International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Cardiovascular Disease Prediction Using Machine Learning Techniques

Divya Annepu[1], Gowtham G[2]

[1,2]B.Tech. Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

---

Abstract - *A In recent years, the leading cause of death for both men and women is cardiovascular disease. In India, the death tolls to 24.8% due to heart attacks. Proactive predication of risk of heart diseases will mitigate the situation to a great extent. This can be achieved by automating the prediction of heart diseases by saving time and effort. The recent development in medical supportive technologies based on data mining, machine learning plays an important role in predicting heart diseases. All the available useful medical information of patients is collected, well-structured and trained into the dataset. For better understanding of medical data to prevent heart diseases, mining is used to identify patterns which are hidden and previously not discovered due to unknown relationships. In this project, we achieved heart disease prediction using python, Random Forest classifiers are used which lead to an accuracy of 97.56%.*

Key Words: **Heart Disease, Random Forest, Classification, Datasets, User Entries, Prediction**

1.INTRODUCTION

Hospitals are generating huge volume of data regarding their patients. With the advancement made in data analysis over big data, the hospital data is useful in building disease predictive models. Data mining techniques can predict the hidden pattern lying in the voluminous hospital data and helps us to build an effective medical diagnosis system. One or other form of heart disease is found to be the major reason for the death of a patient [1]. Irrespective of the region, country and age group, heart disease is the leading death factor. Heart related diseases needs continuous monitoring and treatment based on it. But for rural people frequent medical check-ups are not easily accessible and viable. For the people who are suffering from serious heart disease this condition is a life threatening situation. According to a 2010 survey, for every $6 spent on health care, $1 is for Cardiovascular diseases. Coronary Heart Disease(CHD) is the leading cause of death of 370,000 people annually. But the cost associated with their treatment is estimated to be around $444 billion(US) dollars. The chances of survival are more when it is predicted before an emergency situation occurs. Also, it is observed from the data that the survival of sudden out of hospital heart attack is very low. This paper surveys the different kinds of heart diseases predictive modeling developed based on machine learning, data mining and artificial intelligence techniques.

There are various range of heart diseases apart from heart attack which are collectively called as Cardiovascular diseases. There are many reasons for the development of heart diseases such as smoking, blood sugar, obesity, depression, high

cholesterol, poor diet and genetically descendant. There are many types of heart diseases such as angina, arrhythmia, congenital heart disease, fibrillation, coronary artery disease, heart failure, fibrillation. When a person is under heart attack, the tests to be done are CPR, bypass surgery, Value disease treatment, Cardio, Pace makers, heart transplant and so on. The prediction of heart diseases helps us in treating the patient before the patient reaches heart failure.

Most common heart diseases are as follows[2]:

Angina: A part of heart muscles does not get proper supply of oxygen and nutrients. The main reason may be the muscle spasms in arteries due to cholesterol accumulation in its path.

Coronary Artery disease: When there's no supply of enough oxygen and blood in the coronary arteries, the condition leads to the coronary artery disease. It happens without any warning.

Heart Attack: A part of heart muscle get damaged or die which results in blockage of blood flow. It is reversible which is why it is important to immediate medical help.

Heart Failure: When heart doesn't pump enough blood to meet with the body's requirements the heart failure occurs. It indicates that heart is not squeezing as it should.

Arrhythmia: Irregular heartbeats such as slow, fast and skipping of beats due to irregularity of heart's electrical system. Improper sequence in the heart's the electrical system is known as Arrhythmias.

2. LITERATURE SURVEY

A brief survey is done over the existing works done for heart disease prediction using data mining and machine learning techniques.

Latha et al [3] have developed a neuro-fuzzy based heart disease prediction model named Co-Active Neuro Fuzzy Inference System(CANFIS). They used the Cleveland data source which has 13 attributes related to heart. Further, to optimize the membership functions, momentum coefficient values, learning rate Genetic Algorithm is used.

