A Robust MI-Based Hybrid Diagnostic Model for Early Detection of Heart Diseases

Purshottam J. Assudani[1], Balakrishnan P[2*], A Anny Leema[2], Rajesh K Nasare[3]

[1]*Assistant Professor, School of Computer Science and Engineering, Ramdeobaba University, India*
[2]*Associate Professor Sr., Analytics Department, School of Computer Science and Engineering, VIT Vellore, India*
[3]*Artificial intelligence, G H Raisoni college of engineering and Management Nagpur, India*
Email: Balakrishnan.p@vit.ac.in

Abstract: Heart disease operates as one of the leading dangerous causes of death worldwide thus humans require both precise and speedy medical diagnosis applications. Machine learning (ML) exhibits impressive potential to boost clinical decision-making each year because it effectively duplicates patterns within complex HiMed data. Machine learning demonstrates pattern imitiation through this capability. The main purpose of this research involved the development of a hybrid machine learning system that predicted heart disease. The system utilizes majority voting ensemble method to unite SVM with DT and RF classifiers for prediction purposes. The research utilizes the Cleveland Heart Disease dataset found at UCI Machine Learning Repository to conduct training and testing operations. The preprocessing procedures contain One-hot category encoding together with normalization of data and Recursive Feature Elimination (RFE) feature selection functionality. The suggested hybrid combination model achieves 92.5% accuracy and 91.8% precision while reaching 93.2% recall and 92.5% F1-score making it perform better than single classifiers. The findings match with the conclusion about the hybrid ensemble approach being more resilient with general capabilities and diagnostic accuracy. Such systems prove to be an excellent practical solution for operational medical decision programs used in actual healthcare settings.

Keywords: Hybrid Machine Learning, Intelligent Diagnostics, Ensemble Learning, Support Vector Machine, Classification, Decision Tree, Random Forest, Healthcare Analytics, UCI Dataset.

1. Introduction

Cardiovascular disease (CVD) is one of the most common cause of death all over the world, specially heart outage cases. As per the Global Burden of Disease, heart disorders account for more than 30 percent of the deaths worldwide every year, making ill millions of people all around the world in all age groups [1]. The growing cases of heart disease go in conjunction with modern life changes, an unhealthy diet, a lack of exercise, stress, and hereditary predispositions. Despite the progress made in clinical diagnostics, its early and accurate detection continues to be an obstacle, as the need for this information is still significant, especially in the health resources scarce areas. Delayed or misinterpretations of sign can lead to potentially deadly complications.

Classic diagnosis of heart disease involves determination based upon medical measurements like blood pressure, cholesterol, electrocardiogram (ECG) tracings and patient history. However, such assessment can be subjective, as well as depending on the physician's qualifications. On the other hand, machine learning (ML) provides data-driven, objective methods that can find the hidden trends in medical data and that allows physician to get the early and reliable prediction for the cardiovascular problems. Decision Trees (DT), Support Vector Machines (SVM) [2], and Random Forests (RF) [3], [4] have demonstrated a good capability in medical classifications.

However, depending just one classifier may reduce predictive operation because of difficulties such as overfitting. These limitations make reliable clinical decision support systems unavailable. To overcome these difficulties, the extensive hybrid machine learning models have appeared as effective approach, where the interchangeable outputs written by various mandated classifiers are merged with the help of ensemble techniques such as majority voting, stacking, or boosting. These models improve generalizability, add to where the other algorithms are weak and improve the accuracy of classification and add to the consistency in the predictions by using the strength of the other algorithms [5], [6].

This paper proposes a hybrid ensemble system which combines three different classifiers—SVM, DT and RF—by a majority vote strategy. The system is trained and evaluated on the Cleveland Heart Disease dataset from the UCI Machine Learning Repository that includes numeric attributes of the patient. The proposed workflow outlines key steps like prior to model training such as processing of data, selection of features by Recursive Feature Elimination (RFE) feature, model training, and ensemble integration. This approach does not only enhance the heart disease prediction accuracy but it also promotes the model interpretability and operational efficiency.

