HEART DISEASE PREDICTION USING MACHINE LEARNING

Tirtharaj Tamang
[1]Student
[1] Master Of Technology In Computer Science & Engineering
[1]DR. B. C. Roy Engineering College, Durgapur, India.

Abstract - Across the globe, cardiovascular illnesses are now among the leading causes of death. This worrying problem has been greatly exacerbated by changes in eating, working, and lifestyle patterns in both developed and developing countries worldwide. Early identification of the first symptoms of cardiovascular disorders and ongoing medical care can assist lower the number of patients who are becoming sicker and ultimately the death rate. The number of heart disease cases is rising quickly every day, making it crucial and worrisome to anticipate any such illnesses in advance. This diagnosis is a challenging task that requires accuracy and efficiency. The study article primarily focuses on identifying patients who has given a variety of medical characteristics, are more likely to suffer heart disease. Using the patient's medical history, we developed a heart disease prediction algorithm to determine the likelihood of a heart disease diagnosis. Utilizing various machine learning methods, including logistic regression and KNN, we were able to predict and categorize the patient with heart disease. A very useful method was employed to control the model's ability to increase any person's heart attack prediction accuracy. When compared to the previously used classifiers, such as naive bayes etc, the accuracy of the suggested model's use of KNN and Logistic Regression to predict signs of heart disease in a specific individual was quite satisfying. Thus, by utilizing the provided model to determine the likelihood that the classifier can correctly and precisely diagnose cardiac illness, a sizable amount of pressure has been released. The Given heart disease prediction system optimizes medical care and reduces the cost. We get important information from this experiment that will aid in the prediction of heart disease patient's.

Index Terms- Heart Disease is one of the most significant causes of mortality in today's world.

I.INTRODUCTION

Due to a number of contributing risk factors, including diabetes, high blood pressure, excessive cholesterol, an irregular pulse rate, and many more, it is challenging to diagnose heart disease. A number of data mining and neural network techniques have been used to determine the severity of cardiac disease in humans. Numerous techniques, including the Genetic Algorithm (GA), Decision Trees (DT), K-Nearest Neighbour Algorithm (KNN), and Naive Bayes (NB), are used to classify the severity of the condition, because heart illness has a complicated character, it needs to be treated properly. Failing to do so could damage the heart or result in an early death. To identify different types of metabolic syndromes, data mining and the perspective of medical research are employed. Heart disease prediction and data analysis benefit greatly from data mining and classification. Decision trees have also been used to forecast the accuracy of cardiac disease related events. Many techniques have been used to the established data mining approaches for the prediction of heart disease in order to abstract knowledge. Many readings have been done in this study to create a prediction model that relates two or more procedures in addition to employing distinct techniques.

These combined new approaches are referred to as hybrid approaches. We introduce neural networks utilizing heart rate time series. In order to determine the precise state of the patient with regard to heart disease, this method makes use of a variety of clinical records for prediction, including Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC), and Second degree block (BII). Thirty percent of the data is utilized for classification and seventy percent of the dataset is used for training with a radial basis function network (RBFN). We also discuss the use of Computer Aided Decision Support Systems (CADSS) in research and medicine. It has been demonstrated in earlier research that the healthcare sector can forecast sickness more accurately and in less time by utilizing data mining techniques. We suggest using the GA to diagnose cardiac disease. For the purposes of tournament selection, crossover, and mutation—which yields the newly suggested fitness function—this approach makes use of efficient association rules deduced from the GA.

