

Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms

Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik

Department of Computer Science and Engineering, Dr. SPM International Institute of Information Technology, Naya Raipur, India

Email: divya16100@iiitnr.edu.in, anjali16100@iiitnr.edu.in, akashd16101@iiitnr.edu.in,

aditya16101@iiitnr.edu.in, srinu@iiitnr.edu.in

Abstract—New technologies like Machine learning and Big data analytics have been proven to provide promising solutions to biomedical communities, healthcare problems, and patient care. They also help in early prediction of disease by accurate interpretation of medical data. Disease management strategies can further be improved by the detection of early signs of disease. This early prediction, moreover, can be helpful in controlling the symptoms of the disease as well as the proper treatment of disease. Machine learning approaches can be used in the prediction of chronic diseases, such as kidney and heart diseases, by developing the classification models. In this paper, we propose a preprocessing extensive approach to predict Coronary Heart Diseases (CHD). The approach involves replacing null values, resampling, standardization, normalization, classification, and prediction. This work aims to predict the risk of CHD using machine learning algorithms like Random Forest, Decision Trees, and K-Nearest Neighbours. Also, a comparative study among these algorithms on the basis of prediction accuracy is performed. Further, K-fold Cross Validation is used to generate randomness in the data. These algorithms are experimented over “Framingham Heart Study” dataset, which is having 4240 records. In our experimental analysis, Random Forest, Decision Tree, and K- Nearest Neighbour achieved an accuracy of 96.8%, 92.7%, and 92.89% respectively. Therefore, by including our preprocessing steps, Random Forest classification gives more accurate results than other machine learning algorithms.

Index Terms—Random Forest, Decision Tree, K-Nearest Neighbour, Coronary Heart Disease

I. INTRODUCTION

The enormous advancements in the health departments such as smartwatches and fitness bands have led to a revolutionary change in the detection of day-to-day activities of an individual such as the heart rates, calories burned, etc. The inventions of smart devices such as Continuous Glucose Monitor (CGM), Smart Cholesterol Monitoring System has reduced the chances of occurrence of diseases. These devices in our daily life keep track of the activities and assist in decisions making for healthcare. Even after these advancements in health departments, people are unaware of the risks and symptoms associated with chronic diseases. Therefore, the prediction of such diseases is now a major concern not only for individuals but for mankind. Among all the human-related diseases, chronic diseases are the riskiest and major health problems, according to Lancet Study on Global Burden of Disease Study 2013 [1]. Moreover, in a survey by McKinsey [2], it was found that in several countries like China, chronic diseases are the major cause of death and

in the US, the government spends annually around 2.7 trillion USD for the treatment of chronic diseases.

The continuous technological improvements helped the researchers to develop new methodologies based on Artificial intelligence and Machine learning. The rising health issues have also lead to an increase in the generation of big data, and for utilizing this Big data it is required to develop an automatic computer-based system that can be used to predict those diseases by deploying machine learning algorithms which can work efficiently for various challenges occur in the datasets.

Along with the other chronic healthcare problems such as obesity, smoking, diabetes, etc., cardiovascular diseases (CVDs) are also found to be a huge risk factor. Population aging, especially in developed countries, is highly correlated with the CVDs, means the old age people are more prone to CVDs [3]. According to the World Health Organisation report in 2014 [4], cardiovascular diseases including heart attacks, chronic heart failure, cardiac arrest, etc. led to the death of 17.5 million people. Most heart diseases cannot be detected by primitive Electrocardiography (ECG) process [6], so for that many preventive sensors or devices are invented such as Phonocardiogram (PCG), Electromyogram (EGM), etc.

In this work, we analyze and estimate the use of different machine learning algorithms in the prediction of coronary heart disease by combining all the attributes in the dataset to develop the classification models. Random Forest, K-Nearest Neighbours, and Decision Tree classification models for coronary heart disease risk prediction are developed. We also apply K-fold Cross-Validation for the algorithms. Results show that these models can efficiently predict the risk of heart disease, to conclude whether an individual is prone to suffer from CHD within a span in 10 years.

