Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

A Study to Improve Heart Disease Prediction Using Cross-Validation and Percentage Split Techniques

Madhurima Mondal[1], Dr. Narendra Chaudhari[2], Dr. Amjan Shaik[3]

[1] *Research Scholar, Department of Computer Science & Engineering, Mansarovar Global University, India*
[2] *Research Guide, Dr. Narendra Chaudhari, Department of Computer Science & Engineering, Mansarovar Global University, India*
[3] *Professor of Comp. Sc. and Engg. & Dean-R&D Cell, St.Peter's Engineering College (SPEC), Maisammaguda, Medchal, India*

Worldwide, heart disease claims a lot of lives and puts a lot of strain on healthcare systems. The mortality rate and patient care may be greatly improved with early prediction and diagnosis of cardiac disease. This study examines and develops a prediction system for the analysis and prediction of the likelihood of heart disease using three data mining classification algorithms: Random Forest, Decision Tree, and Naïve Bayes. The primary goal of this substantial study is to determine which classification method is most suited to accurately distinguish between normal and abnormal individuals. The Random Forest method routinely achieves better results in classification accuracy, precision, and recall compared to Naïve Bayes and Decision Tree algorithms, according to the experimental data. Model performance increases as cross-validation folds rise, and Random Forest technique outperforms other methods when training data is divided using a larger fraction of the dataset. According to the data, the Random Forest classifier is the best at predicting the occurrence of heart disease. This study's results will improve patient care and decrease the likelihood of misdiagnosis in the medical field by making it easier for healthcare providers to forecast the occurrence of cardiac disease.

Keywords: Data Mining, Classification, Prediction, Heart Disease, Recall

1. Introduction

Every year, heart disease claims the lives of millions of people throughout the globe. Heart disease and other cardiovascular disorders account for over 31% of all fatalities worldwide, according the World Health Organization (WHO). The growing number of heart disease cases is putting a heavy strain on healthcare systems, highlighting the need of prompt diagnosis and precise prognosis in avoiding deadly consequences. Conditions that impact the heart are collectively known as heart disease. These include coronary artery disease, heart attacks, congestive heart failure, arrhythmias, and congenital heart abnormalities. Timely medical intervention may significantly lower death rates and enhance patients' quality of life by identifying persons at high risk of developing heart disease. When it comes to effectively and reliably forecasting cardiac illness, however, traditional clinical diagnostic approaches sometimes encounter constraints. This has prompted the use of data mining and other cutting-edge technology to improve the accuracy and reliability of heart disease prediction.

When applied to healthcare, data mining has shown to be an effective method for discovering previously unknown relationships, patterns, and insights in massively complicated datasets. Data mining's use for medical domain tasks such as pattern recognition, clustering, classification, and predictive analysis has skyrocketed in the last few years. Data mining's primary goal in healthcare is to sift through mountains of patient records in search of previously unseen correlations, trends, and patterns that might inform clinical decision-making. Through the analysis of patient health records, laboratory results, medical histories, and lifestyle variables, data mining algorithms have the potential to greatly improve the accuracy of diagnosis and prognosis when it comes to heart disease. Data mining is a vital tool for the prognosis and prevention of heart disease because of its capacity to reveal important patterns from medical information.

To identify people at high risk of developing heart disease, data mining methods including classification, clustering, association rule mining, regression, and predictive modeling are used. In order to forecast the probability of cardiovascular disease, these methods examine massive amounts of patient data, which includes demographics, sex, height, weight, cholesterol, smoking status, family medical history, and physical activity levels, among other health indicators. Data mining methods may assist healthcare providers in identifying high-risk patients by combining real-time health information with past patient data. This enables them to prescribe treatments or preventative actions based on the individual's risk level.

Classification is one of the most popular data mining approaches for predicting the occurrence of cardiovascular disease. This method entails dividing patients into two groups: those with and those without the illness. Predicting cardiac illness has been shown to be a very successful task for classification algorithms like Decision Trees, Naïve Bayes, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). These algorithms construct a prediction model to correctly categorize new or unknown data by using existing patient data. For instance, decision trees simplify the process of understanding and interpreting choice outcomes by using a tree-like structure to depict decision options and their potential outcomes. In contrast, Naïve Bayes is a very effective probabilistic classifier for predicting the probability of heart disease using input characteristics; it is based on Bayes' theorem. Similarly, ANNs are great at predicting cases of heart disease because they understand complicated non-linear patterns in medical data by simulating the way the human brain functions.