We leverage the popular Cleveland dataset, gathered from a UCI machine learning repository, for experimental validation. When compared to some of the well-known supervised learning methods, we will see later on how significant our results turn out to be. We present Particle Swarm Optimisation (PSO), the most potent evolutionary method, and find several heart disease-related rules. All things considered, the accuracy has improved because the rules have been applied randomly with encoding approaches. Many symptoms, including age, sex, pulse rate, and many more, are used to indicate heart disease. Neural Network based machine learning technique is presented, with observable improvements in accuracy and dependability. Most people agree that neural networks are the best method for predicting conditions like brain and heart illnesses. There are thirteen attributes in the suggested strategy that we employ to predict heart disease. The outcomes demonstrate a higher degree of performance in works like when compared to the current methods. In recent years, the Carotid Artery Stenting (CAS) has also gained popularity as a therapy modality in the medical community. Elderly patients with heart disease who experience major adverse cardio- vascular events (MACE) are prompted by the CAS. Their assessment assumes paramount significance. We use an Artificial Neural Network (ANN) to create findings, and it performs well in heart disease prediction. We present neural network algorithms that integrate predicted values from several previous methods along with posterior probability. When compared to earlier studies, this model's accuracy level of up to 89.01% is impressive. To enhance the performance of heart disease, a neural network (NN) is employed in all studies with the Cleveland heart

dataset. Recent advancements in machine learning (ML) approaches have also been observed in the context of the Internet of Things (IoT). It has been demonstrated that using machine learning algorithms on network traffic data, IoT devices connected to a network may be accurately identified. Meidan et al. gathered and annotated network traffic data from nine different PCs, cellphones, and IoT devices. They trained a multi-stage meta classifier via supervised learning. The classifier can discriminate between traffic produced by IoT and non-IoT devices in the first stage. Every IoT device is connected to a particular IoT device class in the second stage. Deep learning, with its multilayer structure, is a suitable method for edge computing environments and a potential way for accurately extracting in formation from raw sensor data from Internet of Things devices deployed in complicated contexts.

II. LITERATURE SURVEY

The fields directly relevant to this study have a wealth of related work. The goal of ANN's introduction was to provide the medical industry's highest level of prediction accuracy. To forecast cardiac illness, artificial neural networks (ANNs) use back propagation multilayer perception (MLP). The generated results are shown to be improved when compared to the results of existing models within the same domain. NN, DT, Support Vector machines (SVM), Naive Bayes, and other machine learning techniques are utilised to find patterns in the patient data from the UCI laboratory that certains to heart disease. With these algorithms, the accuracy and performance of the outcomes are compared. In comparison to the other current methodologies, the suggested hybrid approach yields results for the F- measure of 86.8%.

The introduction of Convolutional Neural Networks (CNN) classification without segmentation during the training phase, this approach takes into account the cardiac cycles with different start positions from the Electrocardiogram (ECG) signals. In the patient's testing phase, CNN can produce features with different positions. Previously, a significant quantity of data produced by the medical sector was not utilised efficiently. The novel methods offered here reduce costs and enhance heart disease prediction in a simple and efficient manner. The several research approaches taken into consideration in this work for the deep learning (DL) and machine learning (ML)-based classification and prediction of heart dis- ease are quite accurate in demonstrating the effectiveness of these techniques

Overview of Method and Results:

In Hybrid Random Forest with Linear Model (HRFLM) we utilise a computational method to identify the heart disease risk factors on the UCI Cleve- land dataset using the three association rules of mining: apriori, predictive, and tertius. Based on the evidence that is now available, it may be concluded that women are less likely than men to have heart disease. A precise diagnosis is crucial for heart disorders. However, reliable diagnosis and prediction are beyond the capabilities of the conventional methods. Thirteen clinical characteristics and an ANN with back propagation are used as input by HRFLM. The results are assessed in comparison with conventional approaches. The danger becomes extremely high, and certain characteristics are taken into account to provide a precise diagnosis of the illness. Due to the intricacy and nature of cardiac disease, an effective treatment strategy is necessary.