Rest of the paper is organized as follows: Section II describes the related work. In Section III, proposed work is described including dataset description, preprocessing, analytics and modeling. Section IV includes experimental analysis which describes the performance measurement parameters for different algorithms. After that Results in Section V which includes performance comparison of the algorithms and at last paper is concluded in Section VI.

II. RELATED WORK

Different researchers practice different ways of involving machine learning and data-mining approaches to solve health-related issues. They have used various approaches to classify and predict chronic diseases.

Data mining techniques were proved to efficiently predict chronic kidney disease. Weka [5] is a user interface which provides data processing cycle such as data preprocessing, classification and other data mining technique to a user. It has a large collection of data to which various algorithm can be practiced on it. This tool illustrates out that Random Forest gives the best result among various algorithm like Naive Bayes, J48, etc.

Heart Sounds can also be used for the detection of chronic heart failure [6]. The method involves filtering of audio signals, segmentation for feature extraction, and machine learning. It also described the stacking process of ML algorithms, having three phases: segment based ML phase, recording based feature extraction phase, and recording based ML phase.

Different authors have adopted different approaches for developing classification models. Data can also be captured by ADL (activities of daily living) with wearables [7]. The proposed framework used this data and implemented supervised and unsupervised machine learning algorithms and batch level processing including cohort segmentation. Different ML algorithms like KNN, SVM, Random Forest, etc. were implemented on two different datasets and compared based on accuracy and model building time. Maximum accuracy in NHANES and Framingham Datasets were obtained in Random Forest and SVM modeling respectively.

In [8], NHANES and FHS datasets were used. In the NHANES dataset, feature selection methods lead to an improvement in the performance based on information theory ranking and in FHS dataset, this was done by grouping based on subdivision filter variable method. The model, also, showed the trade-off between accuracy and execution time. Random Forest and KNN gave better results in confusion matrix and classification accuracy but they were unable to satisfy the boundaries of creation time. Meanwhile, the decision tree showed good results in both aspects.

In [9], the study of the effect of class imbalance in data on performance for multilayer perceptron was carried out. Different learning rates were used to evaluate the performance of multilayer perceptron and analyze the dataset by three sampling algorithms, among which the Resample method provided the best accuracy results than others. Also, a comparative study was done based on accuracy metrics and execution time. Spread Sub Sample algorithm has the least execution time.

The target was to see the importance of features on the classification result [10]. Preprocessing and normalization were performed over the dataset. Next, to measure the correlation between features, the correlation matrix was obtained. Further, the classification was carried out in three stages: first, L1-based feature selection; second, AUC (Area Under Curve) based comparison of the performance of five algorithms, namely,

Decision Tree, Neural Network, Logistic Regression, Support Vector Machine, and Naive Bayes over both normalized and original data were presented; third, performance of classifiers was compared based on sensitivity, accuracy, specificity, and AUC. Except for Neural Network, other classifiers showed almost the same results.

In [17] an novel approach is proposed that uses analytic time-frequency flexible wavelet transform (ATFFWT) and fractal dimension (FD) to detect epileptic seizures. By using ATFFWT the EEG signals are decomposed into subbands and then FD is calculated on each subbands. For training the model these FDs are given as input to Least Squares SVM. At last, cross validation is performed to deal with model overfitting.

III. PROPOSED WORK

This section illustrates various resources and approaches that are used in this work. Primarily, the description of dataset is provided to understand how to work on it, followed by the preprocessing steps involved. Finally, the internal working and understanding of the analytical models used are explained.

A. Dataset Description

We have practiced a dataset which is a subset of Framingham Heart Study (FHS) dataset, it is made publicly available through Framingham Heart Institute [7]. The available section of FHS dataset used in this paper contains records of 4240 participants. The dataset is generated by long term study on a population of Framingham, Massachusetts. The study is based on the cause and origin that lead to cardiovascular heart disease and it comes under one of the best public health disease management domain [8]. The Framingham Heart study focused mainly to retrieve the risk factors that have an effect on the health of a person in perceiving a coronary heart disease. The dataset contains 16 different features that affect Coronary Heart Disease.

