Comparative Analysis of Classification Models for Healthcare Data Analysis

Abdul Hafeez Babar
Mehran University of Engineering and Technology
Jamshoro, Pakistan
Email: hafeezbabar76 [AT] gmail.com

Naeem Ahmed Mahoto
Mehran University of Engineering and Technology
Jamshoro, Pakistan

***Abstract*— The classification is one of the important directions of supervised learning. The prediction of highly sparse data is a challenge and an open issue. In this regard, this study conducts comparative analysis of eight machine-learning algorithms for classifying healthcare data (i.e., Heart Diseases). The eight classifiers used in this study are: 1) Naïve Bayes (NB), 2) Single Conjunctive Rule Learner (SCRL), 3) Radial Bias Function (RBF), 4) Decision Tree (DT), 5) K-Nearest Neighbor (k-NN), 6) Multilayer Perceptron (MLP), 7) Random Forest (RF), and 8) Support Vector Machine (SVM). In order to obtain better classification outcomes, ensemble-learning methods such as bagging, boosting, decorate, voting, random sub space and dagging have also been used in conjunction with considered classification models. The experimental results have been validated using 10-fold cross validation method. It has been revealed in results that SVM performed better in both cases: i) Simple classification model, and ii) Classification model with ensemble-learning methods. The accuracy of SVM, in both cases, achieved 86.13% being the top classifier among the considered models. The RBF produced second higher accuracy 83.82% and third MLP as 83.5%. The study indicates that the classification models in conjunction with ensemble-learning methods can significantly enhance the predictive outcomes and scalability of classification schemes, which is of practical importance when used for healthcare data.**

Keywords- Classification, Heart Disease, Machine Learning.

I. INTRODUCTION

Coronary illness has been the most critical reason for death on the planet amid the previous 10 years [1]. This has taken incredible consideration of research group; for example, Heart Disease Monitoring System has been presented in [2].

Wellbeing is a fundamental component of people, which in the long run construct dynamic social orders. The advanced technological changes have made arrangements of electronic information at an extensive scale. Health related information can be used to help in getting data with respect to medical problems, for example, medical issue patterns, infection chance variables, results of treatment or general wellbeing intercessions, social insurance cost and its utilization. The healthcare research has delivered surprising disclosures, for example, the development of new medicines, enhancement of social insurance rules, change in human services administrations.

A huge amount of electronic health record (EHRs) has been made available from healthcare facilities due to the modern technological paradigm. The data mining is one of the sophisticated fields that offer to uncover the hidden patterns and also allows building, predictive models. Data mining in the healthcare improves the quality of patients care and decreases the healthcare service costs. The healthcare industry produces an enormous amount of data that's too hard to be examined by conventional methods.

The Data Mining software application includes various methodologies that have been developed by both medical and heart disease research center. Every year due to the heart diseases, World Health Organization (WHO) has estimated 12 million people died worldwide [1]. In 2012, WHO has estimated 17.5 million people died from Cardio Vascular Diseases (CVD) [1]; WHO estimated, by 2030, almost 23.6 million will die due to the cause of heart diseases [1]. Several research attempts have been made to find the performance of machine learning techniques determining correlations among various attributes of patients and heart- related diseases. This field is still an open challenge due to high-dimensionality and diversified nature of medical dataset.

The main objective of this study is conducting comparative analysis of machine-learning algorithms, when predicting healthcare data that is highly diversified. The considered machine-learning algorithms: Decision Tree (DT), Naïve Bayes (NB), Single Conjunctive Rule Learner (SCRL), Radial Bias Function (RBF), Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) have been applied for the experimental activities. In addition, ensemble-learning methods have also been applied for the classification models in order to achieve higher prediction outcomes. The evaluation of each considered classification model has been measured using well-established matrices: accuracy, precision, recall, and f-measures with the help of cross validation method.

The rest of the paper organization is as follows. Section 2 reports related work; Section 3 addresses classification models being used in this study. The performance of each considered method has been reported in section 4. The experimental results are described in section 5; whilst conclusions are drawn in section 6.

II. RELATED WORK

Several attempts are made to evaluate the performance of classification methods for healthcare data, particularly, Heart Disease [3]. In study [4], a comparison of three different classification algorithms - Neural Network, Support Vector

Machine and Multilayer Perceptron, have been reported for Coronary Disease dataset. The result showed Support Vector Machine (SVM) as able to give better accuracy results than Neural Network and Multilayer Perceptron.

