WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Prediction using Machine Learning

Anubhav Bansal[1], Anand Sinha[1], Shashank Dhoundiyal[1], Ms. Upasna Joshi[2]

[1]Student, Department of Computer Science and Engineering, Delhi Technical Campus, Greater Noida, U.P., India
[2]Professor, Department of Computer Science and Engineering, Delhi Technical Campus, Greater Noida, U.P., India

DOI: https://doi.org/10.55248/gengpi.5.0224.0561

ABSTRACT

Cardiovascular Diseases (CVDs) are major death prone diseases across the whole world. In India 272 per 100000 people die due to CVDs which is much more than that of whole world average of 235 while if we talk about United States 1 person expires in every 36 seconds from CVDs. The rapid increase in the number of deaths over past few decades are due to change in the lifestyle of every human being which is a very serious concern. So, as per the need of society and medical science a system is required which can diagnose such diseases accurately in time for proper treatment. Many industrial experts have designed such systems with the help of machine learning algorithms and these systems are helping professionals from medical fields as well as individuals to diagnose the diseases accurately at early stage. We have used a dataset of 70000 people with 11 different attributes to build our model with the help of machine learning algorithms such as Decision Tree, Logistic Regression and Random Forest. This paper is about the techniques and machine learning algorithms used to analyze the effectiveness of such models.

Keywords: Cardiovascular Diseases; Machine Learning Algorithms; Decision Tree; Logical Regression; Random Forest

1. Introduction

The work we have done majorly drives us to solve the problem of the increasing heart disease in this modern world. The World Health Organization (WHO) says heart disease is the leading cause of death worldwide, claiming 17.9 million lives every year. The human heart is the most vital component of the human body.

A heart's function is to pump enough blood to keep the brain and other organs supplied with oxygen and nutrients. Any irregularity or disturbance in our heart can be classified as heart disease. Alcohol or caffeine affect our heart if we consume excessively. There are different parameters that affect our heart health like Cholesterol, Alcohol Intake, Smoking, Physical Activity, etc.

Any problem in the heart may lead to serious problems so early detection of this disease may even save your life.

Even with this much advancement in our health care system still, the number of death is increasing with a steady increase of around 111% in the past 30 years. Recent studies say the heart is the prime source of death. Even heart disease is the most manageable and preventable disease, we are not able to prevent deaths but if they are predicted accurately within correct timing we may have the possibility of saving more lives. Every year healthcare industry collects large amounts of data we will use that data for training our machine learning models and predicting heart disease at an early stage.

Our work helps to detect CVDs at an early level and prevent any major repercussions.

The principal focus of this paper is to provide instruments to doctors so heart disease can be detected at an early stage. As a result, it will be easier to deliver appropriate medication to patients while avoiding serious effects. Machine Learning has a very important role in detecting discrete patterns and analysing the given data.

Medical experts can use a heart disease prediction system to help them anticipate the CVD of the patients.

Using machine learning algorithms such as Decision Tree, Logistic Regression and Random Forest, a heart disease prediction system can more accurately predict whether or not a patient will be diagnosed.

In brief, due to the rise in CVD over the past few years, we have worked on a heart disease prediction system that will help to early detection of heart disease which will save more lives and help to decrease deaths due to heart disease

2. Literature Survey

A lot of researchers from the past few decades are using different machine learning algorithms and datasets so we can diagnose the heart diseases efficiently. It is an ongoing process from many years and different accuracy have been attained by different researchers with the help of their models. Following are the researches.

Apurb Rajdan and et. al 2020.; have studied many machine learning and data mining techniques such as Naïve Bayes, Random Forest, Logistic Regression etc., and used UCI dataset to carried out their work. Their research concludes that the accuracy obtained by Random Forest was highest.

Jaymin Patel and et. al 2016.; have proposed a system which uses different decision tree classification algorithm and WEKA tool to predict the heart disease more accurately. They used Random Forest algorithms, J48, Logistic Model Tree and datasets of 303 instances and 76 attributes. Their research concludes that they get accuracy of 56.77% from J48 algorithm hence their success is marginal.

V.V. Ramalingam and et. al 2018.; they build the model to predict the heart disease using machine learning techniques such as SVM, KNN, Naïve Bayes, Random Forest, Decision Tree. Their research concluded that Random Forest and Ensemble Model performed extremely well.

