# Heart Disease Prediction Method using Hybrid Classifier

**Saloni Kapoor, Ashwinder Tanwar**

***Abstract*: The data mining is the approach which can extract useful information from the data. The following research work that has been described is related to the heart disease prediction. The prediction analysis is the approach which can predict future possibilities based on the current information. For the heart disease prediction the classifier that is designed in this research work is hybrid classifier. The hybrid classifier is combination of random forest and decision tree classifier. Moreover, the heart disease prediction technique has three steps which are data pre-processing, feature extraction and classification. In this paper, random forest classifier is applied for the feature extraction and decision tree classifier is applied for the generation of prediction results. However, random forest classifier will extract the information and decision tree will generate final classifier result. We have proposed a hybrid model that has been implemented in python. Moreover, the results are compared with Support Vector Machine (SVM) and K-Nearest Neighbor classifier (KNN).***

***Keywords* : Hybrid classifier, Random Forest classifier, Decision Tree Classifier, Heart Disease Prediction.***

## I. INTRODUCTION

The process through which hidden and unknown patterns are identified is called data mining. In order to extract the hidden patterns and relationships from huge databases, the machine learning algorithms, database technology and statistical analysis are combined with each other [1]. The data mining approach, gain the popularity in health care with the need of efficient information extraction. The competent function of heart is the essence of life [2]. The remaining body parts of human are also affected if the functioning of heart is improper. The functioning of heart is affected due to various factors like blood pressure, sugar level etc. The risk of heart disease can be increased due to several factors. Today, most of the deaths are being caused due to heart diseases [3]. An accurate result of diseases can be achieved by applying prediction. As the name suggests, the relationships among independent and dependent variables are discovered through prediction.

A historical heart disease database is used to identify and extract the hidden knowledge related to heart disease [4]. The heart disease can be diagnosed by answering complex queries and therefore the intelligent clinical decisions can be made by the practitioners [5]. The historical records of patients who have heart disease can be analyzed with the approach of classification [6]. The probability of the test samples are calculated and test sample will be classified into closed class [7]. The parameter setting or domain knowledge is not required here and it is easy to handle the high dimensional data [8]. The results that are easy to be read and interpreted are generated here. Only these classifiers provide the feature of accessing the detailed profiles of patients. Including the internal nodes [9], branches and leaf nodes, a tree like structure is created in decision tree. Here, the attribute value is denoted by branches, a test on an attribute is denoted by each branch and the predicted classes are represented by a leaf node. Highly accurate results are known to be achieved using another classifier named neural networks [10]. It is possible to train neural network with heart diseases database using a feed forward neural network model with variable learning rate along with back propagation learning algorithm that includes momentum [11].

A maximum margin classification algorithm which is based on the statistical learning theory is known as Support Vector Machine (SVM). The non-linear as well as linear data is classified through this classifier. A non-linear transformation mechanism is used to convert the training data into non-dimensional data. Further, the transformed data is divided into two various classes using the best hyper-plane searched by SVM [12].A voting based classified is the combination of multiple classifiers which are executed parallel and output of each classifier can be exploited [13]. Various machine learning classifiers are first combined and then a majority vote is used for predicting the class labels. With the help of this integration, the weaknesses of individual classifiers can be balanced and a set of equally well performing model is generated here.

This classifier is performed in two different manners. The majority of class labels that are predicted by every individual classifier are represented by the predicted class label for a specific sample in the majority/hard voting type [14]. However, the class label is returned as argmax of the sum of predicted probabilities in case of soft voting. Weights parameter is used to assign particular weights to each classifier. For each classifier, the predicted class probabilities are gathered, multiplied by the class weight and then averaged when the weights are available [15]. Then, from the class label that has the highest average probability, the final class label is derived.

