Heart Disease Prediction using Machine Learning Techniques: A Review

[1]Kolli Punith Kumar, [2]Prof. Anoop Singh, [3]Dr. Varsha Namdeo
[1]Research Scholar, [2]Assistant Professor, [3]Associate Professor & HOD
[1&2&3]Department of Computer Science Engineering,
[1&2&3]RKDF Institute of Science & Technology, SRK University, Bhopal, India

Abstract : Heart disease is one of the most critical human diseases in the world and affects human life very badly. In heart disease, the heart is unable to push the required amount of blood to other parts of the body. Accurate and on time diagnosis of heart disease is important for heart failure prevention and treatment. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. Therefore classify the healthy people and people with heart disease, non-invasive-based methods such as machine learning are reliable and efficient. This paper studied a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Various results show that each technique has its unique strength to prediction parameters.

IndexTerms **- Heart Disease, Prediction, Data mining, Machine Learning, Decision Trees, Naïve Bayes.**

I. INTRODUCTION

The heart disease (HD) has been considered as one of the complex and life deadliest human diseases in the world. In this disease, usually the heart is unable to push the required amount of blood to other parts of the body to fulfill the normal functionalities of the body, and due to this, ultimately the heart failure occurs [1]. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. In most of the researchers study, integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [2]. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data [3]. Unfortunately, these data are rarely used to support clinical decision making.

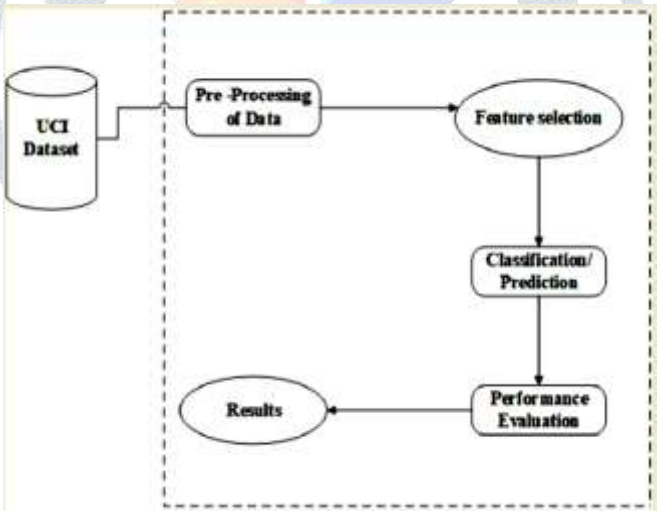

Figure 1: Basic process of Prediction

Figure 1 shows the basic operation of heart disease prediction using classification or data mining approach. The rate of heart disease in the United States is very high [2]. The symptoms of heart disease include shortness of breath, weakness of physical body, swollen feet, and fatigue with related signs, for example, elevated jugular venous pressure and peripheral edema caused by functional cardiac or noncardiac abnormalities [3]. The investigation techniques in early stages used to identify heart disease were complicated, and its resulting complexity is one of the major reasons that affect the standard of life [4]. The heart disease diagnosis and treatment are very complex, especially in the developing countries, due to the rare availability of diagnostic apparatus and shortage of physicians and others resources which affect proper prediction and treatment of heart patients [8]. The accurate and proper diagnosis of the heart disease risk in patients is necessary for reducing their associated risks of severe heart issues and improving security of heart [6]. The European Society of Cardiology (ESC) reported that 26 million adults worldwide were diagnosed with heart disease and 3.6 million were diagnosed every year. Approximately 50% of heart disease people suffering from HD die within initial 1-2 years, and concerned costs of heart disease management are approximately 3% of health-care financial budget [7].