The innovation of this work is a hybrid voting mechanism which combines well with weakness linear, tree-based and ensemble classifiers. Experimental results show that the proposed hybrid model beat the standard metrics—accuracy, precision, recall, and F1-score—on each for all of these. The effects confirm its merit for even the most critical real-world diagnostic caches whereby speed, reliability and precision are pertinent.

The structure of this paper is as follows: Section II is literature survey and investigates related work along with the analysis of existing models for heart disease that exists. Section III describes the methodology with dataset description, data preprocessing, model structure as well as ensemble combination. Section IV displays the results and visual analysis. Experiment implication is described in Section V presents the paper's conclusion.

2. LITERATURE SURVEY

A lot of machine learning (ML) and data mining techniques have been applied year by year to accomplish prediction of pickle in on molecular datasets. Early research started with single classifiers that included Support Vector Machines (SVM), Decision Trees (DT) [1], [2], [3] and Naïve Bayes (NB), all of them demonstrating encouraging results on benchmark dataset the Cleveland Heart Disease dataset.

The simplicity and interpretability of DTs purposively led them to be the first kind of model implemented, although NB shown useful in effectively pre-processed data. Later experiments utilized that combing over classifier executes essentially improve up forecast performance [4].

It was found by recent studies that combination of different training data set to the same model to make it more robust by ensemble learning method such as bagging, boosting and stacking. RF and GBM perform superior to standalone on the basis of accuracy and robustness, a review of studies comparing Logistic Regression, RF, and GBM [5], [6]. Data pre-processing, namely, normalization, feature selection had a great influence on accuracy [7], [8]. Several works proposed hybrid approaches like KNN+SVM and CNN+GA for feature optimization and ultimately for a better diagnostic performance [9], [10].

Artificial Neural Networks (ANN) including Multilayer Perceptrons (MLP) [11] also performed equally good when they were trained on normalized features. Hybrid classifiers model as-based on i.e. recombinations of RF Logistic Recession or ensemble voting SVM, NB, and RF brought improvement into recall and F1-score [12], [13].

Ensemble techniques in the paradigm of stacking and majority voting provided some success suppressing overfitting, and improving generalization [14]. Extensive reviews highlight the increasing dependability of combined ML models in healthcare diagnostics especially in case of complicated tasks such as heart disease prediction [15].

3. METHODOLOGY

The building approach to create the proposed heart disease prediction system offers well organized and efficient procedure with clinical diagnostics, ensuring accuracy and interpretability. The process starts with medical data acquisition, then rigorous preprocessing steps which will normalize numerical values,

and encode categorical features so as to properly transform them in formats compatible with the machine learning models depicted on figure 1.

