INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Prediction of Heart Disease using Ensemble Learning

P Muthulakshmi[1]*, M Parveen[2], P Rajeswari[3]

1 Assistant Professor, Cauvery College for Women (Autonomous), [Affiliated to Bharathidasan University], Trichy, India
2 Professor, Cauvery College for Women (Autonomous), [Affiliated to Bharathidasan University], Trichy, India
3 Associate Professor, Cauvery College for Women (Autonomous), [Affiliated to Bharathidasan University], Trichy, India

OPEN ACCESS

Received: 26-11-2022
Accepted: 12-03-2023
Published: 19-05-2023

Citation: Muthulakshmi P, Parveen M, Rajeswari P (2023) Prediction of Heart Disease using Ensemble Learning. Indian Journal of Science and Technology 16(20): 1469-1476. https://doi.org/10.17485/IJST/v16i20.2279

* **Corresponding author**.

muthulakshmi.cs@cauverycollege.ac.in

Funding: Grant obtained under the scheme of Seed Money for Research projects from Cauvery College for Women (Autonomous), Tiruchirappalli

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To propose a Bagging ensemble method to predict heart disease at early stages. The main focus of this research is to increase the prediction accuracy in a model. **Methods:** The proposed system is experimented with by using the Cleveland datasets collected from the UCI repository. The dataset consists of 14 attributes. In this dataset we applied different machine learning algorithms such as Decision tree, Naïve Bayes, Random Forest and SVM along with the proposed ensemble learning classifier. The entire dataset is trained upon the Pearson correlation coefficient selected features under the k-fold cross-validation setup. Final outcome is obtained by aggregating the prediction accuracy. **Findings:** The performance of the proposed method was validated using prediction accuracy and compared with the Machine learning algorithms and the ensemble models. The proposed method attains a higher classification accuracy of 95.33% than all other methods. **Novelty:** A novel ensemble method has been proposed with a better accuracy in early predicting of the heart disease.

Keywords: Ensemble Model; Machine Learning; Prediction; Accuracy; Kfold cross validation

1 Introduction

Presently, there are approximately 30 million Indians who suffer from heart ailments. This not only includes the urban and economically well-off but also underprivileged individuals. Another surprising fact about heart diseases in India is that 25% of heart-related deaths occur in people aged less than 40 years. Major causes of heart diseases include lack of awareness, illiteracy, lack of accessibility, high treatment costs, and many more. The growth of heart diseases is dependent on several interlinked factors, such as stress, bad eating habits, tobacco use, alcohol consumption, aging, and many more. So, feasible and accurate prediction of heart-related diseases is very important.

Machine Learning is described as part of Artificial Intelligence (AI) in which a model acquires knowledge from past experience, without being explicitly programmed. Medical data are used in different ML classifiers for classification or forecasting of

diseases[1]. Heart disease has become one of the most critical medical topics in recent years as the mortality toll from the cardiovascular disease continues to climb. Prediction aids in the early detection of disease and the most effective treatment are required[2]. Many algorithms for predictive learning are available (e.g., linear and logistic regression, classification and regression trees, learning vector quantization (LVQ), support vector machines (SVM), boosting, and deep neural networks)[3]. More machine learning models need to be studied using various recent databases and used to obtain the best model for early-stage disease prediction at a low cost. Therefore, an attempt is made to bridge the experts' knowledge and experience in order to create a system that equitably supports the diagnosis process[4].

Ensemble learning is a technique of machine learning which helps in improving the performance of our system by using multiple classifiers. By combining multiple classifiers, the performance of the model increases to a great extent as compared to the individual classification model. Thus, using ensemble learning enhances the accuracy of prediction for detecting heart disease[5]. Feature selection plays a crucial role in the development of machine learning algorithms. Understanding the impact of the features on a model, and their physiological relevance can improve the performance[6].

Thangam et al[7] proposed a significant feature selection method using the EKSFM technique. In which the method extracts a small set of features from the features of multidimensional vectors by preserving its data characteristics as such and is mainly used for dimensionality reduction[7].