In the medical field, data mining techniques are helpful in corrective circumstances. DT, NN, SVM, and KNN are other considerations for data mining techniques. Out of all the techniques used, the SVM results show promise for improving disease prediction accuracy. To detect arrhythmias such as bradycardia, tachycardia, atrial, and atrial-ventricular flutters, among many others, a nonlinear method incorporating a module for heart function monitoring is presented. The accuracy of the results based on ECG data can be used to measure the performance efficiency of this method. Accurate disease diagnosis and potential abnormality prediction in patients are made possible by ANN training. Heart disease can be identified and predicted using a variety of data mining techniques and prediction methods, including KNN, LR, SVM, NN, and Vote. This research proposes a hybrid technique employing LR and NB along with the innovative method Vote. The suggested method's studies, which yielded 87.4% accuracy in heart disease prediction, were carried out on the UCI dataset. The evaluation of the Probabilistic Principal Component Analysis (PPCA) approach is suggested, utilising three distinct data sets from UCI: Cleveland, Switzerland, and Hungarian. The technique minimizes the feature dimension by extracting the vectors with high covariance and vector projection. A radial basis function is given the feature selection with minimizing dimension, enabling kernel-based SVM. The approaches' respective outcomes for the UCI data sets of Cleveland, Switzerland, and Hungarian are 82.18%, 85.82%, and 91.30%. The primary innovative contribution of this research is the introduction of a hybrid method that com- bines Linear Regression (LR), Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Networks (ANN) using rough set techniques.

The set of critical qualities was effectively reduced by the proposed strategy. The remaining attributes are then fed into the ANN. The effectiveness of the hybrid approach's development is illustrated through the usage of datasets related to heart disease. It is suggested to use multilayer perception of NN for the prediction of heart disease. In order to determine if a patient has heart disease or not, this method employs 13 clinical attribute features as the input and is trained using back propagation, yielding remarkably precise results. In order to more precisely forecast cardiac disease, we also present the Apriori algorithm with Support Vector Machine(SVM) and contrast it with nine other classification techniques. In comparison to other approaches now in use, the classification method's results have demonstrated a greater degree of accuracy and performance in the prediction of heart dis- ease. A significant factor in the prediction of heart disease is feature selection. A back propagation neural network (ANN) is suggested for improved disease prediction. The application of ANN yields extremely precise and accurate results. Recurrent Fuzzy Neural Network (RFNN), a genetic method incorporating fuzzy neural network, is presented for the diagnosis of heart disease.

It is suggested to use SVM and ANN for heart disease prediction. For the testing time and accuracy premise, this methodology employs two ways. For additional analysis, the suggested model divides the data records into two classes using both SVM and ANN. The introduction of the Back Propagation Neural Network (BPNN) with classification approach generates the gene sequence for hypertension first, and then the precise gene sequence. With varying quantities of samples, the effectiveness of the BPNN ap- proaches has been evaluated both during the training and testing phases. This method's accuracy has increased as the quantity of records has increased.

Proposed Method HRFLM:

In this work, we classified the cardiac illness in the Cleveland UCI repository using a R studio rattle. It offers a user-friendly visual depiction of the dataset, the workspace, and the predictive analytics development process. The machine learning method begins with pre-processing the data, then moves on to feature selection using DT entropy, classification of modelling performance evaluation, and better accuracy results. The process of se- lecting features and modelling is repeated for different combinations of qualities. Every model's performance is tracked, with each iteration and performance based on 13 features and ML approaches. Section A provides an overview of the pre-processing of the data, Section B covers feature selec- tion by entropy, Section C describes the classification process using ma- chine learning approaches, and Section D presents the performance of the results.

Pre-processing of Data:

Data about heart disease is pre-processed following the acquisition of different records. There are 303 patient records in the dataset overall, with 6 records having partially missing data. The remaining 297 patient records are used for pre-processing after those 6 records were eliminated from the dataset. For the properties of the provided dataset, binary classification and multi-class variables are introduced. To determine whether cardiac disease is present or not, the multi-class variable is employed. If the patient has heart illness, the value is set to 1, otherwise it is set to 0, indicat- ing that the patient does not have heart disease. Preprocessing of the data is done by translating diagnosis values from medical records. After 297 patient records were pre-processed, the results showed that 137 of the records had a value of 1, indicating that heart disease was present, and 160 records had a value of 0, indicating that heart disease was absent.