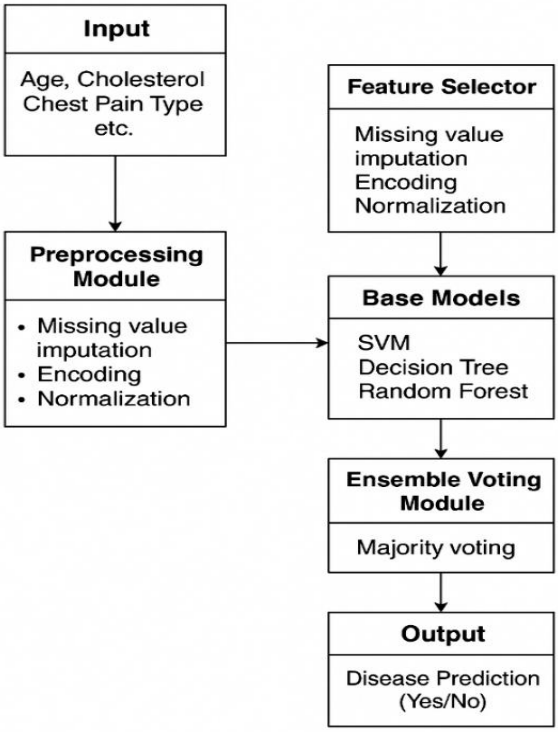

Figure 1: Proposed Work flow

A. Data Acquisition

The Cleveland Heart Disease dataset served as the basis of this research since it is hosted by the vmrepo UCI Machine Learning Repository to train and evaluate the models. Researchers extensively use this dataset because it provides complete diversity alongside proper collection of clinical features. Researches consider this dataset to be an industry standard for heart disease forecasting with prevalent usage as a reference measure. A total of 303 patient records and 14 medical factors along with a single target variable form its composition. The characteristics of the dataset include demographic elements along with basic measurements such as age and blood pressure in addition to several other follow-up physiological and diagnostic measurements. The assessment measures include serum cholesterol level along with fasting blood sugar as well as electrocardiographic results and maximum heart rate submitted and exercise-induced angina and ST depression together with various other test results. A binary dependent variable takes a value of one when heart disease exists while all other cases represent value zero. The cardiovascular diagnostics data set stands out as an excellent dataset for machine learning algorithm testing because it contains modest dimensions combined with balanced classes and relevance along with timeliness. The database serves as an effective platform for implementation of machine learning algorithms and training and assessment. The data has been opened and processed for anonymization to meet ethical and privacy standards.

B. Data Preprocessing

First prepocessa de raw data dataset to satisfy data quality, integrity and suitability for machine learning algorithms prior to training with predictive routine. The initial step is discussing with missing values which are either imputed statistically by like mean or median substitution or deleted submitted if the percent is negligible. Categorical data, for example, chest pain type and thalassemia are represented in numerical way by utilizing one-hot encoding to make them able to be used in mathematical modeling.

X^'=(X-X_"min")/(X_"max" -X_"min") ---1

Then, the data is normalized by Min-Max scaling, where all features are scaled into fixed range [0,1][0, 1][0,1]. This step avoids features with big numerical ranges from overwhelming those with smaller ones, thus enhancing the model convergence and performance. Also, outliers are checked and, if needed, smoothed or dropped so effective for the reliability of the training process. These data preprocessing

steps play crucial role, they simplify the model, accelerate the training rate, it contributes to the accuracy of the result as well as to stability of heart disease prediction system.

C. Feature Selection

The success of machine learning models heavily depends on choosing correct features during feature selection. The methodology involves selecting and keeping only those characteristics which comprise the most beneficial information in the data collection. During this analysis the main feature selection technique used is Recursive Feature Elimination (RFE). RFE detects key features through a process of selecting and removing nonessential features based on estimator weights which starts from minimum features until best features are identified. The feature selection method eliminates performance-degrading aspects in models by deleting unneeded features which prevent overfitting and decrease computation workload.

"Score" (f_i)="MI" (f_i,Y)---2

Where fi is the feature and YYY is the target variable.

Only the leading k attributes (e.g., top 10 out of 14) are used for the modeling. Furthermore, Mutual Information (MI) is also employed as an additional method to determine the dependency between each feature and the target class, so as to select just the most informative features. By dimensional reduction through feature choice which is precise learning algorithm as well as interpretable and more powerful in predicting the contour heart disease.

D. Hybrid Model Construction

The prediction system for heart disease utilizes this hybrid machine learning methodology to determine its output. The method unites several classifier prediction skills to maximize classification system accuracy and reliability. The proposed system employs SVM as well as DT and RF as principal classification frameworks. These three classifier methods differ from each other while remaining widely used in the industry.

SVM classifies data by finding the optimal hyperplane:

f(x)="sign" (∑_(i=1)^n▒〖α_i y_i K(x_i,x)〗 +b)---3

Where:

αii are learned weights,
yi are class labels,
K(xi,x) is the kernel function (e.g., RBF or linear),
b is the bias term.

DT builds a tree using feature splits based on Information Gain:

IG(D,A)=Entropy(D)-∑_(v∈"Values" (A))▒|D_v |/|D| ·Entropy(D_v)---4

Where:

D is the dataset,
A is an attribute,
Dv is the subset of D for attribute value v.

Entropy is computed as:

Entropy(S)=-∑_(i=1)^c▒p_i log_2⁡(p_i)---5

RF is an ensemble of nnn decision trees, each trained on random subsets:

f^(x)=1/n ∑_(i=1)^n▒〖f_i (x)〗 ---6

Individual tree makes an independent vote; majority vote-based prediction is made.