The predication analysis is the approach which is applied to predict future possibilities based on historical information. The predication analysis techniques have the three phases which are preprocessing,

**Revised Manuscript Received on July 10, 2019.**

**Saloni Kapoor**, University of Engineering, Chandigarh University, Mohali, India.
**Ashwinder Tanwar**, University of Engineering, Chandigarh University, Mohali, India.

feature extraction and classification. The predication analysis approaches are broadly classified into supervised and unsupervised learning techniques. This research work is related to supervised learning for the heart disease predication analysis. In the previous research work, the technique of SVM classification is applied for the heart disease predication. It is analyzed that SVM classifier has high complexity due to which execution time is high and accuracy is less for the prediction analysis.

## II. II. LITERATURE REVIEW

Somayeh Nazari, et al. (2018) proposed [16] a hybrid method to predict the possibility of introducing heart disease in an individual within Clinical Decision Support System (CDSS). The proposed method was designed by integrating Fuzzy Analytic Hierarchy Process and Fuzzy Inference System. By eliminating the previously emerging issues, the proposed method designed a very accurate CDSS. It was seen through the evaluations that large number of resources and huge costs were saved by using the proposed CDSS.

Sarath Babu et .al (2017) introduced [17] a technique called data mining, which is the mechanism of discovering new set of information from huge amount of data. It is used to analyze large volume of data and the patterns were extracted to convert the irrelevant information into useful information. This collected information is fed into several classifiers each of which performs some specific tasks. These techniques are used to predict heart diseases at their early stage. It shows very effective performance in order to achieve the correct and perfect diagnose for the heart related diseases. There are certain advantages of this approach such as the diseases can be predicted at their very initial stages and can be diagnosed correctly and properly on time. Therefore, the researcher concluded that, this method is very useful in preventing heart related issues.

Tülay KarayÕlan et.al (2017) proposed back propagation algorithm for the prediction of heart diseases with the help of artificial neural networks [18]. It has some clinical features in which neural networks are used as input and is trained along with this proposed back propagation technique. It can predict the heart related diseases with an accuracy of 95%. The already proposed methods were not sufficient for the early prediction of heart diseases. The advancement of technology will leads to the use of machine learning techniques for the prediction of cardiac diseases at their initial stages. Therefore, the researcher draws the conclusion that the proposed approach has almost 100% accuracy in prediction heart related diseases at their early stages. It gives better results in comparison to the other techniques.

Tahira Mahboob et.al (2017) introduced various learning practices which assist the detection of innumerable heart diseases [19]. As the cardiac diseases treatment is very expensive and unaffordable to any normal individual so, these types of advanced technology are developed to overcome this problem. These techniques are also useful in early stage predictions. It avoids any other future sufferings by making slight changes in daily routine. Hence, the author concludes that the predicted approach has several advantages and is very useful.

Procheta Nag et.al (2017) proposed [20] a very effective technique which is very useful in the prediction of heart diseases at the initial stage. The researcher has developed a prototype called Acute Myocardial Infarction (AMI). Heart attack having various symptoms like chest pain, breathing problem, palpitation, vomiting and continuous sweating. Therefore, the researcher draws the conclusion that the advancement of computer technology in medical and health region provides useful aids and people are becoming more dependent on these technologies. The results of data mining are very beneficial and are used for the better assistance to many physicians as lot of data is related to diseases.

Priyanga et.al (2017) proposed an intelligent and efficient technique called naïve bayes for the prediction of heart related issues [21]. The data is collected from the given attributes and then they are implemented as web based applications. The methods which are proposed by the previous authors do not show effective results in terms various parameters. The existing schemes are not able to detect various type of disease at the initial stage. Therefore, the researcher concludes that the approach classified had low cost and is extensively tested by experienced cardiologists. The research mainly focuses on detection of heart disease using UCI dataset.

