A Literature Survey of Machine Learning Techniques for Classification and Prediction of Heart Disease

BHAVIKA CHANDRABHAN GUPTA[1], JITENDRA DANGRA[2]

[1, 2] *Laxmi Narayan College of Science and Technology, Indore*

Abstract-* *Data mining strategies have previously been used by a number of investigators in order to locate ailments. It is essential to keep in mind that not all methods of sickness prediction are developed in the same way. There is a possibility that the accuracy of disease prediction may be enhanced. In this article, we provided an overview of the many different methods for the classification of data that are presently being used. These algorithms, in a way, are symbolic representations of themselves. The classification of data is a common activity that requires a lot of processing power. In addition to this, we have established a foundation for classifying data. In the course of this activity, we will be analysing and comparing the most significant algorithms among the multitudes that are accessible nowadays. This study presents a literature survey of Machine learning techniques for classification and prediction of heart disease

Indexed Terms-* *Data Mining, Disease Prediction, Decision Tree, KDD

I. INTRODUCTION

In affluent nations, around 20% of persons over the age of 75 and over 5% of those under the age of 35 suffer from heart disease [1, 2]. Heart disease is both a dangerous and prevalent condition. Heart failure accounts for around three to five percent of all hospitalizations. According to the results of their clinical work, physicians have found that heart failure is the primary reason patients are admitted to hospitals. The costs are significant, accounting for as much as 20 percent of overall health expenditure in wealthy nations.

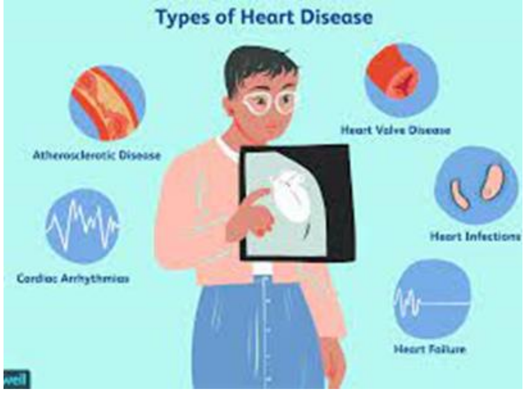

Figure 1- Types of Heart Disease

The inner workings of the heart may be affected in a broad variety of ways depending on the variety of cardiac disorders that a person has. Therefore, it is feasible to categorise any sickness connected to the heart as a cardiovascular disease [3], and some of these disorders will be discussed in more depth in the section that follows. Coronary heart disease is the most prevalent kind of heart disease in every region of the globe (CHD). It is also known as coronary artery disease in certain circles (CAD). This disorder is characterised by the presence of fatty deposits in the capillaries and veins of the circulatory system. As a direct result of this, the internal organs of the heart do not get the adequate amounts of oxygen and blood that they need and suffer from a lack of circulation. It is shown in figure 1 and figure 2.

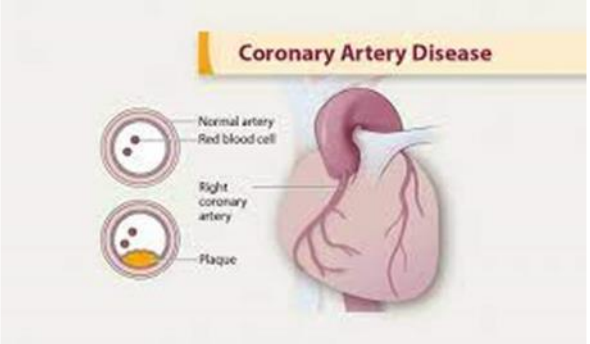

Figure 2- coronary artery disease

In order to design an all-encompassing plan for illness management, it is necessary to conduct an in-depth study of a mountain of data. Examples of common applications of artificial intelligence include illness early diagnosis, severity evaluation, and prognosis. This will minimise the pace at which the sickness spreads, enhance the quality of life for patients, and cut down on the expenses of providing healthcare. This is one example of how the study of machine learning may be put to use.

