Machine Learning Based Heart Disease Prediction

Ambika Sekhar
Department Of Electronics and Communication
Sree Buddha College of Engineering,Pattoor
Alappuzha,India
ec.ambikas@sbcemail.in

Amrutha Babu
Department of Electronics and Communication
Sree Buddha College of Engineering,Pattoor
Alappuzha,India
amrutha16111999@gmail.com

Jayalekshmi V.K.
Department of Electronics and Communication
Sree Buddha College of Engineering,Pattoor
Alappuzha,India
lekshmivijayan111@gmail.com

Adithya Udayan
Department of Electronics and Communication
Sree Buddha College of Engineering,Pattoor
Alappuzha,India
adithyaudayan245@gmail.com

2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS) | 978-1-6654-6792-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICNGIS54955.2022.10079736

***Abstract*— In most health related industry, large types of data are frequently generated. Machine learning algorithms help to check risk for heart disease from person's data. Analysing ECG signal at initial stage helps to detect and prevent heart disease. Machine learning algorithms like KNN, RF, SVM and DT are used to make use of tremendous data to predict disease earlier. From the classification results KNN algorithm best predicted the disease with an accuracy of 96%. Then the model is deployed as a web application to make available it to users. By using the model mortality rate due to heart disease can be reduced by providing an option to get better treatment as early as possible**.

Keywords— KNN, Random Forest, SVM, Disease prediction

I. INTRODUCTION

Health is the supreme and long lasting wealth that everyone needs. So monitoring health has a major role in everyone's day to day life. Diagnosing health diseases at the early stages can help to prevent future complications with proper treatment. Equipment like CT, MRI, PET, and others may quickly detect problems inside the body or beneath the skin. Less common diseases such as heart stroke, heart attack can be avoided at the early stages easily if it is possible to diagnose it at early stage.In diseases, heart disease is the most commonly occurring disease and it is main cause of sudden death nowadays. Unawareness of t symptoms of disease of heart is the main cause of death and other medical complications. In India there are almost three crore heart patients and 2 lakh open heart surgeries are done in every year. Mortality rate all over the world is nearly 17.3 million people every year. Early heart disease prediction is essential to reduce the mortality rate. Early diagnosis pave the path for early treatment thereby mortality rate is reduced. Since huge volume of data is available in today's era. Due to availability of huge in biomedical and healthcare communities, early disease prediction by accurate study of medical data patient care and community services is possible.. In this scenario there is a huge need of a disease prediction system that predicts disease at home.

Review can be done using Machine learning methodologies which process large volumes of data with high accuracy and efficiency. Various supervised machine learning techniques find the hidden pattern in data during training and helps to predict the presence or absence of heart disease when a new data input comes.

C.Youn ,M. Chen, Y.Li,D.Wu, and Y.Zhang with a wearable 2.0 system [3]. Smart washable clothing is present in system. They believed that this approach can further improve the QoS and QoE of the future generation health care system. Chen worked in the area of IoT based data collection system. This work helped him to invent a new sensor based smart washable cloth. As a result, doctors find it easy to capture the physiological conditions of the patient. The main issue hidden in the existing system is later discovered. They are negative psychological effects, sustainable big psychological data collection etc. Y.G Jiang, B.Qian,X.Wang, and N.Cao,H.Li, and proposed and designed a risk prediction system and its corresponding model by using the help of patient data. The actual issue faced by the patient is solved [4]. Nuzhat F. Shaikh ,Ajinkya Kunjir and Harshal Sawant suggested a better clinical decision making system. Based on historical data collected from the patients, diseases are predicted. They use pie charts and 2D/3D graphs for visualization purposes [5]. C.Dharuman,S.Leoni Sharmila, and P.Venkitesan put forwarded a comparison of different types of machine learning techniques like Fuzzy logic ,decision tree and Fuzzy Neutral Network. Apart from other machine learning algorithms Fuzzy Neutral Network results an accuracy of 91% in classifying the liver disease data set [7]. The limiting factor of this paper is that they could not use large data set. Medical data is growing in a tremendous manner. So it is necessary to classify those data is considered a challenging one. CNN-MDRP algorithm was proposed byShraddha Subhash Shirsath for predicting diseases. Here she used a large volume of unstructured and structured hospital data. CNN-UDRP used only structured data with the help of machine learning algorithm [6]. But in case of CNN-MDRP, it checked on unstructured and structured data. So that prediction process was fast as compared to CNN – UDRP. Still they are using bigger data is so challenging. Ramandeep Kaur, Er.Prabhsharn Kaur said that the data set may contain unnecessary, duplicate information. In such a situation, all the data should undergo a preprocessing technique to achieve better results [8].