## III. PROPOSED METHODOLOGY

This research work is related to heart disease prediction. The designed model is based on the hybrid model which is combination to two classifiers which are random forest and decision tree. The random forest classifier is used for the feature extraction and decision tree is used for the classification. The random forest classifier works like the base classifier and decision classifier works like the meta classifier. The proposed methodology has the following steps:-

1. Input dataset and pre-processing:- In the first phase, the dataset is collected from the UCI repository. The dataset is pre-processed to remove missing and redundant values. The collected dataset has the balance data which can be processed easily for the heart disease prediction

2. Feature Extraction:- In the second phase, the features of the dataset are extracted for the classification. In the feature extraction phase, the relationship is established between the target set and attribute set. The technique of random forest classifier is applied in this phase. The random forest classifier will be the base classifier for the feature extraction. An algorithm designed to build a predictor ensemble using a set of decision trees that grow in randomly chosen subspaces of data is called random forest algorithm. It is easy and fast to implement this algorithm. Highly accurate predictions are generated by it and very large number of input variables can be handled by it.

A small group of input coordinates are chosen randomly at each node for splitting to generate a tree in the collection initially. Also, the features within the training set which calculate the best slip can be used secondly for generating the tree.

For maximizing the size of tree

without pruning, CART methodology is used. In order to resample the training data set every time a new individual tree is grown, the subspace randomization mechanism is blended with bagging. The randomized base regression trees {r_n (x,Θ_m,D_(n)),m≥1} collectively generate a random forest. Here, for a randomized variableΘ, the i.i.d outputs are denoted by Θ_1,Θ_2,…. The aggregated regression estimate is generated by combining these random trees.

$$\bar{r_n}(X, D_n) = E_\Theta[r_n(X, \Theta, D_n)] \quad [14]$$

Here, the expectation with respect to random parameter on X and data set D_n is denoted by E_Θ. The estimate in the sample omits the dependency and instead of (r_n )¯(X,D_n ) one can write (r_n )¯(X) as well.

3. Model Building and Prediction Analysis: - In the last phase, the input dataset will be divided into training and test phase. The training set will be more than 50 percent and rest of the part will be the test set. The dataset will be trained using the decision tree classification and final prediction is generated of the test set. The decision is hierarchical data structures which represents the data using a divide and conquer strategy is called decision tree. The categorical labels are used instead of non-parametric classification for discussing the decision trees. They can also be used to perform regression. Determining the labels for new examples is the aim of decision tree within classification. The instances are represented as feature vectors in the decision tree classifiers. The tests for feature values are denoted as nodes, the labels as leaves and for each value of feature at every node, one branch must be available. Entropy is used as a measure to define the information gain in this classifier. The impurity level of an arbitrary collection of examples is defined by entropy. For instance, if a collection S is considered which includes both positive and negative examples of any target set, the entropy is defined as:

$$Entropy(S) \equiv -p_\oplus log_2 p_\oplus - p_\ominus log_2 p_\ominus \quad [11]$$

## IV. RESULT AND DISCUSSION

The hybrid model is designed in this work for the heart disease prediction. The hybrid classification model is the combination of random forest and decision tree classifiers. The data is collected from UCI repository and description of the data is given in the table 1. The performance of the proposed model is analyzed in terms of accuracy and execution time. The results of the hybrid model are compared with the SVM and KNN classifier for the result validation.

Table 1: Dataset Description

| | | | | | |
|---|---|---|---|---|---|
| Data Set Characteristics: | Multivariate | Number of Instances: | 303 | Area: | Life |
| Attribute Characteristics: | Categorical, Integer, Real | Number of Attributes: | 75 | Date Donated | 1988-07-01 |
| Associated Tasks: | Classification | Missing Values? | Yes | Number of Web Hits: | 890323 |