Clustering, a data mining approach that groups data points into clusters according to their similarities, is another important tool utilized in the prediction of heart disease. Many clustering algorithms are used in the field of cardiac disease prediction to classify patients according to shared medical symptoms or traits. These algorithms include Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and K-Means. Healthcare practitioners may enhance healthcare outcomes by concentrating on focused diagnosis and treatment after identifying clusters of high-risk patients. Public health officials may use clustering to pinpoint trends in the incidence of cardiovascular disease across various demographic subsets, which in turn allows for the development of more targeted initiatives for disease prevention and management.

Heart disease prediction also makes use of Association Rule Mining (ARM), a robust data mining approach that seeks for intriguing correlations or connections among variables in massive datasets. Heart disease risk factors and patterns may be discovered using ARM approaches like the Apriori algorithm and the FP-Growth algorithm. As an example, association rule mining may show that those who are overweight, have high blood pressure, and high cholesterol are more likely to have heart disease. Healthcare providers may better understand the major risk factors for cardiovascular disease and create individualized treatment programs for their patients if these relationships are identified.

Another popular data mining technique for predicting the occurrence of cardiac problems is predictive modeling, which involves creating models from past data in order to foretell how things will turn out. With supervised learning algorithms, predictive models may take known inputs (e.g., age, blood pressure, cholesterol levels, etc.) and produce a known output (e.g., presence or absence of heart disease). These models are able to anticipate the results of new data by learning from the results of previous data. Some of the most well-known algorithms for predictive modeling in the field of cardiovascular disease prediction include Decision Trees, Logistic Regression, Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). These prediction models can only be as good as the data used to train them, the features chosen, and the algorithms used.

An essential part of data mining for the prediction of heart disease is data preparation. When it comes to healthcare, raw data is often inconsistent, missing information, and noisy, all of which may impact how well predictive models work. To make sure the predictions are correct, data pretreatment methods including cleaning, integration, transformation, and reduction are used to enhance the data quality. Eradicating discrepancies, duplicate entries, and missing information is what data cleaning is all about. A complete dataset may be created by data integration by combining information from several sources, including labs, clinics, and hospitals. To guarantee consistency across distinct variables, data transformation entails standardizing and normalizing the data. By eliminating superfluous characteristics and reducing the dataset's dimensionality, data reduction approaches boost the performance and precision of models.

Data mining is also essential for assessing the efficacy of models used to forecast the occurrence of cardiovascular disease. To measure how well predictive models work, researchers utilize a variety of performance indicators. These include recall, accuracy, precision, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). While accuracy evaluates the model's general correctness, precision counts the number of right predictions relative to the total number of positive predictions. In contrast to recall, which evaluates the model's capacity to detect all positive instances, the F1-score strikes a compromise between the two. The area under the receiver operating characteristic (ROC) curve is a popular metric for assessing a model's ability to discriminate between positive and negative instances. Researchers can find the best algorithm for predicting heart illness and tweak its parameters for improved accuracy by analyzing the results of predictive models.

Early diagnosis, lower mortality rates, individualized treatment programs, and better patient outcomes are just a few of the many advantages that may be achieved via the use of data mining methods in the prediction of cardiac illness. Healthcare professionals may improve resource allocation, identify high-risk patients, and make evidence-based therapeutic choices via data mining. The public health sector may also benefit from data mining's ability to reveal trends and patterns in cardiac disease statistics, which can guide the development of more efficient intervention and preventive programs. Achieving real-time monitoring and early prediction of heart illness via the integration of data mining methods with electronic health records (EHRs), wearable devices, and mobile health apps will help healthcare systems reduce the burden of cardiovascular diseases.

Data mining for the prediction of heart disease has several advantages, but it also has certain limits and confronts some difficulties. The accessibility and accuracy of healthcare data is a big obstacle. Predictive models may not be as accurate in developing nations due to the prevalence of healthcare data that is inconsistent, missing, and fragmented. Concerns about patient information security and privacy also act as roadblocks to data mining's widespread use in healthcare. For data mining technologies to earn public confidence and be widely used, it is vital to protect patient confidentiality and provide safe data storage and transfer. It might be difficult for healthcare providers to comprehend and implement model predictions in clinical practice due to the predictive models' complexity and interpretability. So, to aid in clinical decision-making, future studies should concentrate on creating explainable AI (XAI) models that provide predictions that are both interpretable and transparent.