978-1-6654-6792-6/22/$31.00 ©2022 IEEE

II. WORKING OF THE PREDICTION SYSTEM

By analyzing the dataset, the research hopes to predict whether the individual is at risk of heart disease. On a data collection containing patient data, machine learning methods will be used to make this prediction. We use real-world hospital data to test the prediction model. Heart Disease data is gathered from the kaggle website and used as an input data in this model. The architecture of the prediction system is depicted in Figure 1. Numerous sorts of data, mostly structured, semi-structured, or unstructured, can be gathered from various sources, such as hospitals. After the data is acquired, it is cleaned to eliminate any missing values and to maintain the same level of granularity. After that, the cleaned data is subdivided into two categories: test data and training data .Using different algorithms the system is trained and using test data testing of the prediction system is done. The algorithm with highest predictive accuracy is selected, and then the model can be deployed.

.

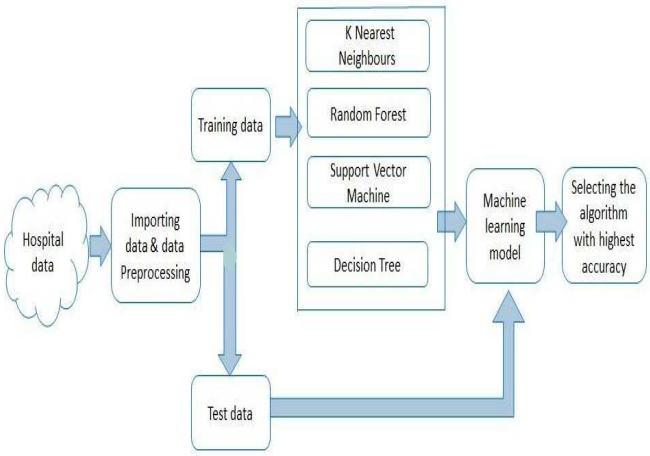

Fig. 1. Architecture of prediction system

The language used for making the machine learning model is Python.Importing libraries is the first step required for high performance calculation, data visualization and data model analysis.

The Heart Disease dataset from the UCI repository was used in our experiment. The information is used to predict whether or not a person has Cardiovascular disease. are included in the Dataset consist of total of 303 individuals' data. The dataset has 14 columns, 5 of which have numerical values and 9 of which have category values.

TABLE I. DATASET BALANCE CHECK

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	1	14	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	129	1	2.6	2	2	7	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	0
37	1	3	130	250	0	0	187	0	3.5	3	0	3	1

A sample of the dataset is shown in table1:

The attributes in the datasets are

A. Age

The threat of cardiovascular disease is approximately is very highin todays scenario. Reports shows that 82 percent of people who die due to coronary heart disease are 65 or older. Also, after age 55 risk of stroke gets doubled in every decade.

B. Persons Sex

Another essential component is the person's sex. Men have a higher chance of heart problems than women. However, a risk in woman is the same as a man's after menopause. In addition, females with diabetes have a higher risk of developing heart disease than males with diabetes.

C. Angina

Indicates the type of chest discomfort a person is experiencing as a result of a less O_2 rich blood to the heart muscle.

