Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

Newfangled Approach for Early Detection and Prevention of Ischemic Heart Disease using Data Mining

Dhara B. Mehta
Dept. of Computer Engineering
Atmiya Institute of Technology and Science
Gujarat, India
dharamehta133@gmail.com

Nirali C. Varnagar
Dept. of Computer Engineering
Atmiya Institute of Technology and Science
Gujarat, India
ncvarnagar@aits.edu.in

***Abstract*— The significant advances in biotechnology and health science have led to produce large amount of data. As the heart disease causing the major problem because it is very difficult to identify disease based on symptoms. For that we need a lot of experience and knowledge. Finding the exact cause of disease can help to patient cure at early stage of risk level. Ischemic heart Disease is one of the most common causes of death nowadays. In this research work, we have applied data mining classification techniques on the stroke dataset and proposed a recommendation system in which a person can come to know their risk level of IHD. By taking proper cure and treatment can help to survive more years if they have high risk of IHD. We have used five data mining algorithms Logistic Regression, Decision Tree, K nearest Neighbor, Naïve Bayes and SVM on Ischemic Stroke Dataset and got the highest accuracy 97.91% using Support Vector Machine algorithm.**

Keywords—Ischemic Heart Disease; CVD; Prediction; Data Mining; Classification; Accuracy.

I. INTRODUCTION

The significant advances in biotechnology and health sciences have led to a significant production of data, such as high throughput genetic data and clinical information. As the health diseases causing the major problem because it is very difficult for identifying the disease, based on symptoms, for that we need lots of experience and also knowledge. Finding the exact cause of disease in a patient requires Doctors expertise. Generally Doctors use their knowledge to cure a particular ailment based on the symptoms shown by the patient. However, this information may sometimes not be helpful to detect possible disease suffered by the patient. Currently the researches are made to overcome this problem[12] but this works are not actually being used by the patients. We are propose a data mining algorithm for early detection and prevention of ischemic heart disease from clinical records using newly data mining techniques that will be beneficial to patient as well as doctor.

II. ABOUT ISCHEMIC HEART DISEASE

Ischemic Heart disease is one of the most common causes of cardiovascular death. In one of the survey it is found that 48% people die due to ischemic heart disease among total deaths caused by any disease. Ischemia is actually a condition in which blood flow is restricted or reduced in all over the body. This happens when the heart does not pump well due to blood clots or plaque in blood vessels. The human brain needs enough amounts of oxygen and nutrition. When brain doesn't get oxygen then it leads to Ischemic Stroke. Ischemic stroke doesn't give the prior warning to the patient that's why it is also known as silent heart attack and it leads to death. Below Fig. 1 shows the reduced flow of blood due to fat or plaque generated in blood vessel.

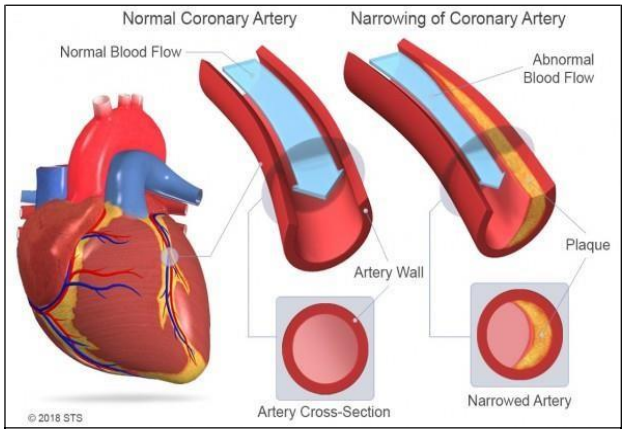

Fig. 1. IHD due to Narrowed Artery [13]

Ischemic Heart disease is predominant cause and is responsible for more than 80% of CVD deaths. The Global Burden of Disease study estimated that around 272 per 100,000 populations in India are higher than that of 235 per 100,000 populations globally [1]. Some aspects of CVD infestation in India are lack of disease knowledge, hypertension, early age of disease onset, more amount of smoking and low intake of fruits and vegetables.