An effective predictive machine learning techniques for heart disease dataset with several classifiers available in WEKA and RapidMiner data mining tool have been addressed in [5]; resulting better accuracy for SVM classifier [5]. Likewise, an accuracy of 80.41% in terms of classification between two classes (absence or presence of heart disease) have been discussed in [6]. The study [7] has developed models for heart disease prediction using Stream Associative classification and Association rules and compared to predictive rules mined with decision trees.

Decision List, K-NN and Naïve Bayes for classification of heart disease have been used and compared the accuracy of models. Naïve Bayes gives the 52.33% of accuracy as better classifier [8]. Another study [9] focused on three popular data mining classification algorithms: Decision Tree, Naïve Bayes, and K-NN, and compared accuracy of highly disperse Cleveland Heart Disease Database. Further, the dataset divided into three different cases and applied each classifier in the disperse datasets. Finally observed that K-NN classifier performed better than two classifiers (i.e. Decision Tree and Naïve Bayes) [9]. The three popular data mining classification algorithms - CART, ID3 and Decision Table have been reported in [10]; the accuracy of each models for the Cleveland Heart Disease Database used 10-fold cross validation. The results showed that CART outperformed other considered methods [10]. The study in [11] considered 10 different classification algorithms - Naïve Bayes, Decision Tree, Decision Stump, K-NN, Random Forest, Rule Induction, CHAID, Neural Network and SVM. The outcomes revealed that Naïve Bayes and SVM performed better for prediction and detection of heart disease [11].

This study considers Decision Tree (DT), Naïve Bayes (NB), Single Conjunctive Rule Learner (SCRL), Radial Bias Function (RBF), K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM) for the heart disease dataset. The reason behind using these algorithms is that almost all possible branches of supervised learning approaches are considered. Thus, the experimental results convers broader spectrum of supervised learning algorithms for the diverse healthcare data (i.e., heart disease). Further, this study also combines ensemble methods with considered classification methods to achieve better accuracy.

III. CLASSIFICATION MODELS FOR HEALTHCARE DATA ANALYSIS

A. Dataset

***The Cleveland Heart Dataset*:** Cleveland Heart dataset has been obtained from the online available repository (UCI Machine Learning Repository). The dataset contains 76 attributes, which possess numeric value. However, after preprocessing and targeted data portion only 14 related attributes are used in this study. These attributes are listed below in Table 1.

TABLE I. HEALTHCARE DATA DESCRIPTION.

Attribute	**Description**
Age	Age in years as input
Sex	it has values (1=Male or 0=Female)
Cp	Chest Pain Type (1=typical type, 2=typical type angina, 3=non-angina pain, 4=asymptomatic)
Trestbps	Resting blood pressure in mm Hg
Chol	Serum Cholesterol is in mg/dl
Fbs	Fasting blood sugar has two values as input (value 1=FBS>120 mg/dl and value 0-=FBS<120 mg/dl)
Restecg	Resting electrographic results has three values as input (value 0=normal, value 1=having ST-T wave abnormality, value 2=showing definite left ventricular hypertrophy)
Thalach	The maximum heart rate achieved by the patient
Exang	Exercise-induced angina has two values as input (value 1= Yes and value 0=No)
Oldpeak	The ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment takes three values as input (value 1=unsloping, value2=flat, value 3=down sloping).
Ca	Number of major vessels colored by fluoroscopy take three values as input (value 0-3)
Thal	Defect type take three values as input (value 3=normal, value 6=fixed defect and value 7=reversible defect)
Age	Age in years as input
Sex	it has values (1=Male or 0=Female)
Cp	Chest Pain Type (1=typical type, 2=typical type angina, 3=non-angina pain, 4=asymptomatic)

Data Portioning: Because of restricted measures or attributes of considered dataset; this study uses K-Fold Cross Validation system for the evaluation of classification results. All the attributes are in the original training data set are used for both training as well as validation. It works as a single hold-out method in which dataset is divided into K subsets, and each time, one of the K subsets is used as the test set and the other K-1 subsets are put together to form a training set. Finally, an average of all K trails is computed. The variance of the resulting estimate is reduced as K is increased.

B. Classification Models/Methods

Decision Tree (DT): Decision Tree has a flowchart structure, which parts information into root hub, branches, and leaf. At each splits in the tree, every input attributes are estimated for their impact on the predictable attributes. It is a simple, fast and easy to evaluate, clarify and implement. It requires no experiential knowledge or boundary setting.

