A Novel Approach For Heart Disease Prediction Using Machine Learning Techniques

Miss. Kalyani Ubale[1], Dr. P. N. Kalavadekar[2]
[1, 2] Dept of Computer Engineering
[1, 2]Sanjivani College Of Engineering, Kopargaon

Abstract- *Heart disease is one of the complex diseases and many people suffered from the disease globally. In the recent years, one if the most significant issue is death because of heart disease. It is necessary to design a system that will correctly predict the presence of heart disease. In this paper, we try to propose an efficient and accurate system for prediction of heart disease and the system is based on Machine learning techniques. The use of machine learning techniques, results in improving the accuracy for prediction of heart disease. A heart disease dataset is classified by using Naive Bayes and Random Forest Machine Learning algorithms that are precisely used for disease prediction. The prediction model is introduced with some classification techniques and the different combinations of features. We try to produce an enhanced performance with high accuracy level through the prediction model for heart disease with the use of Machine Learning algorithms like Random Forest, Naive Bayes and feature selection and extraction. The results shown that Random forest algorithm is best suited for prediction of heart disease and produced an accuracy level of 70% and the error rate is only 20%.*

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 a globally growing health issue all around the world. The 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

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]. 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]. An approach based on back propagation neural network to model heart disease diagnosis. 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 the work have shown the good performance of the proposed algorithm compared to similar approaches of the state of the art [4]. The need for an efficient and accurate prediction for heart disease is on high demand. The various techniques involving feature extraction and classification of the heart diseases resulting in accurate prediction [5]. 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]. 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]. 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]. 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]. 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]. 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. In the research paper, the three independent neural networks models were used for prediction and to construct the ensemble model. SAS base software can be used in many machine intelligence applications [11]. The systems 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 is done from the considered dataset;

1) In the first phase, we pre-process the UCI dataset and classify all the attributes required for further processing. We considered the 13 attributes to learn the severity of the heart disease.
2) In the second phase, we apply feature extraction techniques available in machine learning and algorithms like Naive Bayes, Random forest and decision tree.
3) In the third phase, we 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 is applied to each clustered dataset to estimate its performance.
4) In the fourth phase, classified information is applied to clustered dataset to estimate its performance using various performance evaluation measures like accuracy, precision, and recall.

IV. METHODOLOGY

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

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;

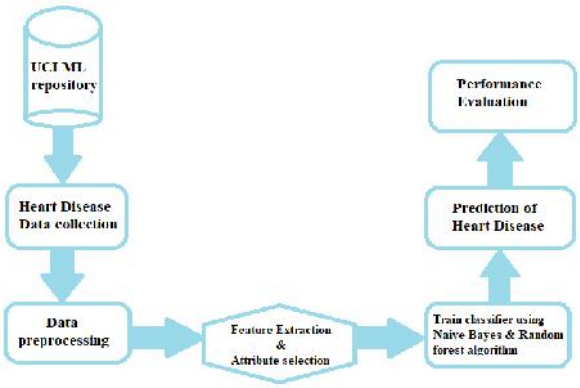

Figure 1 : System Architecture

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.

A. Dataset Description

The data set is taken from UCI ML repository which is available online. The system is validated using heart disease data set from Cleveland. In this dataset, there are total 14 attributes, such as Age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression, slope of the peak, 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).

Sr. No	Attribute	Type	Description	Values for the attributes considered
1	Age	Numeric	Patient's age in completed	Age in years
2	Sex	Nominal	Patient's Gender	Male-1 Female-0
3	Cp	Nominal	The type of Chest pain categorized into 4 values	1. Typical angina, 2. Atypical angina, 3. Non- anginal pain
4	Trestbps	Numeric	Level of blood pressure at resting mode	in mm/Hg at the time of admitting in hospital
5	Chol	Numeric	Serum cholesterol	in mg/dl
6	FBS	Nominal	Blood sugar	levels on fasting > 120 mg/dl; levels on fasting <= 120
7	Thalach	Numeric	The accomplishment of the maximum rate of heart	Exercise induced angina: YES NO
8	Exang	Nominal	Angina induced by exercise	0 depicting 'no' and 1 depicting 'yes'
9	Oldpeak	Numeric	Exercise-induced ST depression in comparison with	Values between 0 to 6
10	Slope	Nominal	ST segment measured in terms of the slope during peak exercise depicted in three values	1. Unsloping, 2. Flat and 3. Downsloping
11	Ca	Numeric	Fluoroscopy coloured major vessels	Numbered from 0 to 3.
12	Thal	Nominal	Status of the heart illustrated through three distinctly	Normal numbered as 3, fixed defect as 6 and reversible defect as 7.
13	Resting	Nominal	Results of electrocardiogram while at rest are represented in 3 distinct values	0- Normal state 1- Abnormality in ST-T wave and 2- any probability or certainty of LV hypertrophy by Estes' criteria