TABLE I
ATTRIBUTES OF THE DATASET AND THEIR INTERPRETATION

Attribute	Interpretation
gender	Female : 0; Male : 1
age	Age at the examination time
education	1: high school 2: high school or GED 3: college or vocational school 4: college
currentSmoker	0 = nonsmoker; 1 = smoker
diabetes	0 = No; 1 = Yes
totChol	Total cholesterol inside patient's body (mg/dL)
sysBP	Systolic Blood Pressure (mmHg)
diasBP	Diastolic Blood Pressure (mmHg)
cigsPerDay	Number of cigarettes smoked per day (average)
BPMeds	Is the person on BP medicines
prevalentStroke	If the person had any prevalent stroke
prevalentHyp	Any beneath prevalent
BMI	Body Mass Index : Weight (kg) /Height(meter-squared)
heartRate	Beats/Min (Ventricular)
glucose	Amount of glucode in mg/dL
TenYearCHD	Risk of developing CHD (Yes : 1; No: 0)

Table I provides an interpretation of different attributes in the dataset. The dataset consists of many inconsistent and discrepant values and can lead to incorrect results. Thus, proper care needs to be taken while treating these values for better performance. Therefore, the dataset is preprocessed before model creation.

B. Preprocessing

Preprocessing is a method to obtain complete, consistent, interpretable data. The data quality affects the mining results that are obtained using machine learning algorithms. Quality data results in a quality decision. Therefore, the FHS dataset is integrated using the following preprocessing steps.

- Irrelevant features can decrease the performance of the model and reduces the learning rate. Therefore, feature selection plays a major role in preprocessing in which those features are selected that contributes the most in predicting the desired results. In the FHS dataset, using an automatic feature selection would have eliminated important features as well. Therefore, an analytical approach gives better performance.
- The mean is the most probable value that tends to occur in any attribute. Also, mean preserves the extremes of an attribute, therefore, missing values in the FHS dataset are replaced by the attribute mean, as shown in equation (1).

$$Attribute\ Mean = \frac{\sum_{i=0}^l (attribute\ value)_i}{l} \quad (1)$$

where, l is the total number of values in an attribute

- Class imbalance of dataset is a major problem in data mining applications. Most of the machine learning algorithms fail to perform well on a dataset where classes are imbalanced [9].
- Sampling is an effective method to balance an imbalanced dataset. Sampling is of two types: oversampling and undersampling. Undersampling involves removing instances from the majority class to balance the class distribution. Oversampling involves replicating instances from minority class to balance the class distribution. Fig 1. illustrates the resampling mechanism.
- The target class in the dataset predicts the risk of coronary heart disease (CHD). The instances with the risk of an individuals those are more likely to suffer from CHD is 15.2% (644 out of 4240 entries) and that of individuals those are not suffering from CHD is 84.8% (3596 out of 4240 entries). In order to balance this class distribution we used random oversampling to replicate the instances in the minority class, that is, individuals suffering from CHD.

Fig 2 depicts the graphical representation of the steps in sequential order that are used in the proposed work.

C. Analytics and Modeling

This section explains the supervised algorithms of Machine learning that are used in this work. It briefs about the analytical approach and internal working of Random Forest, Decision

