|| Volume 5 || Issue 12 || April 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH & ENGINEERING TRENDS
Multidisciplinary Journal
Double-Blind Peer Reviewed Refereed Open Access International Journal

EFFECTIVE HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Miss. K. S. Ubale, Dr. P. N. Kalavadekar
Computer Engineering, Sanjivani College Of Engineering, Kopargaon
kalyaniubale110@gmail.com

----------***----------

Abstract: Heart disease can be considered as one of the complex diseases and globally many people suffered from the disease. In the recent years, a death because of heart disease has become a significant issue. So, it is necessary to design a system that will correctly diagnose heart disease. In this paper, an efficient and accurate system to diagnose heart disease is proposed and the system is based on Machine learning techniques resulting in improving the accuracy in the prediction of heart disease. A cardiovascular dataset is classified by using several state of the art Machine Learning algorithms that are precisely used for disease prediction. The prediction model is introduced with the several classification techniques and the different combinations of features. We try to produce an enhanced performance with high accuracy level through the prediction model for cardiovascular disease with the use of Machine Learning techniques like Random Forest, Naïve Bayes and SVM.

Keywords—*Machine learning, heart disease prediction, cardiovascular disease (CVD), feature selection, prediction model.*

----------***----------

I INTRODUCTION

Coronary Heart Disease (CHD) is one of the common form of disease affecting the heart and also an important cause for premature death. Data mining is involved in discovering various sorts of metabolic syndromes, from the point of view of medical sciences. Classification techniques in data mining play a significant role in data exploration and prediction. In predicting the accuracy and events related to CHD classification technique such as decision trees has been used. [2]

Heart disease is still a globally growing health issue. Limiting human experience and expertise in the health care system, in manual diagnosis leads to inaccurate diagnosis, and the information about various illnesses is either inadequate or lacking in accuracy as they are collected from the various types of medical equipment. Since the correct prediction of a person's condition is of great importance, for diagnosing and treating illness the use of equipping medical science with intelligent tools can reduce doctors' mistakes and financial losses. [3]

Recently, various algorithms and several software tools have been proposed by the researchers for developing an effective medical decision support systems. Moreover, new algorithms and new tools are continued to develop and are represented day by day. Many researchers investigated to develop intelligent medical decision support systems to improve the ability of the physicians as diagnosis of heart disease is one of the important issue. Neural network is one of the widely used tools for predicting heart disease diagnosis. [4]

As people are showing interests in their health recently, development of medical domain application has been one of the most active research areas. The detection system for heart disease based on computer-aided diagnosis methods, where the data are obtained from some other sources and are evaluated based on computer-based applications is one of the examples of the medical domain application. Earlier, the use of computer was only to build a knowledge based clinical decision support system which uses knowledge from medical experts and transfers this knowledge into computer algorithms manually. This process is time consuming and wholely depends on medical experts' opinions which may be subjective. To handle this problem, machine learning techniques have been developed to gain knowledge automatically from examples or the raw data. [6]

Classification procedure is an important task for expert and intelligent systems. The development of new algorithms of classification which improve the accuracy or true positive rates could have an influence on many life problems such as diagnosis and prediction in medical domain. Multi-criteria decision making (MCDM) methods are expected to search the best alternative according to some specified criteria. Each criterion has a value relative to each of the alternative. There are only two sets: a set of criteria and a set of alternatives. [7]

In this work, we introduce an effective technique for predicting heart disease using various machine learning

techniques. The main objective of this research is to improve the performance accuracy of heart disease prediction system. Many studies have been conducted so far, that results in restrictions of feature selection for algorithmic use. Here we conduct experiments used to identify the features using machine learning algorithms. Our proposed system and the method has stronger capability to predict heart disease compared to other existing methods.