Martin Gjoreski [4] have implemented a chronic heart failure detection system using heart sounds. The proposed method involves filtering, segmentation, feature extraction and stacking of ML classifiers. The data set is collected from a total of 152 heart sounds obtained from 122 different subjects out of which 23 were previously diagnosed with heart abnormalities by the physician. The accuracy obtained from the proposed technique was 96% and in addition, it detects 87% of "unhealthy" instances with a precision of 87% . The heart sounds which were collected with the help of digital stethoscope proves that chronic heart failure can be detected.

I Ketut et al [5] proposed heart disease diagnosis system using K-Nearest Neighbors. A real clinical medical record of 450 data sets consisting 15 important parameters such as sex, age, chest pain, shortness of breath and so on were selected. The expected result for the system was to predict what type of heart disease the patient was suffering. The results out of experiments shows that KNN was the best algorithm with 75.11% accuracy (without parameter) and 74.89%, 74.44% and 73.11% accuracy with parameter weighting with OneR Attribute Evaluator, SVM Attribute Evaluator, and ReliefF Attribute Evaluator, respectively. The other classifier Naïve Bayes yields only 50.44% and SVM achieved 45.11% accuracy.

Abdallah et al [6] have created a platform which consists of an electronic device connected with a smart phone for acquiring the electrocardiogram(ECG) signals around the clock. In addition, inter-beat(R-R) interval analysis instantly alerts the pre-programmed emergency service number whenever there was an abnormal values in signals more than the threshold. Then the patient's ECG report, heart rate and patient's location was sent to the nearby doctor by the system. The hardware system consists of three leads, a micro controller and a smartphone with users choice as to either being connected or disconnected from the entire system.

Sushmita Manikandan [7] has proposed a Naïve Bayes based predictive heart disease system. They used a binary classifier system supporting patients with the graphical user interface through web. The proposed system used a dataset of having 13 predator variables with one binary response variable which were taken from UCI's machine learning repository. Initially, the data from medical records converted into structured data. For data pre-processing, Rapid Miner was used and the performance of the proposed system was compared with K-Nearest Neighbor, Decision Trees and Random Forest. Anaconda v2.7 packages were used to construct the classifier. The results proved that Gaussian Naïve

Bayes gave the highest accuracy of 81.25% than other approaches. In the upcoming classification algorithms the accuracy can be improved.

Anitek et al m[8] have proposed an automatic detection of cardiac abnormalities in a convenient, non-invasive and efficient way. In their work, a ML based smart device classifies normal and abnormal heart sounds and was useful for real time diagnosis of cardiac diseases. The audio samples were extracted using Mel frequency Cepstral Coefficient(MFCC). Then, machine learning based supervised classification was performed which gained an accuracy of 97.50% and sensitivity as 100%.

Jagdeep et al [9] have implemented a machine learning repository to test on different data mining techniques on heart prediction system. The dataset of 313 instances with 14 different attributes were collected from Cleveland heart diseases dataset from the University of California Irvine (UCI). The implementation involves 2 associative algorithms such as *Apriori* and *FP-Growth* with a selection of 10 best rules from each method for training dataset. The accuracy obtained using K-Nearest neighbor technique using IBK is 99.19% and higher than hybrid techniques such as J48, ZeroR, Naive- Bayes, OneR.

Tahira et al [10] proposed a technique by using various machine learning practices for detecting innumerable heart diseases. 13 attributes were incorporated in total with 50 instances. The approaches used were Hidden Markov Models, Support Vector Machine, Feature Selection, Computational intelligent classifier, prediction system, data mining techniques and genetic algorithm. The proposed technique has an accuracy of 94.21%, a ROC of 0.981, RMSE of 0.2568, Precision of 0.953; showing significant improvement when compared to the performance of K- Nearest Neighbor, Artificial Neural Networks and Support Vector Machines algorithms.

2.1 Inference from Survey

The surveyed work is compared with their accuracy value in Table 1. From the analysis it is observed that ML classifiers based predictive systems shows better accuracy than other systems.