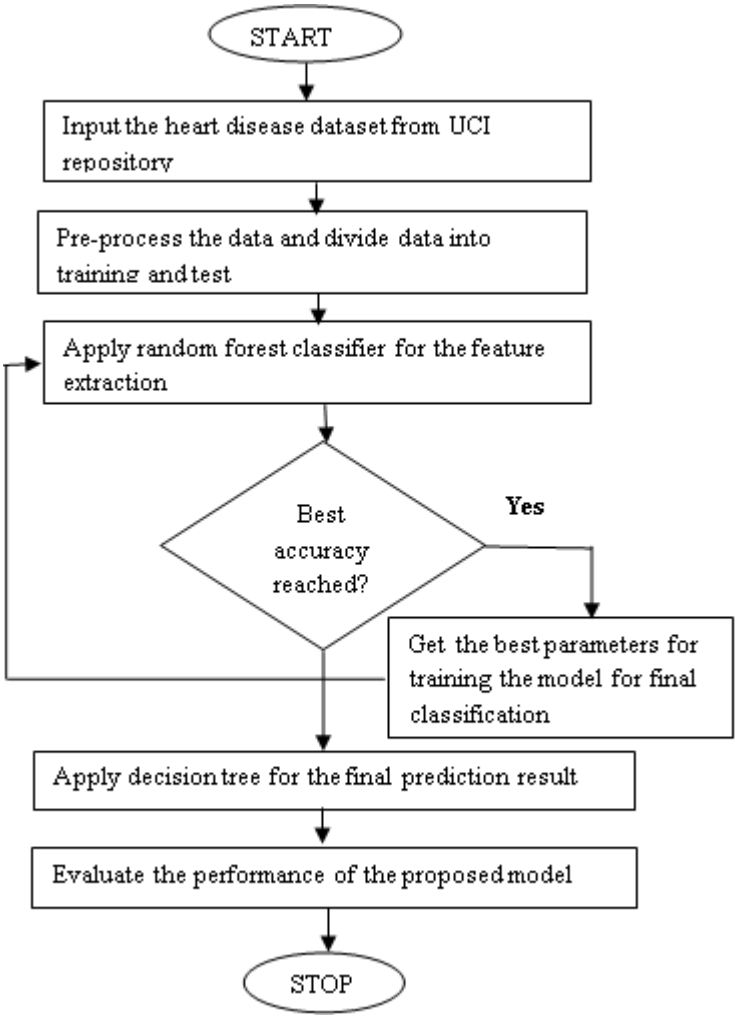


Fig 1: Proposed Methodology

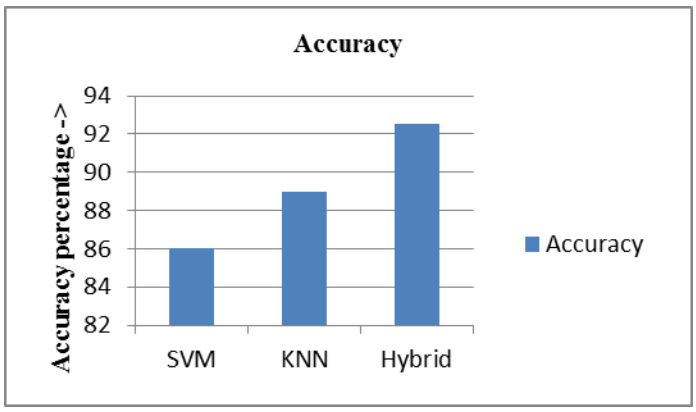


Fig 2: Accuracy Comparison

As shown in figure 2, the accuracy of SVM, KNN and

hybrid models are compared for the performance analysis. It is analyzed that hybrid model has maximum accuracy which is approximately 92 percent. The hybrid classifier is combination of random forest and decision tree classification methods.

As shown in figure 3, the execution time of the hybrid classifier is least as compared to SVM and KNN classifier. The SVM and KNN classifiers are complex as compared to hybrid classifier due to which hybrid classifier has least execution time.