II LITERATURE SURVEY

A.S.Abdullah and R.R.Rajalaxmi presented an analysis on how data mining plays an important role in the identification and prediction of various sort of metabolic syndromes and hence various sorts of diseases can be discovered. Decision tree classification algorithm has been used to assess the events related to CHD [2]. A.H.Alkeshuosh, M.Z.Moghadam, I.Al Mansoori, and M.Abdar presented how Classification rule mining is one of the most important tasks in data mining community. PSO-based algorithm for classification rule mining is presented. The algorithm is compared with the Decision tree based on C4.5 algorithm in UCI Repository of Machine Learning Databases. The experimental results show that the PSO algorithm achieved higher predictive accuracy and much smaller rule list than C4.5 [3]. N.Al-milli proposed an approach based on back propagation neural network to model heart disease diagnosis . In this research paper, a heart disease prediction system is developed using the neural network. The system used 13 medical attributes for heart disease predictions. The experiments conducted in this work have shown the good performance of the proposed algorithm compared to similar approaches of the state of the art [4]. C.A.Devi, S.P.Rajamhoana, K.Umamaheswari, R.Kiruba, K.Karunya, and R.Deepika has mentioned that the need for an efficient and accurate prediction for heart disease is on high demand. This paper deals with various techniques involving feature extraction and classification of the heart diseases resulting in accurate prediction [5]. P.K.Anooj, presented a weighted fuzzy rule-based clinical decision support system (CDSS) for computer-aided diagnosis of the heart disease. The proposed clinical decision support system proposed for risk prediction of the heart patients contains two steps such as: (1) generation of weighted fuzzy rules and (2) developing of a fuzzy rule-based decision support system [6]. L.Baccour proposed ATOVIC, a classification method based on fused TOPSIS and VIKOR methods of multi-criteria decision making. The obtained fused MCDM method is revised to be useful and suitable for classification. ATOVIC is applied to CLEVELAND data set to predict presence of heart disease. The results of experiments are compared with different classifiers and ATOVIC is shown to be promising and efficient [7]. C.A.Cheng and H.W.Chiu have done research and the preliminary results demonstrate that ANN can be used to build an accurate model that can serve as a reference of communication when neurologists refer patients and before patients are treated by cardiologists [8].

H.A.Esfahani and M.Ghazanfari presented the ability of a new data mining technique investigated for early diagnosis of heart disease. This data mining technique uses a fusion strategy in which three classifiers including neural network. Rough Set and Naive Bayes have been combined by a weighted majority vote. The ensemble classifier was evaluated on a dataset of 303 patients [9]. F.Dammak, L.Baccour, and A.M.Alimi introduced and applied crisp TOPSIS to crisp data set with different methods of weight such as Entropy, Standard Deviation etc. The comparison of results in both crisp and IF TOPSIS shows that the choice of weight formula influence results, and the latter can be different [10]. R.Das, I.Turkoglu, and A.Sengur proposed a system which shows how SAS enterprise miner 5.2 can be used to construct a neural networks ensemble based methodology for the diagnosis of heart disease. To diagnose heart disease in a fully automatic manner, experiments were conducted on the heart disease dataset. The three independent neural networks models were used to construct the ensemble model. SAS base software can be used in many machine intelligence applications [11]. S.K.Mohan, C.S.Thirumalai, G.Srivastava proposed the system were Machine learning techniques are used to process raw data and provide a new and novel discernment towards heart disease. The proposed HRFLM hybrid approach is used by combining the characteristics of Random Forest (RF) and Linear Method (LM) to achieve higher accuracy. HRFLM proved to be quite accurate in the prediction of heart disease [1].

III PROPOSED METHODOLOGY

In our proposed model we are using UCI dataset which contains various attributes that is used for analysis of diseases. The attributes are fetched and data is pre-processed for classifying the attributes required for our system model. There are 13 attributes in the data set, but two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes in the dataset are considered important as they contain vital clinical records which will provide the insights of a patients health. Clinical records used in the system are vital to diagnosis and for learning the severity of heart disease.

In our system, several (ML) techniques are used namely Naïve Bayes and Random Forest algorithms. The experiment is repeated with the ML techniques using all the 13 attributes present in the dataset. The clustering of datasets is done on the basis of the variables and criteria of the Decision Tree (DT)

features. Then, the classifiers is applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the results based on their low rate of error. The performance is further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on the dataset.

Our system model is divided into four phases where all the processing and prediction of heart disease will be done from the considered dataset;

1) In the first phase, we will pre-process the UCI dataset and classify all the attributes required for further processing. We will be considering the 13 attributes to learn the severity of the heart disease.

2) In the second phase, we will test and train the dataset using the various machine learning algorithms and feature extraction techniques like Naive Bayes, Random forest, decision tree and HRFLM.

3) In the third phase, we will train the classifiers and use them for prediction of the heart disease by considering the severity of the various attributes mentioned in the dataset and further all the classified information will be applied to each clustered dataset to estimate its performance.

4) In the fourth phase, classified information will be applied to clustered dataset to estimate its performance using various performance evaluation measures like accuracy, precision, and error.