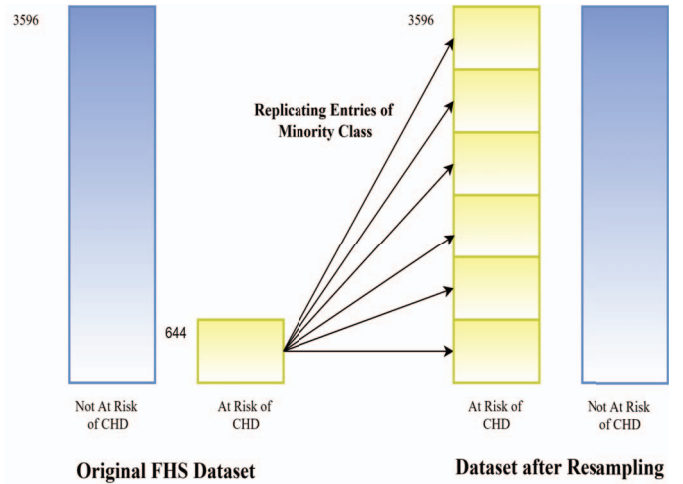


Fig. 1. Resampling of data set

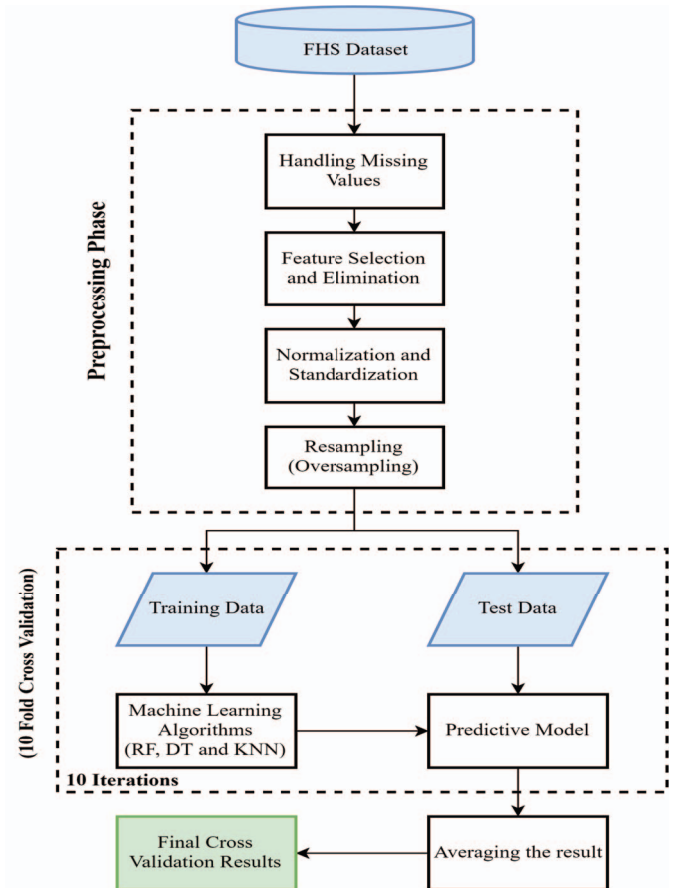


Fig. 2. Flowchart of proposed work.

Tree and KNN to create a prediction model. Though, there are also other such powerful machine learning algorithms like Convolutional Neural Network (CNN) and Naive Bayes etc. But these algorithms didn't fit well as they gave lesser accuracy and performance measures comparatively in our experimental analysis. Therefore, we choose Random Forest, Decision Tree and KNN over them.

- **Random forest (RF)** is a supervised machine learning algorithm. As the name predicts, it is a forest of randomly generated decision trees. It basically uses an approach bagging, where various learning models are combined to improve the overall results. To perform the bagging operation, it produces manifold decision trees and synthesizes them together to obtain a refined result. It is one of the finest machine learning algorithms. It uses a random subset of features by splitting a node to obtain the best feature that contributes the most to build the model. The result even is increased further by adding random threshold values to each feature. Random Forest algorithm is also used to score the features. On the basis of how much impurity a feature adds to the model, it decides the relative feature importance. Also, RF is robust to outlier values.
- **Decision Tree (DT)** is one of the simplest algorithms, yet most effective and useful. It is a tree which comprises of three nodes, first is the chance node, second is decision node and at the last end node. The chance node shows the probable outcomes of a particular node whereas the decision node is a node on which a decision is to be made based on the outcome. The end node is the last node of the tree which gives the final result of a path. Decision Tree starts from a node known as the Root node and gets split off into various branches or nodes. The splitting of this root node is done on the basis of probabilities. Each node extracts some information about features of data and each link represents a decision rule taken on the nodes. The Tree is mapped or drawn on two bases: Gini index and entropy rule. It is one of the simplest, easy to understand and finest predictive model.
- **K-Nearest Neighbour (KNN)** is also a supervised classification algorithm. It predicts the target class on the basis of how similar that particular data is from other provided training data labels to the model. This can be understood as, the characteristic(features) of that data, whose target label needs to be predicted, is compared with features of existing data (except the target class). The resemblance of this data with any class decides which class it will belong to. KNN uses the approach of comparing unclassified data with classified data by calculating the distance between features of data points (using Euclidean distance, Manhattan distance, etc.). First, the model collects unclassified data. It then calculates the distance of each feature of that data from features of classified data. By doing so, it selects K small distances. Then, it counts the class that appears the most among these K observations. Finally, it

classifies that data with the class that appeared the most.