D. Resting Blood Pressure

Gives measure of a persons blood pressure in millimeters of mercury (mmHg) (unit).

E. Serum Cholestrol

Shows the serum cholesterol in milligrams per deciliter (unit). Bad" cholesterol or Low-density lipoprotein (LDL) cholesterol narrows arteries.

F. Fasting Blood Sugar

Shows the outcome of a comparison of a person's fasting blood sugar levels with 120mg/dl. Body's7. If enough insulin is not produced by the pancreas or does not respond to insulin adequately, blood sugar levels rise. This also raises your chances of having a heart attack.

G. Resting ECG

Displays the results of resting electrocardiograph. ECG records the electrical activity of heart at rest, an abnormal ECG can signal a medical emergency, like heart attack.

H. Max heart rate achieved

Displays an individual's maximum heart rate. The risk of cardiovascular disease increases as the heart rate rises.

I. Exercise induced angina

Displays if chest pain occurs or not while doing exercise. Symptom of serious conditions, like attack in heart and less serious issues like muscle strains and asthma can be revealed by chest pain while exercise.

J. Peak exercise ST segment

The presence of ST segment changes like depression or elevation on the ECG suggests presence of Coronary Artery Disease (CAD).

K. Fluoroscopy-colored main vessels(0-3)

presents the value as float or integer.

L. Thal

Showss the presnce of thalassemia

M. Heart disease diagnosis

Provides information on whether or not the individual has heart disease.

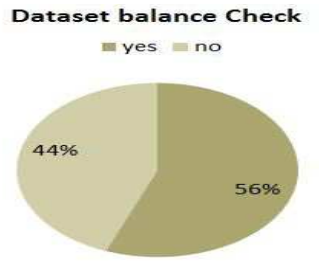

Fig. 2. Dataset balance check

As seen in Figure 2, 56 percent of the individuals in the dataset have cardiac disease. Heatmaps are coloured maps that assist visualise data in a two-dimensional format. Figure 3 depicts a heatmap demonstrating correlations between dataset attributes as well as how they interact with one another. The type of chest pain (cp), exercise-induced angina (exang), and ST depression generated by exercise relative to rest can all be seen on the heatmap. Heatmaps are coloured maps that assist visualise data in a two-dimensional format. Figure 3 depicts a heatmap demonstrating correlations between dataset attributes as well as how they interact with one another. We can see from heatmap that the chest pain type (cp), Exercise-induced angina (exang), exercise-induced ST depression (oldpeak), and the slope of the peak exercise ST segment are all terms used to describe the number of main vessels coloured by fluoroscopy (0–3), exercise-induced angina (exang), exercise-induced angina (exang), exercise-induced angina (exang), exercise-induced angina (exang), exercise-induced angina (exang), exercise-induced angina (exang), exercise-induced angina (exang (slope),and thalassemia (thal) are all substantially connected with heart disease (target). We also see that cardiac illness and maximum heart rate are inversely proportional.