III METHODOLOGY

The various steps involved in our proposed system are as follows;

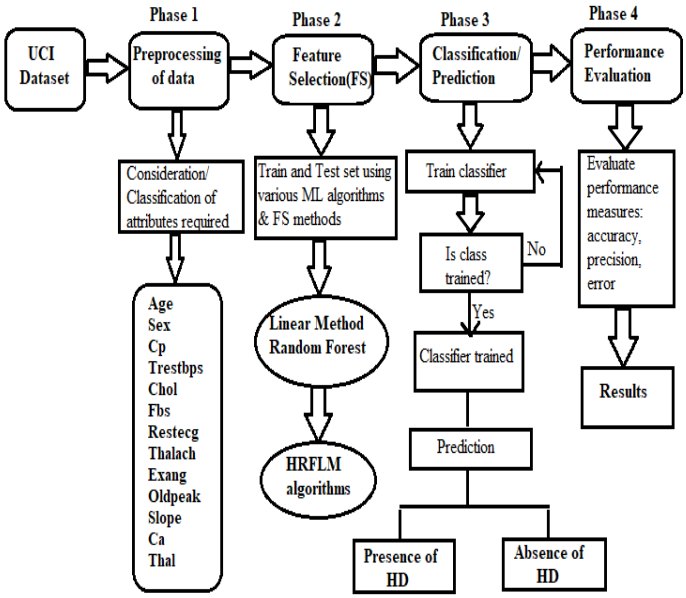

Figure 1 : System Architecture

1) DATA GATHERING

Download heart disease dataset from UCI repository https://archive.ics.uci.edu/ml/datasets/heart+disease

Read the dataset and retrieve attributes to predict heart disease.

The detailed system architecture of the proposed system application can be given and described as follows;

2) PREPROCESSING

After extracting the attributes, it needs to preprocess the records to remove stop words and duplication of data, using proper binary classification techniques to convert records into values which can be further used for predicting heart disease.

3) FEATURE SELECTION

From among the 13 attributes of dataset, 2 attributes are used to identify patients' information and the remaining 11 attributes will be used for diagnosis to learn the severity of heart disease.

4) CLASSIFICATION/TRAINING

Once input is ready, train the extracted attributes for prediction of heart disease by clustering the datasets on the basis of variables and criteria of decision tree features. The performance will be further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features.

5) TESTING

Predict the performance of classifier for error optimization on the dataset.

IV ALGORITHM

A) RANDOM FOREST

The random forest ensemble classifier builds several decision trees and incorporates them to get the best result. It mainly applies bootstrap aggregating or bagging, for tree learning. For a given data, $X = \{x1, x2, x3, \ldots, xn\}$ with responses $Y = \{x1, x2, x3, \ldots, xn\}$ which repeats the bagging from $b = 1$ to B.

Pseudocode for Radom forest algorithm:

Precondition: A training set $S := (x1, y1), \ldots, (xn, yn)$, features F, and number of trees in forest B.

```
1  function RandomForest (S , F)
2          H ← □
3          for i ϵ 1,…, B do
4              S(i)  ← A bootstrap sample from S
5              hi  ← RandomizedTreeLearn(S(i) , F)
6              H  ← H U {hi}
7          end for
8          return H
9  end function
10 function RandomizedTreeLearn(S , F)
11         At each node:
12           f ← very small subset of F
13          Split on best feature in f
14         return The learned tree
15 end function
```

C) NAIVE BAYES

Naïve Bayes uses Bayesian rule to distinguish samples of two classes. Based on this theory, each quantity has a statistical distribution by which a test sample can be categorized. This learning model applies Bayes rules through the independent features. In this algorithm every instance of data D is allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability $P(Xf) = priority \in (0 : 1)$.

$$P(X_{f1}, X_{f2}, \ldots, X_{fn} | c) = \prod_{i=1}^{n} P(X_{fi} | c)$$

$$P(X_f | c_i) = \frac{P(c_i | X_f)\ P(X_f)}{P(ci)}$$

At last, the testing data is categorized based on the probability of association.

$$c_{nb} = \arg\max P(c_k) \prod_{i=1}^{n} P(X_{fi} | c)$$

Pseudocode for Naive Bayes algorithm:

Input : Training dataset T,

F = *(f1, f2, f3,…,fn)* //value of the predictor variable in testing dataset

Output : A class of testing dataset

1. Read the training dataset T;

2. Calculate the mean and standard deviation of the predictor variables in each class;

3. Repeat

Calculate the probability of *fi* using the gauss density equation in each class;

Unlike the probability of all predictor variables *(f1, f2, f3,…fn)* has been calculated.