IV. EXPERIMENTAL ANALYSIS

A. Parameters Used

Performance evaluation of the proposed work is done based on the following measures:

Confusion Matrix is a matrix that is used to evaluate the performance of a model. The four terms associated with the confusion matrix which is used to determine the performance matrices are :

True Positive (TP): An outcome when the positive class is correctly predicted by the model

True Negative (TN): An outcome when the negative class is correctly predicted by the model

False Positive (FP): An outcome when the positive class is incorrectly predicted by the model

False Negative (FN): An outcome when the negative class is incorrectly predicted by the model

Accuracy: is the ratio of a number of correct predictions given by the model to the total number of instances.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (2)$$

Precision: Precision in this work measures the proportion of individuals predicted to be at risk of developing CHD and had a risk of developing CHD.

$$Precision = \frac{TP}{(TP + FP)} \quad (3)$$

Recall/Sensitivity: Recall, in this work, measures the proportion of individuals that were at a risk of developing CHD and were predicted by the algorithm to be at risk of developing CHD.

$$Recall = \frac{TP}{(TP + FN)} \quad (4)$$

Specificity: Specificity here measures the proportion of individuals who were not at risk of developing CHD and were predicted by the algorithm to be not at risk of developing CHD.

$$Specificity = \frac{TN}{(TN + FP)} \quad (5)$$

F1 Score: F1 Score is the harmonic mean of precision and recall.

$$F1Score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \quad (6)$$

ROC(Receiver Operator Characteristic): It is a probability curve indicating the capability of a model to distinguish

between classes. The ROC curve shows trade-off between True Positive Rate (TPR) and the False Positive Rate (FPR). According to [13], AUC (Area Under the Curve) closer to 1 would be able to perfectly differentiate the two classes in the case of binary classification. Therefore, AUC closer to 1 is better predictive measure.

$$TPR = \frac{TP}{(TP + FN)} \quad (7)$$

$$FPR = \frac{FP}{(FP + TN)} \quad (8)$$

B. Results

In the proposed work, 10-Fold cross-validation is performed for the machine learning algorithms like RF, DT, and KNN that are used for analysis. Different performance measures as mentioned in the parameters section are calculated and compared.

The accuracy and AUC for RF, DT and KNN for the 10-folds iterations are illustrated in Table II. It can be observed that the average accuracy and AUC of RF classifier come out to be more than that of DT and KNN.

TABLE II
10-FOLD RESULTS

Fold No.	RF		DT		KNN	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
1	96.80%	1	93.33%	0.92	91.52%	0.91
2	97.36%	0.99	91.38%	0.92	92.08%	0.91
3	94.86%	0.99	91.94%	0.92	93.05%	0.91
4	96.67%	0.99	91.94%	0.92	93.47%	0.91
5	96.38%	0.99	92.22%	0.91	92.91%	0.91
6	96.52%	0.99	92.08%	0.92	92.63%	0.91
7	97.49%	0.99	94.15%	0.92	94.15%	0.91
8	97.63%	0.99	92.61%	0.91	93.03%	0.91
9	97.21%	0.99	92.2%	0.92	93.03%	0.91
10	97.07%	0.99	92.61%	0.92	92.2%	0.91

The ROC curves for 10 folds of RF classifier with a mean AUC of 0.99 is shown in Fig 3. AUC closer to 1 depicts a better model [13]. A model is better if it predicts true more often, that is, TPR is higher. Therefore, curve passing through top left corner gives a better predictive model as in case of RF.