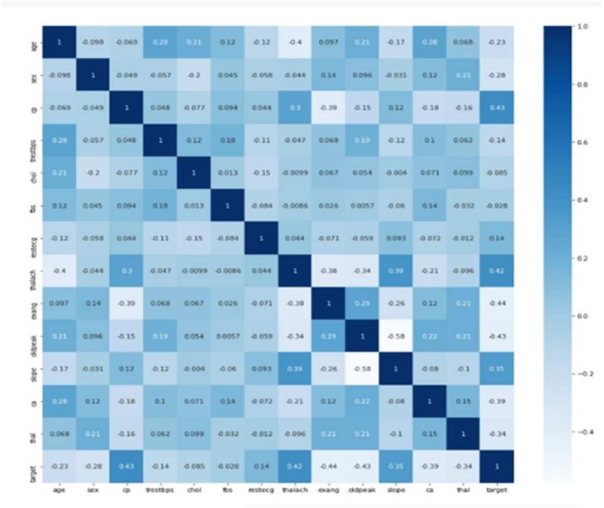

Fig. 3. Heatmap

The raw data is converted into a clean data set using data preparation. When data is acquired from many sources, it is usually in raw format, which is often partial or missing, and may contain inaccuracies. To avoid this, data is pre-processed. First, all null or incorrect values in the dataset were removed during data processing. After that, a training set and a test set are created from the cleaned dataset. The dataset is divided into 80:20 or 70:30 or ratios, meaning that 80% or 70% of the data is used for training model and 30% or 20% of the data to test the model. The last step in the preparation of data for machine learning is feature scaling.This technique is chosen to standardize the independent variables of a dataset within a particular range. Feature scaling can perform in two ways: Standardization or Normalization. Here we use the standardization method.

III. METHODOLOGY

Algorithms used for classification are listed below

A. K-Nearest Neighbours(KNN)

The Supervised Learning approach is used in K-Nearest Neighbour algorithms. The K-NN algorithm assigns the new data to the group which has similarity to the existing categories and compares the new data to existing data. [2].

B. Random Forest Algorithm

One technique used for supervised learning is Random Forest which follows ensemble learning, which is the process of using numerous classifiers to improve the model's performance by solving complex problem.

C. Support Vector Machine Algorithm

One most commonly used Supervised Learning approaches is Support Vector Machine (SVM). The SVM method is used to find the best line or decision boundary for dividing n-dimensional space into classes so that subsequent data points can be placed in the right category easily.

D. Decision Tree Algorithm

A supervised learning technique is the Decision Tree. In this tree-structured classifier branches check decision rules, internal nodes illustrate properties of dataset and each leaf node gives the conclusion.

IV. RESULTS AND IMPLEMENTATION

Out of 303 entries in the dataset, 70% of the data is taken for training and 30% is used to test the system. For KNN algorithm, generated error rate for different values of K (figure4) on which for K=11 we got 96% accuracy

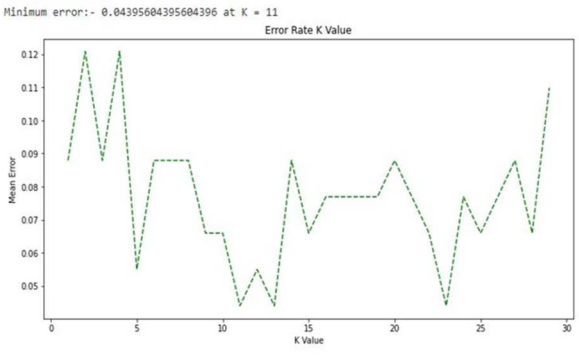

Fig. 4. Error rates for different values of K

For random forest algorithm we calculated the scores for different values of n_estimators. From the figure 5, n_estimators of 12 the random forest classifier has the highest score.

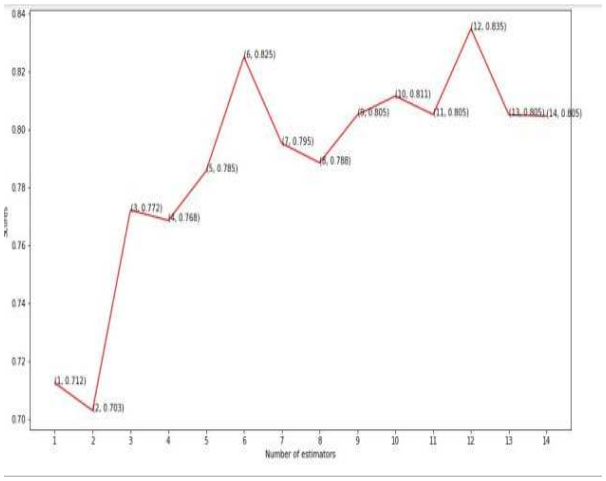

Fig. 5. Scores of Random forest classifier for different values of estimators

From the line graph for decision tree algorithm the maximum score (figure 6) for max_features of 9.