Youness Khourdifi and Mohamed Bahaj 2019.; have prepared a model using ML algorithms optimized by ant colony optimization and particle swarm optimization. The algorithms they used are KNN, SVM, Naïve Bayes, Random Forest, ANN optimized by PSO and ACO approaches. They have also used fast correlation-based feature selection method and their research conclude that they get 99.65% accuracy using the optimized model proposed by FCBF, PSO and ACO.

Pavan Kumar and et. al.; they have prepared a model using Decision Tree, KNN, and K-Means algorithms and they concluded that they get the highest accuracy by Decision Tree.

Fahd Saleh Alotaibi built model using Logistic Regression, Random Forest, Naïve Bayes, Decision Tree, SVM and rapid miner tool and found out that the Decision Tree algorithm has the highest accuracy among others.

S. K. Srivatsa and et. al.; they have designed a model which diagnosed heart disease in diabetic patients using algorithms like Naïve Bayes and SVM and WEKA. They collected data of 500 patients from Chennai Research Institute out of which 142 have disease and 358 patients do not have it. Their research concluded that the Naïve Bayes provides them the accuracy of 74% whereas SVM provides them the highest accuracy of 94.60%.

Chethan C and et. al.; they have built their model using Naïve Bayes classification and Support Vector Machine (SVM). For the analysis they used Mean Absolute Error, Sum of squared error and RMS error and their research concluded that SVM is more accurate than the naïve bayes.

Theresa Princy. R and et. al,; they used algorithms like Naïve Bayes, KNN, Decision tree, Neural Network for predicting the heart disease and their research concluded that the KNN and ID3 gives them the best accuracy for the given number of attributes.

3. Requirement Analysis

3.1 Python

Python language was developed by Guido van Rossum during 1985-1990 which is a high-level, multipurpose, and interactive, object-oriented programming language designed to be more readable and easy to learn.

Python's object-oriented approach is designed to help programmers create a dynamic and straight-forward code for small and large projects.

Python is well-known among developers for its role in Artificial Intelligence (AI) systems, data science, Machine Learning (ML) and data analysis.

Its open source environment makes it easy to learn. A large number of libraries are used for activities such as web development, text processing, and statistics.

It is easy to distribute software, allowing independent software to be built and operated using Python. Software can be edited from start to finish using Python as the only programming language. It is a combination of developers as some planning languages need support in other languages before the project can be fully completed.

3.2 Machine Learning

Machine Learning or ML is an intelligent programming technique which adds the ability of automatically learning and improving from trial-error and their results without being explicitly coded to do that. This is mostly better suited for programs that use non-uniform data like image / text / problems with big number of parameters such as making predictions.

Machine learning needs a model which is trained to do a specific task, like making a prediction or classifying some input.

In short, ML is a branch of artificial intelligence which classifies patterns out of raw data with the help of algorithm or model. Its key feature is to make a computer program learn from experience without explicitly coding for it or human intervention.

4. Proposed Methodology

The following parameters are set for this prediction model.

1. The prediction model will not assume any prior data about the user's record it is classifying.
2. The model must be rugged to run with a large database with thousands of data

This model predicts the heart disease by scouting the three mentioned classification algorithms and doing an accuracy test. Objective of this model is to accurately predict if the user might be having a heart disease.

