Heart Disease Prediction System Using SVM

Sneha Dange[1], Priyanka Gaikwad[2], Rutuja Sheral[3], Pratiksha Shewale[4]
Prof. Dr. Sulochana Sonkamble[5]
[1,2,3,4.5]*JSPM Narhe Technical Campus, Pune*

Abstract—**Coronary Heart Disease (CHD) is the most common type of heart disease, killing over 370,000 people annually. Every year about 735,000 Americans has a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack. This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors.**

The healthcare domain is one of the prominent research fields in the current scenario with the rapid improvement of technology and data. It is difficult to handle the huge amount of data of the patients. There are a lot of procedures for the treatment of multiple diseases across the world. Machine Learning is an emerging approach that helps in prediction, diagnosis of a disease. This system depicts the prediction of disease based on symptoms using machine learning. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output. Machine Learning algorithms such as Support Vector Machine, is employed on the provided dataset and predict the disease. Its implementation is done through the python programming language. The research demonstrates the best algorithm based on their accuracy. The accuracy of an algorithm is determined by the performance on the given dataset.

Index Terms— **heart disease, machine learning, real-time, support vector machine**

I. INTRODUCTION

At present, when one suffers from particular disease, then the person has to visit to doctor which is time consuming and costly too. Also, if the user is out of reach of doctor and hospitals it may be difficult for the user as the disease cannot be identified. So, if the above process can be completed using an automated program which can save time as well as money, it could be easier to the patient which can make the process easier. There are other Heart related Disease Prediction System using data mining techniques that analyzes the risk level of the patient. Heart Disease Predictor is a web-based application that predicts the heart disease of the user with respect to the symptoms given by the user. Heart Disease Prediction system has data sets collected from different health related sites. With the help of Heart Disease Predictor, the user will be able to know the probability of the disease with the given symptoms. As the use of internet is growing everyday, people are always curious to know different new things. People have access to internet than hospitals and doctors. People do not have immediate option when they suffer with particular disease. So, this system can be helpful to the people as they have access to internet 24hours.

II. LITERATURE REVIEW

In Predicting presence of heart disease using Machine Learning research paper by Chirag Sharma, Niriksha Shetty, Amol Shinde, Prof. Dhanashri Bhapatra Heart disease is most common now a days and it is a very serious problem. Machine learning provides a best way for predicting heart disease. The aim of this paper is to develop simple, light weight approach for predicting presence of heart disease using Machine learning. Machine learning can be implemented in heart disease prediction. In this paper different machine learning techniques have been used and it compares the result using various performance metrics. This study aims to perform comparative analysis of heart disease detection using publicly available dataset collected from UCI machine learning repository. There are various datasets available such as Switzerland dataset, Hungarian dataset and Cleveland dataset. Here Cleveland dataset is used which is having 303 records of patients along with 14 attributes are used for this study and testing. These datasets are pre-processed by removing all the noisy and missing data from the dataset. And then the pre-processed dataset is used for analysis. In this study six different machine learning techniques were used for comparison based on various performance metrics. Then the one with a good accuracy is taken as the model for predicting the heart disease. A GUI is developed for the prediction of heart disease.

In Real time machine learning for early detection of heart disease using big data approach research paper by Abderrahmane Ed-daoudy and Khalil Maalmi Over the last few decades, heart disease is the most common cause of global death. So early detection of heart disease and continuous monitoring can reduce the mortality rate. The exponential growth of data from different sources such as wearable sensor devices used in Internet of Things health monitoring, streaming system and others have been generating an enormous

amount of data on a continuous basis. The combination of streaming big data analytics and machine learning is a breakthrough technology that can have a significant impact in healthcare field especially early detection of heart disease. This technology can be more powerful and less expensive. To overcome this issue, this paper proposes a real-time heart disease prediction system based on apache Spark which stand as a strong large-scale distributed computing plat-form that can be used successfully for streaming data event against machine learning through in-memory computations. The system consists of two main sub parts, namely streaming processing and data storage and visualization. The first uses Spark MLlib with Spark streaming and applies classification model on data events to predict heart disease. The seconds uses Apache Cassandra for storing the large volume of generated data.