II. LITERATURE SURVEY

Bhuvaneswari and Kalaiselvi [9] conducted research about the Naive Bayes classifier and its applications in the realm of medicine. Back propagation Neural Network (BNN) and Naive Bayesian analysis, two of the most used data mining classification methods, were used to generate priors, also known as the likelihood of the item among all other items based on previous experiences (NB). The concept of probability serves as the cornerstone around which Bayesian analysis is built. They used Bayes' principles to draw conclusions about the posterior based on the prior. They discovered that, depending on the particulars of the probability model, Naive Bayes classifiers may be utilised to train quite well in a supervised learning environment. This was discovered despite the fact that these classifiers are not particularly sophisticated.

Voting was used by Shouman et al. [10] in order to integrate a variety of classifiers in order to improve upon the performance of individual classifiers. KNN was used in this study to assist medical professionals in recognising heart issues. In addition to this, they looked at the possibility of including a vote component into the KNN diagnostic approach to see whether it would enhance its performance. A process known as Voting was used to compile the decisions made by a number of different classifiers. The concept of voting with multiple classifiers is predicated on the notion of dividing the data used for training into equal parts and constructing a classifier for each subset of the data. The findings shown that voting did not result in an improvement in the accuracy of the K-nearest Neighbor method for detecting heart disease.

Senthil Kumar [11] introduced many aspects of fuzzy logic, including fuzzification, Advanced Fuzzy Resolution Mechanism, and defuzzification, all of which are included in the phrase. The process of converting discrete values into fuzzy ones is referred to as the "fuzzification" of such values. Using a fuzzy resolution technique with five levels, each of which is represented by its own set of nodes, the examination of heart disease makes use of this approach. According to the data that they came up with, the new approach that uses anticipated value would perform better than the older way when accuracy is used as a parameter in the detection of heart disease. Confirmation of the findings was accomplished by the use of validation on the Cleveland heart disease dataset. MATLAB was used during the construction of the fuzzy resolution mechanism. Through the process of defuzzification, we are able to transform a fuzzy set into a group of concrete numbers.

Back propagation is a technique that is used in the MLFFNN approach that was developed by Priya and her colleagues [12]. They were able to identify, with the use of a genetic algorithm, that the maximum number of neurons that might remain hidden could be accomplished. They discovered that improving accuracy with a bigger consistent rate utilising optimised hidden neurons produced the greatest results when they employed ANNs to predict stroke illness as part of their planned study. This was accomplished by employing ANNs. The group came up with an algorithm to analyse a patient's medical history and assess, on the basis of that information, whether or not the patient was having a stroke. The medical histories of three hundred individuals were reviewed. 180 people out of the total were sick. We used 196 training data points and 104 assessment data points in order to carry out the study that they recommended.

The RIPPER classifier, Decision Tree, Artificial Neural Networks (ANNs), and Support Vector Machine were the data mining classification approaches that were investigated and analysed by Kumari and Godara [13]. (SVM). In order to assess the sensitivity, specificity, and accuracy of each data mining approach that was under consideration, they used a lift chart in conjunction with an error rate. We looked at the structure as well as the analysis of these strategies.

The techniques of data mining that are based on categorization, such as Association rule, Decision tree, Nave Bayes, and Artificial Neural Network, were the focus of extensive research carried out by Srinivas and colleagues [14]. [14] Massive amounts of clinical data are collected by the healthcare business, but this data is seldom "mined" to uncover previously unknown patterns or correlations. Both the Naive Creedal Classifier 2 (NCC2) and the One Dependency Augmented Naive Bayes classifier were used in order to cleanse the data (ODANB). It was an erroneous hypothesis that aimed to provide strong classifications with scant or inadequate data, and Naive Bayes was an addition that was grafted on to that hypothesis.

It is easy to overlook that finding new patterns and connections in data is one of the most significant components of data analysis; nonetheless, this is one of the most crucial aspects. It is possible to predict the risk that a person will develop heart disease by using information about the individual's health, such as their age, gender, blood pressure, and blood sugar levels. Because of this, key facts, such as extraordinary patterns and connections between health markers connected to heart disease, were able to be established. A classifier model of advancing feature inclusion mixed with back-elimination is determined by Shilaskar et al. [15] for a variety of datasets, including those relevant to arrhythmia, cardiac disease, and ECGs. This model is used for determining classification. On the basis of empirical data, it was shown that the choice of features improved categorization processes while simultaneously reducing intakes. The arrhythmia dataset that is supplied results in a performance boost of 78% while using just 19% of the features that were initially used. The performance on the subsequent batch of data is 85% better, and there are just four features to take into account as opposed to eight. Research done in the past has shown that reduct features increase the efficiency of classifiers.