Naïve Bayes (NB): The Naïve Bayes classifier works according to use of Bayes Theorem. This classifier assumes each attribute of the considered dataset as an independent, which is key point for prediction. This classification helps in prediction of larger dataset, specially, textual dataset.

Single Conjunctive Rule Learner (SCRL): SCRL classifier technique centers to derive an arrangement of standards from a dataset, which embody all generalized data inside the undertaken dataset. Different standards or rules can occur in this model and generally from the most predominant to the most particular.

K-Nearest Neighbors (K-NN): K-NN classifier is a simplest classification method to classify the dataset attributes based on their resemblances. In K-NN, the similar instances with a high amount of occurrences and close to another are called neighbours. Therefore, the classifier sets the groups that contain the same instance with the nearest neighbours and put them into the classifier groups. K-NN makes the prediction promptly by computing the resemblance between the input data set and each training attribute instances.

Multilayer Perceptron (MLP): Multilayer Perceptron is an enlargement from the simple perceptron in which extra hidden layers (both the additional input and output layer, not connected externally) are added. Further on a hidden layer can be used. The existence of these layers allows an Artificial Neural Network (ANN) too imprecise a variety of non-linear functions. This classifier has an unmistakable engineering and straightforward calculation. Hence, it is a standout amongst the most renowned neural system models.

Radial Bias Function (RBF): In contrast with Multilayer Perceptron (MLP), Radial Basis Function (RBF) network uses a moderately different approach in terms of the number of hidden layers, an output layer, local and global approximation, and a number of parameters.

Random Forest (RF): The random forest is a machine learning classifier that comprises of several decision trees that randomly selected subsets of training sets and use averaging to improve the predictive accuracy and control over-fitting. Random forest increases several classification trees without pruning. Then each decision tree classifies a test sample and random forest assigns a class, which have maximum occurrence among these classifications.

Support Vector Machine (SVM): Support Vector Machine is used to search a decision boundary between two classes that is famous for away from any point in the training data. SVM fosters a hyperplane or a set of hyperplanes in endless dimension space. The hyperplane acts as a separator of the two classes and SVMs are binary classifiers in nature.

Ensemble Learning Methods: Ensemble Learning Method uses several machine learning classification algorithms to obtain effective prediction performance results than a single classifier model. The main idea to use the ensemble learning method is to make a predictive model by combining multiple models into a single once and improve the classification accuracy. The six different models - *Bagging, Boosting, Decorate, Voting, RandomSubSpace, and Dagging* have been considered in this study. ***Bagging*** is the collection of prediction of entirely the identical type of vote. ***Boosting*** works similar to bagging aside from the execution of the past model affects the new model. ***Decorate*** obtain higher accuracy than Boosting on small training sets and acquire comparable performance on large training sets and also used for building diverse ensembles of classifiers. ***Voting*** is the easiest way of combining the prediction from multiple machine learning classifiers. ***RandomSubSpace*** builds a decision tree based classifier that enables the highest accuracy on training data and enhances on generalization accuracy as it growths in complexity. ***Dagging*** generates a number of disjoining, separate folds out of the data and prepares each chunk of data of the supplied base classifier. In the Dagging predictions are made via greater numbers of votes [12].

IV. PERFORMANCE MEASURE

The performance of each classifier and combination of ensemble methods has been measured using evaluation metrics such as accuracy, precision, recall, and F-measure. Specifically, the performance of the classifiers is measured to differentiate between actual and predicted class/label.

Accuracy: accuracy is defined as the ratio of correctly predicted data to the total number of data and is calculated as follows

$$Accuracy = \frac{TP \mid TN}{P \mid N} \tag{1}$$

TP (True Positive): It denotes the number of records classified as true while they were actually true.

TN (True Negative): It denotes the number of records classified as false while they were actually false.

P (Positive): It denotes the total number of actual positive data records.

N (Negative): It denotes the total number of actual negative data records.

Precision: This is the positive predictive value measure of resultant significance.

$$Precision = \frac{TP}{TP \mid FP} \tag{2}$$

Recall: This is the sensitivity.

$$Recall = \frac{TP}{P} \tag{3}$$

F-Measure: it is the harmonic mean of precision and recall.

$$F - measure = \frac{2}{{}^{1}/_{precision} + {}^{1}/_{recall}} \tag{4}$$

V. EXPERIMENTAL RESULTS

The considered dataset comprises of 14 different attributes with 303 patient records. The evaluation has been carried out using popular 10-fold cross-validation technique. The experiments are performed on Intel Core-2 Duo processor at 2.4GHz with 4GB RAM (Random Access Memory) using WEKA (data mining software tool).