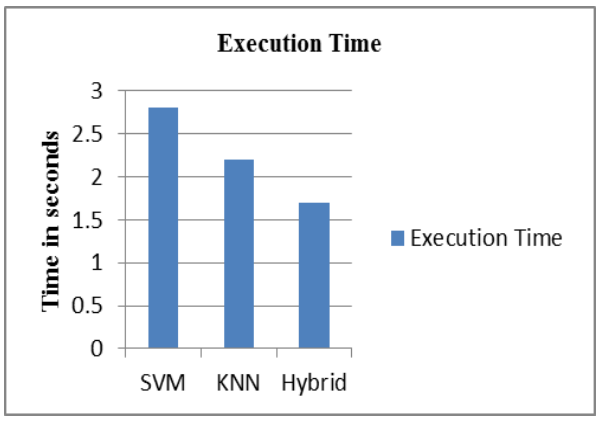


Fig 3: Execution time

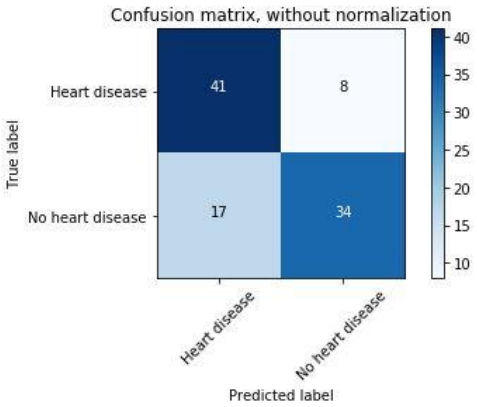


Fig 4: Confusion Matrix

As shown in figure 4, the confusion matrix of the proposed technique is drawn which shows the true and predicted value. The x axis shows the prediction value and y axis shows the true value for heart disease prediction.

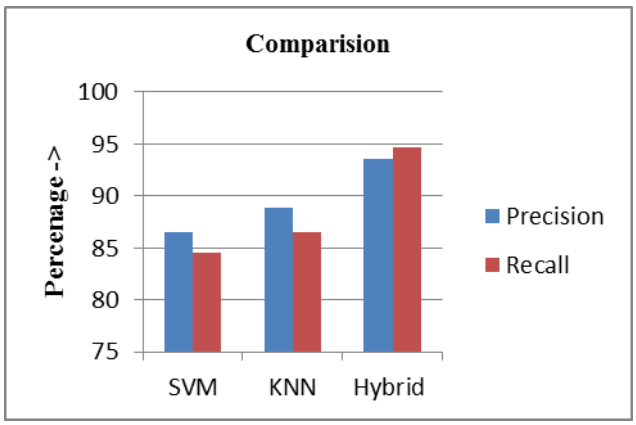


Fig 5: Precision-recall Comparison

As shown in figure 5, the precision and recall values are compared of SVM, KNN and hybrid classifiers. It is analyzed that hybrid classifier has maximum precision-recall values as compared to SVM and KNN.

## V. CONCLUSION

In this work, it is concluded that prediction analysis is the approach which predict future possibilities based on current data. The hybrid model designed in this work is the combination of random forest and decision tree classifier. The proposed model is implemented in python and results are validated by comparing it with SVM and KNN classifier. The hybrid classifier has maximum accuracy around 92.4 percent as compared to SVM and KNN. In future, the clustering algorithm will be applied with the hybrid classifier method for the data division.

