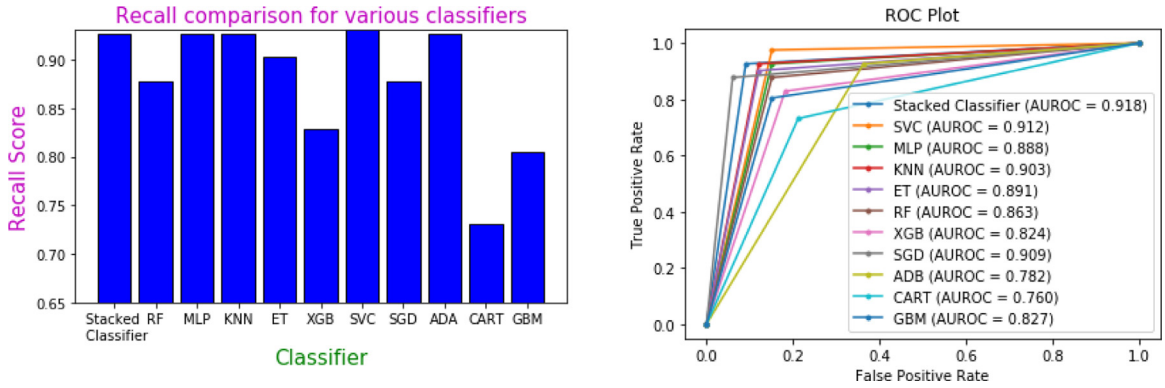

Fig. 5. (a) Comparison of recall scores; (b) Comparison of AUC-ROC values.

Herein, we used 10 different classifiers with different strengths, such as instance-based (e.g. KNN), probabilistic (e.g.NB), and a few ensemble (e.g. XGB and GB) classifiers. Considering those different methods for prediction, we stacked various classifiers to take advantage of their differences in strengths. Using MLP as the meta-learner, we obtained results with 92% accuracy. The proposed stacked classifier outperformed the traditional machine learning classifiers better in terms of overall parameter comparison with a precision of 92.6%, sensitivity of 92.6%, and specificity of 91%. The proposed model exhibits the advantages of combining weak learners and using their heterogeneity to strengthen overall prediction results.

Heart diseases often result in poor quality of life or even death. Therefore, early treatment could help to save many lives if CVDs are predicted on time. However, it is not manually feasible for cardiologists to analyze the large amount of data acquired for a patient to make a timely treatment plan. Hence, primary screening by machine learning-based systems is a promising solution. Such systems need to be reliable and efficient for diagnosing patients by predictive analysis. The aim and use of such predictive models in the healthcare sector will help save lives and make sure no patient with heart disease is left undiagnosed. In this work, our method achieved good accuracy with high sensitivity in predicting patients with heart diseases.

The demonstrated high sensitivity indicates that the model has fewer false-negative results compared to traditional approaches. In other words, our method will ensure that no patient with heart disease is misdiagnosed or classified as negative for the disease. Importantly, this will allow the cardiologist to quickly establish the appropriate plan of action.

For future work, we plan to evaluate and test the proposed model on various other datasets. The limited amount of data, instances, and number of attributes is the main issue facing machine learning approaches. In the future, more work and research can be done if we are able to acquire a greater quantity of good-quality medical data by collaborating with hospitals and other data-producing entities.

Data availability

Data will be made available on request.

References

[1] K Shailaja, B Seetharamulu, M.A. Jabbar, Machine learning in healthcare: A review, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology, ICECA, 2018, pp. 910–914, http://dx.doi.org/10.1109/ICECA.2018.8474918.