II. LITERATURE SURVEY

S. Mohan, et al.,[2019] Heart disease is one of the most noteworthy reasons for mortality on the planet today. Expectation of cardiovascular disease is a basic test in the territory of clinical information investigation. Machine learning (ML) has been demonstrated to be viable in helping with settling on choices and expectations from the enormous amount of information delivered by the medicinal services industry. it is have likewise observed ML methods being utilized in late improvements in various territories of the Internet of Things (IoT). Different investigations give just a look into foreseeing heart disease with ML methods. In this work, it is propose a novel strategy that targets finding critical highlights by applying machine learning strategies bringing about improving the exactness in the forecast of cardiovascular disease. The forecast model is presented with various blends of highlights and a few known classification strategies. it is produce an upgraded exhibition level with a precision level of 88.7% through the expectation model for heart disease with the half and half random timberland with a straight model (HRFLM). [1]

W. Chang, et al.,[2019] The adjustments throughout individuals' life cadence and improvement in material levels that occurred as of late expanded the quantity of individuals experiencing hypertension on the planet. In this way, as a cardiovascular complexity of hypertension, the commonness of hypertensive heart disease has expanded every year, it has truly imperiled the wellbeing of human life, and the compelling expectation of hypertensive heart disease has become an overall issue. This work utilizes the recently proposed XGBSVM crossover model to anticipate whether hypertensive patients will create hypertensive heart disease inside three years. The last trial demonstrates that through this model, hypertensive patients can gain proficiency with their danger of hypertensive heart disease inside 3 years and then experience focused on preventive treatment, along these lines lessening the mental, physiological and monetary weight. This work affirms that the machine learning can be effectively applied in the biomedical field, with solid true hugeness and research esteem. [2]

T. S. Brisimi, et al.,[2018] Urban living in present day huge urban communities has huge unfriendly consequences for wellbeing, expanding the danger of a few incessant diseases. it is center around the two driving bunches of constant diseases, heart disease and diabetes, and create information driven strategies to anticipate hospitalizations because of these conditions. it is base these expectations on the patients' medicinal history, later and progressively inaccessible, as depicted in their Electronic Health Records (EHRs). it is figure the expectation issue as a double classification issue and think about an assortment of machine learning techniques, including kernelized and meager Support Vector Machines (SVMs), scanty calculated relapse, and random backwoods. To find some kind of harmony among precision and interpretability of the expectation, which is significant in a medicinal setting, it is propose two novel techniques: K - LRT, a probability proportion test-based strategy, and a Joint Clustering and Classification (JCC) strategy which recognizes shrouded quiet groups and adjusts classifiers to each bunch.[3]

A. Mdhaffar, et al.,[2017] This work exhibits a novel wellbeing investigation approach for heart disappointment forecast. It depends on the utilization of complex occasion handling (CEP) innovation, joined with factual methodologies. A CEP motor procedures approaching wellbeing information by executing edge based investigation rules. Rather than having to physically set up limits, our novel factual calculation consequently registers and updates edges as indicated by recorded chronicled information. Test results show the benefits of our methodology as far as speed, exactness, and recall.[4]

A. Khan et al.,[2017] Machine learning and classification algorithms facilitate to design “Intrusion Detection Models” which might classify the network traffic into intrusive or traditional traffic. This paper discusses some usually used machine learning techniques in Intrusion Detection System and conjointly reviews a number of the prevailing machine learning IDS proposed by researchers at different times. in this paper an experimental analysis is performed to demonstrate the performance analysis of some existing techniques in order that they will be used further in developing Hybrid Classifier for real data packets classification. The given result analysis shows that KNN, RF and SVM performs best for NSL-KDD dataset.[5]

D. Tay, et al.,[2015] Myocardial dead tissue (MI) is one of the main sources of death in many created nations. Thus, early identification of MI occasions is basic for powerful protection treatments, possibly lessening avoidable mortality. One methodology for early disease forecast is the utilization of hazard expectation models created utilizing machine learning strategies. One significant part of these models is to furnish clinicians with the adaptability to tweak (e.g., the forecast range) and utilize the hazard expectation model that they esteemed generally advantageous for their patients. In this manner, in this work, it is create MI expectation models and research the impact of test age and forecast goals on the presentation of MI hazard expectation models. The cardiovascular wellbeing study dataset was utilized in this investigation. Results demonstrate that the forecast model created utilizing SVM calculation is fit for accomplishing high affectability, explicitness, and adjusted exactness of 95.3%, 84.8%, and 90.1%, separately, over a period length of 6 years. Both example age and forecast goals were found not to significantly affect the exhibition of MI hazard expectation models created utilizing subjects matured 65 and above. This suggests chance expectation models created utilizing distinctive example age and forecast goals is an attainable methodology. [6]

