Heart Disease Prediction with Machine Learning

Jaideep Phatak[1], Kaustubh Tambe[2], Trisha Sawant[3], Srikant Bagewadi[4]

[1,2,3]*B.E. Department of Computer Engineering, PVPPCOE Maharashtra, India*
[4]*Assistant Professor Department of Computer Engineering, PVPPCOE Maharashtra, India*

---***---

Abstract - *Heart disease is contemplated as one of the vital causes of death throughout the globe. It cannot be easily predicted by the medical practitioners as it is a heavy task which command expertise and higher understanding for prediction. The healthcare province is still „information rich" but „knowledge deficient". There is a lot of data accessible within the healthcare systems on the internet. However, there is a absence of effective analysis tools to find hidden connection and patterns in data. An automated system in medical identification would increase medical efficiency and lower costs. This web application plan to predict the occurrence of a disease based on data assembles from medical research especially in Heart Disease. The goal is to draw out the hidden patterns by put in data mining techniques on the dataset, which are notable to heart diseases and to predict the existence of heart disease in patients where the existence is valued on a scale. The prediction of heart disease demands a huge size of data which is too difficult and huge to process and analyses by traditional techniques. Our objective is to find out the satisfactory machine learning technique that is computationally well organized as well as precise for the prediction of heart disease. Data mining combines Statistical analysis machine learning and database technology to find hidden patterns and relationships from huge databases.*

Key Words: Heart Diseases, Machine Learning, Prediction, Heart Problem Prediction,

1. INTRODUCTION

The highest impermanence of both India and abroad is mainly because of heart disease. According to World Health Organization (WHO), heart related diseases are in charge for the taking 17.7 million lives every year, 31% of all worldwide deaths. Hence, this is essential time to check this death rate by identifying the disease correctly in the beginning stage. We can use data mining technologies to invent knowledge from the datasets. The invented knowledge can be used by the healthcare managing directors to improve the standards of service. The invented knowledge can also be used by medical practitioners to reduce the number of unfavorable drug effect, to suggest less high cost therapeutically equivalent alternatives. Predict patient's future conduct on the given history is one of the important applications of data mining skill that can be used in healthcare industry. A major provocation facing healthcare organizations (hospitals, medical centers) is the planning of class services at affordable costs. Quality service indirect diagnosing patients correctly and administering handling that are productive. Poor clinical decisions can lead to catastrophic consequences which are therefore intolerable. Hospitals must also reduce the cost of clinical tests. They can achieve these results by employing suitable computer-based information and/or decision keep up systems. Healthcare data is huge It cover patient data, resource management data, and transformed data. Healthcare organizations must have the capacity to analyses data. Handling documentation of millions of patients can be added , and computerized and data mining techniques may help in respond several important and critical questions connected to health care. Clinical decisions are frequently made based on doctors" instinct and experience sooner than on the knowledge-rich data unseen in the database. This practice leads to undesirable biases, errors and immoderate medical costs which affects the class of service provided to patients. Wu, et al proposed that mixture of clinical decision support with computer-based patient data could lower medical errors, increase patient safety, decrease undesired practice variation, and upgrade patient outcome. This proposal is promising as data modeling and analysis tools, e.g., data mining, have the likely to generate a knowledge-rich habitat which can help to notably improve the quality of clinical decisions.

2. LITREATURE SURVEY

Intelligent Heart Disease Prediction System Using Data Mining Techniques: The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network.

Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction: An Android based prototype software has been developed by integrating clinical data obtained from patients admitted with IHD (Ischemic Heart Disease). The clinical data from 787 patients has been analyzed and correlated with the risk factors like Hypertension, Diabetes, Dyslipidemia (Abnormal cholesterol), Smoking, Family History, Obesity, Stress and existing clinical symptom which may suggest underlying non-detected IHD. The data is mined with data mining technology and a score is generated. Risks are classified into low, medium and high for IHD. Analysis of Data Mining Techniques for Heart Disease Prediction: Heart disease is considered as one of the major causes of death throughout the world. It is hard to predict for the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. This paper addresses the issue of prediction of heart disease according to input attributes based on data mining techniques. We have investigated the heart disease prediction using KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka software. The performance of these data mining techniques is measured by combining the results of predictive accuracy, ROC curve and AUC value using a 6 standard data set as well as a collected data set. Based on performance factor SMO and Bayes Net techniques show optimum performances than the performances of K-Star, Multilayer Perceptron and J48 techniques.