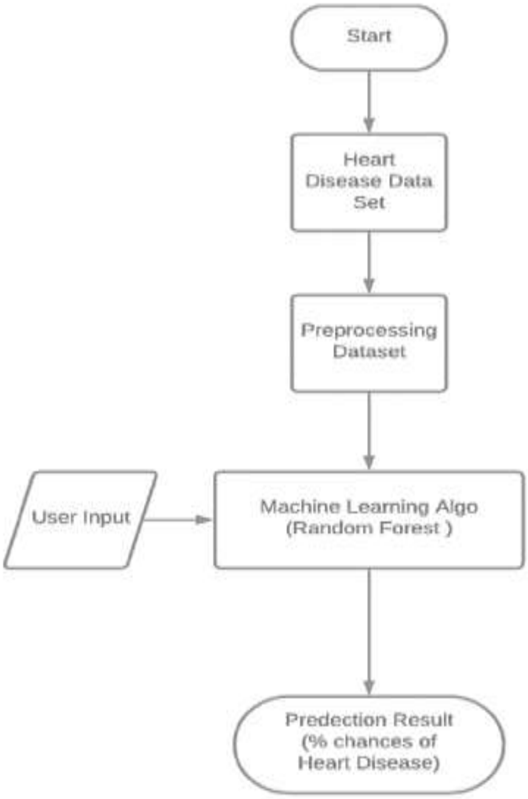

Start
Heart Disease Data Set
Preprocessing Dataset
User Input
Machine Learning Algo (Random Forest)
Predection Result (% chances of Heart Disease)

Fig. 1: Prediction Model

A. Data Collection

The dataset we used is the Cardiovascular Disease Dataset which consists of 12 attributes with a total data collection of 70,000 patients.

The required parameters are inputted into the model by the health professionals/user and the model predicts the probability of having a heart disease.

The necessary fields that are to be inputted into the model are given below.

S. No.	Parameter Description	Different Values of Parameter
1.	Age – the age of the patient	Integer value in days
2.	Height – describes the height of the patient	Integer value in cm
3.	Weight – the weight of the user	Float value in kg
4.	Gender – describes the gender of the patient	1-Women, 2-Men
5.	Systolic blood pressure [ap_hi]	Integral Value
6.	Diastolic blood pressure [ap_lo]	Integral Value
7.	Cholestrol – It represet the cholesterol level of the user	1-Normal, 2-Above Normal, 3-Well Above Normal
8.	Glucose – the glucose level of the patient	1-Normal, 2-Above Normal, 3-Well Above Normal
9.	Smoking – if the user smoke or not	1-Yes, 0-No

10.	Alcohol Intake – if the user consumes alcohol	1-Yes, 0-No
11.	Physical Activity – if the user is physically active or not	1-Yes, 0-No
12.	Target – if the user suffers from heart disease	1-Yes, 0-No

Table 1. Parameters from Dataset

5. Algorithm and Techniques Used

A. Decision Tree

Decision Trees are a famous data mining technique which uses a tree-shape structure to produces results based on inputs. It works as a flowchart where the inner blocks are the attributes from the dataset and the outer nodes represent the outcome. This type of method has the capability of managing non-uniform as well as missing data. Also the classifications are processed without much computation. It maps out the possible outcomes from different binary choices. Typically it starts with a single block, which leads into possible outputs. Each of the block leads to additional blocks, which jumps to other possibilities. Which forms a tree-like shape.

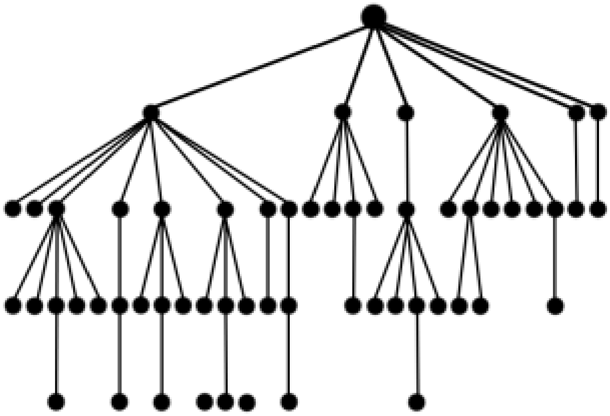
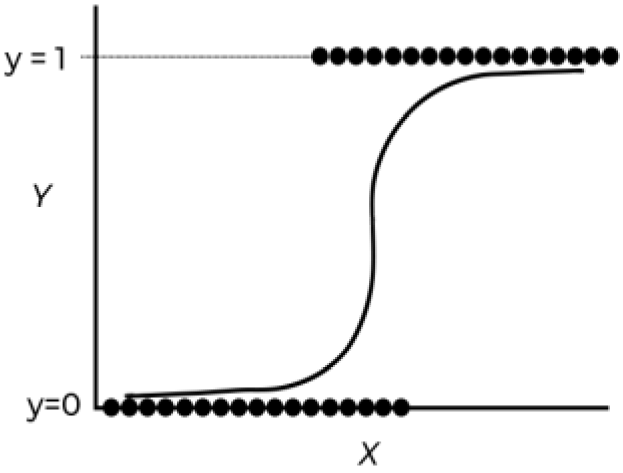


Fig. Decision Tree

B. Logistic Regression

Logistic regression is an algorithm which is mostly considered for classification problems like heart disease detection, diabetes detection etc. and it is also used to predict the probability of any variable. Logistic regression is a linear method but with the help of logistic function predictions are transformed and logistic regression algorithm uses that logistic function which can take any linear equation and can map it between the value 0 and 1. This makes logical regression good for classification.