In this research work, we proposed a prediction approach for knowing the risk of Ischemic Heart Disease (IHD) by using the earlier patient's dataset of Stroke that we got from

https://www.kaggle.com. We have applied some Data mining classification techniques on Ischemic Stroke Dataset which gives the better accuracy. We have also proposed a system which will predict the risk level of Ischemic Stroke for newly entered data of any person it will give the result whether the person is having high risk or low risk of Ischemic Stroke. And if high risk is found then the system will give the recommendation for the prevention of risk by following the given recommendation. So that a patient can survive more years.

III. LITERATURE SURVEY

Mohammad Amin, et. al. [2] proposed a Vote algorithm for the prediction of heart disease using most significant attributes. Dataset containing 13 attributes only 9 attributes were selected for prediction. (i.e, Vote algorithm gives 87.4% accuracy in Heart Disease Prediction)

D.Karthick, et. al. [3] have used Naïve Bayes algorithm for predicting the risk occurrence of CVD. The main aim of this paper is to predict risk factor for the young age people. they have used R studio for the classification and for the classification they have used Naïve Bayes and Random Forest algorithm.

Sarath Babu, et. al. [4] aims to early detection of heart disease and its diagnosis correctly on time and providing treatment with affordable cost. The system uses dour techniques namely Generic algorithm, K means clustering, Mafia algorithm and decision tree. The system gets better accuracy by using K means clustering before decision tree.

Meenal, Niyati, et. al. [5] have aimed to achieve accurate result for the prediction of heart disease. So they have made three combination of three different classifiers. The algorithm used for classification are SVM, Neural Network, decision tree, GLM, Lasso, bmn, rpart2, earth, ctree. The results were arranged in ascending order by accuracy. The combination result again ensemble for the final result. The Hybrid Classifier with Weighted Voting (HCWV) is proposed with highest accuracy by 82.54%.

Rishabh Wadhawan et. al. [6] have aimed to developed a system to extract unknown knowledge from the past dataset of heart disease. The system uses 7 attributes out of 14 attributes from the dataset of UCI repository for heart disease. The author has used visual studio c# for implementation. The system uses K means clustering and Apriori algorithm for classification and it gets 74% accuracy.

IV. RELETED WORK

The dataset of stroke patient is used for training the module. The dataset initially analyzed by their attributes affecting for Ischemic heart disease. There are 12 attributes in the dataset mainly affected for IHD out of them one is target stroke. The details of dataset are given in below table.

TABLE I. ATTRIBUTES DETAILS OF STROKE DATASET

Attributes	Description
Id	Patient's ID
Gender	Gender of patient
Age	Age of patient
hypertension	0- no hypertension, 1- suffering from hypertension
Heart_disease	0- no heart disease, 1- suffering from heart disease
ever_married	0- Yes 1- No
work_type	1- Private, 2- self_emp, 3-Gove_job,
Residence_type	1 – Rural, 2 – Urban
avg_glucose_level	Average glucose level(Measured after meal)
Bmi	Body mass index
smoking_status	1 – never smokes, 2 – formally smokes, 3 - smokes
Stroke(Target)	0- No stroke 1- Suffred Stroke

As ischemic stroke does not gives prior warnings and this attributes seems normal but they are strongly affected attributes in cause of ischemic stroke. The person who smokes many a times a day has high risk of getting ischemic stroke. Same as for the person having more fat in blood have high risk of ischemic heart disease because the flow of blood to the heart is being reduced or restricted due to fat or plaque generated in blood arteries.

A. Classification Techniques

Data mining classification techniques have been used for predicting the result. We have applied five classification techniques.