Ramatenki et al[8] proposed an ensemble method for predicting heart disease and determining the presence or absence of heart disease. Four algorithms were employed, namely KNN, Modified KNN, SVM, and Decision Tree. After the algorithms have been trained, their performance is evaluated and compared by using accuracy, recall, precision, and F1 score measures, of which the ensemble classifier produces better accuracy than all the other algorithms.

Ibomoiye et al.[9] introduced an improved ensemble learning approach for the prediction of heart disease risk. They randomly partition the data set into small subsets using mean value. A homogenous CART model using an Accuracy-based weighted aging classifier ensemble (AB-WAE) is applied to evaluate the accuracy of the system. By comparing the previous machine learning algorithms, this method produces better accuracy than some of the scholarly works. The performance indicators used here are accuracy, precision, sensitivity, and F1 score.

Muhammad et al.[10] developed A Novel Approach based on Significant Feature and Ensemble Learning Model in which Random Forest with Stratified Kfold is applied to tune the parameters. Cross-validation prevents data from overfitting and underfitting. Random Forest employs the bootstrapping technique in conjunction with stratified KFold to produce a similar or different result by taking the parameters as train data, test data, and the number of estimators for each tree. Finally, the testing data goes to a majority voting decision tree sample, either it is true or false. This method produces high accuracy compared to the existing methods.

Benjamin et al.[11] analysed the impact of ensemble learning algorithms on accurate heart disease prediction. The aim was to identify the best ensemble learning classification. The UCI data repository is used to determine the performance of the stacking, Bagging and Boosting ensemble methods. The metrics used here are precision, recall, F-measure and ROC. After performing the comparative study, AdaBoost outperforms well.

Adithya et al.[12] formed an optimal multi-disease prediction framework using hybrid machine learning techniques. They have used genetic algorithm-based recursive feature elimination and the AdaBoost (GAE-RFE) method to predict the disease. while applying this feature selection process to Random forest, Decision tree, XGBoost and AdaBoost. A 10-fold cross validation is employed to test the dataset. The performance of the machine learning algorithm is boosted by the GAE-RFE method. AdaBoost achieves remarkable values of precision, specificity, sensitivity, and F-measure in comparison to benchmark techniques.

Bhanu et al.[13] proposed a soft voting classifier model to predict heart disease. Three benchmark datasets' performance is compared. Data preprocessing is carried out to remove noisy and missing values. The Min-max scalar is used to standardize the data and the voting classifier uses NB, RF, SVM and gradient Boost to build up the model. Resting on the majority voting process, the proposed ensemble method produces better accuracy.

Divyansh Khanna et al.[14] performed a comparative study of classification to predict heart disease. It uses the publicly available Cleveland Dataset with 14 attributes and models the classification techniques SVM, LR and Neural Networks. The results prove that logistic regression and SVM approaches provide better performance, especially with linear kernels. Among neural networks, the GRNN method (Generalized Regression Neural Network) stands out, but the RBF-NN is of little use. The results suggest that the SVM method is a very good technique for accurately predicting cardiac disease, especially considering classification accuracy as a measure of performance.

Mythili T et.al.[15] Proposed a framework for identifying the risk of heart disease using a combination of models. This approach is divided into six modules involving preprocessing, training, testing with individual models, and application of rules. The Cleveland Heart Disease Dataset (CHDD) available on the UCI Repository is used and 13 attributes are utilized

in this work. The model is tested using SVM, DT and Logistic Regression. It is recommended that the classification rule (C-rule) be used for this model. C-Rules are in the form of if-else ladders and provide the simplest and most comprehensible way of expressing knowledge. It is hypothesized that a result with higher sensitivity and specificity but lower accuracy will be attained from the results of this model.

Amin UlHaq et al.[16] developed a hybrid and intelligent machine learning-based predictive system for heart disease diagnosis. The system was tested on the Cleveland Cardiology dataset. Seven well-known classifiers, such as logistic regression, K-NN, ANN, SVM, NB, DT, and random forest, and three feature selection algorithms, Relief, mRMR, and LASSO, to select important features used with for validation, the system used a K-fold cross-validation method. Logistic regression of the classifier with 10-fold cross-validation showed the best accuracy of 89% when selected by the feature selection algorithm Relief. SVM (linear) with feature selection and mRMR algorithm had the best performance in terms of singularity. The highest sensitivity was achieved at 100% by a classifier ANN (MLP) with 16 hidden neurons on selected features in relief. Therefore, we can use the FS algorithm to reduce computation time and improve the classification accuracy of our classifier.