Table III shows the mean accuracy and mean AUC for the 10-fold cross-validation scores of RF, DT and KNN. It also compares the execution time taken by each of the algorithms for model creation and prediction. The observation states that RF achieves the maximum accuracy, that is 96.8%, among RF, DT, and KNN with AUC 0.99. The execution time of RF is

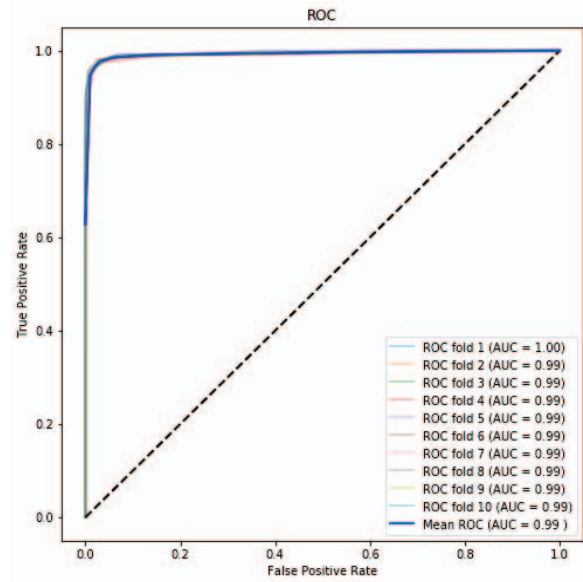


Fig. 3. 10- Fold RF ROC Curve

1.39 seconds falling between DT and KNN. DT has the least execution time of 0.81 seconds, whereas, that of KNN is as high as 1.9 seconds.

TABLE III
PERFORMANCE STATISTICS

Algorithm	Execution Time (Seconds)	Mean Accuracy	Mean AUC
RF	1.3969	96.80%	0.99
DT	0.8138	92.45%	0.92
KNN	1.9029	92.81%	0.91

Fig. 4 compares the ROC curves for RF, DT and KNN giving AUC of 0.99 for RF, 0.92 for DT and 0.91 for KNN. The average AUC of RF is more close to 1, and hence RF is more suitable for the prediction model than DT and KNN since its AUC is more closer to 1.

Table IV shows the results considering different performance measures such as accuracy, precision, recall, specificity, and F1 Score. While the recall is taken into consideration, RF gives recall of 94.4% which is more than that of DT and KNN. Thus, RF outperforms the other algorithms to predict CHD risk among individuals.

TABLE IV
PERFORMANCE STATISTICS

Algorithm	Accuracy	Precision	Recall	Specificity	F1 Score
RF	96.71%	98.94%	94.4%	99%	96.61%
DT	92.1%	98.57%	85.33%	98.78%	91.47%
KNN	91.49%	98.42%	84.21%	98.67%	90.76%

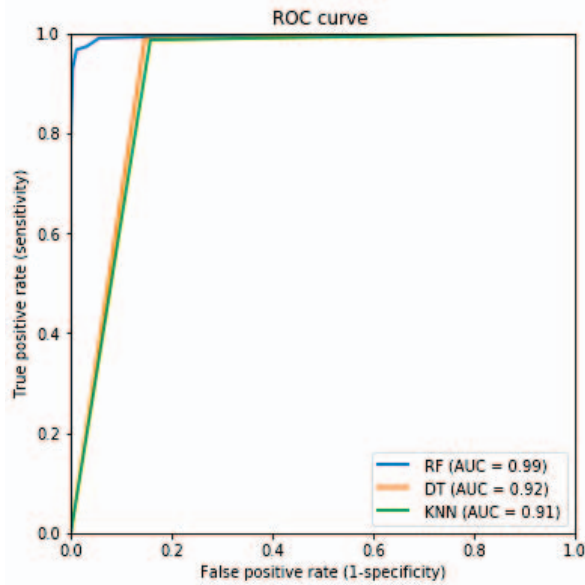


Fig. 4. Comparison ROC Curve

TABLE V
RESULT COMPARISON

Algorithm	Previous work [7]		Proposed work	
	Time Taken (sec)	Accuracy (%)	Time Taken (sec)	Accuracy (%)
Random Forest	2180	90.1	1.3969	96.80
Decision Tree	77.4	90	0.8138	92.45
KNN	88.8	90.1	1.9029	92.81

Table V highlights the comparison of the work done in [7] and the proposed work based on the parameters, time taken by the algorithms and accuracy. The accuracy is more and the time taken by the algorithms is very less in the proposed work in all the three classification algorithms. Moreover, the accuracy of the algorithms in previous work are nearly same while in proposed work, Random Forest has shown much higher accuracy. Therefore, the Random Forest machine learning algorithm outperforms the results in [7] and [8] on the same dataset.