The experiments have been performed into 2 cases: 1) Classification models and 2) Classification models with ensemble-learning methods. The 1st case experiments considered simple classification methods; whilst 2nd case experiments used combination of ensemble methods with classification models. The results of both cases have been reported in the following.

A. Classification Models

The performance measures related to each considered classifier have been shown in Table 2. The result shows that SVM has better accuracy (84.158%) result than other classifiers.

TABLE II. PERFORMANCE MEASURES OF CONSIDERED CLASSIFIERS.

Classification Model	F-Measure	Precision	Recall	Accuracy %
SCRL	0.718	0.734	0.703	69.967
DT	0.801	0.774	0.83	77.778
MLP	0.84	0.873	0.842	82.5083
RF	0.847	0.839	0.855	83.1683
K-NN	0.845	0.848	0.842	83.1683
NB	0.851	0.836	0.867	83.4983
RBF	0.853	0.845	0.861	83.8284
SVM	0.86	0.827	0.897	**84.1584**

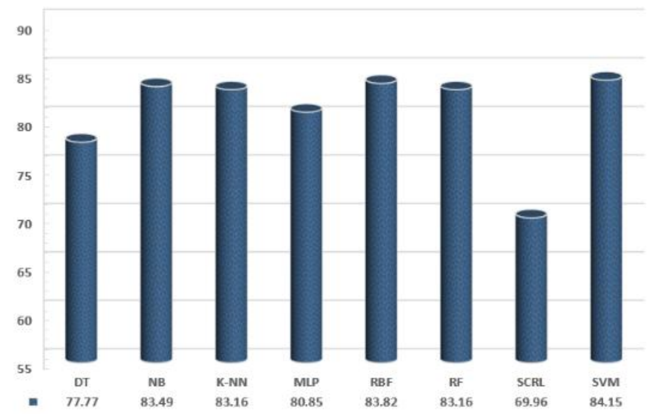

Figure 1. Accuracy of each classification method.

However, K-NN classifier requires number of neighbours for its working mechanism. Several different k values have been used; better results have been obtained when k value equals to 9. Fig. 1 represents the accuracy results of each considered classification method.

B. Classification Models with Ensemble-learning Methods

The second experiment uses ensemble-learning methods with considered classification models. The outcomes of each ensemble-learning methods in combination of classification methods are reported in the following.

Bagging: The experiment showed that bagging improves the performance accuracy of the classifier. As illustrated in Fig 2, Bagging improves the performance of some classifiers. For instance, DT expanded the accuracy level results from 77.77% to 81.18% and SCRL expanded from 69.96% to 78.87. SVM did not enhance, but rather, in any case kept up similar accuracy results.

Boosting: Boosting implies operating a weak classifier. In this trial, Boosting has been connected with all eight considered classification algorithms. Though, it was expected that this combination would function admirably well on the weaker classifiers. As portrayed in Fig 3, the accuracy level of Decision Tree algorithm has been increased from 77.77% to 82.17%, likewise, for SCRL from 69.96% to 81.18%. The rest of the classification models remained nearly the same.

Decorate: Decorate holds higher accuracy than Boosting on small training sets, and acquire comparable performance on large training sets. As illustrated in Fig 4, significant increase in accuracy of few classifiers has been observed; for instance, DT produced from 77.77% to 79.2%, SCRL from 69.96% to 76.56. However, the other classifiers could not produce appealing results.

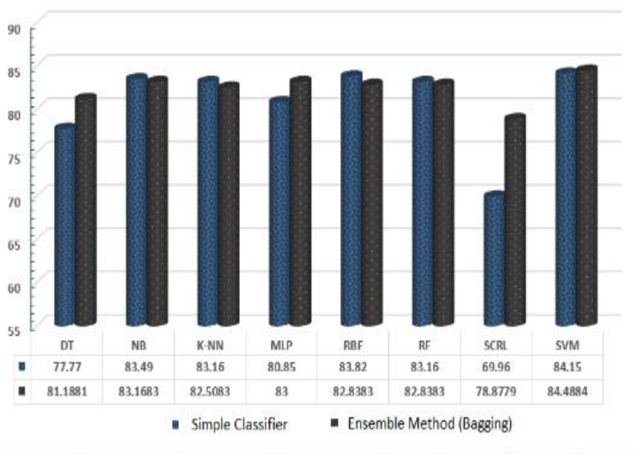

Figure 2. Accuracy of considered classifiers with Bagging method.