4. Calculate the likelihood of each class;

5. Get the greatest likelihood.

V DATASET DESCRIPTION

The data set is taken from Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998).

Finally the system is validated using data sets from Cleveland. In this dataset, there are total 14 attributes, such as Age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels, thal and diagnosis of heart disease are presented. Among all these 13 attributes are taken that feature the heart disease, where only one attribute serves as the output to the presence of heart disease in the patient. The Cleveland dataset contains an attribute named restecg to show the diagnosis of heart disease in patients on different scales, from 0 to 2. In this scenario, 0 represents the absence of heart disease and all the values 1,2 represent patients with heart disease, where the scaling refers to the severity of the disease (2 being the highest).

VI RESULTS AND ANALYSIS

The UCI dataset is loaded and the data becomes ready for pre-processing. The subset of 13 attributes (Age, sex, cp, treetops, chol, FBS, restecg, thalach, exang, olpeak, slope, ca, that, target) is selected from the pre-processed data set of heart disease. The models used for heart disease prediction are Naïve Bayes and Random Forest algorithm. They are used to develop the classification. The evaluation of the model is performed with the confusion matrix. The following measures are used to calculate the accuracy, precision and recall.

Precision = TP / (TP+FP)

Recall = TP / (TP+FN)

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Performance Measures	Naive Bayes	Random Forest
Precision	0.7	0.8
Recall	0.3	0.2
Accuracy	0.6	0.7

Table 1 : Comparison between NB and RF algorithm

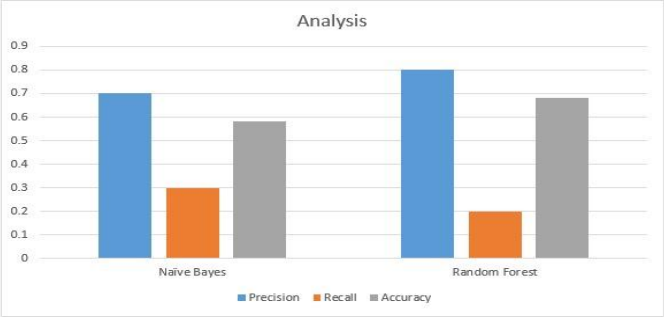

Figure 2 : Results and performance Analysis

VII CONCLUSIONS

Identifying the processing of raw healthcare data of heart information will definitely help in saving of human lives in the long term and early detection of abnormalities in heart conditions. Machine learning techniques will be used in this proposed work to process the raw data which is considered for analysis and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and also very important in the medical field. However, if the disease is detected at the early stages and preventative measures are adopted as soon as possible the mortality rate can be drastically controlled. Further extension of this study will be highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed approach will be used combining the characteristics of Random Forest (RF), Naive Bayes and SVM. Our system will prove to be quite accurate in the prediction of heart disease.

ACKNOWLEGMENT

“Effective heart disease prediction using machine learning techniques” had been a wonderful subject to research upon, which leads ones mind to explore new heights in the field of Computer Engineering. I dedicate my project to my esteemed guide, Dr. P. N. Kalavadekar, whose interest and guidance helped me to complete the work on research paper successfully. I would thank all my department members who have provided facilities to explore the subject with more enthusiasm. Last but not the least, I thank all others, and especially my friends who in one way or another helped me in the successful completion of this project.

REFERENCES

[1] S.K.Mohan, C.S.Thirumalai, G.Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” Special section on smart caching, Communications, Computing and cybersecurity for information-centric Internet of things, July 2019.

[2] A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf.Recent Trends Comput.Methods, Communication. Controls, Apr. 2012, pp. 22–25.

[3] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ‘‘Using PSO algorithm for producing best rules in diagnosis of heart disease,’’ in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306– 311.

[4] N. Al-milli, ‘‘Backpropogation neural network for prediction of heart disease,’’ J. Theor.Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[5] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks based heart disease prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[6] P. K. Anooj, ‘‘Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[7] L. Baccour, ‘‘Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets,’’ Expert Syst. Appl., vol. 99, pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[8] C.-A. Cheng and H.-W. Chiu, ‘‘An artificial neural network

model for the evaluation of carotid artery stenting prognosis using a national-widedatabase,'' in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566–2569.

[9] H. A. Esfahani and M. Ghazanfari, ''Cardiovascular disease detection using a new ensemble classifier,'' in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[10] F. Dammak, L. Baccour, and A. M. Alimi, ''The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,'' in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1–8.

[11] R. Das, I. Turkoglu, and A. Sengur, ''Effective diagnosis of heart disease through neural networks ensembles,'' Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009.