International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.013 ISSN: 2581-4621

Heart Disease Prediction Using Machine Learning

Mrs.Chandana[1], K.Shamala[2], K.Akshaya[3], K.Nitin[4], B.Sravani[5]

[1]Assistant professor, Dept. Of CSE, Hyderabad Institute Of Technology And Management
[2,3,4,5]UG Student,Dept. Of CSE, Hyderabad Institute Of Technology And Management

Abstract**—** In the medical field, diagnosing heart disease is one of the most challenging tasks. It relies on analyzing large sets of clinical and pathological data, making the process complex. As a result, there has been a growing interest among researchers and clinical professionals in developing efficient and accurate methods for heart disease prediction. Early diagnosis of heart disease is crucial due to its potential severity, as timely intervention can significantly impact patient outcomes. Heart disease is a leading cause of death globally, underscoring the importance of predicting it at an early stage. In recent years, machine learning has emerged as a reliable tool in the medical domain, offering valuable support in disease prediction when provided with appropriate training and testing data. A range of machine learning algorithms, encompassing decision trees, support vector machines, random forests, neural networks, and ensemble methods, have been employed to analyze varied datasets containing demographic information, clinical parameters, and medical imaging data. To improve prediction accuracy and interpretability, techniques for feature selection and model

Optimizations have been utilized. The primary objective of this study is to explore various prediction models for heart disease and identify key features using the Random Forest algorithm. Random Forest is a supervised machine learning algorithm known for its high accuracy compared to other methods such as logistic regression. Utilizing the Random Forest algorithm, we aim to predict whether an individual has heart disease or not.

Keywords**—** Heart disease diagnosis, Machine learning, Random Forest algorithm, Prediction, Early detection, Medical data analysis, Feature selection, Accuracy

I. INTRODUCTION

The heart, a muscular organ responsible for pumping blood throughout the body, is central to the cardiovascular system, which also includes the lungs and a network of blood vessels such as veins, arteries, and capillaries. Disruptions in normal blood flow from the heart can lead to various heart diseases, collectively known as

cardiovascular diseases (CVDs), which are a leading cause of mortality worldwide. According to the World Health Organization (WHO), approximately 17.5 million deaths globally are attributed to heart attacks and strokes, with over 75% occurring in middle and low-income countries. Moreover, around 80% of CVD-related deaths are due to stroke and heart attack.

With the advancement of healthcare systems, large volumes of patient data, often referred to as Big Data in Electronic Health Record Systems, are now available. This wealth of data presents an opportunity for designing predictive models for cardiovascular diseases using techniques such as data mining and machine learning. The implementation of a heart disease prediction system using machine learning algorithms can be validated on open-access heart disease prediction datasets. Data mining facilitates exploratory analysis by extracting nontrivial information from vast amounts of evidence. Medical data mining holds promise for uncovering cryptic patterns in clinical datasets, which can aid in healthcare diagnosis.

However, medical data are often distributed, voluminous, and heterogeneous, necessitating organization and integration into a structured medical information system. Classification algorithms, such as Random Forest and Logistic Regression, are utilized in this project to predict various heart-related problems. These algorithms enable the exploration of different heart disease scenarios and contribute to disease prediction in healthcare data. Heart disease prediction using machine learning is driven by the urgent need to tackle the substantial impact of cardiovascular diseases on global public health. CVDs, such as heart attacks and strokes, are major contributors to worldwide mortality, especially in middle and low-income nations.

Discovering patterns and relationships between various risk factors and the likelihood of developing heart disease aids in the development of effective prevention and treatment strategies. The project's goal is to develop a robust predictive model using machine learning algorithms to identify individuals at risk of developing heart disease. By leveraging large volumes of patient data, including demographic information, clinical parameters, and medical imaging data, the project aims to create a tool capable of accurately predicting heart disease outcomes.

Ultimately, the project aims to contribute to the early detection and prevention of heart disease by providing healthcare professionals with a reliable and efficient tool for risk assessment. By identifying individuals at high risk of developing heart disease, healthcare providers can implement targeted interventions and treatment plans, ultimately improving patient outcomes and reducing the global burden of cardiovascular diseases.

Preprocessing of these datasets involves cleaning, normalization, and feature extraction to ensure data quality and relevance for predictive modeling. Utilizing machine learning algorithms, the project seeks to identify the most informative features or variables associated with heart disease risk. This involves employing techniques such as feature selection algorithms, dimensionality reduction, and model optimization to develop accurate predictive models.