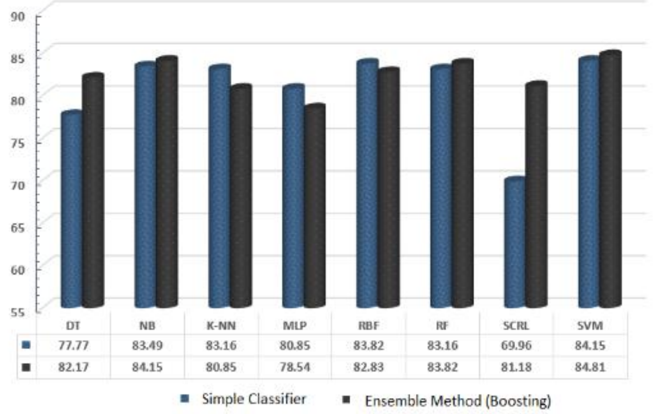

Figure 3. Accuracy of considered classifiers with Boosting method.

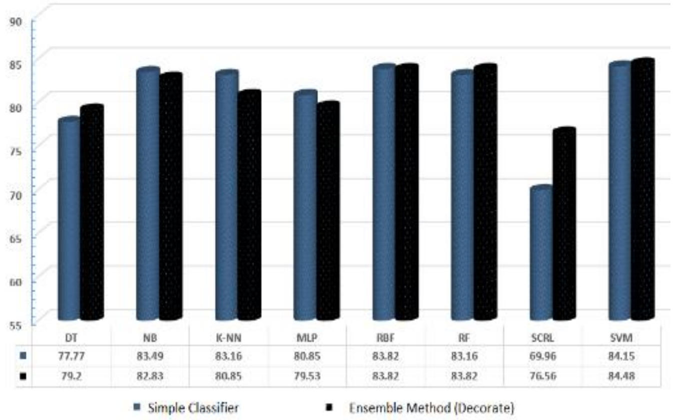

Figure 4. Accuracy of considered classifiers with Decorate method.

Voting: The experimental results obtained while use of voting ensemble-learning method with classification methods are illustrated in Fig. 5

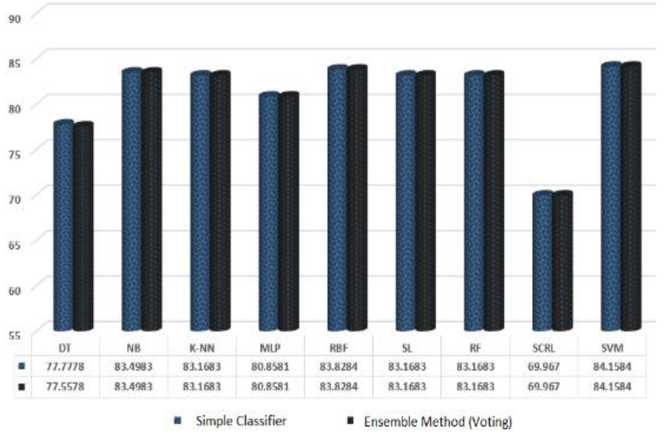

Figure 5. Accuracy of considered classifiers with Voting method.

RandomSubSpace: RandomSubSpace based on complexity, improve the performance accuracy of classifiers. As illustrated in Fig 6. It increased the accuracy of only two classifiers DT and SCRL and not more effective to other classifiers.

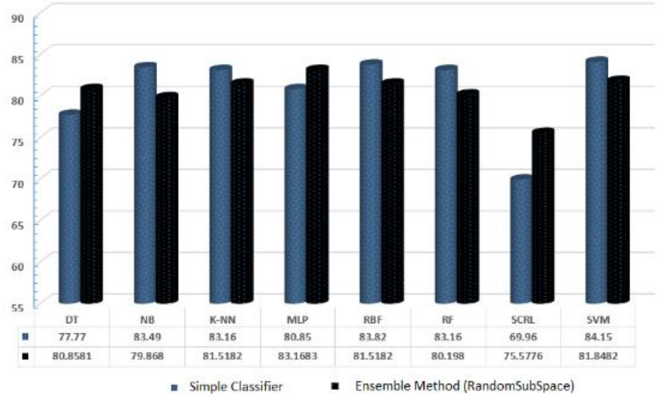

Figure 6. Accuracy of considered classifiers with RandomSubSpace method.