Machine Learning Application Predict the Risk of Coronary Artery Atherosclerosis: Coronary artery disease is the leading cause of death in the world. In this research, we propose an algorithm based on the machine learning techniques to predict the risk of coronary artery atherosclerosis. A ridge expectation maximization imputation (REMI) technique is proposed to estimate the missing values in the atherosclerosis databases. A conditional likelihood maximization method is used to remove irrelevant attributes and reduce the size of feature space and thus improve the speed of the learning. The STULONG and UCI databases are used to evaluate the proposed algorithm. The performance of heart disease prediction for two classification models is analyzed and compared to previous work. Experimental results show the improved accuracy percentage of risk prediction of our proposed method compared to other works. The effect of missing value imputation on the prediction performance is also evaluated and the proposed REMI approach performs significantly better than conventional techniques.

3. PROPOSED SYSTEM

The system will consist of a website, where users will register themselves for getting the report of health of their heart in terms of predictive analysis about their heart disease. User will have to fill a form initially for registration. Then user will get redirected to the profile page where they will have to complete their profile by filling all the information related to their heart. After submitting the health information the patient will be able to have look at the report where they will be knowing the status or risk of their heart in terms of percentage. If the user will have risk greater than 60% then user will be redirected to another form where he will have to enter additional symptoms so that system will give prediction about the category of heart disease from two most common categories i.e. CAD (Coronary Artery Disease) and valvular disease

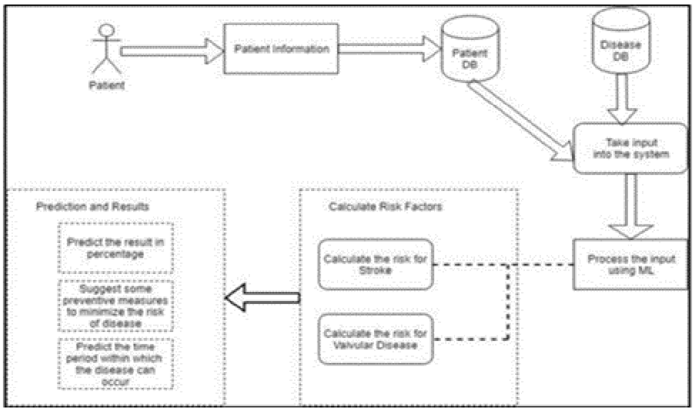

Patient
Patient Information
Patient DB
Disease DB
Take input into the system
Process the input using ML
Calculate Risk Factors
Calculate the risk for Stroke
Calculate the risk for Valvular Disease
Prediction and Results
Predict the result in percentage
Suggest some preventive measures to minimize the risk of disease
Predict the time period within which the disease can occur

Fig 1: Block diagram for heart disease prediction

A. Database

The server will be using MySQL database. The system's database consists of following tables.

Users table – This table will consist of all the user information which includes user's name, e-mail id, phone number, address, etc.

Medical history table – This table will consist of all the health related information of users which is related to heart that includes attributes such as age, gender, resting blood pressure, cholesterol, fasting blood sugar, old peak, etc.

B. Machine learning algorithm

The machine learning algorithm will be used to predict the risk of heart disease in terms of percentage.

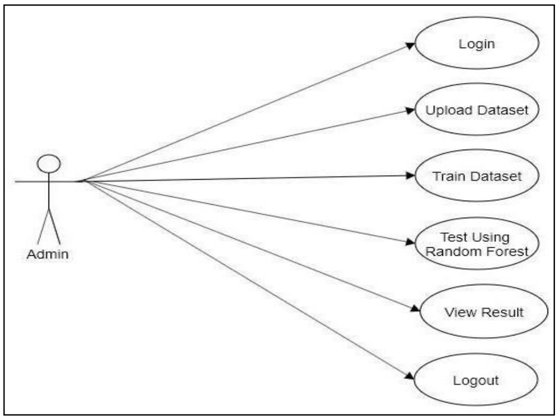


Fig 2: Use Case Diagram

As per the data and information we have gathered, we found that these following tasks must be carried out in order to get much accurate predictions. The tasks that we are going to carry out are as follows.

Data Pre-processing: The dataset we obtained is not completely accurate and error free. Hence, we will first carry out the following operations on it.

