Heart Disease Prediction Using Effective Machine Learning Techniques

Dr. Bharani B R[1], Dr. Manjunatha S [2], Prof. Vijayalakshmi R Y[3], Dr. Preethi S[4]

[1]Associate Professor, Information Science and Engineering, Cambridge Institute of Technology
[2]Associate Professor, Computer Science and Engineering, Cambridge Institute of Technology
[3]Assistant Professor, Information Science and Engineering, Cambridge Institute of Technology
[4]Professor, Information Science and Engineering, Cambridge Institute of Technology

Abstract

In today's era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and female category and this ratio may vary according to the region also this ratio is considered for the people of age group 25-69. This does not indicate that the people with other age group will not be affected by heart diseases. This problem may start in early age group also and predict the cause and disease is a major challenge nowadays. Here in this paper, we have discussed various algorithms and tools used for prediction of heart diseases.

Keywords: Classification, Heart Disease, Decision Tree, Logistic Regression, Random Forest.

1. Introduction

The contents of this paper mainly focus on various data mining practices that are valuable in heart disease forecast with the assistance of dissimilar data mining tools that are accessible. If the heart doesn't function properly, this will distress the other parts of the human body such as brain, kidney etc. Heart disease is a kind of disease which effects the functioning of the heart. In today's era heart disease is the primary reason for deaths. WHO-World Health Organization has anticipated that 12 million people die every year because of heart diseases. Some heart diseases are cardiovascular, heart attack, coronary and knock. Knock is a sort of heart disease that occurs due to strengthening, blocking or lessening of blood vessels which drive through the brain or it can also be initiated by high blood pressure. The major challenge that the Healthcare industry faces now-a-days is superiority of facility.

Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted. Records or data of medical history is very large, but these are from many dissimilar foundations. The interpretations that are done by physicians are essential components of these data. The data in real world might be noisy, incomplete and inconsistent, so data preprocessing will be required in directive to fill the omitted values in the database. Even if cardiovascular diseases are found as the important source of death in world in ancient years, these have been announced as the most avoidable and manageable diseases. The whole and accurate management of a disease rest on on the well-timed judgment of that disease. A correct and methodical tool for recognizing high-risk patients and mining data for timely analysis of heart infection looks a serious want. Different person body can show different symptoms of heart disease which may vary accordingly. Though, they frequently include back pain, jaw pain, neck pain, stomach disorders, and tininess of breath, chest pain, arms and shoulders pains.

There are a variety of different heart diseases which includes heart failure and stroke and coronary artery disease. Even though heart disease is acknowledged as the supreme chronic sort of disease in the world, it can be most avoidable one also at the same time. A healthy way of life (main prevention) and timely analysis (inferior prevention) are the two major origins of heart disease director. Conducting steady check-ups (inferior prevention) Shows outstanding role in the judgment and early prevention of heart disease difficulties. Several tests comprising of angiography, chest X-rays, echocardiography and exercise tolerance test support to this significant issue. Nevertheless, these tests are expensive and involve availability of accurate medical equipment. Heart experts create a good and huge record of patient's database and store them. It also delivers a great prospect for mining a valued knowledge from such sort of datasets. There is huge research going on to determine heart disease risk factors in different patients, different researchers are using various statistical approaches and numerous programs of data mining approaches. Statistical analysis has acknowledged the count of risk factors for heart diseases counting smoking, age, blood pressure, diabetes, total cholesterol, and hypertension, heart disease training in family, obesity and lack of exercise. For prevention and healthcare of patients who are about to have addicted of heart disease it is very important to have awareness of heart diseases. Researchers make use of several data mining techniques that are accessible to help the specialists or physicians identify the heart disease. Commonly used procedures used are decision tree, k-nearest and Naïve Bayes. Other different classification-based techniques used are bagging algorithm, Logistic Regression, SVM (Support Vector Machine).

2. Literature Survey

In recent years, the healthcare industry has seen a significant advancement in the field of data mining and machine learning. These techniques have been widely adopted and have demonstrated efficacy in various healthcare applications, particularly in the field of medical cardiology. The rapid accumulation of medical data has presented researchers with an unprecedented opportunity to develop and test new algorithms in this field. Heart disease remains a leading cause of mortality in developing nations [12,13,14,15,16], and identifying risk factors and early signs of the disease has become an important area of research. The utilization of data mining and machine learning techniques in this field can potentially aid in the early detection and prevention of heart disease.