[2] SS Virani, A Alonso, EJ Benjamin, MS Bittencourt, CW Callaway, AP Carson, AM Chamberlain, AR Chang, S Cheng, FN Delling, L Djousse, MSV Elkind, JF Ferguson, M Fornage, SS Khan, BM Kissela, KL Knutson, TW Kwan, DT Lackland, TT Lewis, JH Lichtman, CT Longenecker, MS Loop, PL Lutsey, SS Martin, K Matsushita, AE Moran, ME Mussolino, AM Perak, WD Rosamond, GA Roth, UKA Sampson, GM Satou, EB Schroeder, SH Shah, CM Shay, NL Spartano, A Stokes, DL Tirschwell, LB VanWagner, CW Tsao, American Heart Association Council on Epidemiology, Prevention Statistics Committee, Stroke Statistics Subcommittee, Heart disease and stroke statistics-2020 update: A report from the American heart association, Circulation 141 (9) e139–e596, http://dx.doi.org/10.1161/CIR.0000000000000757, Epub 2020 Jan 29. PMID: 31992061.

[3] Centers for Disease Control and Prevention (CDC), Deaths:leading causes, 2020, URL: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.

[4] World Health Organization(WHO), Cardiovascular Diseases(CVDs) [Online]. URL: https://www.who.int/health-topics/cardiovascular-diseases.

[5] F Amato, A Lopez, EM Pena-Mendez, P Vanhara, A Hampl, J. Havel, Artificial neura; networks in medical diagnosis, J. Appl. Biomed. 11 (2) (2013) 47–58.

[6] M Motwani, D Dey, D.S Berman, G Germano, S Achenbach, M.H Al-Mallah …, P.J. Slomka, Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis, Eur. Heart J. (2016) ehw188, http://dx.doi.org/10.1093/eurheartj/ehw188.

[7] J Thomas, R.T. Princy, Human heart disease prediction system using data mining techniques, in: 2016 International Conference on Circuit, Power and Computing Technologies, ICCPCT, 2016, pp. 1–5, http://dx.doi.org/10.1109/ICCPCT.2016.7530265.

[8] A Gavhane, G Kokkula, I Pandya, K. Devadkar, Prediction of heart disease using machine learning, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology, ICECA, 2018, pp. 1275–1278, http://dx.doi.org/10.1109/ICECA.2018.8474922.

[9] Na Liu, Jiang Shen, Man Xu, Dan Gan, Er-Shi Qi, Bo Gao, Improved cost-sensitive support vector machine classifier for breast cancer diagnosis, Math. Probl. Eng. 2018 (2018) 3875082, http://dx.doi.org/10.1155/2018/3875082, 13 pages.

[10] B. Pavlyshenko, Using stacking approaches for machine learning models, in: 2018 IEEE Second International Conference on Data Stream Mining & Processing, DSMP, 2018, pp. 255–258, http://dx.doi.org/10.1109/DSMP.2018.8478522.

[11] B Zenko, L Todorovski, S. Dzeroski, A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods, in: Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp. 669–670, http://dx.doi.org/10.1109/ICDM.2001.989601.

[12] David H. Wolpert, Stacked generalization, Neural Netw. (ISSN: 0893-6080) 5 (2) (1992) 241–259, http://dx.doi.org/10.1016/S0893-6080(05)80023-1.