An ANN-based fuzzy inference system that offers a predictive technique for risk information and prediction was developed by Yang et al. [16]. The authors used ANFIS and linear LDA algorithms to develop their system. The inference system makes use of a hybrid classification strategy, and the findings demonstrate that it greatly surpasses the techniques that are considered to be state-of-the-art in terms of both its efficiency and its performance. It is helpful in detecting cardiac issues before they become life-threatening and in preventing them from developing in the first place.

Long et al. [17] presented the idea of an accurate prediction system that is accomplished via the use of a firefly-based algorithm and rough sets. Although a combination of fuzzy and rigorous theoretical approaches may help relieve both of these challenges, reduced uncertainty and decreased dimensionality remain two of the most pressing concerns about datasets related to heart disease. According to the roughest-based fuzzy learning approach, finding the optimal solutions calls for doing just the barest minimum of computer processes. The results obtained with this method are superior to those obtained with support vector machines and ANN when it comes to the prediction of heart illness and the treatment plan used.

Bayasi and colleagues [18] created a method that may predict the occurrence of ventricular arrhythmia. Within the scope of this investigation, we provide a completely integrated electrocardiogram (ECG) signal processor for pain prediction. A particular group of electrocardiogram (ECG) indicators may be used to make an accurate prediction about whether or not a human being has ventricular arrhythmia. It does this by detecting and labelling ECG waves so that it may later analyse and extract fiducial points (PQRST). In order to finish this procedure, we rely on approaches that are both real-time and flexible. These algorithms have the ability to effectively smooth out anomalies in ECG signals, which enables highly sensitive and accurate readings to be obtained. Evaluated by using the database of the American Heart Association, which contains information relevant to the heart. The outcomes of the simulations are shown to be more reliable than the findings obtained using the traditional approaches. In order to do the simulation, an application-specific integrated circuit is utilised (ASIC). This technique is noteworthy since it is the first ASIC implementation to employ ESP for the purpose of predicting ventricular arrhythmia.

In their study [19], Wang and colleagues proposed using a naïve Bayes classifier to determine an individual's likelihood of acquiring cardiovascular illness. [Citation needed] The first thing that has to be done is look at the patient's medical history to see if there are any warning indications of heart disease. The primary goal of this research is to enhance diagnostic and prognostic approaches for cardiovascular disease by achieving higher levels of sensitivity, accuracy, and specificity in the relevant metrics. Diabetes, elevated levels of cholesterol in the blood, and abnormalities in renal and vascular function are some of the primary risk factors that should be prioritised in order to diagnose heart disease in its early stages. Threats are categorised as level 1, level 2, or level 3 depending on their severity. Experiments have shown that this strategy delivers a more accurate forecast of heart illness than other approaches now in use (more than 80 percent). Patients, cardiologists, and other medical experts may all speak to the fact that our study improves upon earlier efforts at forecasting heart disease.

An analysis of the relevant published research indicates a vast number of strategies for the classification and forecasting of patient data. In spite of this, it has been shown that there is potential for development in terms of parameters like as accuracy, precision, error rate, and recall. In addition to it, the records of patients are not always complete. More research is required in the field of data preparation in order to realise the goals of achieving consistent and noise-free input data.

An analysis of the relevant published research indicates a vast number of strategies for the classification and forecasting of patient data. In spite of this, it has been shown that there is potential for development in terms of parameters like as accuracy, precision, error rate, and recall. In addition to it, the records of patients are not always complete. More research is required in the field of data preparation in order to realise the goals of achieving consistent and noise-free input data.

CONCLUSION

Techniques for data mining that are based on decision trees have the potential to be used for illness prognostic prediction. The findings of this research will have significant repercussions for the field of healthcare. It's likely that both patients and doctors might reap potentially enormous advantages from this development. There is still a lot of work that has to be done to improve the accuracy of the classifier by using various data mining classification methodologies such as rule-based inference, expectation maximisation, and so on. In this investigation, a critical analysis of the existing decision tree-based technique was carried out. In essence, an analysis and comparison of the benefits and drawbacks of their job is carried out.