Gupta et al.[17] proposed a Stacking Ensemble-Based Intelligent Machine Learning(SEIML) Model for Predicting Post-COVID-19 Complications. A binary classifier based on a stacking ensemble is modeled with deep neural networks for the prediction of heart diseases, post-COVID-19 infection. The proposed model is validated against other baseline techniques, such as decision trees, random forest, support vector machines, and artificial neural networks. This method produces 93.23% of accuracy.

To improve the prediction accuracy, we have proposed a Pearson correlation coefficient based min-max normalized feature selection method with a bagging classifier. The rest of this paper is organized as follows. Section 2 provides information on the related work along with the proposed methodology, and Section 3 provides an evaluation and comparison with other methods. Finally, conclusions are given in Section 4.

2 Methodology

2.1 Data Preprocessing

Data Preprocessing is a major task in the data analysis process. It takes raw data and transforms it into a formal that can be acceptable by the machine learning algorithm. The most encountered problem in preprocessing is missing values and outliers. We have not found any missing values and outliers in this dataset.

2.2 Feature selection

Feature selection is a method of removing replica, redundant, or noisy characteristics from a feature collection in order to choose a subset among the most relevant features. The key goal is to select the optimal characteristics to help the model work successfully. For selecting the best features, we find a correlation of the attributes with all the other features using the equation below:

$$Pxy = \frac{Cov(x,y)}{\sigma x \sigma y}$$

$$Cov\,(x,y) = E(x-\mu_x)(y-\mu_y)$$

The highly correlated features are identified and they are removed from the dataset. After selecting the extracted features, we performed data standardization. The method chosen to standardize the data is MinMax Scalar technique. It works as follows:

i) Fit the scalar using available data by fit method

ii) Apply the scale to training data using transform method

iii) Prepare a new data for making prediction by using

$$y = \frac{x - min}{max - min}$$

2.3 Splitting the Data

Train test split is a model validation process that allows us to simulate how our model would perform with new data. The first 80 % of the data was assigned to training and the remaining 20% was assigned to test.

2.4 Machine learning classification

The process of categorizing data into groups based on the correlation between different data pieces is referred to as classification. Categorization is used to anticipate cardiovascular problems in this scenario. There are several machine learning models available, however the suggested method can use any of the following techniques or models. To forecast the aim, we used a number of techniques. However, by utilizing an ensemble technique and the concept of hyper tuning, we achieve the best outcomes. The following are the algorithms that were used:

2.4.1 Decision Tree

It is a supervised learning algorithm which can be used for both classification and regression problems. Initially, the root node consists of the training data and finds the best attributes selected based on the Gini Index. Then divides the training set into subsets that contain all possible values and generates the best decision tree node. Using all 13 attributes of our dataset, an accuracy of 81.66% was reached.

2.4.2 Random forest

Random forest is an ensemble learning method that combines multiple algorithms to get better results in classification, regression, and other tasks. Individual classifiers are weak individually, but can produce excellent results when combined with other classifiers. Algorithms start with a decision tree and are given input at the top. Then go down the tree and segment the data into smaller sets based on certain variables. Using all 13 attributes of our dataset, an accuracy of 83.33% was reached.

2.4.3 Naive Bayes

Naive Bayes classifiers are based on Bayes' theorem and classify all values as independent of all other values. This allows us to use probabilities to predict classes/categories based on a given set of traits.

$P(X|y) = (y|X)\ P(X)) / P(y)$

Regardless of its simplicity, the classifier does surprisingly well and is often used due to the fact it outperforms more sophisticated classification methods. Using all 13 attributes of our dataset, an accuracy of 88.33% was reached.

2.4.4. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. It finds a hyper plane in an N-dimensional space that distinctly classifies the data points. All of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. By applying this learning algorithm we got the accuracy of 90%.