[13] UCI Heart Disease Data set. [Online] Available: https://archive.ics.uci.edu/ml/datasets/heart+disease.
[14] Statlong Heart Data set. [Online] Available: https://archive.ics.uci.edu/ml/datasets/statlog+(heart).
[15] L Verma, S Srivastava, P.C. Negi, A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data, J. Med. Syst. 40 (2016) 178, http://dx.doi.org/10.1007/s10916-016-0536-z.
[16] A Chauhan, A Jain, P Sharma, V. Deep, Heart disease prediction using evolutionary rule learning, in: 2018 4th International Conference on Computational Intelligence & Communication Technology, CICT, 2018, pp. 1–4, http://dx.doi.org/10.1109/CIACT.2018.8480271.
[17] S.M Saqlain, M Sher, F.A Shah, et al., Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines, Knowl. Inf. Syst. 58 (2019) 139–167, http://dx.doi.org/10.1007/s10115-018-1185-y.
[18] C Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Inform. Med. Unlocked (ISSN: 2352-9148) 16 (2019) 100203, http://dx.doi.org/10.1016/j.imu.2019.100203.
[19] Juan-Jose Beunza, Enrique Puertas, Ester García-Ovejero, Gema Villalba, Emilia Condes, Gergana Koleva, Cristian Hurtado, Manuel F. Landecho, Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease), J. Biomed. Inform. (ISSN: 1532-0464) 97 (2019) 103257, http://dx.doi.org/10.1016/j.jbi.2019.103257.
[20] Subrat Kumar Nayak, Pravat Kumar Rout, Alok Kumar Jagadev, Tripti Swarnkar, Elitism based multi-objective differential evolution for feature selection: A filter approach with an efficient redundancy measure, J. King Saud Univ. - Comput. Inf. Sci. (ISSN: 1319-1578) 32 (2) (2020) 174–187, http://dx.doi.org/10.1016/j.jksuci.2017.08.001.
[21] L Gao, Y. Ding, Disease prediction via Bayesian hyperparameter optimization and ensemble learning, BMC Res. Notes 13 (2020) 205, http://dx.doi.org/10.1186/s13104-020-05050-0.
[22] Ivan Miguel Pires, Gonçalo Marques, Nuno M. Garcia, Vasco Ponciano, Machine learning for the evaluation of the presence of heart disease, Procedia Comput. Sci. (ISSN: 1877-0509) 177 (2020) 432–437, http://dx.doi.org/10.1016/j.procs.2020.10.058.
[23] Nongyao Nai-arun, Rungruttikarn Moungmai, Comparison of classifiers for the risk of diabetes prediction, Procedia Comput. Sci. (ISSN: 1877-0509) 69 (2015) 132–142, http://dx.doi.org/10.1016/j.procs.2015.10.014.
[24] A. Sethi, One-Hot Encoding vs. Label Encoding Using Scikit-Learn. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/.
[25] R Indrakumari, T Poongodi, Soumya Ranjan Jena, Heart disease prediction using exploratory data analysis, Procedia Comput. Sci. (ISSN: 1877-0509) 173 (2020) 130–139, http://dx.doi.org/10.1016/j.procs.2020.06.017.
[26] J Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: International Conference on Machine Learning, 2006, pp. 233–240.
[27] S Mohan, C Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554, http://dx.doi.org/10.1109/ACCESS.2019.2923707.
[28] A.U Haq, J Li, M.H Memon, M Hunain Memon, J Khan, S.M. Marium, Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection, in: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1–4, http://dx.doi.org/10.1109/I2CT45611.2019.9033683.
[29] M Kavitha, G Gnaneswar, R Dinesh, Y.R Sai, R.S. Suraj, Heart disease prediction using hybrid machine learning model, in: 2021 6th International Conference on Inventive Computation Technologies, ICICT, 2021, pp. 1329–1333, http://dx.doi.org/10.1109/ICICT50816.2021.9358597.
[30] B Kolukisa, et al., Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2232–2238, http://dx.doi.org/10.1109/BigData.2018.8622609.
[31] R Jothiramalingam, A Jude, R Patan, M Ramachandran, J.H Duraisamy, A.H. Gandomi, Machine learning-based left ventricular hypertrophy detection using multi-lead ECG signal, Neural Comput. Appl. 33 (9) (2021) 4445–4455.