Data Cleaning: NA values in the dataset is the major setback for us as it will reduce the accuracy of the prediction profoundly so, we will remove the fields which does not have values. We will substitute it with the mean value of the column. This way, we will remove all the values in the data set.

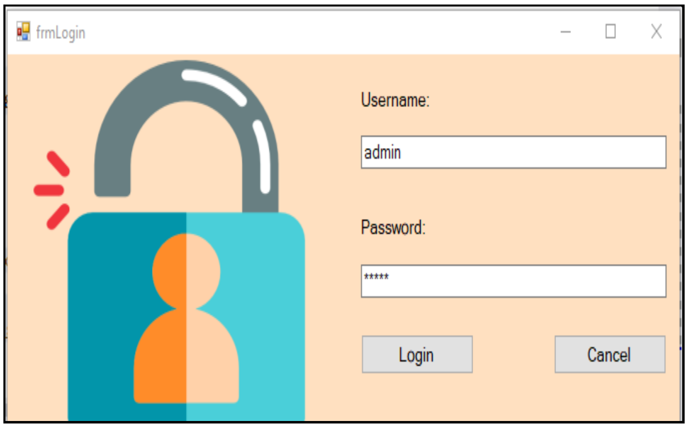
Feature Scaling: Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance. So we will scale the various fields in order to get them closer in terms of values. e.g. Age has just two values i.e. 0,1 and cholesterol has high values like 100. So, in order to get them closer to each other we will need to scale them.

Factorization: In this section, we assigned a meaning to the values so that the algorithm doesn't confuse between them. For example, assigning meaning to 0 and

1 in the age section so that the algorithm doesn't consider 1 as greater than 0 in that section.

Support Vector Machine: Support vector machine (SVM) is supervised learning method that analyzes data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories; an SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The points are separated based on hyper plane that separate them. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. In the project, we have used this algorithm to classify the patients into groups according to the risk posed to them based on the parameters provided. It was observed that: Naïve Bayes had 60% accuracy, logistic regression had 61.45% and SVM had 64.4%. Hence SVM was selected as the most efficient algorithm for the web application.

4. OUTPUT AND EXPLANATION



Username: admin

Password: *****

Fig 3: Login Page

This is the Login page of the project. User has to login first with specific Username and Password.

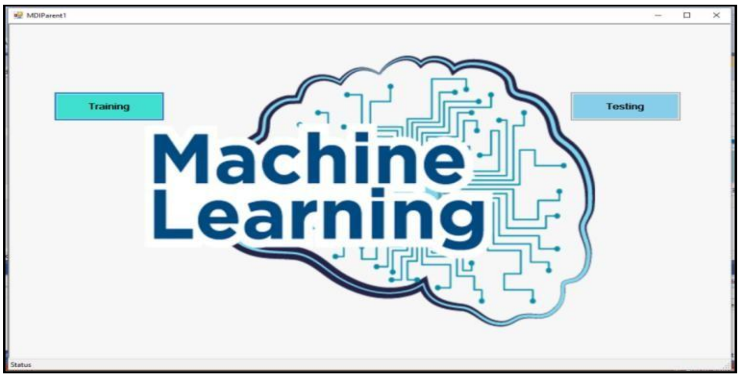


Fig 4: Training and Testing Home Page

After successfully login we see this page. This is the home page of our project. In that it shows the two buttons. One is Training and second one is Testing Buttons.