II. LITERATURE SURVEY

Heart disease is a global health concern, causing significant morbidity and mortality. Machine learning techniques are being used to predict heart disease, leveraging healthcare data. Supervised Machine Learning algorithms, such as Random Forest, are used for data classification and prediction, resulting in better accuracy and timely intervention.

Alsharqi et al. (2021) conducted a systematic review of the use of machine learning techniques for predicting cardiovascular disease risk. They found that machine learning algorithms, such as logistic regression, decision trees, and artificial neural networks, can be used to develop predictive models for cardiovascular disease with high accuracy.

Abawajy et al. (2020) developed a deep learning approach for diagnosing heart disease using a combination of convolution and recurrent neural networks. They found that their approach achieved high accuracy in diagnosing heart disease using clinical data. The Heart Disease Dataset provides information about various demographic and clinical variables that are associated with heart disease, including age, sex, blood pressure, cholesterol levels, electrocardiographic results, and exercise induced angina.

Abdi et al. (2021) conducted a systematic review and meta-analysis of the use of machine learning for predicting cardiovascular disease risk using electronic health record data. They found that machine learning algorithms can provide accurate predictions of cardiovascular disease risk using routinely collected clinical data from electronic health records.

Kumar, Sharma, and Verma, published in the International Journal of Computer Applications in **2020,** examine the application of machine learning techniques in predicting heart disease. It discusses the utilization of algorithms like decision trees, k-nearest neighbors, and naive Bayes classifiers for this purpose. Additionally, it explores methods such as feature selection and model optimization to enhance the accuracy of heart disease prediction models.

In **2021, Wang, Wong, and Goldberg** concentrate on machine learning methods for forecasting cardiovascular disease risk through electronic health records (EHRs). It assesses the efficacy of various algorithms in predicting heart disease outcomes while addressing the hurdles linked with harnessing EHR data for predictive modeling. This comprehensive analysis sheds light on the potential and limitations of utilizing machine learning in leveraging EHRs for cardiovascular disease risk prediction.

Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. International journal on recent and innovation trends in computing and communication, 2(10), 3003-3008. Heart disease remains one of the leading causes of morbidity and mortality globally. With the increasing availability of healthcare data, data mining techniques have become essential in predicting and diagnosing heart disease.

III. METHODOLOGY

In this research, the process begins with gathering data and selecting relevant attributes. The collected data is then pre-processed to fit the required format. Next, the data is divided into two sets: one for training and the other for testing..

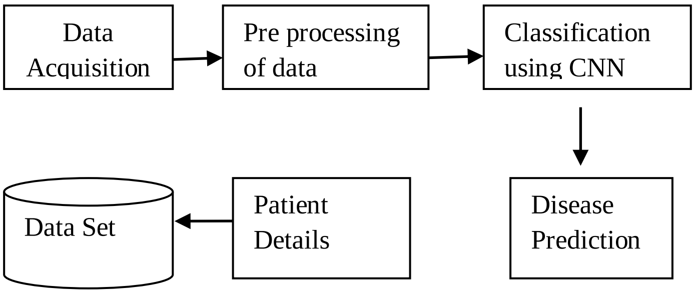

Figure: SYSTEM ARCHITECTURE

EXISTING SYSTEM:

Heart disease prediction traditionally relies heavily on clinical judgment and experience of healthcare professionals rather than utilizing the vast amount of data available in medical databases. This reliance on intuition may lead to biases, errors, and increased medical costs, ultimately affecting the quality of patient care. Despite advancements in medical technology, misdiagnosis of heart disease remains a significant concern, with studies indicating that a substantial portion of medical patients have experienced errors or missed diagnoses. Machine learning techniques offer a promising avenue for improving heart disease prediction by leveraging the wealth of data stored in electronic health records (EHRs), medical imaging databases, and other healthcare repositories. These algorithms can analyse large datasets containing patient demographics, clinical parameters, genetic information, and lifestyle factors to identify patterns and correlations indicative of heart disease risk. Medical Misdiagnoses are a serious risk to our healthcare profession. If they continue, then people will fear going to the hospital for treatment. We can put an end to medical misdiagnosis by informing the public and filing claims and suits against the medical practitioners at fault.

PROBLEM STATEMENT:

Traditional methods to predict heart disease are unreliable because they require manual analysis and only consider a few pieces of information. This heart disease prediction project can cause delays in diagnosing and treating the disease. Also, these methods don't provide real-time monitoring or personalized risk assessment, which is a big problem.