The purpose of the study described by Narain et al. (2016) [17] is to create an innovative machine-learning-based cardiovascular disease (CVD) prediction system in order to increase the precision of the widely used Framingham risk score (FRS). With the help of data from 689 individuals who had symptoms of CVD and a validation dataset from the Framingham research, the proposed system—which uses a quantum neural network to learn and recognize patterns of CVD—was experimentally validated and compared with the FRS. The suggested system's accuracy in forecasting CVD risk was determined to be 98.57%, which is much greater than the FRS's accuracy of 19.22% and other existing techniques. According to the study's findings, the suggested approach could be a useful tool for doctors in forecasting CVD risk, assisting in the creation of better treatment plans, and facilitating early diagnosis.

In a study conducted by Shah et al. (2020) [18], the authors aimed to develop a model for predicting cardiovascular disease using machine learning techniques. The data used for this purpose were obtained from the Cleveland heart disease dataset, which consisted of 303 instances and 17 attributes, and were sourced from the UCI machine learning repository. The authors employed a variety of supervised classification methods, including naive Bayes, decision tree, random forest, and k-nearest neighbor (KKN). The results of the study indicated that the KKN model exhibited the highest level of accuracy, at 90.8%. The study highlights the potential utility of machine learning techniques in predicting cardiovascular disease, and emphasizes the importance of selecting appropriate models and techniques to achieve optimal results.

3. Proposed System

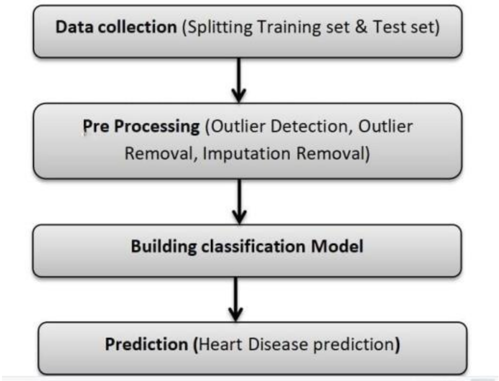

Data collection (Splitting Training set & Test set)
Pre Processing (Outlier Detection, Outlier Removal, Imputation Removal)
Building classification Model
Prediction (Heart Disease prediction)

Fig 1: Dataflow model for machine Learning Model

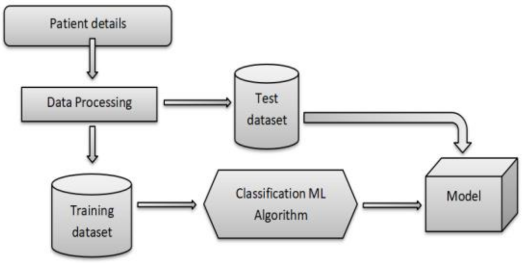


Fig 2: Architecture of Proposed Model

In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions. Training the Dataset:

- The first line imports iris data set which is already predefined in sklearn module. Iris data set is basically a table which contains information about various varieties of iris flowers.
- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- Then we encapsulate load_ data() method in data_ dataset variable. Further we divide the dataset into training data and test data using train_ test_ split method. The X prefix in variable denotes the feature values and y prefix denotes target values.
- This method divides dataset into training and test data randomly in ratio of 67:33. Then we encapsulate any algorithm.

- In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

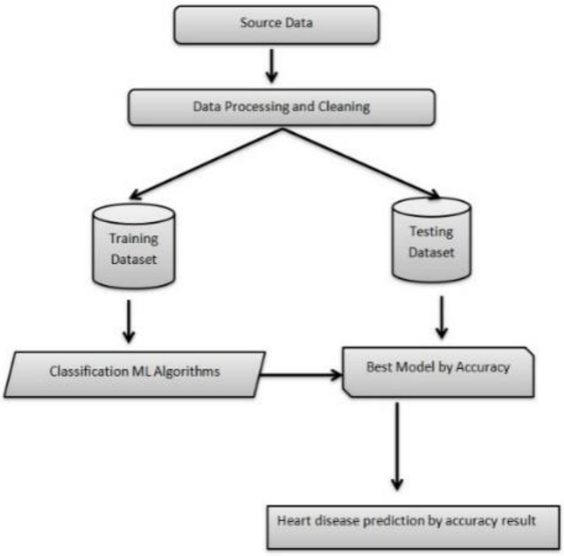

Source Data
Data Processing and Cleaning
Training Dataset
Testing Dataset
Classification ML Algorithms
Best Model by Accuracy
Heart disease prediction by accuracy result

Fig 3: Workflow Diagram

4. Model Selection

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results. Usually Data Scientists use different kinds of Machine Learning algorithms to the large data sets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. Supervised learning: Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into **Regression** and **Classification** problems.