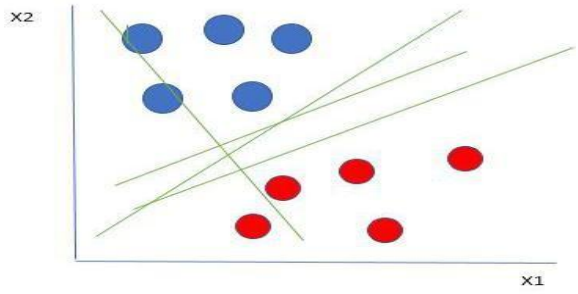
In Application of Machine Learning in Disease Prediction research paper by Pahulpreet Singh Kohli,Shriya Arora. The application of machine learning in the field of medical diagnosis is increasing gradually. This can be contributed primarily to the improvement in the classification and recognition systems used in disease diagnosis which is able to provide data that aids medical experts in early detection of fatal diseases and therefore, increase the survival rate of patients significantly. In this paper, we apply different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The results of the study strengthen the idea of the application of machine learning in early detection of diseases.

In Efficient Heart Disease Prediction System using Decision Tree by Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma Cardiovascular disease (CVD) is a big reason of morbidity and mortality in the current living style. Identification of Cardiovascular disease is an important but a complex task that needs to be performed very minutely, efficiently and the correct automation would be very desirable. Every human being cannot be equally skillful and so as doctors. All doctors cannot be equally skilled in every subspecialty and at many places we don't have skilled and specialist doctors available easily. An automated system in medical diagnosis would enhance medical care and it can also reduce costs. In this study, we have designed a system that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health. The rules can be prioritized based on the user's requirement. The performance of the system is evaluated in terms of classification accuracy and the results shows that the system has great potential in predicting the heart disease risk level more accurately.

III. ALGORITHM

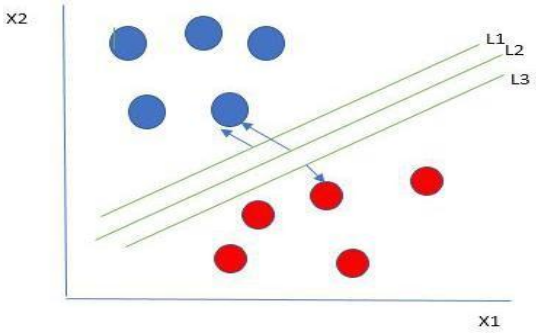
Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Let's consider two independent variables x1, x2 and one dependent variable which is either a blue circle or a red circle.



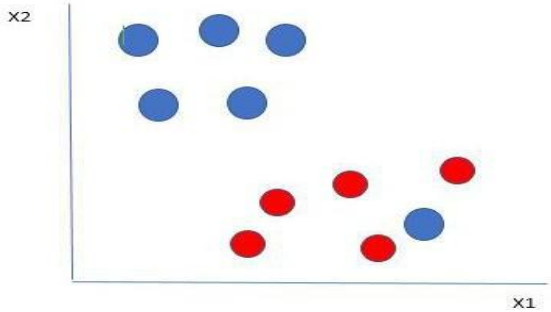
From the figure above its very clear that there are multiple lines (our hyperplane here is a line because we are considering only two input features x1, x2) that segregates our data points or does a classification between red and blue circles. So how do we choose the best line or in general the best hyperplane that segregates our data points.