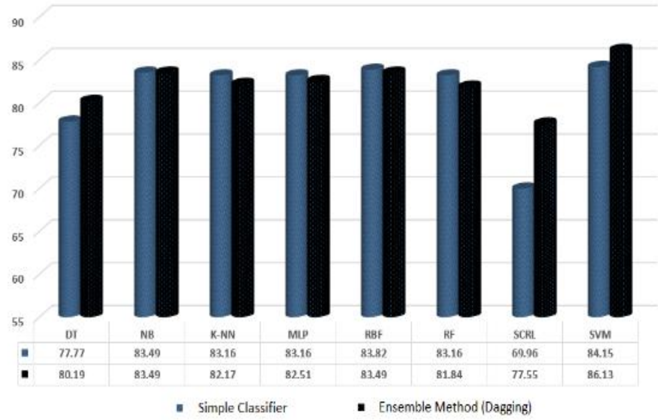

Figure 7. Accuracy of considered classifiers with Dagging method

Dagging: In this technique with the majority vote, the SVM has the better result from 84.15% to 86.13%. The other classifier results little bit changed as shown in Fig. 7.

VI. CONCLUSION

The performance of machine learning algorithms has been compared, when applied for medical dataset (i.e., Heart Diseases). This study considered classification methods of almost every family of supervised learning technique. The considered classifiers include SVM, MLP, Decision Tree, Single Conjunctive Rule Learner, Naïve Bayes, Radial Bias Function, K-Nearest Neighbor (KNN), and Random Forest. In order to obtain better classification outcomes, ensemble-learning methods including Bagging, Boosting, Decorate, Voting, Random Sub Space and Dagging have also been used to measure the accuracy of each considered classification method. The 10-Fold cross validation approach has been used for the experimental analysis, which showed that SVM outperformed in both cases: 1) Simple classification method, and 2) Classification method in combination of ensemble method. The accuracy of SVM has been obtained 86.13% as the top classifier among the considered methods. The RBF produced second higher accuracy 83.82% and third MLP as 83.5%. These results may provide insights about classification methods and their behavior, when used for healthcare data that is highly dimensional.

REFERENCES

[1] "World Health Organization. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs310.pdf

[2] G. Sannino and G. De Pietro, "A smart context-aware mobile monitoring system for heart patients," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, 2011, pp. 655-695.

[3] Cleveland Heart Disease Dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] P. Naraei, A. Abhari, and A. Sadeghian, "Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data," in *Future Technologies Conference (FTC)*, 2016, pp. 848-852

[5] R. K. Y. A. K. Shrivas, "An Effective Prediction Factors for Coronary Heart Disease using Data Mining based Classification Technique," *International Journal on Recent and Innovation Trends in Computing and Communication,* vol. 5, 2017.

[6] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in *Computer and Communication Technology (ICCCT), 2010 International Conference on*, 2010, pp. 741-745.

[7] A.Dhanasekar and R. Mala, "Analysis of Association Rule for Heart Disease Prediction from Large Datasets," *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 5, 2016.

[8] A. JARAD, R. KATKAR, A. R. SHAIKH, and A. SALVE, "Intelligent Heart Disease Prediction System With MongoDB," *future,* vol. 13, p. 14.

[9] A. Shaikh, N. Mahoto, F. Khuhawar, and M. Memon, "Performance evaluation of classification methods for heart disease dataset," *Sindh University Research Journal-SURJ (Science Series),* vol. 47, 2015.

[10] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications,* vol. 17, pp. 43-48, 2011.

[11] A. S. Ebenezer, S. J. Priya, D. Narmadha, and G. N. Sundar, "A novel scoring system for coronary artery disease risk assessment," in *Intelligent Computing and Control (I2C2), 2017 International Conference on*, 2017, pp. 1-6.

[12] R. E. Banfield, L. O. Hall, K. W. Bowyer, D. Bhadoria, W. P. Kegelmeyer, and S. Eschrich, "A comparison of ensemble creation techniques," in *International Workshop on Multiple Classifier Systems*, 2004, pp. 223-232.

.