PROPOSED SYSTEM:

Based on the above review, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms has performed extremely well in some cases but poorly in some other cases. Alternating decision trees when used with PCA, have performed extremely well but decision trees have performed very poorly in some other cases which could be due to over fitting. Random Forest and Ensemble models have performed very well because they solve the problem of over fitting by employing multiple algorithms (multiple Decision Trees in case of Random Forest). Models based on Naïve Bayes classifier were computationally very fast and have also performed well. SVM performed extremely well for most of the cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and over fitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data.

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible.

MODULES:

There are four Modules to predict the heart disease using machine learning below are the steps :

1. DATA PREPROCESSING:

This file contains all the pre-processing functions needed to process all input documents and texts. First we read the train, test and validation data files then performed some preprocessing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

2. FEATURE:

Extraction In this file we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used methods like simple bag-of-words and n-grams and then term frequency like tf-tdf weighting.

3. CLASSIFICATION:

Here we have built all the classifiers for the breast cancer diseases detection. The extracted features are fed into different classifiers. We have used Logistic Regression and Random forest classifiers. After fitting all the classifiers, 2 best performing models were selected as candidate models for heart diseases classification. We have also used Precision-Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers.

4. PREDICTION:

Our finally selected and best performing classifier was algorithm which was then saved on disk with name final_model.csv once you close this repository, this model will be copied to user's machine and will be used by prediction.py file to classify the Heart diseases. It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of truth**.**

RESULT AND DISCUSSION:

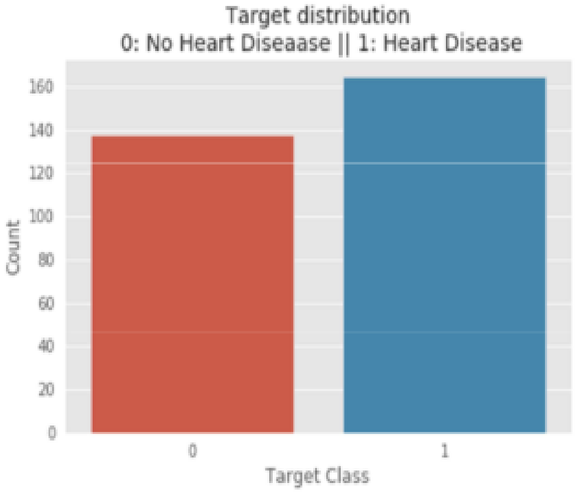

Figure: TARGET CLASS

Before going to study the performance of considering machine learning algorithms in this research, analysis of the features of the heart disease dataset will be focused on here. The total number of observations in the target attributes is 1025, where not having heart disease 499(denoted by 0) and having heart disease 526 (represented by 1), So, the percentage of not having heart disease is 45.7%, and the percentage of having heart disease is 54.3%.

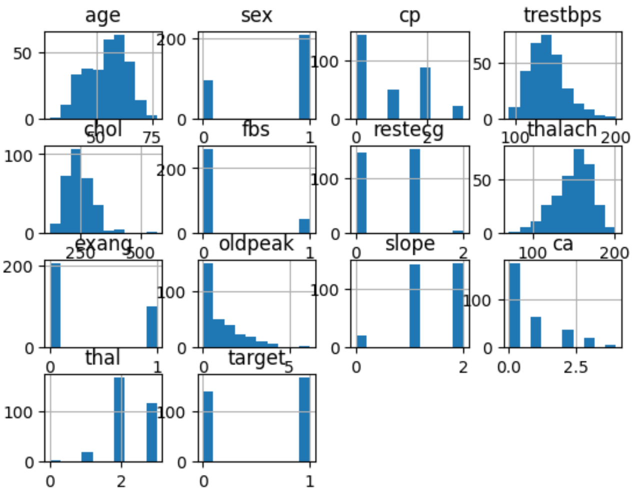

Figure: HISTOGRAM OF ATTRIBUTES

In heart disease prediction using machine learning, histograms of attributes play a crucial role in data exploration and pre processing. By visualizing the distribution of key clinical features such as age, cholesterol levels, resting blood pressure, and maximum heart rate, histograms help identify patterns and anomalies within the dataset. These visualizations can highlight skewness, outliers, and the overall spread of the data, which are essential for feature engineering and normalization processes. Understanding the distribution of these attributes aids in refining predictive models and improving their accuracy in diagnosing heart disease.