Selecting the best hyper-plane: One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes



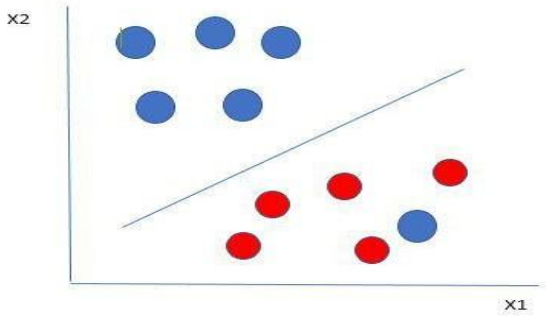
So, we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin. So, from the above figure, we choose L2.

Let's consider a scenario like shown below:



Here we have one blue ball in the boundary of the red ball. So how does SVM classify the data? It's simple!

The blue ball in the boundary of red ones is an outlier of blue balls. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers.



IV. METHODOLOGY

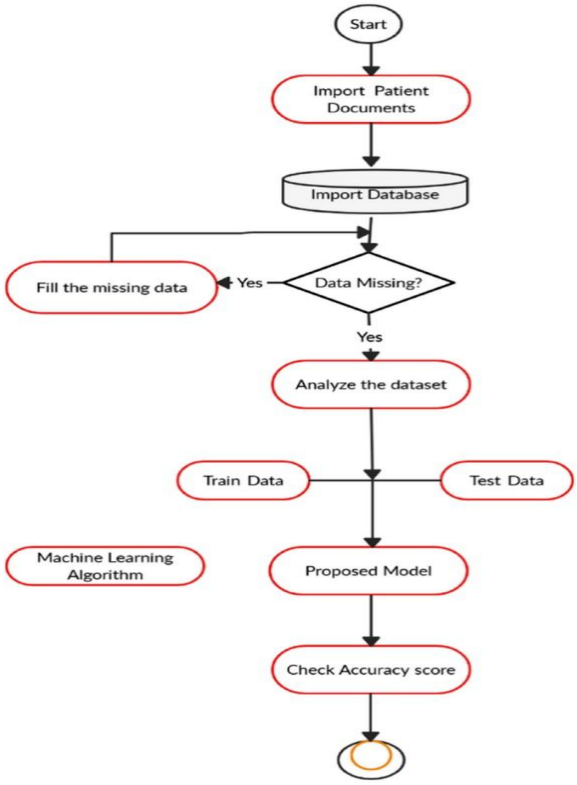


Fig. 1 Flow of system

A. Select Dataset

A data set (or dataset) is a collection of data which is usually presented in tabular form. There are many datasets available at the UCI machine learning repository. Some of the datasets are Hungarian dataset, Switzerland dataset and Cleveland dataset. The attributes are as follows–

Table 1. Attributes

Sno	Attributes	Description
1	Age	Age in years
2	Sex	Male or Female
3	Cp	Chest pain type
4	Thestbps	Resting blood pressure
5	Chol	Serum cholesterol
6	Restecg	Resting electrographic results
7	Fbs	Fasting blood sugar
8	Thalach	Max. heart rate achieved
9	exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Ca	No. of major vessels colored
13	Thal	Defect type

B. Data Preprocessing

It is a process of removing all the noisy and missing data from the data set.

C. Train and build machine learning model for heart disease detection

In this step the dataset is divided into two parts: training dataset and testing dataset. Training dataset contains 60% and testing dataset contains 40% which are selected randomly.

D. Input Details

User input details such as Age, sex, cp, Trestbps, Chol, Fbs, Exang, Thalach, old peak, slope, ca, thal, restecg, class are the 13 attributes

E. Prediction of heart disease

A GUI is developed in python by using tkinter to generate a simple dialog box which takes input for all the values necessary for evaluation. After the input is taken from the user, prompt appears which decides whether a person has a presence of heart disease or not.

V. EXPERIMENTATION RESULTS AND DISCUSSION

For the evaluation process, confusion matrix, accuracy score, precision, recall, sensitivity, and F1 score are used. A confusion matrix is a table-like structure in which there are true values and predicted values, called true positive and true negative. It is defined in four parts: the first one is true positive (TP) in which the values are identified as true and, in reality, it was true also. The second one is false positive (FP) in which the values identified are false but are identified as true. The third one is false negative (FN) in which the value was true but was identified as negative. The fourth one is true negative (TN) in which the value was negative and was truly identified as negative. The table is shown in Figure 8.