D. R. Patil et al.,[2014] In medicinal field the finding of heart disease is most troublesome undertaking. It relies upon the cautious examination of various clinical and neurotic information of the patient by therapeutic specialists, which is entangled procedure. Because of progression in machine learning and data innovation, the scientists and restorative specialists in huge degree are keen on the advancement of robotized framework for the expectation of heart disease that is profoundly exact, compelling and accommodating in early determination. In this work it is available an expectation framework for heart disease utilizing multilayer perceptron neural system. The neural system in this framework acknowledges 13 clinical highlights as information and it is prepared utilizing back-spread calculation to anticipate that there is a nearness or nonappearance of heart disease in the patient with most elevated exactness of 98% similar to different frameworks. The exactness therefore got with this framework shows that it is preferable and productive over different frameworks. [7]

R. Wijaya, et al.,[2013] In this work talked about the improvement of heart disease forecast utilizing machine learning (for this situation the Artificial Neural Network or ANN). There are 13 factors that can decide heart disease as indicated by Miss Chaitrali paper. Forecast of an individual's heart disease one year ahead is performed by contemplating the model heart rate information. Information is taken by utilizing device, for example, brilliant mirror, shrewd mouse, advanced mobile phones and savvy seat. Heart rate information were gathered through the Internet and gathered in a server. Learning in this framework is performed for a time of one year to get enough information to make forecasts. Prescient of future heart disease in one year can build an individual's familiarity with heart disease itself. The framework is likewise expected to decrease the quantity of patients and the quantity of passings from heart disease.[8]

Table 1: Summary of literature survey

Sr No	Author Name	Year of Publication	Proposed Work	Outcome
1	S. Mohan	IEEE, 2019	A novel method aims at finding significant features by applying machine learning	An accuracy level of 88.7%
2	W. Chang,	IEEE, 2019	The newly proposed XGBSVM hybrid model	An accuracy level of 82%.
3	T. S. Brisimi	IEEE, 2018	A JCC method which identifies hidden patient clusters	Validate algorithms on large data sets from the Boston Medical Center.
4	A. Mdhaffar	IEEE, 2017	A novel health analysis approach for heart failure prediction.	The merits of our approach in terms of speed, precision, and recall.
5	A. Khan	IEEE, 2017	The adherence of patients with HF.	The highest detection accuracy is 82 and 91%
6	D. Tay,	IEEE, 2015	It is develop MI prediction models.	Accuracy of 92.3%, 84.8%, and 90.1%,
7	D. R. Patil	IEEE, 2014	A prediction system for heart disease using multilayer perceptron neural network	Accuracy of 88% comparative to other systems.
8	R. Wijaya	IEEE, 2013	The development of heart disease prediction machine learning in this case the ANN.	Expected to reduce the number of patients and the number of deaths from heart disease.

III. VARIOUS METHODS

The data mining has four main techniques namely

1. Classification,
2. Clustering,
3. Regression, and
4. Association rule.

Data mining techniques have the ability to rapidly mine vast amount of data. Data mining is mainly needed in many fields to extract useful information from a large amount of data. The fields like the medical field, business field, and educational field have a vast amount of data, thus these fields data can be mined through those techniques more useful information. Data mining techniques can be implemented through a machine learning algorithm. Each technique can be extended using certain machine learning models.

There are many types of classification algorithms and machine learning, such as decision trees, naive bayes, linear discriminant analysis, k-nearest neighbor, logistic regression, neural networks, and support vector machines.

1. K-Nearest Neighbors Algorithm (KNN)

The K-Nearest Neighbors is an algorithm for supervised learning and is a classification algorithm that takes a bunch of labeled points and uses them to learn how to label other points. This algorithm classifies cases based on their similarity to other cases. In K-Nearest Neighbors, data points that are near each other are said to be neighbors. K-Nearest Neighbors is based on this paradigm. Thus, the distance between two cases is a measure of their dissimilarity.