Table -1: Comparison of ML Algorithms

Year	Author	Algorithm	Accuracy
2018	Amita Malav et al	ANN and K-means	93.52%
2018	I Ketut Agung Kassem et al	KNN Naïve Bayes SVM	75.11% 50.44% 45.11%
2017	Martin Gjoreski et al	Machine Learning classifiers	96%
2017	Sushmita Manikandan	Gaussian Naïve Bayes	81.25%
2017	Anitek Bhattacharya et al	Machine Learning	97.50%
2017	Tahira Mahboob et al	Various ML techniques	94.21%

2016	Jagdeep Singh et al	J48, ZeroR, Naïve Bayes, OneR and KNN	99.19%
2015	Purushottam et al	MV algorithm	86.75%
2016	Akash Mukherjee et al	ANN	96.2%
2012	Chaitrali	Neural Networks	100%
2011	Jyoti Soni et al	Naïve Bayes Decision Tree KNN	96.5% 99.2% 88.3%
2008	Latha et al	Neuro – Fuzzy	More than 90%

3. PROBLEM STATEMENT

She existing model deals with patients reaching doctors for medical diagnosis and examining them using variety of equipment's during this reach to the doctors and to the hospital can worsen the health conditions which includes death. It involves lots of cardiovascular equipment's which results in lots of expenditure which is out of reach for common people. It's strenuous for rural people to reach the hospitals for regular check ups. Using a platform where they can find out their current stage of heart functioning might help them in getting their required treatment at the right time at least cost. you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

4. SOLUTION

Using a GUI we give in specific details to calculate the level of severity of heart diseases. Our main objective is to classify whether a person is a heart patient or not by using machine learning classification technique such as random forest. Using data sets and attributes such as age, sex, chest pain etc., we classify the patient. In order to achieve this, many researches were done previously using various machine learning techniques where as we reduced the number of attributes and tried to maintain the accuracy.

5. METHODOLOGY

In this study, we discuss about the modules of the project which mainly involves training of the dataset and then the model, testing the model and predicting the output.

5.1 Training the dataset

Algorithms learn from knowledge. They realize relationships, develop understanding, create decisions, and valuate their confidence from the training data they're given and therefore the better the training data is, the higher the model performs so as to accomplish our goal of achieving a highly accurate model, we have a tendency to arrange to use a huge annotated API, which might be used as training data, that might not simply facilitate to identify the proper label for the testing data set entity on, however additionally establish relationship between entities.

5.2 Training the model

This step happens to be a really crucial section of the complete process. During this step, we have a tendency to train the classifier in such a fashion that the classifier will identify and label a test data that has been given by the input on the idea of the information it has learned from the training data set. This step involves random shuffling the training data so that the classifier will establish relationship between entities and in turn increase its accuracy of labelling.

5.3 Training the trained model

Once the classifier has been trained upon it properly, we need to make sure that the model is indeed working well, and not generating any error, thus testing plays a major role towards the end of the project. If the model is generating any error, we need to rework on the training set and classifier to see to it that the error is completely eliminated and is recognizing the entities accurately.

5.4 Saving the model

Once the classifier has been trained properly on the training data set, and is performing as desired, we need to save the model so that it can be used as and when required and we then test the saved model.

6. PROPOSED METHOD

In this study, we used a Cleveland clinic dataset which contains 9 attributes such as age, gender, resting blood pressure(in mm Hg on admission to the hospital), serum cholesterol in mg/dl, fasting blood sugar, Rest ECG results, maximum heart rate achieved during ECG, chest pain during exercise, chest pain type. 75% of the dataset is used to train the model using the classifier and 25% is used for testing. A GUI is constructed using Visual Studio Code to get the details of the user and then is subjected to the model which classifies the patient into four stages depending on the criticality of the disease or as if the patient is out of danger. A random forest classifier here creates few set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. In this model we used sklearn classifiers which requires data cleaning. Hence the data is preprocessed before it was trained. As random forest classifier is an ensembled algorithm it tends to give more accurate result because of the principle, “Number of weak estimators when combined forms a strong estimator.”