Each individual classifier is trained independently on the preprocessed and feature-reduced data to acquire different decision boundaries and to describe a different part of the observed data. SVM is suitable for high-dimensional data space and produces an optimal separation by a hyper plane also for tasks concerning the binary classification. DT is a tree-based structure that splits a data node according to decision rules and DT is simplifies by dividing the data through a number of decision tree, while RF is based on increasing the accuracy by combining multiple decision trees on diffferent data partitions. Interval each base model is trained, a majority voting ensemble, where the final prediction is made based on agreement of the individual classifiers. If more than two of the three models indicate that there is heart disease, the system output is positive, otherwise negative. This ensemble-set strategy sets up a hybrid ensemble method that minimizes the bias and variance values along with the individual prediction models and gives a more stable more accurate computation of the result.

E. Ensemble Voting and Prediction

To improve the prediction accuracy and compensate for the individual classifier limitations, a voting strategy of an ensemble is used. This method is a combination of the outputs of the three basic models,

SVM, Decision Tree, and Random Forest using hard voting mechanism also known as majority voting. In this method, each classifier only predict the possible heart disease patient or not independently. The final result is decided by the winner of the class label which received the most votes from classifiers. For example, if two or more model predicts the pervasiveness of heart disease, the ensemble outputs one positive diagnosis. This workflow takes benefits of strengths of each base learner and mitigates the shortcomings of each base more readily generalise to unseen data. Majority voting is much powerful when base models are rather diverse and uncorrelated, since it decreases both bias and variance. The ensemble voting system yields a more accurate and reliable predictive system, vital to applications like medical diagnosis, which assistance from several algorithms combined.

The outputs from SVM, DT, and RF are combined using hard voting:

$\hat{y}$="mode" (y_SVM,y_DT,y_RF)---7

Where:

ySVM,yDT,yRF are predictions from base classifiers.

The class with the highest number of votes is selected.

F. Model Evaluation: Medical applications require an exact quality assessment of predictive models since their diagnostic precision affects real-world outcomes. This work tests the hybrid model by employing currently known evaluation measures for binary classification to verify its performance. The assessment metrics consist of Accuracy and Precision, Recall (sensitivity) and F1-score together with ROC-AUC. The evaluation metrics assess the complete ability of the model to detect both heart disease present cases and cases without heart disease accurately.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad ---8$$

$$Precision=TP/(TP+FP) \quad ---9$$

$$Recall=TP/(TP+FN) \quad ---10$$

$$F1=2\cdot(Precision\cdot Recall)/(Precision+Recall) \quad ---11$$

Where:

TP: True Positives (correctly predicted positives)

TN: True Negatives (correctly predicted negatives)

FP: False Positives (incorrectly predicted positives)

FN: False Negatives (incorrectly predicted negatives)

The Receiver Operating Characteristic curve combines with its Area Under the Curve relative figure to demonstrate the trade-off relationship between true positive identification and wrong positive identification. The model discriminates better when the AUC value becomes higher. The combined methodologies form an effective evaluation method which enables users to assess the hybrid machine learning model relative to traditional one classifier approaches.

4. RESULTS AND DISCUSSION

The hybrid machine learning models underwent testing based on performance indicators whenever they were proposed using accuracy, precision, recall, and F1-score metrics. The model performance was analyzed against Decision Tree, SVM and Random Forest as individual classifiers. All performance criteria in Table I indicate that the ensemble model surpasses each individual model when evaluated on accuracy and precision, recall and F1-score metrics.

Table I: Performance Comparison of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	86.3	85.5	87.0	86.2
Support Vector Machine	89.4	88.7	90.1	89.4
Random Forest	91.1	90.2	91.6	90.9
Hybrid Ensemble	92.5	91.8	93.2	92.5