1) Logistic Regression: Logistic Regression algorithm is use for describing the relationship between the data. The logistic regression algorithm is based on the following formula.

$$P = \frac{1}{(1+e^{-x})}$$

2) Decision Tree: Decision tree is a greedy algorithm that follows top down approach [10]. The algorithm selects the attribute with highest information gain nearest neighbor.

3) K Nearest Neighbor (KNN): KNN classifies test data directly by using train data[8, 10]. Firstly it calculates K value, which is the number of nearest neighbors. Here we took the value of k=4. The nearest neighbor equation is based on Euclidian distance given below:

4) Naïve Bayes: Naïve Bayes Classification uses conditional independency[9]. It assumes that one attribute of dataset is independent to the other attribute. Here we have used Bernoulli naïve bayes classification technique.

5) Support Vector Machine (SVM): SVM is a supervised machine learning algorithm used for classification as well as regression [7, 11]. The data is scattered initially on a plain graph. SVM algorithm gives better result when the data has to be separated into two classes. SVM algorithm uses linear and nonlinear kernel functions for making hyperplane. The hyperplane differentiates the data by its category. When the maximum margin is found between a data point and hyperplane, it gives more accurate result.

V. Proposed Methodology And Implimentation

A. Description

We took the dataset from Kaggle which is an online community for dataset and machine learning [14]. The dataset contains 800 records and 12 attributes. Out of 12 attributes 11 attributes are used for prediction and the remaining one attribute is target attribute as Stroke. First we applied preprocessing steps. In that we converted all the values into nominal form. The dataset was splitted into two parts as training dataset and testing dataset. Initially the module will be trained by training dataset and then it will test for the testing dataset. Table II. Shows the dataset splitting which was best fit for the dataset.

TABLE II. Train and Test Split Dataset

	Training Data	Testing Data
Dataset	70%	30%

B. Preprocessing

Preprocessing of the dataset is needed for making it into understandable form. The data mining module for classification must be able to understand the formation of dataset. Here in this dataset, the data values were in the form of word and expression. So we need to change it into nominal form. Attributes like heart_attack, ever_married, work_type, residencial_type, smoking_status and stroke have values in the form of expression. For example we changed the values of these attributes. For example, the value for no stroke is replaced by 0 and suffered stroke is replaced by 1 to make the result in the binary form.

C. Workflow Chart

The proposed model works on prediction of risk of Ischemic stroke. The system first takes the dataset. Then preprocessing of data will be done. The labeled data will be given to the classification algorithms and then it will generate the result. The API is developed for users of IHD risk prediction system. In which user can enter the data and the risk factor will be given as an output. The proposed recommendation system is to give recommendation for preventing risk of Ischemic Heart Disease. Fig 2 represents the architecture of proposed system.