7. EXPERIMENTAL SETUP

7.1 Dataset

The [Heart disease data set] consists of patient data from Cleveland, Hungary, Long Beach and Switzerland. The combined dataset consists of 14 features and 303 samples with many missing values.
The features used in here are,

1. age: The patients age in years
2. sex: The patients gender(1=male; 0=female)
3. cp: Chest pain type,
   *Value 1: typical angina
   *Value 2: atypical angina
   *Value 3: non-anginal pain
   *Value 4: asymptomatic
4. trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
5. chol: Serum cholestoral in mg/dl
6. fbs: Fasting blood sugar > 120 mg/dl? (1=true; 0=false)
7. restecg: Resting electrocardiographic results
   *Value 0: normal
   *Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   *Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: Maximum heart rate achieved
9. exang: Chest pain(angina) after exercise? (1=yes; 0=no)
10. thal: Not described

*Value 3=normal
*Value 6=treated defect
*Value 7=reversible defect

11. num: Target
*Value 0: less than 50% narrowing of coronary arteries(no heart disease)
*Value 1,2,3,4: >50% narrowing. The value indicates the stage of heart diseases.

Dataset creators

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

7.2 User Interface

User interface is constructed using visual studio code and the input is taken from the user and the converted into the preprocessing data which is then classified.

7.3 Tools Used

7.3.1 Scikit-learn

Scikit-learn(sklearn) is known for its simplicity and effective tools for data mining and data analysis. It can be accessed by anyone and reused in various contexts. Classification can be done using this tool for identifying an object to which category it belongs.

7.3.2 Visual Studio Code

Visual Studio code is a source code editor that can be used with a variety of programming languages. It includes debugging, embedded Git control, syntax highlighting, intelligent code completion, snippets and code refactoring. It has various major programming language support such as Javascript, CSS, HTML etc., which can be downloaded for free.

7.3.3 Flask

Flask is a web framework written in Python and is used to develop web applications . Flask is designed to provide a fast and easy way which can scale up to complex web applications.

8. RESULT & ANALYSIS

The outcome we intend to obtain is to build a classifier which can differentiate between a heart disease patient and a normal patient. A model that works well with real time data entries. On calculating the accuracy using our proposed model we could obtain 97.56%.

Table -2: Data Summary

Total number of datasets	303
Total available attributes	14
Chosen number of attributes	9

9. OUTPUT

The screenshots of the output are as below:

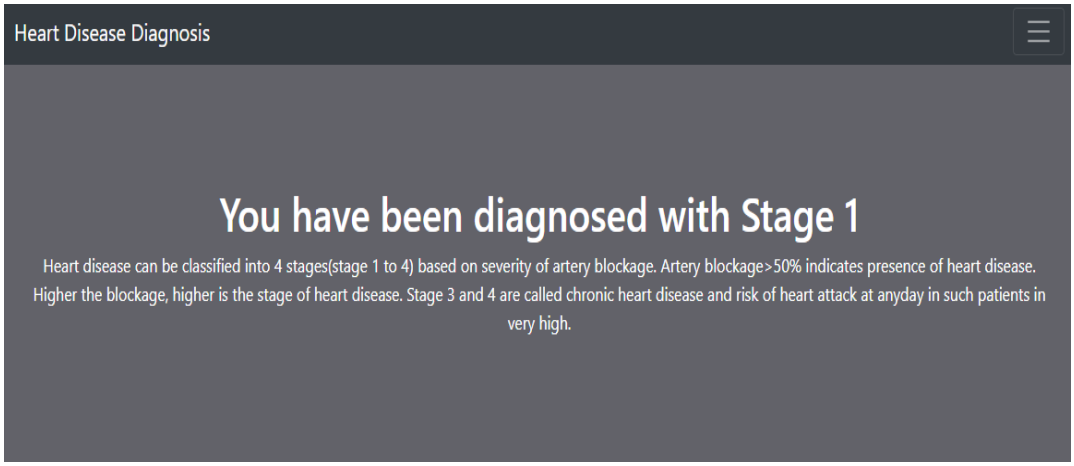

Fig 1: User Interface

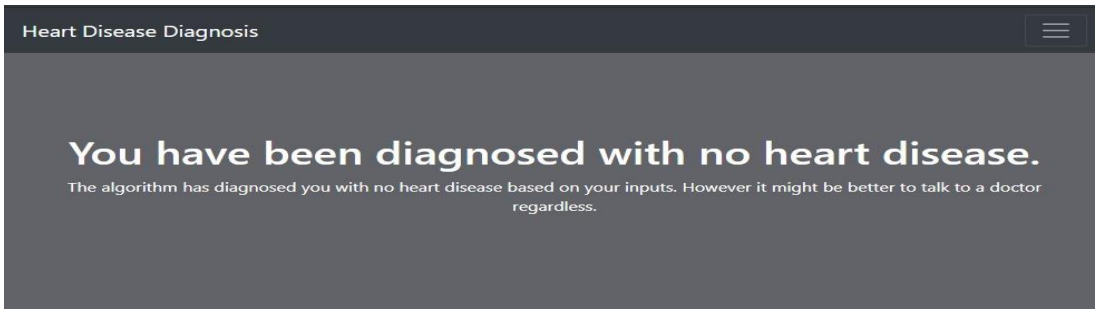

Fig 2: Output 1

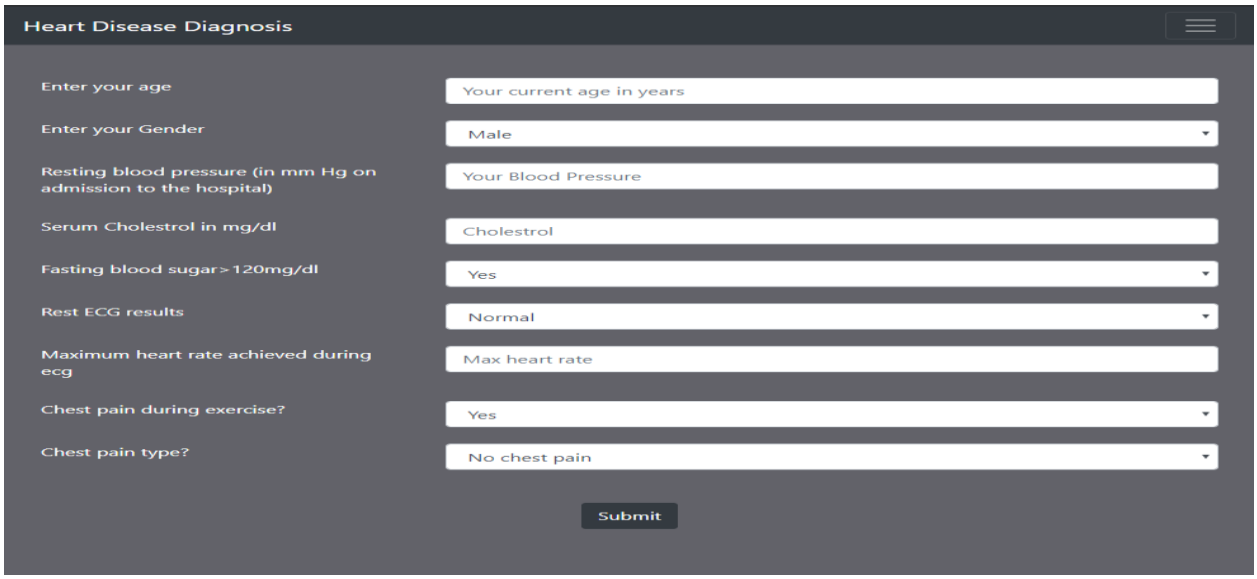

Fig 3: Output 2

10. CONCLUSION AND FUTURE WORK

From the above work, it is observed that the implementation of ML based techniques for heart disease identification is improving the accuracy and reducing the cost factor. Almost, all the work identifies the possibility of heart failure without any major medical infrastructure equipment but with intelligent ML techniques. Using random forest we built a platform which can be used to classify the heart patients with an accuracy of 97.5%. In future, an interactive platform can be built between the doctors and the patients to communicate about risks and factors.

11. REFERENCES

1. M. Ati, "*Knowledge capturing in autonomous system design for chronic disease risk assessment,* "Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on, pp. 62-66, 2014.

2. Sowmiya, C., and P. Sumitra. "*Analytical study of heart disease diagnosis using classification techniques*." In Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017 IEEE International Conference, pp. 1-5. IEEE, 2017.

3. Parthiban, Latha, and R. Subramanian. "*Intelligent heart disease prediction system using CANFIS and genetic algorithm.*" International Journal of Biological, Biomedical and Medical Sciences 3, no. 3 (2008).

4. Martin Gjoreski, Anton Gradiš̌ek, Matjaž̌ Gams, Monika Simjanoska, Ana Peterlin, Gregor Poglajen et al, "*Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers*", 13th International IEEE Conference on Intelligent Environments, 2017.

5. I Ketut Agung Enriko, Muhammad Suryanegara, Dadang Gunawan al, "*Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records*", 4th International Conference, June 2018.

6. Abdallah Kassem, Mustapha Hamad, Chady El Moucary and Elie Fayad, "*A Smart Device for the Detection of Heart Abnormality using R-R Interval*", 28th IEEE International conference on Microelectronics(ICM), 2016

7. Sushmita Manikandan, "*Heart Attack Prediction System*", International Conference on Energy, Communication, Data Analytics and Soft Computing, ICCET 2017