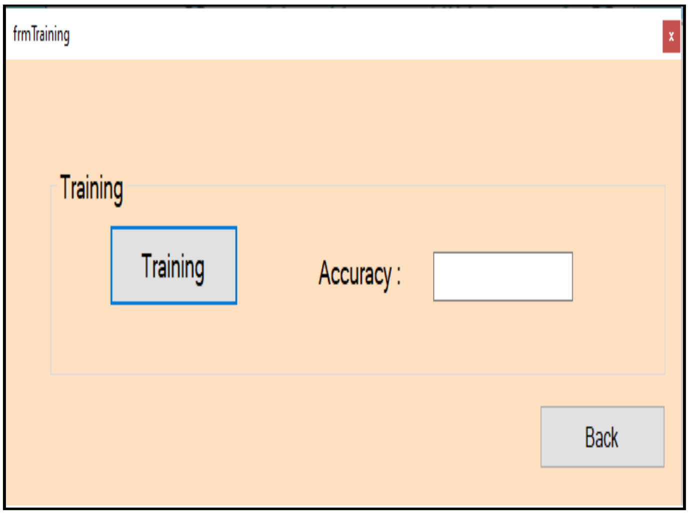
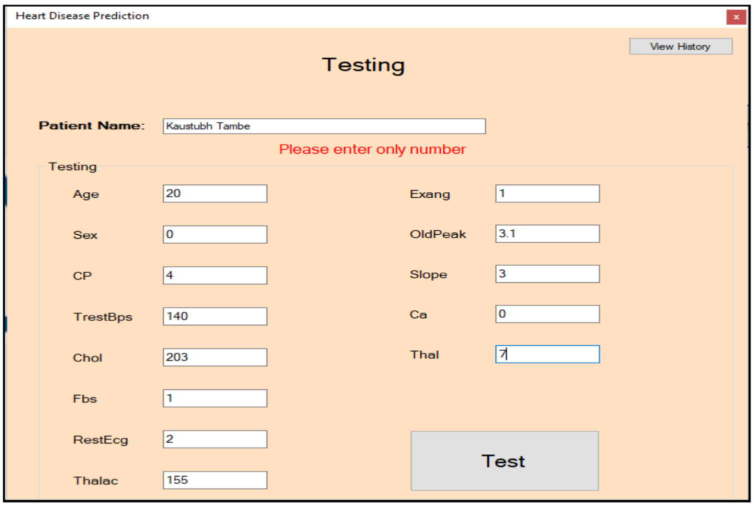


Fig 5: Training Form

After clicking on the traning button on the home page, this page will pop up which is Training form where we check the training accuracy of the data set.



Patient Name: Kaustubh Tambe

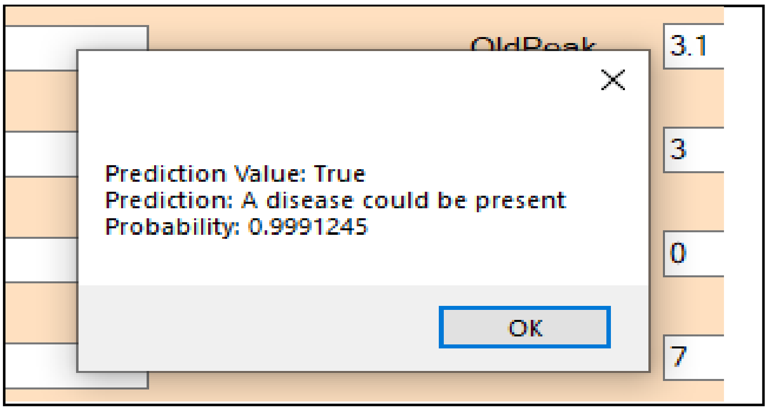
Please enter only number

Testing

Age	20	Exang	1
Sex	0	OldPeak	3.1
CP	4	Slope	3
TrestBps	140	Ca	0
Chol	203	Thal	7
Fbs	1		
RestEcg	2		
Thalac	155		

Fig 6 : Testing Form

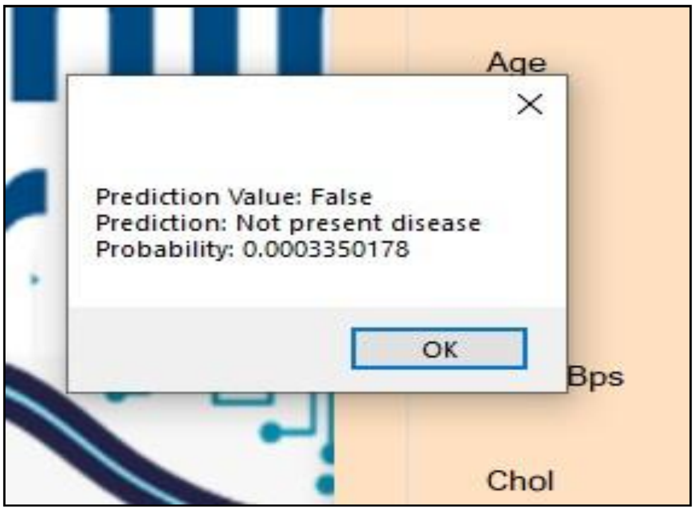
After clicking on testing button of home page this page will appear here user have to provide the certain information like age, cp etc



Prediction Value: True
Prediction: A disease could be present
Probability: 0.9991245