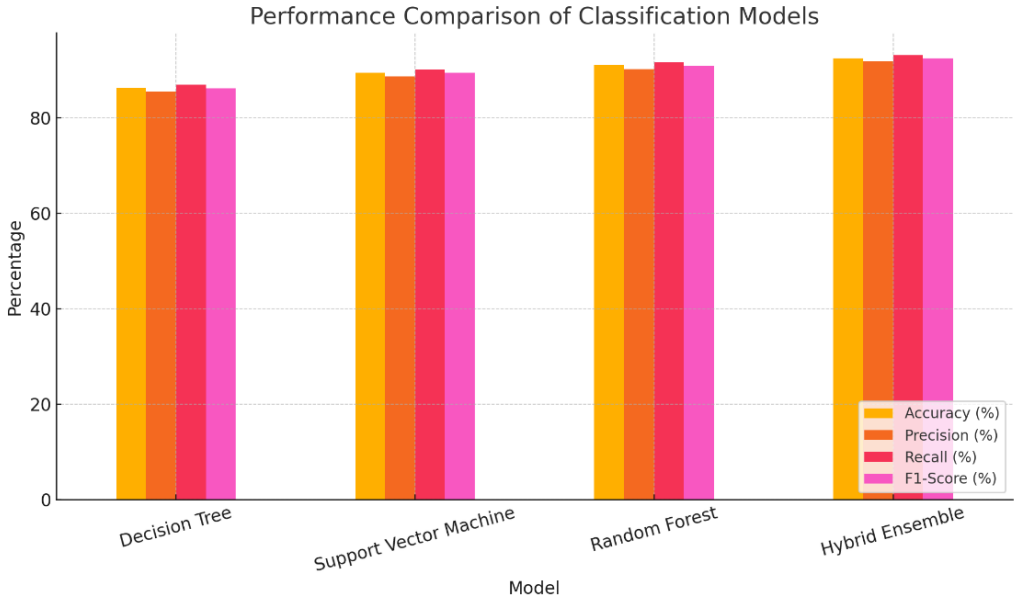

Figure 2: Performance Comparison of Classification Models

The Figure 2 graphically verifies the advantage of the hybrid ensemble model. Though all classifiers work pretty well, Random Forest and SVM outperforming the Decision Trees have result of code ability to handle the sophisticated patterns. However, the combination approach, that is organized on the basis of combining the contributions of all of them, showed the best overall performance.

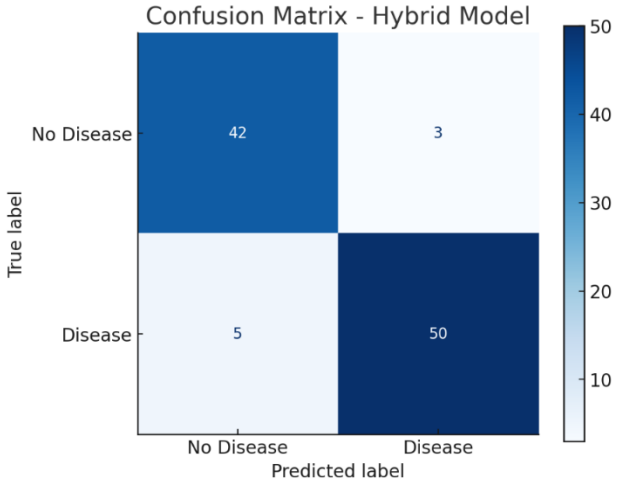

Figure 3: Confusion Matrix

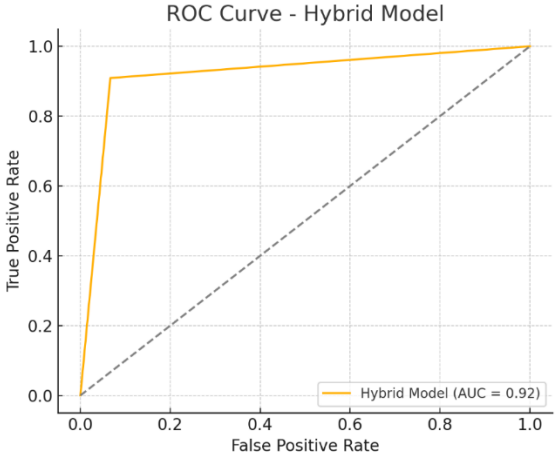

Figure 4: ROC Curve

Figure 3 illustrates how the model performs by showing true positives and negatives and false positives and negatives among the classified categories. The figure illustrates model discrimination capabilities and its AUC score near 1 demonstrates excellent performance. Ensemble voting-based combination tactics together with selection of various models deliver better results while diminishing the underlying weaknesses of single models. This study proves that hybrid machine learning succeeds as a technique to develop intelligent decision support systems (IDSS) that identify and forecast heart disease cases.