Table 1: Features considered from the Cleveland Dataset for Heart Disease Prediction

B. Preprocessing

Heart disease data will be pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records will be removed from the dataset and the remaining 297 patient records will be used in pre-processing. The multiclass variable and binary classification will be introduced for the attributes of the given dataset. The multi-class variable will be used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value will be set to 1, else the value will be set to 0 indicating the absence of heart disease in the patient. The pre-processing of data will be carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value 0 indicating the absence of heart disease.

C. Feature selection

There are total 13 attributes in the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records of the patient. Clinical records are vital to diagnosis and learning the severity of heart disease. Machine learning techniques are used namely, NB, DT, RF. The experiment will be repeated with all the ML techniques using all 13 attributes. The subset of 13 attributes that are; Age, sex, cp, treetops, chol, FBS, restecg, thalach, exang, olpeak, slope, ca, that and target is selected from the pre-processed data set of heart disease. In the system, Naive Bayes classifier algorithm and Random Forest algorithm are used which creates the feature set of the dataset we considered for prediction of heart disease. For creating the feature set the Cleveland dataset contains an attribute named restecg to show the diagnosis of heart disease in patients on different scales, from 0 to 2. These values determine the severity of the heart disease and also define and calculate the probability of all the records of patients considered for prediction of heart disease.

D. Classification/training

The clustering of datasets will be done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers will be applied to each clustered dataset in order to estimate its performance. The best performing models will be identified from these results based on their low rate of error. The performance will be further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. In the proposed system, Random forest algorithm and Naive Bayes algorithm is used for classifying and clustering of the dataset. The implementation of random forest algorithm and naive bayes algorithm creates the respective decision trees and then this data is further used to do the classification of heart disease. In the system two classes are used for effective prediction of heart disease. The value "0" indicates the absence of heart disease and values "1" and "2" indicates the presence of the heart disease, where the scaling refers to the severity of the disease (2 being the highest). The same results can be shown by using the two values, as Yes and No, that shows whether patient is suffering from heart disease or not.

E. Testing

Testing is the final step where model performance is evaluated. In the HeidiSQL datasets, the UCI heart disease data file is used for predicting the performance of model. Confusion matrix can be used to check the accuracy of model. Accuracy is most important performance indicator used to measure performance of random forest and naive bayes algorithm.

V. SYSTEM IMPLEMENTATION STEPS

In data gathering, first we need to access UCI ML repository and download the dataset for the heart disease. Then, perform preprocessing on dataset to make categories data into classes. This includes the various attributes and values. Feature selection includes selection of the appropriate features. Extract required attributes by applying various ML techniques. Once input is ready, the classifier needs training for predicting the heart disease. In testing it calculates the performance measures and makes correct prediction of heart disease.

Algorithm for Proposed Effective Heart Disease Prediction System:

1: Begin
2: Download UCI Heart Disease Dataset from UCI ML repository
3: The pre-processing of heart disease dataset using preprocessing methods
4: Features selection using standard state of the art and proposed
RM and NB algorithm
5: Train the classifiers using training dataset
6: Validate using testing dataset
7: Computes performance evaluation metrics
8: End

VI. SYSTEM ANALYSIS

A. Evaluation Metrics

Several standard performance metrics such as accuracy, precision and error in classification are considered for the computation of performance efficacy of the system. To identify the significant features of heart disease, these three performance metrics are used which helps in better understanding the behavior of the various combinations of the feature-selection. ML technique focuses on the best performing model compared to the existing models. The performance of every classifier is evaluated individually and all results are adequately recorded. 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)