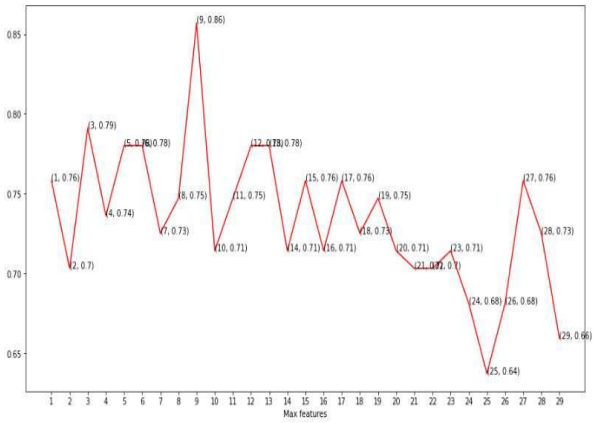

Fig. 6. Scores of Decision Tree Algorithm

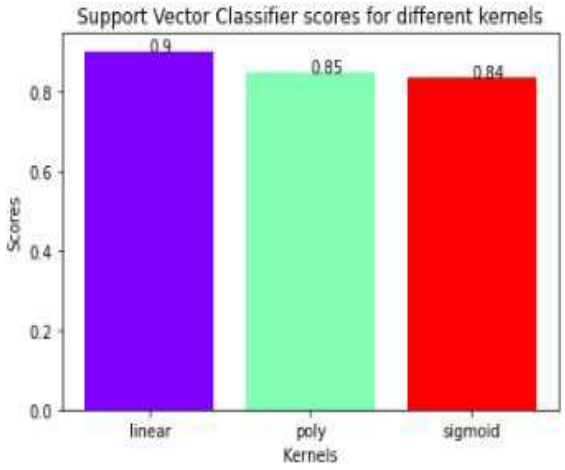

Fig. 7. Scores of Support Tree Vector for different Kernels

Proposed machine learning model trained and tested using different algorithms with the values providing highest accuracy and generated classification reports. Based on precision, F1 score and recall accuracy, the tested algorithms are compared to choose the right one

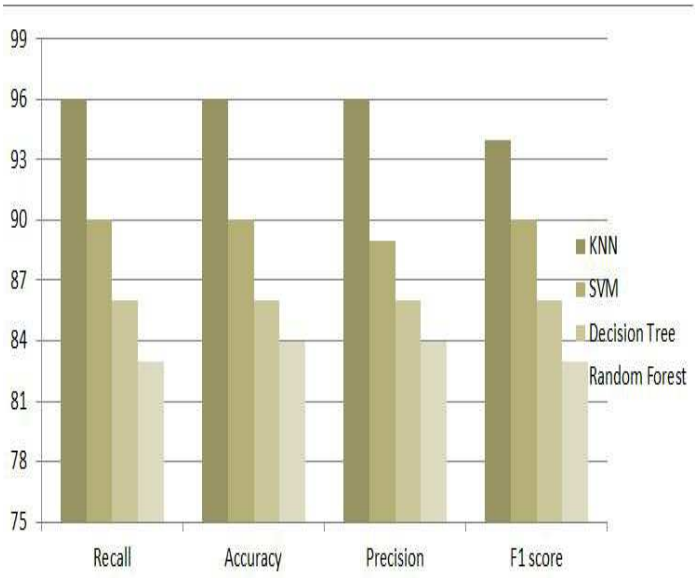

Fig. 8. Comparison of algorithms based on Recall,Accuracy,Precision and F1 score

From the figure 8,its proven that the KNN algorithm has highest value of F1 score, precision, accuracy, , recall and for K value of 11.Figure 8 represents our developed web application for predicting heart disease based on KNN algorithm.