A **regression** problem is when the output variable is a real or continuous value, such as “salary” or “weight”. A **classification** problem is when the output variable is a category like filtering emails “spam” or “not spam”

Unsupervised Learning: Unsupervised learning is the algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. In our dataset we have the outcome variable or Dependent variable i.e Y having only two set of values, either M (Malign) or B(Benign). So we will use Classification algorithm of supervised learning.

Modules:

- Data validation and pre-processing technique (Module-01)
- Exploration data analysis of visualization and training a model by given attributes (Module-02)
- Performance measurements of logistic regression and decision tree algorithms (Module-03)
- Performance measurements of Support vector classifier and Random forest (Module-04)
- Performance measurements of KNN and Naive Bayes (Module-05)
- GUI based prediction of heart disease (Module-06).

5. Sequence Diagram

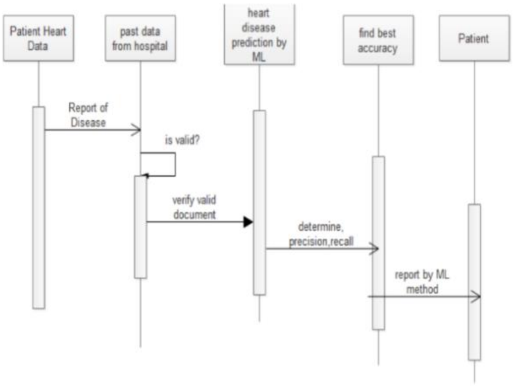


Fig 4: Sequence Diagram

UML sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior within your system. Other dynamic modeling techniques include activity diagramming, Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development. disease prediction. As maximum types of dataset will be covered under this system, doctor may get to know about the disease exactly using ML algorithms, it helps the doctor in decision making weather patient has heart disease or not.

Use Case Diagram:

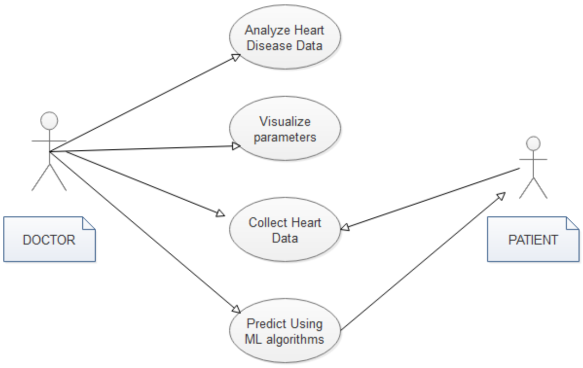


Fig 5: Use Case Diagram

Use case diagrams are considered for high level requirement analysis of a system. So, when the requirements of a system are analysed the functionalities are captured in use cases. So, we can say that uses cases are nothing but the system functionalities written in an organized manner. Now the second things which are relevant to the use cases are the actors. Actors can be defined as something that interacts with the system. The actors can be human user, some internal applications or may be some external applications. Functionalities to be represented as a use case, Actors and Relationships among the use cases and actors and the name of a use case is very important.

6. Conclusion

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally, we predict the heart disease using machine learning algorithm with different results. This brings some of the following insights about heart disease prediction. As maximum types of dataset will be covered under this system, doctor may get to know about the disease exactly using ML algorithms, it helps the doctor in decision making weather patient has heart disease or not.