In these equations,

- **True Positive (TP):** means the patients with heart disease that are correctly classified are considered.
- **True Negative (TN):** means the patients without any heart disease that are correctly classified
- **False Positive (FP):** means the patients without any heart disease but misclassified as having heart disease
- **False Negative (FN):** means the patients having heart disease but misclassified as not having heart disease

Table 2: Comparative analysis of NB and RF algorithms using performance measures

Algorithms used for implementation	Performance Measures		
	Precision	Recall	Accuracy
Naive Bayes	70%	30%	60%
Random Forest	80%	20%	70%

B. Results

Results of heart disease prediction system are given below. The system is implemented using Java language and eclipse IDE. Java packages are installed and used to implement the system. Standard dataset of heart disease from UCI is downloaded and preprocessed. A personal laptop, which has configuration of DESKTOP-V0A6I9C, Intel(R) Core(TM) i3-4000M CPU with AMD Radeon R2 Graphics @ 2.40GHz, 4GB memory is used to platform this experiment and GPU acceleration is not used. The proposed system is developed by implementing Random forest and Naive Bayes algorithm and performance of both algorithms is evaluated and analyzed after correct prediction of heart disease. Dataset is used for both classification and prediction of heart disease. The performance of both the algorithms is evaluated and compared to find the best performing model.

In the process data option on the dashboard of our system, first user needs to enter his health records supporting heart disease. Then, he should click on submit. The user will see the result for prediction of heart disease if present system shows value 1 or 2. If the user or patient is not suffering from heart disease, system displays value 0 showing absence of heart disease. The results generated by the system are shown in figure 2;

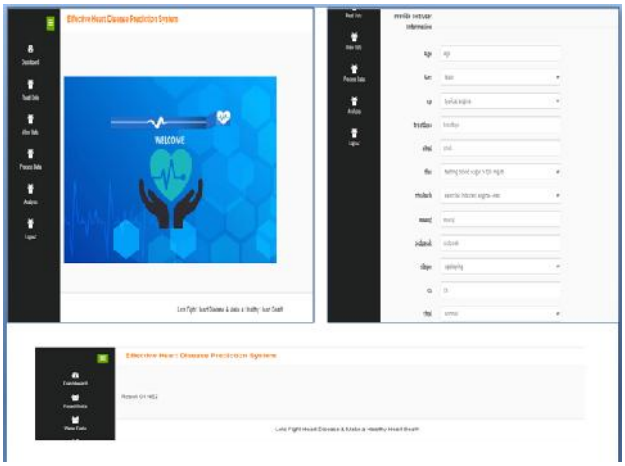

Figure 2: Results generated by our heart disease prediction system

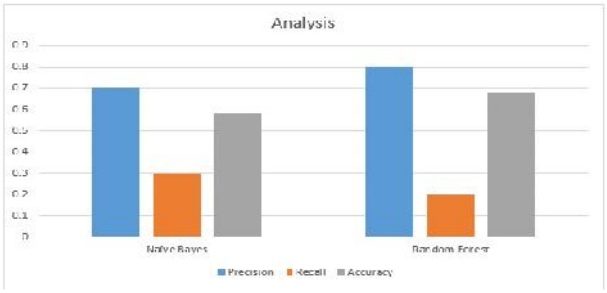

Figure 3 : Results and performance Analysis

VII. CONCLUSIONS

Heart disease prediction is challenging but also very important in the medical field. If the heart disease is detected at the early stages and preventative measures are adopted as soon as possible the death rate can be controlled drastically. 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 system is used by combining the characteristics of Random Forest (RF) and Naive Bayes (NB) algorithm. RF proves to be quite accurate in the prediction of heart disease. In this system, we proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of heart disease. The prediction model is introduced with different combinations of features and known classification techniques. We produce an enhanced performance level with an accuracy level of 70% through the prediction model for heart disease with the random forest classifier model.

VIII. ACKNOWLEGMENT

"A novel approach for heart disease prediction using machine learning techniques" had been a wonderful subject to research upon in the field of Computer Engineering. I thank my esteemed guide, Dr. P. N. Kalavadekar, whose interest and guidance helped me to complete the work on research paper successfully. I would thank my department members and Dr. D. B. Khirsagar for their valuable help for providing facilities to explore the subject with more enthusiasm.

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] 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] 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] 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.