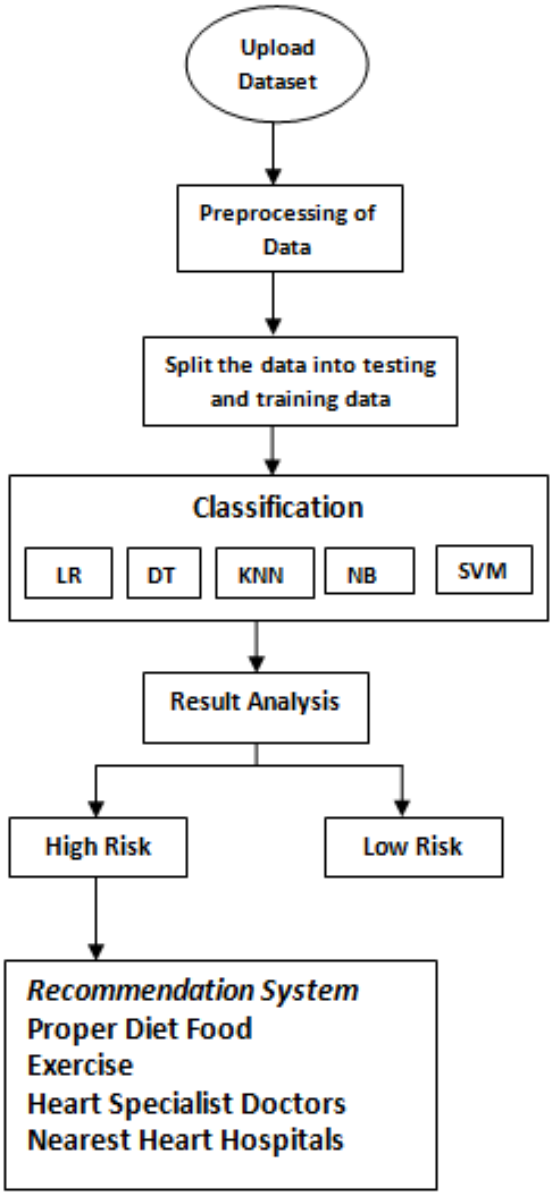

Fig. 2. Workflow of System Architecture

D. Implementation

The dataset was collected from kaggle for Ischemic stroke. The implementation is done based on prediction of Ischemic heart disease risk. We have done the implementation on Python 3.6 version. Python libraries like Scikit-learn provides mechanism for machine learning algorithms. We have used this library and data mining techniques were used which gives the result as below. Flask restful API used for generating result of model for any users. Fig 3 shows the User Interface of Proposed Recommendation system. After entering all the values, user will get the result as either high risk or low risk of Ischemic Heart Disease.

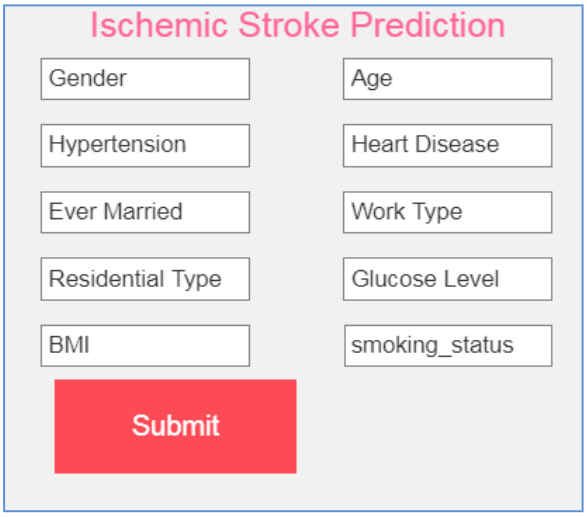

Fig. 1. User Interface of IHD Pridiction System

E. Result Analysis

Initially, we examined the dataset of Ischemic Stroke and then applied data mining classification techniques. The result analysis of classification techniques were analyzed by their accuracy. The accuracy is the measure of correctness of the predicted values to the standard values.

$$\text{Accuracy}=\frac{\textit{The number of correctly predicted samples}}{\textit{Total number of standard samples}}$$

Result analysis of data mining algorithms for Ischemic heart Disease, where the dataset of Ischemic Stroke was used for prediction. All the algorithms give better accuracy. Decision tree algorithm gives 94.11% accuracy and Naïve Bayes gives 96.67% accuracy. We got higher accuracy in Logistic regression, K nearest neighbor and Support Vector Machine which is 97.50%, 97.81% and 97.91% respectively. Table 2 shows accuracy results of classification techniques.

TABLE III. ACCURACY OF CLASSIFICATION TECHNIQUES

Classification Technique	Accuracy
Logistic Regression	97.50%
Decision Tree	94.11%
K nearest Neighbor	97.81%
Naïve Bayes	96.67%
Support Vector Machine	97.91%

As the more accurate result, the user of IHD prediction system comes to know the exact outcome of IHD risk. Figure shows the graphical representation of accuracy values for classification techniques. In the graph, we can see that Support Vector Machine gives highest accuracy.