		Predicted value: P	Predicted value: N
True value	P	TP	FN
	N	FP	TN

Fig. 2 Confusion Matrix

In Figure 2, P = positive, N = negative, TP = true positive, FN = false negative, FP = false positive, TN = true negative.

Then for checking how well a model is performing, an accuracy score is used. It is defined as the true positive values plus true negative values divided by true positive plus true negative plus false positive plus false negative. The formula is:

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

After accuracy there is specificity which is the proportion of true negative cases that were classified as negative; thus, it is a measure of how well a classifier identifies negative cases. It is also known as the true negative rate. The formula is:

Specificity=TN/(TN+FP)

Then there is sensitivity in which the proportion of actual positive cases got predicted as positive (or true positive). Sensitivity is also termed as recall. In other words, an unhealthy person got predicted as unhealthy. The formula is:

Sensitivity=TP/(TP+FN)

VI. CONCLUSION

For this project we have selected 13 attributes from Cleveland UCI library. In this system we are predicting Heart Disease using Support Vector Machine (SVM) algorithm. This system will seek medical attention as soon as possible to avoid missing treatment time. This application will help in faster diagnosis, also reduces medical error and easy to apply. But this project should only be considered

as an assistive tool and must be completely relied upon only after it has been validated in the future by integrating more practical data. This system successfully predicts heart disease with 98.50% accuracy score.

REFERENCES

[1] International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 04 | Apr 2021

[2] Abderrahmane Ed-daoudy, Khalil Maalmi, "Real-time machine learning for early detection of heart disease using big data approach", IEEE, 2019

[3] Anil Parajuli, "Disease Predictor", Project Report, August 2016

[4] Aditi Gavhane, Gauthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), "Prediction of heart disease using machine learning", International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record

[5] Purushottam, Prof. Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System using Decision Tree", International Conference on Computing, Communication and Automation (ICCCA2015)

[6] Akash Chahuan, Aditya Jain, Purushottam Sharma, Vikas Deep "Heart Disease Prediction using Evolutionary Rule Learning", International Conference on "Computational Intelligence and Communication Technology" (CICT 2018)

[7] C.Sowmiya, Dr. P. Sumitra, "Analytical Study of Heart Disease Diagnosis Using Classification Techniques", IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing 2017

[8] Sana Bharti, Dr. Shaliendra Narayn Singh, "Analytical Study of Heart Disease Prediction Comparing with Different Algorithms", International Conference on Computing, Communication and Automation (ICCCA2015)

[9] Monika Gandhi, Dr. Shaliendra Narayn Singh, "Predictions in Heart Disease Using Techniques of Data Mining", 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015)

[10] Aswathisasidharan 2021 Jan 22. Support Vector Machine Algorithm [English]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[11] Garima Choudhary, Shailendra Narayan Singh, "Prediction of cardiovascular disease using data mining technique", International Conference on information systems and computer networks, IEEE, 2019.

[12] Pranav Motarwar, Ankita Duraphe, G Suganya, M Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning", International Conference on Emerging Trends in Information Technology and Engineering,2020. © 2021 IJRAR December 2021, Volume 8, Issue 4 www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138) IJRAR21D2105 International Journal of Research and Analytical Reviews (IJRAR) www.ijrar.org 34

[13] Rahul Katarya, Polipireddy Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey", Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE.

[14] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", International Conference on Electrical and Electronics Engineering, 2020.

[15] AliveKor, [Online]Available: https://www.alivecor.com/how-it-works

[16] Abderrahmane Ed-daoudy and Khalil Maalmi, "Real time machine learning for early detection of heart disease using big data approach", IEEE, 2019