2. Review of Literature

Shetgaonkar, Pratiksha et al., (2021) There is no more important organ than the heart. Our ability to pump blood effectively and in excellent health is fundamental to our existence. This illness ranks high among the most deadly in the modern age. Heart disease is one of the most pressing issues in contemporary public health. People often consider it to be the most lethal disease in the world. Identifying the onset of cardiac disease at an early stage may be challenging for medical professionals. Predictions made by modern healthcare, in particular, are very accurate because to a wealth of confidential data and information at their disposal. The goal of data mining is to find valuable patterns and insights in massive datasets by using state-of-the-art artificial intelligence techniques. The purpose of this article is to estimate the likelihood of cardiovascular disease or heart illness using three artificial intelligence-based techniques: Decision Tree, Naïve Bayes, and Neural Network. A set of altered parameters will be used to assess each approach, with the goal of improving their accuracy. We will next use a number of metrics pertaining to accuracy to assess the methods' efficacy. Evaluation of the risk of coronary heart disease using the gold standard method in both sexes is the next stage. Foreseeing a patient's illness in its early stages with this technology may help medical practitioners hasten treatment.

Fadnavis, R et al., (2021) Healthcare systems have a plethora of data at their fingertips. However, analytics tools that may be used to uncover and benefit from intriguing, non-traditional patterns inside data are few. Data mining has shown to be an invaluable resource in several domains, facilitating the discovery of previously unseen trends and patterns, the detection of suspect data, and, finally, the improvement of decision-making. This research explores the potential of data mining classification approaches for the prediction of cardiac disease. The effectiveness of algorithms such as Decision Tree and Naïve Bayes is evaluated so that patients might be warned of potential cardiac problems. Our data comes from the Cleveland set, which has fourteen different variables.

C. B. Gokulnath et al., (2019) In order to facilitate further action, cardiac illness detection has emerged as a difficult problem that can provide a computerized assessment of the severity of heart disease. So, the healthcare system throughout the globe is preparing for a great deal more focus on the detection of cardiac disease. When it came to accurately diagnosing cardiac illness, optimization methods were crucial. Proposing an optimization function based on support vector machine (SVM) is the purpose of this work. The genetic algorithm (GA) makes use of this objective function to determine which traits are most predictive of cardiovascular disease. The GA-SVM experimental findings are contrasted with those of other feature selection methods, including GA, filtered subset, Chi squared, one attribute based, Filtered attribute, Consistency subset, Relief, CFS, Info gain, and Gain ratio. An evaluation of the SVM classifier's excellent performance is carried out using receiver operating characteristic analysis. Using data acquired from the Cleveland heart disease database, the suggested framework is shown in a MATLAB environment.

Devi, S. et al., (2016) In order to better diagnose and predict the prognosis of cardiac sickness, we want to research a variety of healthcare data mining approaches and technologies. By using data mining methods, medical professionals may create a model that can assess the likelihood of cardiovascular disease based on a patient's medical information. Healthcare decision-makers may benefit from data mining classification approaches such as Neural Networks, Naive Bayes, Decision Trees, and Support Vector Machines. Decisions might be made more quickly and with more accuracy if these algorithms were integrated or hybridized. By sifting through massive databases in search of patterns, data mining is a powerful new tool for unearthing hidden insights that might inform forecasts or other useful decisions. In an attempt to reduce healthcare costs and diagnosis time while increasing service quality and accuracy, activists have suggested adopting modern data mining techniques to uncover relevant information. This approach has the potential to reliably forecast the occurrence of cardiac disease. Adding more input variables, such as controllable and uncontrollable risk factors, might lead to more precise findings. Additional advancements in this technology are possible. A multitude of input attributes are used by it. Prediction may also make use of other data mining approaches including clustering, time series, and association rules. Text mining may also be used by healthcare organizations to extract structured data from databases.