[32] S Mohan, C Thirumalai, G. Srivastava, Effective heart diseaseprediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554.
[33] Amelec Viloria, Yaneth Herazo-Beltran, Danelys Cabrera, Omar Bonerge Pineda, Diabetes diagnostic prediction using vector support machines, Procedia Comput. Sci. (ISSN: 1877-0509) 170 (2020) 376–381, http://dx.doi.org/10.1016/j.procs.2020.03.065.
[34] C Boukhatem, H.Y Youssef, A.B. Nassif, Heart disease prediction using machine learning, in: 2022 Advances in Science and Engineering Technology International Conferences, ASET, 2022, pp. 1–6, http://dx.doi.org/10.1109/ASET53988.2022.9734880.
[35] M.A Khan, F. Algarni, A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS, IEEE Access 8 (2020) 122259–122269, http://dx.doi.org/10.1109/ACCESS.2020.3006424.
[36] K.M. Almustafa, Prediction of heart disease and classifiers' sensitivity analysis, BMC Bioinf. 21 (2020).
[37] C Krittanawong, H.U.H Virk, S Bangalore, et al., Machine learning prediction in cardiovascular diseases: a meta-analysis, Sci. Rep. 10 (2020) 16057, http://dx.doi.org/10.1038/s41598-020-72685-1.
[38] X Liu, et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis, Lancet Digit. Health 1 (2019) e271–e297.
[39] Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, M. Essam Khalifa, Feature analysis of coronary artery heart disease data sets, Procedia Comput. Sci. (ISSN: 1877-0509) 65 (2015) 459–468, http://dx.doi.org/10.1016/j.procs.2015.09.132.
[40] Jacqueline Kazmaier, Jan H. van Vuuren, The power of ensemble learning in sentiment analysis, Expert Syst. Appl. (ISSN: 0957-4174) 187 (2022) 115819, http://dx.doi.org/10.1016/j.eswa.2021.115819.
[41] Sibo Prasad Patro, Gouri.Sankar Nayak, Neelamadhab Padhy, Heart disease prediction by using novel optimization algorithm: A supervised learning prospective, Inform. Med. Unlocked (ISSN: 2352-9148) 26 (2021) 100696, http://dx.doi.org/10.1016/j.imu.2021.100696.
[42] Rout Ranjeet Kumar, et al., Feature-extraction and analysis based on spatial distribution of amino acids for SARS-CoV-2 protein sequences, Comput. Biol. Med. 141 (2022) 105024.
[43] P Govindarajan, R Soundarapandian, AH Gandomi, R Patan, P Jayaraman, R. Manikandan, Classification of stroke disease using machine learning algorithms, Neural Comput. Appl. 32 (3) (2020) 817–828.
[44] Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, Comput. Biol. Med. (ISSN: 0010-4825) 136 (2021) 104672, http://dx.doi.org/10.1016/j.compbiomed.2021.104672.
[45] M.A Khan, F. Algarni, A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS, IEEE Access 8 (2020) 122259–122269, http://dx.doi.org/10.1109/ACCESS.2020.3006424.
[46] J Wang, et al., A stacking-based model for non-invasive detection of coronary heart disease, IEEE Access 8 (2020) 37124–37133, http://dx.doi.org/10.1109/ACCESS.2020.2975377.
[47] M Shehab, L Abualigah, Q Shambour, M.A Abu-Hashem, M.K.Y Shambour, A.I Alsalibi, A.H. Gandomi, Machine learning in medical applications: A review of state-of-the-art methods, Comput. Biol. Med. 145 (2022) 105458.
[48] J.P Li, A.U Haq, S.U Din, J Khan, A Khan, A. Saboor, Heart disease identification method using machine learning classification in E-healthcare, IEEE Access 8 (2020) 107562–107582, http://dx.doi.org/10.1109/ACCESS.2020.3001149.
[49] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang, An improved ensemble learning approach for the prediction of heart disease risk, Inform. Med. Unlocked (ISSN: 2352-9148) 20 (2020) 100402, http://dx.doi.org/10.1016/j.imu.2020.100402.
[50] Charles E. Metz, Basic principles of ROC analysis, Semin. Nucl. Med. (ISSN: 0001-2998) 8 (4) (1978) 283–298, http://dx.doi.org/10.1016/S0001-2998(78)80014-2.