$f(x)= 1/1+e^{-(b0-b1x)}$

Fig. Logistic Function

C. Random Forest

Random forest is an algorithm which is mostly considered for regression problems and classification problems.

In Random Forest, we use random sampling with a replacement that means after selecting a row we are putting it back into the data, and here are the rest of the Delta cells This process to create new data is called bootstrapping.

This data is trained using a decision tree on bootstrapped data sets predictions form this data is combined and majority voting is used as a result this process of combining results from multiple models is called aggregation so in the random forest we first perform bootstrapping then aggregation and in the jargon is called bagging. Random forest is called random because we have used random processes bootstrapping and random feature selection, bootstrapping ensure that we are not using the same data for every tree so in a way it helps our model to be less sensitive to the original training data

The random feature selection helps to reduce the correlation between the trees if you use every feature then most of your trees will have the same decision nodes and they will act very similarly that will increase the variance

Another benefit of the end of the selection some of the trees will be trained on less important so they will give bad predictions but there will also be some trees that give bad predictions in the opposite direction so they will balance out.

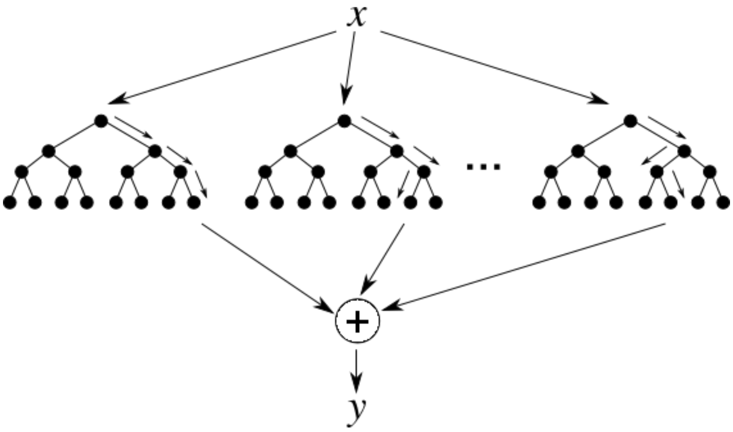


Fig. Random Forest

6. Result and Analysis

Predicting heart disease is done using the Decision Tree Algorithm, Logistic Regression Algorithm and the Random Forest Algorithm.

The results we obtained by performing performance analysis of various algorithms Precision Score (P), Recall Score (R), Accuracy Score and F-measure (F^m).

[3]*Precision (P): It is the right number of positive results.*

[4]*Recall (R): It is the number of real positive results.*

[5]*F-Measure (F^m): It checks the accuracy.*

[3]$(P) = (T^p) / (F^p + T^p)$

[4]$(R) = (T^p) / (Fn + T^p)$

[5]$(F^m) = (2 * Recall * Precision) / (Recall + Precision)$

F^p: False Positive- The patient is not having this disease and is diagnosed positive.

T^p: True Positive- The patient is having this disease and is diagnosed positive.

T^n: True Negative- The patient is having this disease and is diagnosed negative.

F^n: False Negative- The patient is not having this disease and is diagnosed negative.

The previously-processed data was used train the model and the algorithms mentioned were studied and performed. The confusion matrix is used to classify the accuracy success of the model.

The confusion matrix from our research is shown below. The second table is the accuracy for Decision Tree Algorithm, Logistic Regression Algorithm and Random Forest Algorithm.