Methaila, Aditya et al., (2014) Data mining has seen extensive application due to its effectiveness in well-publicized industries such as e-commerce, marketing, and retail. This is especially true of the healthcare industry, which is still in its infancy. The healthcare industry boasts of having "information rich" data, but not all of it is really mined to discover patterns and make sound judgments. Patterns and relationships that have not been found sometimes go unnoticed. We could find a solution with the aid of current data mining modeling techniques. For this study, we want to use data mining classification modeling techniques such as decision trees, naïve bayes, neural networks, weighted association apriori, and MAFIA algorithms in order to predict the frequency of heart disease. Age, sex, blood pressure, and glucose levels are some of the medical variables used to forecast the probability of cardiovascular disease.

Dangare, Chaitrali & Apte, Sulbha. (2012) Although healthcare organizations often have access to a wealth of "information rich" data, not all of it is sufficiently mined to reveal previously unknown patterns and enable informed decision-making. Medical research, especially that pertaining to the prognosis of heart disease, makes use of advanced data mining methods to unearth information in databases. Heart disease prediction systems with a larger set of input qualities are the focus of this research. To calculate the probability of heart illness, the algorithm takes into account 13 medical characteristics, including sex, blood pressure, cholesterol, and other similar words. We have utilized 13 characteristics for prediction thus far. Two additional characteristics, namely smoking and obesity, were included in this study. The Heart disease database is used to examine the data mining classification approaches, namely Decision Trees, Naive Bayes, and Neural Networks. We evaluate these methods by comparing their accuracy-based performance. Our research shows that naive bayes, decision trees, and neural networks all achieve 100% accuracy, whereas decision trees and neural networks each achieve 99.62% accuracy. Our results demonstrate that Neural Networks outperforms the other two classification models when it comes to predicting the occurrence of heart disease.

3. Research Methodology

Predicting patients' risk of cardiovascular disease is as easy as following the steps shown in Figure 1, which details the research methods needed to construct the classification model needed for this purpose. Using the model as a foundation, any machine learning approach may be used to forecast the occurrence of cardiac disease. Predictions are generated by training a classifier with the data, which in turn creates a classification model. This model is then used to input fresh, unknown records and make predictions.

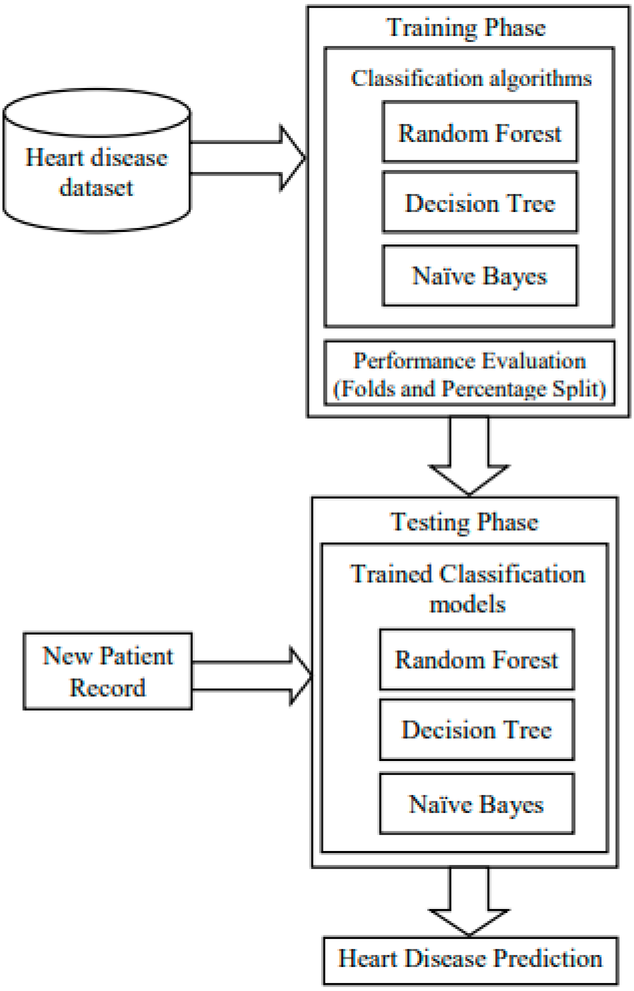

Figure 1: Framework of study

This study's approach comprises a performance assessment of the three categorization algorithms, with an emphasis on evaluations utilizing percentage split and cross validation. To do cross validation on the heart disease dataset, the training and testing sets are partitioned using a number of folds, say ten, and then each fold is recursively used for testing and training by substituting it with the other set. To implement a percentage split, we divide the data into two sets, one for training and one for testing. For example, we may utilize 80% of the data for training and 20% for testing. The training portion of this study involves building a classification model utilizing the heart disease dataset and three classification algorithms: Naïve Bayes, Decision Tree, and Random Forest. Precision, recall, f-measure, ROC area, and PRC area are the measures used in the study effort.

