Hybrid Classification Method for Heart Disease Prediction

Kanika Sharma[1], Er. Poonam Chaudhary[2]
[1]Mtech Scholar, [2]Assistant Professor
[12]Sirda Institute of Engineering Technology, Sunder Nagar

Abstract- Prediction analysis is the data mining and machine learning which can predict future possibilities on the basis of current information. Prediction analysis has three phases which are pre-processing, feature extraction and classification. SVM classifier was used previously for heart disease detection. SVM classification method gave low accuracy for heart disease prediction. In this paper, hybrid classification method is designed which is the combination of decision tree and random forest classifier. The proposed method is implemented in Python and results are compared in terms of accuracy and execution time. It is analyzed that proposed method performs well in terms of all parameters in comparison to the existing method.

Keywords- Hybrid classification, Heart disease prediction, SVM

I. INTRODUCTION

There are large amount of data being stored in files, databases and many other appliances, as the confidential and the private data should not be stored anywhere else. So, it is very important to discover and propose a system in which all the data and information [1] can be stored safely and securely. Sometimes it becomes very difficult for the users to extract and use only relevant data from this large data. So, in order to overcome from this situation, Data Mining is used [2]. Data Mining is the process of selecting, choosing and extracting only that data which is useful and relevant for that particular instant of time. It allows the user to access their data anytime and from anywhere. There are large amount of relevant and irrelevant data being stored on the databases and many other areas. It has given rise to the term Data Mining, which can be further useful in decision making process. It is the phenomenon of extracting useful and important data from the large amount of data being stored almost everywhere on the internet [3]. It mostly deals with the already stored and gathered data for any other purpose instead of data mining purpose. This shows that the main objective of data mining is to collect and choose the relevant data from the previously stored data. There are enormous types of data sets being used in data mining. Among other data mining techniques, association is known to be the best technique. It helps in transacting the similar kind of data from one particular image to another. Here, depending upon the relationship, a pattern is discovered [4]. For instance, in predicting the presence of diabetes in the body, association has been useful. The relationships among various attributes were identified through this analysis. It is possible to identify the several risk factors related to particular diseases using which the kind of disease can be predicted. The approach in which objects are clustered together based on their similarity is known as clustering. Due to its similar characteristics, an automatic method is required to perform clustering [5]. For defining the process in which objects are assigned to predefined classes, the classes are assigned to objects through clustering. Clustering helps in predicting heart diseases as well. It clusters the similar kinds of risk factors and list patients that are likely to have heart diseases. A classic approach that is based on machine learning is known as classification. Here, every item present in the data set is classified into predefined groups or classes for data classification [6]. There are several mathematical techniques which are applied to perform classification among which few will be discussed in this research further. Prediction approach helps in discovering the relationship that exists in between the dependent and independent variables. For predicting the profit of future, several applications have been applying prediction. So, in these applications, sale is considered as independent variable and profit as dependent variable. Heart diseases are one of the major causes of death nowadays. Smoking, consumption of alcohol in large quantity, cholesterol, and pulse rate are the reason of heart diseases [7]. The heart is the operating system of human body, if it will not function properly then it will directly affects the functioning of the other body parts. Some of the major factors leads to the heart diseases are family history, high blood pressure, high rate of cholesterol, age, poor diet and many more. The stretching of blood vessels will increases the blood pressure which will further cause the cardiac rest. Smoking is one of the major causes of heart diseases; almost 40% of the population is dying because of this. Because it limits the supply of oxygen in our body and prevents the proper flow of blood and tightens the blood vessels [8]. Different types of data mining techniques are employed for the prediction of data mining. The factors that are responsible for causing heart diseases are identified and represented by k in the KNN algorithm. For the patients that suffer from heart, the classification report is deployed using decision tree. The probability of heart diseases can be predicted by applying naïve bayes algorithm. Last but not the least, the neural networks are used to minimize the errors occurred at the time of prediction. By using all these techniques, the records are classified as well [9] as maintained regularly. The activity of every patient is properly checked, if there is any change, and then the level of risk is informed to the patients. With the help of all these classifiers the doctors are able to predict the heart diseases at the very initial stage.