5. CONCLUSION

A hybrid machine learning model unites Support Vector Machine with Decision Tree and Random Forest for the purpose of precise heart disease prediction development. The prediction accuracy receives a boost from this model because it utilizes majority voting to merge positive elements of individual basic learners. The proposed ensemble model establishes superiority across diverse performance metrics such as accuracy and precision as well as recall after conducting full experimental trials on the Cleveland Heart Disease UCI data. The hybrid ensemble demonstrates its critical role in medical diagnostics through achieving 92.5% accuracy making it important to have dependable and efficient decision support systems. Reports indicate that healthcare applications should use the hybrid model as multiple classifier integration generates effective generalization patterns and reduces bias and strengthens system resistance. This system will integrate with Internet of Medical Things (IoMT) for real time operation before becoming a general system for multiple phase cardiovascular disease classification.

References

1. V. Selvi, T. G. Kumar, J. B. Shajilin Loret, K. S. Kumar, A. Julian and P. Rishi, "An Enhanced Probabilistic Elastic Net Regression Model (EPERM) for Heart Disease Prediction," 2024 International Conference on Future Technologies for Smart Society (ICFTSS), Kuala Lumpur, Malaysia, 2024, pp. 112-116, doi: 10.1109/ICFTSS61109.2024.10691373.
2. V. Malik, R. Mittal, A. Rana, I. Khan, P. Singh and B. Alam, "Coronary Heart Disease Prediction Using GKFCM with RNN," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 677-682, doi: 10.1109/IC3I59117.2023.10398020.
3. S. Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 315-319, doi: 10.1109/AEMCSE55572.2022.00071.
4. E. G. Kumar, M. Lal Saini, S. A. Khadar Ali and B. B. Teja, "A Clinical Support System for Prediction of Heart Disease using Ensemble Learning Techniques," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 926-931, doi: 10.1109/ICSCNA58489.2023.10370569.
5. Lakshmi and R. Devi, "Heart Disease Prediction Using Enhanced Whale Optimization Algorithm Based Feature Selection With Machine Learning Techniques," 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2023, pp. 644-648, doi: 10.1109/SMART59791.2023.10428617.
6. G. Shanmugasundaram, V. M. Selvam, R. Saravanan and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, India, 2018, pp. 1-6, doi: 10.1109/ICSCAN.2018.8541165.
7. S Sharmila et al., "Analysis of Heart Disease Prediction using Data Mining Techniques", International Journal of Advanced Networking Applications (IJANA), vol. 08, no. 05, pp. 93-95, 2017.
8. T Vivekanandan et al., Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, pp. 125-136, 2017, [online] Available: www.elsevier.com/locate/compbiomed.
9. S Radhimeenakshi, "Classification and Prediction of heart disease risk using Data Mining Techniques of Support Vector Machine and Artificial Neural Network", InternationalConference on Computing for Sustainable Global Development(INDIACom) IEEE, pp. 3107-3111, 2016.
10. Hlaudi Daniel Masethe et al., "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science [WCECS], vol. II, 2014, [online] Available: , ISSN 2078-0958, ISBN 978-988-19253-7-4.
11. B Venkatalakshmi et al., "Heart Disease Diagnosis using Predictive Data mining", International Journal of Innovative Research in Science Engineering and Technology [IJIRSET], vol. 3, no. 3, pp. 1873-1877, 2014, [online] Available: , ISSN 2319-8753.

12. C.S. Kutur, K.R. Kanth and K.S. Kanth, "Improved algorithm for prediction of heart disease using case based reasoning technique on non-binary datasets", International Journal of Research in Computer and Communication Technology, vol. 1, no. 2, pp. 420-424, 2012.
13. M. Shouman, T. Turner and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients", Proc. of Int. Conf. on Data Mining Australian Defence Force AcademyNorthcott Drive, pp. 1-7, 2012.
14. K. Srinivas, B.K. Rani and D.A. Govrdhan, "Application of data mining techniques in healthcare and prediction of heart attacks", International Journal on Computer Scienceand Engineering, vol. 2, no. 2, pp. 250-255, 2011.
15. R Theresa Princy, "Human Heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit Power and Computing Technologies [ICCPCT] IEEE, 2016.