Dataset Description

The data used in this study came from the Stat-Log dataset housed at the University of

California, Irvine. All all, there are thirteen characteristics. This study makes use of a heart disease dataset that has 270 complete cases. Typical angina, atypical angina, non-anginal pain, and asymptomatic heart disease are some of the common uses for this dataset. Irrespective of the kind of sickness, the goal of this study is to forecast cardiac problems. The patient's age, which might be anywhere from 29 to 65 years old, is represented by this number data type characteristic. A property that may be used to determine the kind of pain is the Cp, which can take values between 1 and 4. Resting blood pressure (trestbpd) is a number between ninety-two and one hundred, and fasting blood sugar (fbs) is a number between one and zero, expressing the Boolean values true and false, respectively. Resting electrocardiogram (resting ECoG) results are shown as a three-case scale from 0 to 2. Somewhere between eighty-two and one hundred and fifty beats per minute is the thalach. A Boolean value representing exercise-induced angina is the exang. Whether cardiac disease is present is denoted by a yes or no in the dataset's target class, disease.

4. Results and Discussion

Classification Using Cross Validation

Divide the inventive sample into k subsamples at random via k-fold cross-validation. Next, divide the validation data set into 10 equal subsamples at random; save one subsample for further use. Out of the 10 subsamples, 1 is kept as validation data, also known as testing data, for evaluating the model, while the other 9 are used as training data for training the classification algorithm. The cross-validation procedure is then iterated a total of ten times (the folds), with a single instance of validation data taken from each of the ten subsamples. It is possible to get a single estimate by combining or averaging the 10 outcomes from the data folds after they have been shuffled and swapped. Table 1 displays the experimental findings of the heart disease categorization that was performed utilizing numerous folds of cross validation.

Table 1: Classification of heart disease using Cross Validation

Algorithm	No. of folds	Metrics				
		Precision	Recall	F-measure	ROC Area	PRC Area
Naïve Bayes	2	0.313	0.58	0.405	0.498	0.506
	5	0.313	0.58	0.405	0.488	0.501
	8	0.313	0.58	0.405	0.481	0.497
	10	0.313	0.58	0.405	0.482	0.498
Decision Tree	2	0.778	0.778	0.772	0.789	0.747
	5	0.785	0.788	0.778	0.821	0.797
	8	0.778	0.778	0.772	0.802	0.751
	10	0.77	0.77	0.77	0.821	0.774
	2	0.790	0.789	0.789	0.86	0.842
	5	0.805	0.804	0.806	0.864	0.847

Random Forest	8	0.802	0.799	0.799	0.865	0.841
	10	0.81	0.809	0.809	0.862	0.848

Table 1 show that folds significantly contribute to better metric values for f-measure, recall, and accuracy. The random forest approach is more effective than the Decision Tree and Naïve Bayes algorithms, even if the performance has been enhanced with the use of several folds.

Classification Using Percentage Split

The data is divided between training and testing sets using a percentage split. In fact, it partitions the data, setting aside x% for wisdom and the remaining data for testing. It comes in handy when your algorithm takes a long time to run. Consider a 60%-40% split, for example. A test set, which is an element of the original data, will be used to assess the classification findings. With a 60% split, we can say that training and testing both get 40% of the data. Based on this, the classification model is developed and the experiment is run. Using various splits of percentage splitting, Table 2 presents the experimental findings of heart disease categorization. Using the first, you can train the classifier, and using the second, you can test it.