.

II. LITERATURE REVIEW

Min Chen, et.la (2017) proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. The data was collected from a hospital and contained both structured and unstructured data types. For making predictions regarding the chronic disease, different machine learning algorithms were modified in this study [10]. In this technique, a latent factor model was utilized for rebuilding incomplete data type occurring in the collected data. Several comparative analyses performed on earlier and the proposed approach depicted that that none of the earlier accessible techniques were able to handle both types of data collected from medical domains. The proposed approach showed a forecasting accuracy rate of 94.8% along with the higher convergence pace. This pace was superior to several other similar existing algorithms.

Tülay Karayilan, et.al (2017) proposed a novel heart disease forecasting model. The proposed model was identified as Multilayer Perceptron Neural Network [11]. The proposed model utilized Cleveland dataset. The thirteen medical data achieved from Cleveland Dataset as input was utilized by the neural network system. The back propagation algorithm was utilized for the training of proposed system to foresee the occurrence of heart disease in the body of patient. In the last few years, several analyses were performed to predict heart disease. The outcomes of these analyses varied equal to nearly 100% accuracy rate. The proposed approach showed an accuracy rate of 95%. This several studies performed in this domain considered this accuracy level fabulous. In the future, the proposed system can be modified as a hybrid model using other classification algorithms for obtaining further precise heart disease analyses.

Ms. Tejaswini, et.al (2017) presented an analysis executed by the world health organization [12]. This survey was conducted globally for the prediction of heart disease. This deadly disease caused more than 12 million causalities annually. Hence, precise prediction of this disease was essential. In this study, big data approach was utilized for heart disease as this approach utilized Hadoop Map reduce platform. Modified K-Means algorithm was used to perform clustering while decision tree classifier was used for classification process. The tested results determined that that proposed technique provided optimum results for the prediction of the heart disease than other existing techniques by providing improved healing procedure and enhanced medical decision making.

Marjia Sultana, et.al (2016) concentrated on the problem related to the forecasting of heart disease on the basis of several input features. The heart disease was a chronic disease and extended worldwide [13]. The prediction of this disease was a complex job and required proficiency and advanced understanding for forecasting. The concealed information was extracted by data mining approach. This approach was imperative in decision making process. In this study, a test was carried out by means of different data mining methods for discovering an additional precise method to predict heart disease. This study utilized two data sets in separate way, i.e. one for individual data mining method. The tested results depicted that the Bayes Net and SMa classification model showed optimal performances among other examined five classification models to predict heart disease.

M. A. Jabbar, et.al (2016) stated that cardiovascualr heart disease was a deadly heart disease. Because of this disease, various fatalities occurred globally [14]. The diagnosis process of this disease was complex as it required accurate monitoring all over the time. Thus, there was the need to invent an intelligent decision support system. The discovery of this decision support system was necessary to foresee heart disease. In this study, the employment of data mining methods in medical domain was reviewed. In the data mining, the conditional independence supposition of conventional technique was composed by this model. In this study, Hidden Naïve Bayes approach had been used to classify and predict heart disease. The tested results depicted that proposed approach showed better performance in terms of optimum accuracy.

Theresa Princy, et.al (2016) presented a survey of various classification methods. These techniques predicted the heart disease danger level of every person on the basis of age, gender, Blood pressure, cholesterol, pulse rate, etc [15]. In this study, different data mining methods and classification models reviewed for competent and effectual analysis of heart disease. Therefore, different accuracy was shown by different technologies according to considered amount of features. The danger level of heart disease was identified with the help of KNN and ID3 algorithm. These algorithms also offered accuracy rate for several amount of features. The amount of features can be decreased in nearby future. In the same way, accuracy would be increased by means of several other algorithms in future as well.

III. RESEARCH 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.

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.