7. Outputs

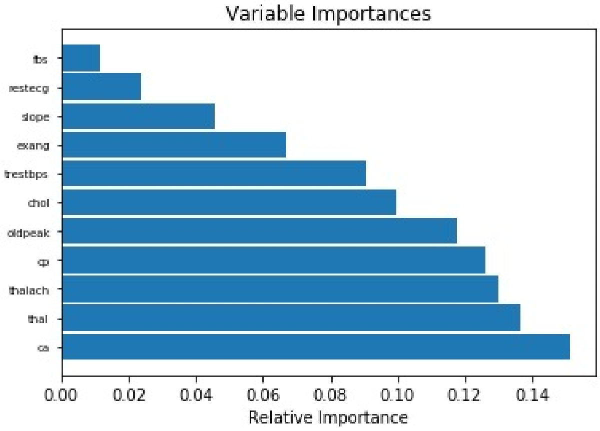


Fig 6: Output of Random Forest Algorithm

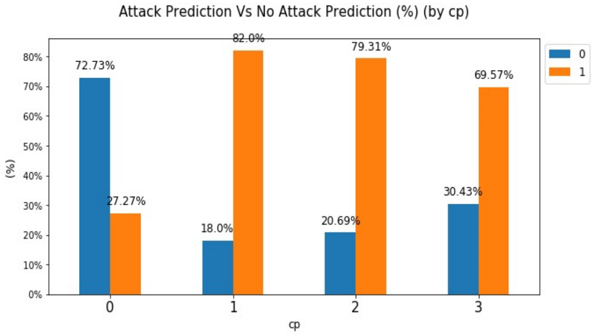


Fig 7: Output of Heart Disease Attack percentage

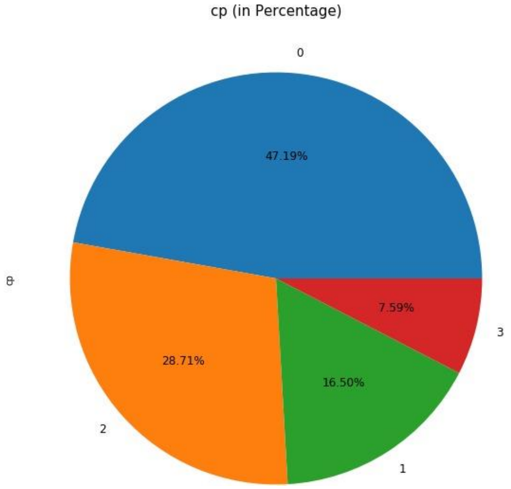
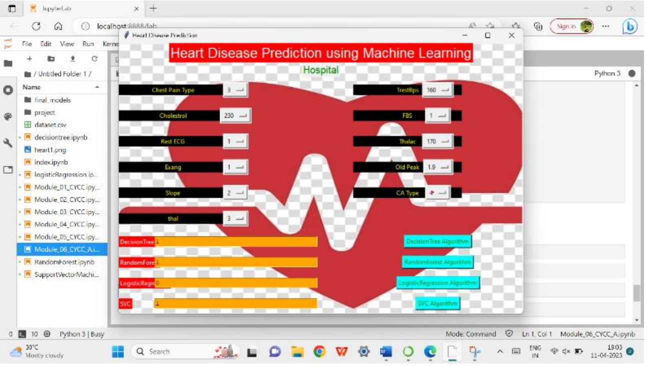


Fig 8: Output of Constrictive Pericarditis



Heart Disease Prediction using Machine Learning

Hospital

Chest Pain Type 3

Cholestrol 230

Rest ECG 1

Exang 1

Slope 2

thal 3

TrestBps 160

FBS 1

Thalac 170

Old Peak 1.9

CA Type 4

DecisionTree 1

RandomFore 1

LogisticRegr 0

SVC 1

Fig 9: Final Output

References

1. V. Krishnaiah, G. Narasimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review" IJCA 2016.
2. K. Sudhakar, Dr. M. Manimekalai "Study of Heart Disease Prediction using Data Mining", IJARCSSE 2016.
3. Naganna Chetty, Kunwar Singh Vaisla, Nagamma Patil,"An Improved Method for Disease Prediction using Fuzzy Approach", ACCE 2015.
4. Vikas Chaurasia, Saurabh Pal, "Early Prediction of Heart disease using Data mining Techniques", Caribbeanjournal of Science and Technology,2013
5. Shusaku Tsumoto," Problems with Mining Medical Data", 0-7695- 0792-1 I00@ 2000 IEEE.
6. Y. Alp Aslandoganet. al.," Evidence Combination in Medical Data Mining", Proceedings of the international conference on Information Technology: Coding and Computing (ITCC'04) 0-7695-2108-8/04©2004 IEEE., "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Rev
7. LathaParthiban and R.subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", ssInternational Journal of Biomedical and Medical Sciences, Vol.3, Page No.3.
8. Niti Guru, Anil Dahiya, NavinRaipal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol.8, No.1, January-June 2007.
9. SellappanPalaniappan, RafiaAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (IJCSNS), Vol.8, August 2008.
10. Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904.
11. Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022.
12. Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332.

13. Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. Circulation 2019, 139, e56–e528
14. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.
15. Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. Circulation 2015, 131, e29–e322
16. Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.
17. Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. J. Occup. Health 2016, 58, 216–219.
18. Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. Procedia Comput. Sci. 2016, 85, 962–969.
19. Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. Int. J. Comput. Appl. 2011, 17, 43–48
20. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554.
21. Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.
22. Fayez, M.; Kurnaz, S. Novel method for diagnosis diseases using advanced high-performance machine learning system. Appl. Nanosci. 2021.
23. Hassan, C.A.U.; Iqbal, J.; Irfan, R.; Hussain, S.; Algarni, A.D.; Bukhari, S.S.H.; Alturki, N.; Ullah, S.S. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors 2022, 22, 7227.
24. Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. Sustainability 2022, 14, 14208.
25. Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. Int. J. Comput. Appl. 2020, 176, 46–54.
26. Kaggle Cardiovascular Disease Dataset. Available online: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset (accessed on 1 November 2022).