Table 2: Classification of heart disease using Percentage Split

Algorithm	Range of percentage to split	Metrics				
		Precision	Recall	F-measure	ROC Area	PRC Area
Naïve Bayes	60-40	0.315	0.58	0.405	0.5	0.509
	70-30	0.329	0.566	0.418	0.5	0.50
	75-25	0.331	0.574	0.425	0.5	0.514
	80-20	0.355	0.596	0.46	0.5	0.521
Decision Tree	60-40	0.716	0.715	0.68	0.8477	0.817
	70-30	0.689	0.692	0.692	0.834	0.816
	75-25	0.712	0.714	0.710	0.810	0.79
	80-20	0.657	0.664	0.665	0.817	0.816
Random Forest	60-40	0.756	0.72	0.776	0.731	0.683
	70-30	0.744	0.745	0.740	0.779	0.71
	75-25	0.727	0.727	0.726	0.767	0.775
	80-20	0.725	0.715	0.718	0.741	0.708

In order to get the findings shown in Table 2, we randomly stratified the dataset and divided it into training and testing halves. The Random Forest approach has greater values for the f-measure, recall, and accuracy, according to the metrics. In conclusion, as compared to the Decision Tree and Naïve Bayes classification algorithms, the Random Forest performs better when it comes to predicting the occurrence of heart disease.

5. Conclusion

The results showed that no matter the cross-validation fold or percentage split, the Random Forest algorithm always beat the Naïve Bayes and Decision Tree algorithms. When it came to predicting the occurrence of heart disease, the Random Forest model performed better due to its greater accuracy and predictive power. The significance of using suitable data assessment methodologies and strong machine learning algorithms for trustworthy illness prediction is emphasized by the results of this research. By facilitating the early detection and efficient treatment of cardiac illness, this discovery has the potential to have a substantial impact on the healthcare business, helping to lower the death rate. To further improve the accuracy of predictions and the performance of models, future studies might investigate how to combine deep learning methods with bigger datasets.

References

1. P. Shetgaonkar and S. Aswale, "Heart Disease Prediction using Data Mining Techniques," Int. J. Eng. Res. Technol. (IJERT), vol. 10, no. 2, pp. 281–286, 2021.
2. R. Fadnavis, K. Dhore, D. Gupta, J. Waghmare, and D. Kosankar, "Heart disease prediction using data mining," J. Phys.: Conf. Ser., vol. 1913, no. 1, pp. 1–6, 2021.
3. C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," Cluster Comput., vol. 22, no. 6, pp. 14777–14787, 2019.
4. Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," Int. J. Intell. Eng. Syst., vol. 12, no. 1, pp. 242–252, 2019.
5. F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 261–268, 2019.
6. A. Kaur, "Heart disease prediction using data mining techniques: A survey," Int. J. Adv. Res. Comput. Sci., vol. 9, no. 2, pp. 569–572, 2018.
7. T. Rajesh, M. Shaik, H. Hafeez, and H. Krishna, "Prediction of heart disease using machine learning algorithms," Int. J. Eng. Technol., vol. 7, no. 2.32, pp. 363–366, 2018.
8. M. Tabassian et al., "Diagnosis of heart failure with preserved ejection fraction: Machine learning of spatiotemporal variations in left ventricular deformation," J. Am. Soc. Echocardiogr., vol. 31, no. 12, pp. 1272–1284, 2018.
9. L. Verma, S. Srivastava, and P. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," J. Med. Syst., vol. 40, no. 7, p. 178, 2016.
10. S. Devi, S. Krishnapriya, and D. Kalita, "Prediction of heart disease using data mining techniques," Indian J. Sci. Technol., vol. 9, no. 39, pp. 1-10, 2016.
11. D. Chandna, "Diagnosis of heart disease using data mining," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 1678–1680, 2014.
12. A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining techniques," Comput. Sci. Inf. Technol., vol. 4, no. 2, pp. 53–59, 2014.
13. K. Rajeswari et al., "Feature selection in ischemic heart disease identification using feed forward neural networks," Int. Symp. Robot. Intell. Sens., vol. 41, no. 1, pp. 1818–1823, 2012.
14. C. Dangare and S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," Int. J. Comput. Appl., vol. 47, no. 10, pp. 44-48, 2012.
15. D. Babaoglu et al., "Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization," J. Expert Syst. Appl., vol. 36, no. 1, pp. 2562–2566, 2009.
16. X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, and D. Steinberg, "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.