Selection and Reduction of Features:

Two characteristics related to age and sex are selected from the data set's 13 attributes in order to identify the patient's personal information. Since the remaining 11 qualities include crucial clinical records, they are regarded as significant. Clinical data are essential for diagnosing cardiac disease and determining its severity. This experiment employs a number of ma- chine learning (ML) techniques, including NB, GLM, LR, DL, DT, RF, GBT, and SVM, as was previously described. Every ML technique was used in the experiment, and all 13 attributes were used.

Modelling of Classification:

Datasets are clustered according to the properties of Decision Trees (DT) and their variables. Subsequently, the classifiers are employed on every clustered dataset to assess its overall performance. Based on their low error rate, the top-performing models are distinguished from the aforementioned results. By selecting the DT cluster with the highest error rate and extracting the appropriate classifier features, the performance is further enhanced. Using this data set, the classifier's performance is assessed for error optimisation.

III. Assessment Findings

Thirteen features are used in the development of the prediction models, and modelling methodologies' accuracy is computed. The accuracy, preci- sion, F-measure, sensitivity, specificity, and classification error are all compared in this table. When compared to other approaches currently in use, the HRFLM classification method yields the highest accuracy.

Talking about HRFLM to Get Better Outcomes:

Based on categorisation procedures, the UCI dataset is further divided into 8 categories. A list of the classification rules. R-Studio Rattle performs additional classification and processing on each dataset. The dataset's classification rule is applied to provide the results. The completion of data pre-processing, the categorisation rules are constructed based on the rule. The three best machine learning techniques for the data are selected after pre-processing, and the results are produced. The optimum classification technique is determined by using the different datasets with DT, RF, and LM. The best two are RF and LM, according to the results. In contrast to the previous datasets, dataset 4 has a significant RF error rate (20.9%). When compared to the DT and RF approaches, the LM method performs the best on the dataset (9.1%). To improve the outcomes, we suggest the HRFLM approach, which combines the LM and RF methods.

Comparing the Suggested Model to Others:

To evaluate how well the suggested model performs in comparison to the current models, benchmarking is required. This technique is used to determine whether or not the suggested strategy is the most effective and increases accuracy. The accuracy is determined by multiplying the number of selected features by the output of the model. HRFLM is not limited in what functionalities it can choose to use. This model's attributes are all chosen to produce the greatest outcomes. The sug- gested approach is applied to all 13 attributes and categorised. When compared to established models, this result unequivocally demonstrates that all of the characteristics chosen and the machine learning approaches applied are successful in correctly predicting patients' heart disease.

IV. CONCLUSIONS

Understanding how to interpret unprocessed medical data about the heart can save lives in the long run and aid in the early identification of irregulari- ties in cardiac diseases. In this work, raw data was processed using ma- chine learning techniques to produce a unique and novel diagnosis of heart disease. Predicting heart disease is a difficult but crucial task in medicine. However, if the condition is identified early and preventative measures are implemented as soon as feasible, the death rate can be significantly reduced. In order to focus the research on real-world datasets rather than merely theoretical approaches and simulations, it would be highly desired to extend this work further. The suggested hybrid HRFLM approach com- bines the benefits of the Linear Method (LM) and Random Forest (RF) techniques. When it came to predicting heart disease, HRFLM turned out to be fairly accurate. This research can be carried out in the future using various combinations of machine learning algorithms to improve prediction techniques. To improve the accuracy of heart disease prediction, new feature selection techniques can be created to obtain a wider understanding of the important traits.

V. REFERENCES

[1] *https://ieeexplore.ieee.org/abstract/document/8740989*

[2] *https://towardsdatascience.com/predicting-presence-of-heart-diseases using- machine-learning-36f00f3edb2c*

[3] *https://www.geeksforgeeks.org/7-major-challenges-faced-by-machine-learning-professionals/-*