A. Steps in creating web app using flask library

- The Python pickling of a machine learning model for heart disease prediction has been completed. After processing the data, the method with the highest accuracy is chosen as the best. The pickle paradigm in Python is used to serialise and de-serialize Python object structures. Object is stored to disc using Python. The pickle file in Python contains all of the information needed to recreate the object in another script
- Loading the pickle file into the python script to predict the disease: After model is built, the pickle file is loaded into the python script.
- To enter the user inputs, a form is built. Users must supply certain information as input throughout the forecast phase, as shown in Fig.9. so that the test result can be predicted by our web application
- The data values entered by the user are supplied to machine learning model.
- The predicted results are provided into the HTML file for displaying it to the user.

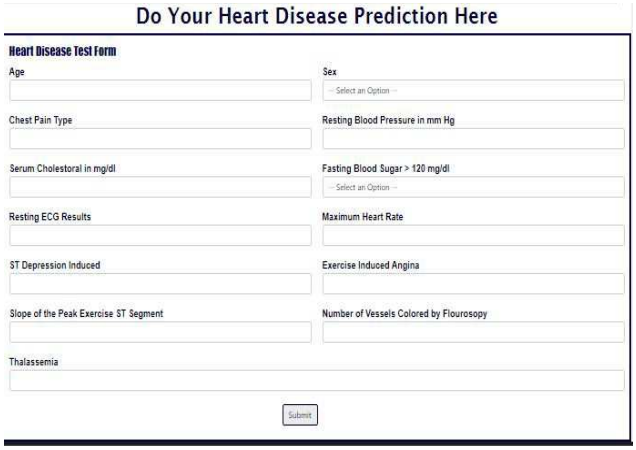

Fig. 9. Developed web application

Figure 9 shows the form for supplying patient details for prediction of disease. Here we have used python programming language and flask library for making the web app.

V. CONCLUSION

Using patient data the machine learning based disease prediction model predicts disease. Various machine learning algorithms helped to make use of the tremendous data available in the medical field to predict disease earlier. The dataset is taken from UCI repository. Initially, the classification algorithms execute the training process which uses the dataset to study predicting the disease .A comparison on the accuracy for particular data set was performed. With the study it is inferred that KNN have highest accuracy of 96% out of all models. So that KNN is selected for prediction of heart disease.

REFERENCES

[1] WHO (World Health Organization): Cardiovascular Diseases - https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[2] Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, MasoudiFA, Spertus JA, Krumholz HM , "Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction", JACC : Heart Failure, vol. 8, Issue 1, January 2020

[3] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017

[4] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction", Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[5] AjinkyaKunjir, HarshalSawant, NuzhatF.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare", IEEE big data analytics and computational Intelligence, Oct 2017 pp.2325.

[6] ShraddhaSubhashShirsath ,"Disease Prediction Using Machine Learning Over Big Data", International Journal of Innovative Research in Science, Vol. 7, Issue 6, June 2018.

[7] S.LeoniSharmila, C.Dharuman and P.Venkatesan,"Disease Classification Using Machine Learning Algorithms -A Comparative Study", International Journal of Pure and Applied Mathematics ,Volume 114 No. 6 2017, 1-10

[8] RamandeepKaur, Er. PrabhsharnKaur."A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques ",(June2016)

[9] KashviTaunk; Sanjukta De; SrishtiVerma; Aleena Swetapadma," A Brief Review of Nearest Neighbor Algorithm for Learning and Classification ", IEEE Access,7 ,1718- 1735, 15-17 May 2019

[10] Martin Gjoreski; Anton Gradišek; BorutBudna; Matjaž Gams; GregorPoglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure from Heart Sound" , IEEE Smart World,19, 85714-85728, 23 January 2020.

[11] Jiaming Chen; Ali Valehi, AbolfazlRazi, "Smart Heart Monitoring, Early Prediction of Heart Problems Through Predictive Analysis of ECG Signals", IEEE Access , 120831 – 120839 , 2019.