2.4.5 Proposed Ensemble Method

Combining several weak models can result in a strong learner. We use ensemble learning, using several different methods to counteract each model's individual weaknesses. When compared to other learning methods such as Decision Tree, Random Forest, SVM and Naive Bayes, the proposed ensemble model MinMax Normalized – Pearson coefficient Bagging Ensemble (MPBE) technique produces better results. The block diagram of the proposed ensemble model is given in Figure 1.

Once the preprocessing and feature selection processes are completed, the processed data is to be divided into an "n" number of samples. Initially assign that the number of estimators as 100. A bagging technique creates multiple subsets from the original data set with equal number of samples. For each of these samples a base classifier model is created. Some base classifiers applied here are Decision tree, Random forest, Naïve Bayes and SVM. Each model is learned in parallel and the final predictions are determined by combining all the predictions made by each model.

The bootstrap technique is implemented to divide the training set S into k samples by random sample with replacement. The same data may appear in all samples or not be at all present in any one of the samples. After training, we need to aggregate the individual model by using the mean score. The initial decision function for k samples is denoted as f_k where k=(1,2,3,... n) is applied to the k number of Classifiers C_k and the final decision of the Ensemble for a given test vector x. k-fold cross validation is applied and the model is fitted. Obtaining all the scores and aggregating them produces a better result than the benchmarked methods. The bagging ensemble method also overcomes the problem of overfitting due to the k fold cross validation.

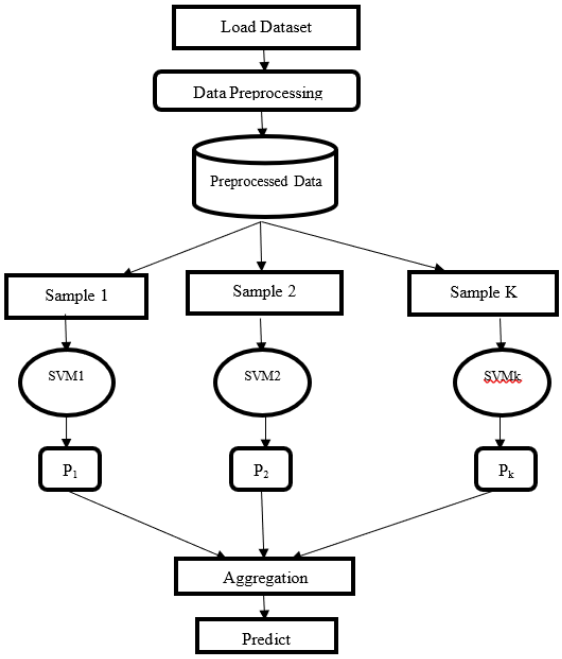

Fig 1. Block Diagram of Proposed MPBE Model

MPBE Algorithm:
Procedure LOAD (heart_disease_data)
Procedure PREPROCESS
Procesure DATA_SPLIT (heart_disease_data) Train_data, Test_data
split (heart_disease_data,lables) return Train_data, Test_data
C1 = DecisionTreeClassifier (Train_data, Train_label, Test_data)
C2 = RandomForestClassifier(Train_data, Train_label, Test_data)
C3= NaiveBayesClassifier (Train_data, Train_label, Test_data)
C4= SupportVectorMachineClassifier (Train_data, Train_label, Test_data)
Procedure Ensemble_Model(Train_data, Train_label, Test_data)
Divide the training into k samples
for each k apply the base classifier model
model.fit(Train_data, Test_data)
Apply k fold cross validation
Find Score
Accuracy= mean(score)

Using the aforementioned Algorithm, implement the suggested MPBE technique. where the data set is initially loaded. The Min-Max scalar scalar is used for data preprocessing and feature selection. After that, training and testing sets of the scaled data are created. The bagging classifier in this instance is made up of many base classifiers. The various classifier models are applied to each training sample and then fitted into the model. The best model output must then be aggregated.

3 Results and Discussion

The UCI Cleveland Dataset with 14 attributes is taken from kaggle.com. The attributes considered here are Age, Sex, Chest pain type, resting blood Pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, Heart rate, Depression Number of major vessels, thalassemia and a target variable.