2. Decision Tree

Decision trees are built using recursive partitioning to classify the data, i.e., by splitting the training set into distinct nodes, where one node contains all of or most of one category of the data. A decision tree can be constructed by considering the attributes one by one.

3. Naïve Bayes

Naïve Bayes classifier is a supervised algorithm which classifies the dataset on the basis of Bayes theorem. The Bayes theorem is a rule or the mathematical concept that is used to get the probability is called Bayes theorem. Bayes theorem requires some independent assumption and it requires independent variables which is the fundamental assumption of Bayes theorem.

4. Logistic Regression vs. Linear Regression

While Linear Regression is suited for estimating continuous values (e.g. estimating house price), it is not the best tool for predicting the class of an observed data point. In order to estimate the class of a data point, we need some sort of guidance on what would be the most probable class for that data point. For this, we use Logistic Regression.

The difference between linear and multiple linear regression is that the linear regression contains only one independent variable while multiple regression contains more than one independent variables. The best fit line in linear regression is obtained through least square method. Linear regression finds a function that relates a continuous dependent variable, y, to some predictors.

5. Support Vector Machine

Support Vector Machine (SVM) is a supervised algorithm that can classify cases by dividing a data set into two or more classes using a separator.

6. Random Forest

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

IV. CONCLUSION

From these studies, it is clear that there are various methods of data analysis of any application. Heart decease dataset is available from UCI Machine Learning Repository. It has been further preprocessed and cleaned out to prepare it for classification process. Decision trees, naive bayes, linear discriminant analysis, k-nearest neighbor, logistic regression, neural networks, and support vector machines are studied in this paper. The fact is that computers cannot replace humans and by comparing the computer-aided detection results with the pathologic findings, doctors can learn more about the best way to evaluate areas that computer aided detection highlights.

REFERENCES

1. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019.
2. W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou and W. Cao, "A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease," in *IEEE Access*, vol. 7, pp. 175248-175258, 2019.
3. T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams and I. C. Paschalidis, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach," in *Proceedings of the IEEE*, vol. 106, no. 4, pp. 690-707, April 2018.
4. A. Mdhaffar, I. Bouassida Rodriguez, K. Charfi, L. Abid and B. Freisleben, "CEP4HFP: Complex Event Processing for Heart Failure Prediction," in *IEEE Transactions on NanoBioscience*, vol. 16, no. 8, pp. 708-717, Dec. 2017.
5. A. Khan and A. Nigam, "Analysis of Intrusion Detection and Classification using Machine Learning Approaches", IJOSCIENCE, vol. 3, no. 10, Oct. 2017. DOI:https://doi.org/10.24113/ijoscience.v3i10.13.
6. D. Tay, C. L. Poh, E. Van Reeth and R. I. Kitney, "The Effect of Sample Age and Prediction Resolution on Myocardial Infarction Risk Prediction," in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1178-1185, May 2015.
7. D. R. Patil and J. S. Sonawane, "Prediction of heart disease using multilayer perceptron neural network," *International Conference on Information Communication and Embedded Systems (ICICES2014)*, Chennai, 2014, pp. 1-6.
8. R. Wijaya, A. S. Prihatmanto and Kuspriyanto, "Preliminary design of estimation heart disease by using machine learning ANN within one year," *2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T)*, Bandung, 2013, pp. 1-4.
9. N. G. B. Amma, "Cardiovascular disease prediction system using genetic algorithm and neural network," *2012 International Conference on Computing, Communication and Applications*, Dindigul, Tamilnadu, 2012, pp. 1-5.
10. K. R. Taylor *et al.*, "AudioGene: Computer-based prediction of genetic factors involved in non-syndromic hearing impairment," *2011 9th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, Sharm El-Sheikh, 2011, pp. 75-79.

11. B. A. Thakkar, M. I. Hasan and M. A. Desai, "Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier," *2010 International Conference on Advances in Recent Technologies in Communication and Computing*, Kottayam, 2010, pp. 101-105.
12. H. Y. Wang, H. Zheng and F. Azuaje, "Evaluation of computational classification methods for discriminating human heart failure etiology based on gene expression data," *2006 Computers in Cardiology*, Valencia, 2006, pp. 277-280.