## REFERENCES


1. Nidhi Bhatla, Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", 2012, IJERT, Vol 1, Issue 8
2. Syed Umar Amin, Kavita Agarwal, Rizwan Beg, "Genetic Neural Network based Data Mining in Prediction of Heart Disease using Risk Factors", 2013, IEEE Conference on Information & Communication Technologies
3. A H Chen, S Y Huang, P S Hong, C H Cheng, E J Lin, "HDPS: Heart Disease Prediction System", 2011, IEEE, Computing in Cardiology
4. M. Akhil Jabbar, B. L Deekshatulu, Priti Chandra, "Heart Disease Prediction using Lazy Associative Classification", 2013, IEEE, International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)
5. Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", IJCA, Volume 47– No.10, June 2012.
6. P. Bhandari, S. Yadav, S. Mote, D.Rankhambe, "Predictive System for Medical Diagnosis with Expertise Analysis", IJESC, Vol. 6, pp. 4652-4656, 2016.
7. Nishara Banu, Gomathy, "Disease Forecasting System using Data Mining Methods", IEEE Transaction on Intelligent Computing Applications, 2014.
8. Shadab Adam Pattekari and Asma Parveen, "Prediction System For Heart Disease Using NaiveBayes", International Journal of Advanced Computational and Mathematical Sciences, ISSN 2230 - 9624, Vol 3, Issue 3, pp 290-294, 2012.
9. Nilakshi P. Waghulde, Nilima P. Patil, "Genetic Neural Approach for Heart Disease Prediction", International Journal of Advanced Computer Research (ISSN (print): 2249-7277, Vol 4 Number-3 Issue-Sept 2014.
10. Upasana Juneja et. al., "Multi Parametric Approach Using Fuzzification on Heart Disease Analysis", IJESRT, Juneja et al., 3(5) ISSN: 2277-9655, Page No.492-497,2014.
11. Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq,"Performance of Data MiningTechniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease", International Journal of Database Management System, Vol.8, No.3, 2016.
12. B. Umadevi, D.Sundar, Dr.P.Alli, "A Study on Stock Market Analysis for Stock Selection – Naïve Investors' Perspective using Data Mining Technique", International Journal of Computer Applications (0975 – 8887), Vol 34– No.3,2011
13. Swathi P, Yogish HK, Sreeraj RS, "Predictive data mining procedures for the prediction of coronary artery disease", International Journal of Emerging Technology and Advanced Engineering, 5(2):339– 42,2015.
14. Das, R. and A. Sengur, "Evaluation of ensemble methods for diagnosing of valvular heart disease", Expert Systems with Applications, 2010, 37(7): p. 5110-5115.
15. Kurgan, L. and K.J. Cios, "Ensemble of classifiers to improve accuracy of the CLIP4 machine-learning algorithm", in Sensor Fusion: Architectures, Algorithms, and Applications VI. 2002, International Society for Optics and Photonics
16. Somayeh Nazari, Mohammad Fallah, Hamed Kazemipoor, Amir Salehipour, "A fuzzy inference- fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases", Expert

Systems with Applications, Volume 95, 1 April 2018, Pages 261-271
17. Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, "Heart Disease Diagnosis Using Data Mining Technique," 2017, IEEE International conference of Electronics, Communication and Aerospace Technology (ICECA)
18. Tülay KarayÕlan, Özkan KÖlÕç, "Prediction of Heart Disease Using Neural Network," 2017, IEEE International Conference on Computer Science and Engineering (UBMK)
19. Tahira Mahboob, Rida Irfan, Bazelah Ghaffar, "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics," 2017, IEEE Internet Technologies and Applications (ITA)
20. Procheta Nag, Saikat Mondal, Foysal Ahmed, Arun More and M.Raihan, "A Simple Acute Myocardial Infarction (Heart Attack) Prediction System Using Clinical Data and Data Mining Techniques," 2017, IEEE 20th International Conference of Computer and Information Technology (ICCIT)
21. Priyanga and Dr. Naveen, "Web Analytics Support System for Prediction of Heart Disease Using Naïve Bayes Weighted Approach (NBwa)," 2017, IEEE Asia Modelling Symposium (AMS)

## AUTHORS PROFILE

**Ms.Saloni Kapoor** is a student of Master's in Computer Science and Engineering at Chandigarh University, Gharuan, Mohali. She has done her Bachelor's in Computer Science from Chandigarh University. Her area of interest are Data Mining and Machine Learning.

**Prof. Ashwinder Tanwar** is an Assistant Professor in Computer Science Department at Chandigarh University,Gharuan, Mohali. He has received B.Tech and M.Tech in Computer Science and Engineering from Kurukshetra University, Kurukshetra. He is Research Scholar in Computer Science and Engineering Department at Uttrakhand Technical University(Dehradun).His research interests include Web application security, Machine Learning, Wireless sensor network and Mobile computing.