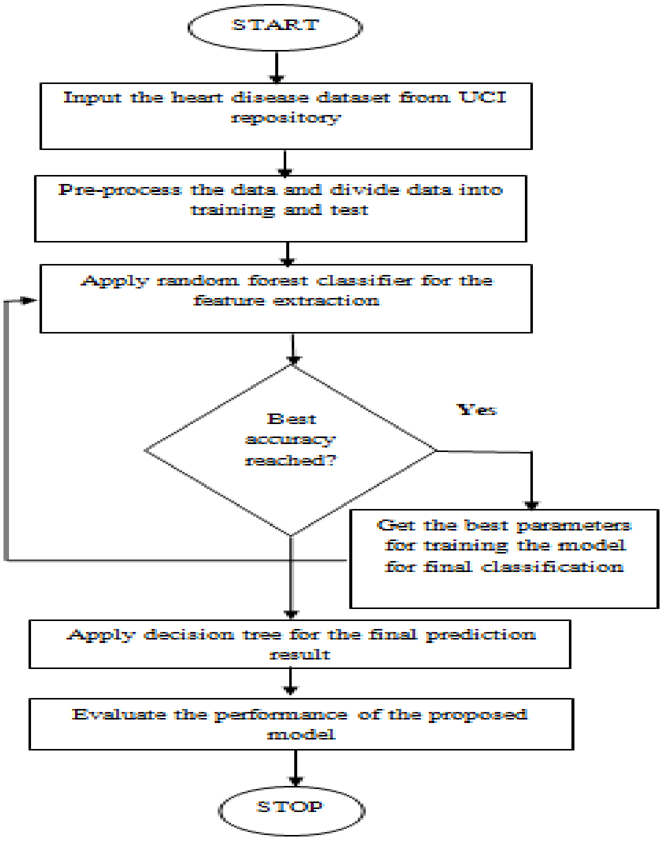
START
Input the heart disease dataset from UCI repository
Pre-process the data and divide data into training and test
Apply random forest classifier for the feature extraction
Best accuracy reached?
Yes
Get the best parameters for training the model for final classification
Apply decision tree for the final prediction result
Evaluate the performance of the proposed model
STOP

Fig.1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The proposed research is implemented in Python and results are evaluated based on certain performance measures.

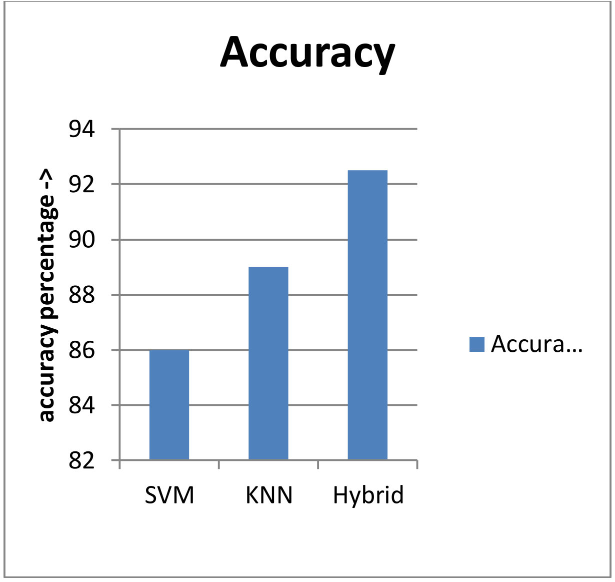

Fig.2: Accuracy Comparison

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 approximate 92 percent. The hybrid is combination of random forest and decision tree classification methods.