V. CONCLUSION AND FUTURE WORK

We propose a preprocessing extensive work where Random Forest is the most compatible contender for prediction model and gives the highest performance measure among K- Nearest Neighbour and Decision Tree. The accuracy, recall, precision, specificity and F1 score of RF on the proposed work are 96.71%, 98.74%, 94.4%, 99%, 96.61% respectively, under execution time of 1.3969 seconds. The Decision Tree, however, gives lesser performance set against that of Random Forest though in quite lesser time (0.8138). The execution time for K- Nearest Neighbour is the highest among all, however, the performance measures are quite similar to that of the Decision

Tree. Thus, in an environment similar to that of the used dataset, if all the features are preprocessed such that they acquire normal distribution, Random Forest is a good selection to obtain a robust prediction model. And, such models provide a valuable assistant to the society for health care management domain.

Further, as an extension to this work, a more real-time and bigger dataset is required to obtain a better training model. Also, an emphasis on refining the preprocessing further will give veracious outcomes.

REFERENCES

- [1] Huse, Hettiarachchi, Gearon E, Nichols M, Allender S, Peeters A, "Obesity In Australia Modi", June 2015.
- [2] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, "The Big Data Revolution in Healthcare: Accelerating Value and Innovation", USA: Center for US Health System Reform Business Technology Office, 2016.
- [3] D. Kumar, "Automatic heart sound analysis for cardiovascular disease assessment," Ph.D. dissertation, University of Coimbra, 2014.
- [4] S. Mendis et al., "Global status report on noncommunicable diseases 2014", World Health Organization, 2014.
- [5] Tilakachuri Balakrishna, B. Narendra, Mooray Harika Reddy, Damarapati Jayasri, "Diagnosis of Chronic Kidney Disease Using Random Forest Classification Technique", Helix Vol. 7(1): pp.873-877, 2017.
- [6] Martin Gjoreski; Monika Simjanoska; Anton Gradisek "Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers", IEEE 13th International Conference on Intelligent Environments, pp. 14-19, 2017.
- [7] Nitten S. Rajliwall; Girija Chetty; Rachel Davey, "Chronic disease risk monitoring based on an innovative predictive modeling framework", IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-8, 2017.
- [8] Rajliwall, Nitten S., Rachel Davey, and Girija Chetty, "Machine learning based models for Cardiovascular risk prediction", IEEE International Conference on Machine Learning and Data Engineering (ICMLDE), 2018.
- [9] Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron", IEEE 41st Annual Computer Software and Applications Conference, pp. 193-198, 2017.
- [10] Maryam Soltanpour Gharibdousti, Kamran Azimi "Prediction of Chronic Kidney Disease Using Data Mining Techniques", Proceedings of the Industrial and Systems Engineering Conference, 2017.
- [11] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, And Lin Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", IEEE Access Volume 5, pp. 8869 - 8879, 2017.
- [12] M. A. Jabbar, Shirina Samreen, "Heart disease prediction system based on Hidden Nave Bayes classifier", IEEE International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1-5, 2016.
- [13] Hajian-Tilaki, Karimollah. "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation", in Caspian Journal of Internal Medicine, Vol.4, pp. 627-635, 2013.
- [14] Gunarathne W.H.S.D, Perera K.D.M, Kahandawaarachchi K.A.D.C.P, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease", IEEE 17th International Conference on Bioinformatics and Bioengineering, pp. 291-296, 2017.
- [15] Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 15, No. 3, pp. 861-868, 2018.
- [16] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics", IEEE Congress on Evolutionary Computation (CEC), pp. 1-9, 2018.
- [17] M. Sharma, Tan, RS. Acharya, U.R. A new method to identify coronary artery disease with ECG signals and time-Frequency concentrated anti-symmetric biorthogonal wavelet filter bank., Pattern Recognition Letters, Vol. 125, pp. 235-240(2019).