Algorithm	**True Positive**	**False Positive**	**False Negative**	**True Negative**
Decision Tree	21	6	79	93
Logistic Regression	62	32	38	68
Random Forest	71	27	29	73

Table. Results from Confusion Matrix

Algorithm	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Decision Tree	0.778	0.210	0.331	57.10%
Logistic Regression	0.660	0.620	0.640	64.89%
Random Forest	0.724	0.710	0.717	71.97%

Table. Analysis of different algorithm

7. Acknowledgement

We are thankful to Delhi Technical Campus and we owe special debt of gratitude to our project coordinator Mrs. Upasna Joshi (Professor), Department of Computer Science Engineering, Delhi Technical Campus, Greater Noida. It is only because of her consistent efforts that we have performed our project successfully. We are also thankful to Dr. Shipra Saraswat (HOD, CSE) for their support and providing us with the necessary facilities to perform our project work successfully.

8. Conclusion

The early diagnosis of heart disease can help in saving the life of high-risk patients, which can be considered as a great achievement in medical science. As we all know that cardiovascular disease is one of the major diseases from past few decades which causes death in huge numbers across the world, early diagnosis of this disease can help doctors as well as patients and will reduce the death rate also. It is also going to reduce the efforts of doctors in terms of testing as well as in the form of resources and it is going to help the patients in monetary way. So we can say that therefore it is a state of victory for both. In our project "Heart Disease Prediction Using Machine Learning" we have used different machine learning algorithms like Logistic Regression, Random Forest and Decision Tree to get the result accurate and precise. We took 11 parameters and after training the model we get training accuracy of 99.98% and validation accuracy 0f 71.97% from Random Forest Algorithm.

Hence after doing all the work we can conclude that this project is useful in everyone's life and more importantly for the healthcare sector.

8. References

[1] Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr. Poonam Ghuli, "Heart Disease Prediction using Machine Learning", *International Journal of Engineering Research & Technology (IJERT)*, Vol. 9, pp.659-662, 2020.

[2] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", *2017 IEEE 2nd International Conference on Big Data Analysis*.

[3] Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique", *International Journal of Computer Science & Communication (IJCSC)*, Vol. 7, pp.129-137, 2016.

[4] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", *22nd IEEE Symposium on Computers and Communication*.

[5] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques", *International Journal of Engineering & Technology (IJET)*, Vol. 7, pp.684-687, 2018.

[6] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", *22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017.*

[7] Youness Khourdifi, Mohamed Bahaj, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", *International Journal of Intelligent Engineering & Systems (INASS),* Vol. 12, pp.242-252, 2019.

[8] Quazi Abidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and Hagit Shatkay et al. "Utilizing ECG-based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification", *DOI 10.1109/TNB.2015.2426213, IEEE Transactions on Nano Bioscience TNB-00035-2015.*

[9] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", *International Journal of Recent Technology and Engineering,* Vol 8, pp.944-950, 2019.

[10] Fahd Saled Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease", *International Journal of Advanced Computer Science and Applications,* Vol. 10, pp.261-268, 2019.

[11] T. Padmapriya and V.Saminadan, "Handoff Decision for Multi-user Multiclass Traffic in MIMO-LTE-A Networks", *2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016) – Elsevier - PROCEDIA OF COMPUTER SCIENCE, vol. 92, pp: 410-417, August 2016.*

[12] S.V.Manikanthan and D.Sugandhi "Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel" *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)* ISSN: 0976-1353 Volume-7, Issue 1 –MARCH 2014.

[13] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling, schemes for heart disease classification," *Applied Soft Computing,* vol. 14, pp. 47–52, 2014.

[14] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *International Journal of Computer Applications,* vol. 47, no. 10, pp. 44–48, 2012.

[15] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical Knowledge driven approach,"

[16] *Expert Systems with Applications,* vol. 40, no. 1, pp. 96–104, 2013.

[17] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', *International Conference on Circuit Power and Computing Technologies,*Bangalore,2016.

[18] Combination data mining methods with new medical data to predicting outcome of coronary heart disease," *in Convergence Information Technology, 2007. International Conference on. IEEE,* 2007, pp. 868–872.

[19] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," pp. 868–872, 2007.

[20] P. V. Ankur Makwana, "Identify the patients at high risk of re-admission in hospital in the next year," *International Journal of Science and Research,* vol. 4, pp. 2431–2434, 2015.

[21] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," pp. 173–177, 2012.

[22] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," pp.108–115, 2008.