The outcome of the prediction is measured through the parameter prediction accuracy. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy=Number of correct predictions/Total number of predictions

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

Accuracy=TP+TN/TP+TN+FP+FN

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

The proposed ensemble method is compared with the standard machine learning algorithms such as Decision trees(DT), Random forest(RF), Naïve Bayes(NB) and Support Vector Machine (SVM). The proposed MPBE produces 95.33% accuracy, which states that it produces a better performance than the existing methods. Table 1 shows the detailed data.

Table 1. Accuracy Comparison

Classification Method	Accuracy (%)
Decision Tree	81.66
Random Forest	83.33
Naïve Bayes	88.33
Support Vector Machine	90.00
Proposed MPBE	**95.33**

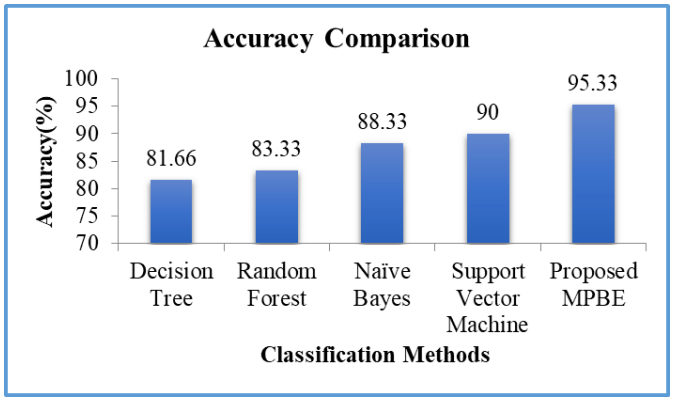

Fig 2. Accuracy Comparison Chart

Figure 2 depicts that compared with the other mentioned algorithm our proposed model provides a better result. The heart disease prediction accuracy using MPBE is said to be improved by 13.67%, 12%,7% and 5.33% compared to DT, RF, NB and SVM respectively. This is achieved by using only the relevant features of the ensemble method.

The proposed ensemble method is also compared with the existing ensemble methods such as AB-WAE [9], Stratified K fold [10], GAE-RFE [12], soft voting classifier [13] models and SEIML [17]. Table 2 shows the detailed data.

Table 2. Accuracy Comparison with existing ensemble methods

Ensemble Model	Accuracy (%)
Stratified K fold	86.12
soft voting classifier	88.24
GAE-RF adaboost	91.90
AB-WAE	93.00
SEIML	93.23
Proposed MPBE	**95.33**

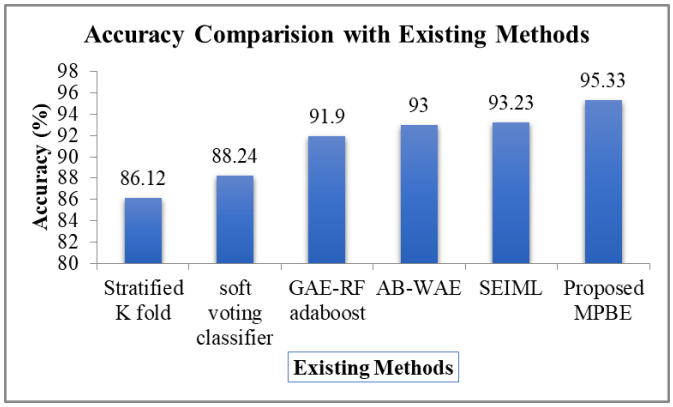

Fig 3. Accuracy comparison with existing methods

Figure 3 depicts that compared with the other mentioned algorithm our proposed model provides a better result. The heart disease prediction accuracy using MPBE is said to be improved by 9.21%, 7.09%,3.43% ,2.33% and 2% compared to Stratified K fold, Soft Voting Classifier, GAE-RF adaboost, AB-WAE and SEIML respectively.

4 Conclusion

Heart disease (CHD) is the major cause of death in most developed countries and in many developing countries. The clinical complications of CHD lead to substantial disability and are a main source of the rising cost of healthcare. Our proposed model produced more accurate result of 95.33% accuracy on the Cleveland dataset obtained from the UCI repository. The suggested model produces better result due the data preprocessing, feature selection methods and also the cross validation performed on the given data set. Our model is compared with benchmarked classification algorithm as well as the existing algorithms. Both result in a better accuracy. For further improvement, one can choose deep learning and Artificial Intelligence.