REFERENCES

[1] Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[2] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.

[3] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.242.

[4] Manne, Ravi, and Sneha C. Kantheti. 2021. “Application of Artificial Intelligence in Healthcare: Chances and Challenges”. Current Journal of Applied Science and Technology 40 (6), 78-89. https://doi.org/10.9734/cjast/2021/v40i631320

[5] R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570-575, doi: 10.1109/ICICCS48265.2020.9121169.

[6] Chen, Joy Iong Zong, and P. Hengjinda. "Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative

Study." Journal of Artificial Intelligence 3, no. 01 (2021): 17-33.

[7] Mehmood, Awais, Munwar Iqbal, Zahid Mehmood, Aun Irtaza, Marriam Nawaz, Tahira Nazir, and Momina Masood. "Prediction of Heart Disease Using Deep Convolutional Neural Networks." Arabian Journal for Science and Engineering 46, no. 4 (2021): 3409-3422.

[8] Karunakaran, P., and Yasir Babiker Hamdan. "Early Prediction of Autism Spectrum Disorder by Computational Approaches to fMRI Analysis with Early Learning Technique." Journal of Artificial Intelligence 2, no. 04 (2020): 207-216.

[9] Bhuvaneswari, R & Kalaiselvi, K 2012, 'Naive Bayesian Classification Approach in Healthcare Applications', International Journal of computer Science and Telecommunication, vol. 3, no. 1, pp. 106-112

[10] Shouman, M, Turner, T & Stocker, T 2012, 'Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients', International Journal of Information and Education Technology, vol. 2, no. 3, pp. 220-224

[11] Senthil Kumar, A 2013, 'Diagnosis of Heart Disease Using Advanced Fuzzy Resolution Mechanism', International Journal of Science and Applied Information Technology, vol. 2, no. 2, pp. 22-30

[12] Priya, K, Manju, T & Chitra, R 2013, 'Predictive Model of Stroke Disease Using Hybrid Neuro-Genetic Approach', International Journal of Engineering and Computer Science, vol. 2, no. 3, pp. 781-788

[13] Kumari, M & Godara, S 2011, 'Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction', International Journal of Computer Science and Technology, vol. 2, no. 2, pp. 304-308

[14] Srinivas, K, Kavitha Rani, B & Govardhan, A 2010, 'Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks', International Journal on Computer Science and Engineering ,vol. 2, no. 2, pp. 250-255

[15] Shilaskar, S. and Ghatol, A. (2013), 'Expert systems with applications feature selection for medical diagnosis : evaluation for cardiovascular Diseases', Journal of expert sy stem with application 4(10), 4146–4153.

[16] Yang, X., Li, M., Zhang, Y. and Ning, J. (2014), 'Cost-sensitive naive bayes classification of uncertain data', Journal of Scientific World 9(8), 1897–1904.

[17] Long, Nguyen Cong, M. H. (2015), 'A highly accurate firefly based algorithm for heart disease prediction', Journal of Expert Systems with Applications 42(21), 8221–8231.

[18] Bayasi, N. and Tekeste (2016), 'Low-power ECG-based processor for predicting ventricular arrhythmia', Journal of IEEE transactions on very large scale integration systems 24(5), 1962–1974.

[19] Zhiyong Wang, Xinfeng Liu, J. G. (2016), 'Identification of metabolic biomarkers in patients with type-2 diabetic coronary heart diseases based on metabolomic approach', 6(30), 435–439.

[20] Hssina, Merbouha, A. and Ezzikouri (2014), 'A comparative study of decision tree ID3 and C4.5', International Journal of Advanced Computer Science and Applications 4(2), 13–19.

[21] Cheng-Hsiung Wenga, Tony Cheng-Kui Huang, R.-P. H. (2016), 'Disease prediction with different types of neural network classifiers', Journal of Telematics and Informatics (4), 277–292.

[22] Zhang, Shuai, Y.-L. S. A. (2017), 'Deep learning based recommender system: a survey and new perspectives', Journal of ACM Computing Surveys 1(1), 1–35.