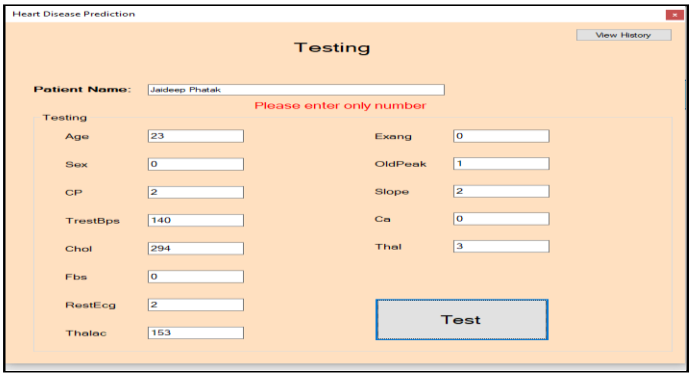
Fig 7: Output of Prediction 1

After providing all the values in the given text box the analyzed prediction depending upon the provided input the corresponding output with the probability will be provided. The above image shows the prediction of input provided.



Prediction Value: False
Prediction: Not present disease
Probability: 0.0003350178

Fig 8: Input for prediction 2



Patient Name: Jaideep Phatak

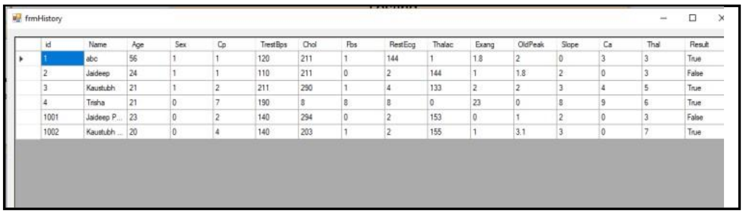
Please enter only number

Testing

Age	23	Exang	0
Sex	0	OldPeak	1
CP	2	Slope	2
TrestBps	140	Ca	0
Chol	294	Thal	3
Fbs	0		
RestEcg	2		
Thalac	153		

Fig 9: Output of Prediction i2

The above image shows the output of the prediction two. With the prediction value, Prediction and the Probability.


frmHistory

	id	Name	Age	Sex	Cp	TrestBps	Chol	Fbs	RestEcg	Thalac	Exang	OldPeak	Slope	Ca	Thal	Result
▸	1	abc	56	1	1	120	211	1	144	1	1.8	2	0	3	3	True
	2	Jaideep	24	1	1	110	211	0	2	144	1	1.8	2	0	3	False
	3	Kaustubh	21	1	2	211	290	1	4	133	2	2	3	4	5	True
	4	Trisha	21	0	7	190	8	8	8	0	23	0	8	9	6	True
	1001	Jaideep P...	23	0	2	140	294	0	2	153	0	1	2	0	3	False
	1002	Kaustubh ...	20	0	4	140	203	1	2	155	1	3.1	3	0	7	True

Fig 10: History

The above image shows the all patient history with all the input provided and the predicted result in the form of True or False.

5. CONCLUSION AND FUTURE WORK

In this project various suggestion and classification methods are executed on the heart datasets to predict the heart diseases. Classification algorithms are used to predict small set of relations between attributes in the databases to build an correct classifier. The main contribution of the present study to attain high calculation accuracy for early diagnoses of heart diseases. The proposed hybrid associative classification is implemented on spicy environment. Finally an skilled system is developed for the end user to check the risk of heart diseases on the basis of assumed parameters and the best associative classification method. The experimental results show that large number of the rules support to the better determines of heart diseases that even support the heart professional in their diagnosis in decision.

6. REFERENCES

[1] Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE, July 2015

[2] M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur and Kushal Ghosh "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design", September 2014.

[3] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin "Analysis of Data Mining Techniques for Heart Disease Prediction", May 2015.

[4] Soodeh Nikan, Femida Gwadry-Sridhar, and Michael Bauer "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis", IEEE, August 2016

[5] Sanjay Kumar Sen Asst. Professor, Computer Science & Engg. Orissa Engineering College, Bhubaneswar, Odisha – India." Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms" International Journal of Engineering and Computer Science. Volume 6 Issue 6, June 2017

[6] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja "Heart disease prediction using machine learning tech : A survey" International Journal of Engineering & Technology, 7 (2.8), April 2018.

[7] Heart Disease Dataset - https://www.kaggle.com/c/heart-disease dated: Sept 2018

[8] K. Srinivas, B. Kavihta Rani, A. Govrdhan "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attack" IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010.