Acknowledgement

The authors are grateful to the Management and Principal for their motivation and constant support. This research has been supported by the grant obtained under the scheme of Seed Money for Research projects from Cauvery College for Women (Autonomous), Tiruchirappalli- 620 018, India.

References

1) Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*. 2022;17(1):1100–1113. Available from: https://doi.org/10.1515/med-2022-0508.
2) Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular Disease Detection using Ensemble Learning. *Computational Intelligence and Neuroscience*. 2022;2022:1–9. Available from: https://doi.org/10.1155/2022/5267498.
3) El-Hasnony IM, Elzeki OM, Alshehri A, Salem H. Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. *Sensors*. 2022;22(3):1184. Available from: https://doi.org/10.3390/s22031184.
4) Absar N, Das EK, Shoma SN, Khandaker MU, Miraz MH, Faruque MRI, et al. The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. *Healthcare*. 2022;10(6):1137. Available from: https://doi.org/10.3390/healthcare10061137.
5) Gupta P, D SD. Improving the Prediction of Heart Disease Using Ensemble Learning and Feature Selection. *International Journal of Advances in Soft Computing and its Applications*. 2022;14(2):37–40. Available from: http://www.i-csrs.org/Volumes/ijasca/2022.02.03.pdf.
6) Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*. 2010;43(1):15–23. Available from: https://doi.org/10.1016/j.jbi.2009.07.008.
7) Thangam M, Bhuvaneswari A. Exponential kernelized feature map Theil-Sen regression-based deep belief neural learning classifier for drift detection with data stream. *International Journal of Advanced Technology and Engineering Exploration*. 2022;9(90):663–675. Available from: https://doi.org/10.19101/IJATEE.2021.874851.
8) Kumar SRS, Fatima AS, Thomas. Heart Disease Prediction using Ensemble Learning Method. *International Journal of Recent Technology and Engineering*. 2020;(9):2277–3878. Available from: https://www.ijrte.org/wp-content/uploads/papers/v9i1/A2997059120.pdf.
9) Mienye ID, Sun Y, Professor Z, Wang. An improved ensemble learning approach for the prediction of heart disease risk". *Informatics in Medicine Unlocked*. 2020;20:100402. Available from: https://doi.org/10.1016/j.imu.2020.100402.
10) Alim MA, Habib S, Farooq Y, Rafay A. Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model. *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. 2020;p. 1–5. Available from: https://doi.org/10.1109/iCoMET48670.2020.9074135.

11) David HBF. Impact of Ensemble Learning Algorithms Towards Accurate Heart Disease Prediction. *ICTACT Journal On Soft Computing*. 2020;(10):3. Available from: https://doi.org/10.21917/ijsc.2020.0296.
12) Gupta A, Singh A. An optimal multi-disease prediction framework using hybrid machine learning. *Kuwait Journal of Science*. 2022;2022:1–13. Available from: https://doi.org/10.48129/kjs.splml.19321.
13) Doppala BP, Bhattacharyya D, Janarthanan M, Baik N. A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques. *Journal of Healthcare Engineering*. 2022. Available from: https://doi.org/10.1155/2022/2585235.
14) Khanna D, Sahu R, Baths V, Deshpande B. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. *International Journal of Machine Learning and Computing*. 2015;5(5):414–419. Available from: http://www.ijmlc.org/vol5/544-C039.pdf.
15) Mythili T, Mukherji D, Padalia N, Naidu A. A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications*. 2013;68(16). Available from: https://research.ijcaonline.org/volume68/number16/pxc3887250.pdf.
16) Haq AU, Li JP, Memon MH, Nazir S, Sun R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*. 2018;p. 1–21. Available from: https://doi.org/10.1155/2018/3860146.
17) Gupta A, Jain V, Singh A. Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications. *New Generation Computing*. 2022;40(4):987–1007. Available from: https://doi.org/10.1007/s00354-021-00144-0.