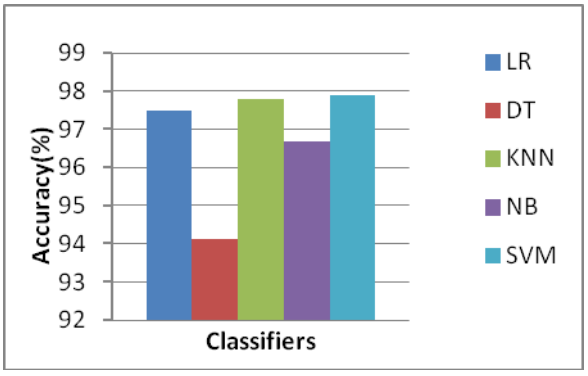

Fig. 2. Classification Accuracy Chart

VI. CONCLUSION

The research work is to develop a simple approach to detect the risk of Ischemic Stroke at early stage. We have used dataset of previously suffered stroke patients and applied data mining classification techniques on it. Support vector machine gives 97.91% accuracy. We proposed a recommendation system for the people in which by entering data values of attributes affected for IHD and system gives result by either high risk or low risk. If risk is high, then system will give the recommendation for diet food and exercise as well as the nearest heart hospital details along with heart specialist doctors. Out main aim of utilizing the actual usage of research can be fulfilled by proposed recommendation system. The current population can be aware about their risk factors and early prediction can help to survive more years.

ACKNOWLEDGMENT

This research work was carried out with the help of faculty members of Atmiya Institutions of Technology and Science (AITS). Special thanks to Professor Nirali Varnagar for their guidance and support throughout the whole dissertation work.

REFERENCES

[1] Dorairaj Prabhakaran, Panniyammakal Jeemon, “Cardiovascular Disease in India Current Epidemiology and Future Directions” Circulation. 2016;133:1605-1620. DOI: 10.1161/CIRCULATIONAHA.114.008729.

[2] Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan, “Identification of Significant Features and data mining techniques in predicting heart disease” 2018 ELSEVIER, Telematics and Informatics

[3] Karthick, B.Priyadharshini, "Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using Classification Techniques within fifty years of age"2018 IEEE 2nd International Conference on Inventive and Control

[4] 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 on Electronics, Communication and Aerospace Technology(ICECA)

[5] Meenal Saini, Niyati Baliyan, Vineeta Bassi, "Prediction of Heart Disease Severity with Hybrid Data Mining" 2017 International Conference on Telecommunication and Networks

[6] Rishabh Wadhawan, "Prediction of Coronary Heart Disease Using Apriori algorithm with Data Mining Classification" www.ijrst.org Volume 3, Issue 1(April 2018)

[7] Mrs.S.Raadhimeenakshi, "Classification and Prediction of Heart Disease Risk Using Data Mining Techniques of Support Vector Machine and Artificial Neural Network" 978-9-3805-4421-2/16/$31.002016 IEEE

[8] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, M. A. Hossain, "Comparative Analysis of Classification Approach for Heart Disease Prediction" IEEE-2018

[9] Narender Kumar, Sabita Khatri, "Implementing WEKA for medical data classification and early disease prediction", In 3rd IEEE Conference on Computational Intelligence and Communication Technology 2017

[10] Theresa Princy R, J. Thomas, "Human Heart Disease Prediction System using System using Data Mining Techniques" 2016 International Conference on Circuit, Power and Computing Technologies(ICCPCT)

[11] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin, "Analysis of Data Mining Techniques for Heart Disease Prediction" 978-1-5090-2906-8/16 2016-IEEE

[12] E. mercy Beulah, et. al., "Applications of Data Mining In Healthcare: A Survey0", Asian Jr. of Microbiol. Biotech. Env. Sc. Vol. 18, No (4): 2016

[13] https://ctsurgerypatients.org/adult-heart-disease/coronary-artery-disease

[14] https://www.kaggle.com/asaumya/healthcareproblem-prediction-stroke-patients/data