8. Anitek Bhattacharya,Mohan Mishra, AnushikhaSingh & Malay Kishore Dutta, "*Machine Learning Based Portable Device for Detection of Cardiac Abnormality*", International IEEE conference on Emerging Trends in Computing and Communication Technologies (ICETCCT 2017).

9. Jagdeep Singh, Amit Kamra and Harbhag Singh, "*Prediction of Heart Diseases Using Associative Classification*", 5th International Conference on Wireless Networks and Embedded System(WECON 2016).

10. Tahira Mahboob, Rida Irfan, Bazelah Ghaffar, "*Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics*", IEEE Internet Technologies and Application(ITA 2017).

11. Ankita Dewan, Meghna Sharma et al, "*Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification*", 2nd International IEEE Conference on Computing for Sustainable Global Development, 2015.

12. Purushottam, Prof. (Dr.) Kanak Saxena and Richa Sharma, "*Efficient Heart Disease Prediction System using Decision Tree*", International IEEE Conference on Computing , Communication and Automation (ICCCA2015).

13. Jia Xin Low, Keng Wah Choo, *"Classification of Heart Sounds Using Softmax Regression and Convolutional Neural Network"*, Association for Computing Machinery, ICCET '18, February 24–26, 2018, Singapore.

14. Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, *"Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction"* International Journal of Computer Applications, March 2011.

15. Dangare, Chaitrali S., and Sulabha S. Apte. *"Improved study of heart disease prediction system using data mining classification techniques."* International Journal of Computer Applications 47, no. 10 (2012): 44-48.

16. Amita Malav, Kalyani Kadam, *"A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K-Means"*, International Journal of Pure and Applied Mathematics 2018.

17. Akash Mukherjee, Raj Manjrekar, Ashish Marde and Prof. Rajesh Gaikwad, *"Heart Disease Prediction Using Artificial Neural Networks"*, IJSRD || National Conference on Technological Advancement and Automatization in Engineering, January 2016.