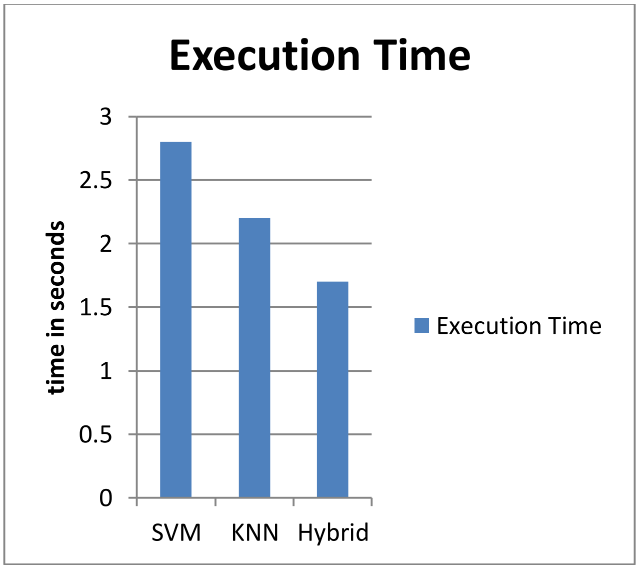

Fig.3: Execution time

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.

V. CONCLUSION

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. 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, KNN classifier. The hybrid classifier has maximum accuracy upto 92 percent as compared to SVM and KNN. In future, the clustering algorithm will be applied with the hybrid classifier method for the data division.

VI. REFERENCES

[1]. Mohammed Mahmood Ali, Khaja Moizuddin Mohammed and Lakshmi Rajamani, “Framework for Surveillance of Instant Messages in Instant messengers and Social networking sites using Data Mining and Ontology”, IEEE- Students' Technology Symposium, vol. 4, issue 1, pp. 23-48, 2014.

[2]. SushantBharti, Ashutosh Mishra. “Prediction of Future possible offender’s network and role of offender’s”,Fifth International Conference on Advances in Computing and Communications, vol. 8, issue 1, pp. 23-48, 2015.

[3]. Dahlia Asyiqin Ahmad Zainaddin and Zurina Mohd Hanapi, Hybrid of Fuzzy Clustering Neural Network over Nsl Dataset for Intrusion Detection System, Journal of Computer Science, Volume 9, No. 3, pp. 391-403, 2013.

[4]. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Din, “Data Mining With Big Data”,IEEE Transactions on Knowledge and Data Engineering, Vol. 26, issue 1, pp. 23-34, 2014

[5]. L. Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm”, International Journal of Biological and Life Sciences, vol. 3, no. 3, pp. 157-160, 2007.

[6]. Anupama Chadha, Suresh Kumar,”An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K”, 2014 International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, vol. 8, issue 1, pp. 6-8, 2014

[7]. Anand Bahety, “Extension and Evaluation of ID3- Decision Tree Algorithm”, ICCCS, ICCC, vol. 4, issue 1, pp. 23-48, 2014.

[8]. Vikas Chaurasia, et al, Carib.j., Early Prediction of Heart Diseases Using Data Mining Techniques, SciTech, Vol.1, issue 4, pp. 208-217, 2013.

[9]. K. Zakir Hussain, M. Durairaj, G. Rabialahani Farzana. “Criminal Behavior Analysis By Using Data Mining Techniques”, IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012), vol. 4, issue 1, pp.30-31, 2012.

[10]. Min Chen, YixueHao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, 2017, IEEE, vol. 15, issue 4, pp- 215-227, 2017.

[11]. TülayKarayilanTülayKarayilan, “Prediction of Heart Disease Using Neural Network”, IEEE, vol. 14, issue 1, pp. 423-468, 2017.

[12]. Ms. Tejaswini U. Mane, “Smart heart disease prediction system using Improved K-Means and ID3 on Big Data”, 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), vol. 8, issue 11, pp. 123-148, 2017.

[13]. Marjia Sultana, Afrin Haider and Mohammad ShorifUddin, “Analysis of Data Mining Techniques for Heart Disease Prediction”, IEEE, vol. 14, issue 1, pp. 123-138, 2016.

[14]. M. A. Jabbar, Shirinasamreen, “Heart disease prediction system based on hidden naïve bayes classifier”, vol. 4, issue 11, pp. 23-48, 2016

[15]. Theresa Princy. R, J. Thomas, “Human Heart Disease Prediction System using DataMining Techniques”, 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], vol. 4, issue 1, pp. 23-48, 2016.