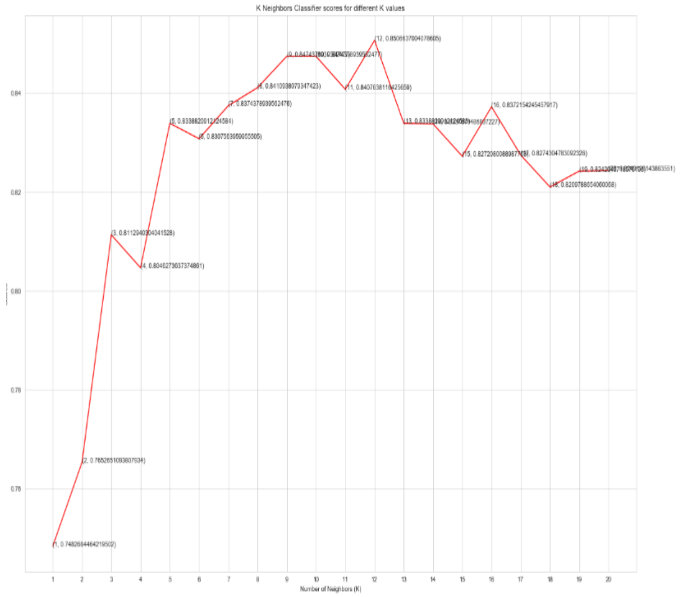

Figure: EARLY AND ACCURATE DETECTION

CONCLUSION:

In this project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Random Forest and Logistic Regression: we have analysed that the Random Forest has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Random Forest by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task. The computational time was also reduced which is helpful when deploying a model. It was also found out that the dataset should be normalized; otherwise, the training model gets over fitted sometimes and the accuracy achieved is not sufficient when a model is evaluated for real-world data problems which can vary drastically to the dataset on which the model was trained. It was also found out that the statistical analysis is also important when a dataset is analysed and it should have a Gaussian distribution, and then the outlier's detection is also important and a technique known as Isolation Forest is used for handling this. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning and ML as well. The algorithm applied by using ANN architecture increased the accuracy which we compared with the different researchers. The dataset size can be increased and then deep learning with various other optimizations can be used and more promising results can be achieved. Machine learning and various other optimization techniques can also be used so that the evaluation results can again be increased. More different ways of normalizing the data can be used and the results can be compared. And more ways could be found where we could integrate heart-disease-trained ML and DL models with certain multimedia for the ease of patients and doctors.

FUTURE SCOPE:

These findings suggest that machine learning algorithms can effectively learn about the disease predictions. We may extend this kind of study to diagnose other diseases. We may also analyze the past history of data and combine other machine learning techniques for better study. Other possible further applications of this study can include such as, cardiovascular disease prediction, diabetic prediction, breast cancer prediction, tumor prediction, and multiple disease predictions.

REFERENCES:

1. SibghaTaqdees,"Heart Disease Prediction"-Department of Software Engineering, Fatima Jinnah Women University (2021)

2. Harshit Jindal, "Heart Disease prediction using Machine Learning Algorithm", Dept. of Electronics and Communications Engineering, Bharti Vidyapeeth's College of Engg.,New Delhi(2020)

3. Kelvin Kwakye. "Machine Learning Based classification Algorithm for CoronaryHeart Disease".(2021)

4. Rubini PE, "A Method for Improving Prediction of Human Heart disease usingMLAlgorithm", CMR Institute of Technology (2021)

5. Pronab Ghosh, "Efficient Prediction of Cardiovascular Disease usingMLAlgorithm with Relief and LASSO algorithm", Lakehead University, Ontario, Canada(2020)

6. Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014), A coronary heart disease prediction model. The Korean Heart Study. BMJ open, 4 (5), e005025.

7. Mai Shouman, Tim Turner, and Rob Stocker. Applying kNearest Neighbour in diagnosis heart disease patients.. International Journal of Information and Education Technology, vol. 2, No. 3, June 2012.

8. Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Arnlov J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33 (9), 2267-72.

9. Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischeme heart disease (heart attact) risk prediction using clinical data and data mining approaches, a prototype design. 19th International conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

10. Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415, 190-8.

11. Wikipedia contributors. (2022, June 21). Logistic regression. In Wikipedia, The Free Encyclopedia. Retrieved 06:36, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Logistic_regressio n&oldid=1094256072.

12. Wikipedia contributors. (2022, June 1). Linear regression. In Wikipedia, The Free Encyclopedia. Retrieved 06:39, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Linear_regression &oldid=1091044459.

13. Wikipedia contributors. (2022, June 4). K-nearest neighbors algorithm. In Wikipedia, The Free Encyclopedia. Retrieved 06:40, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Knearest_neighbors_algorithm&oldid=1091525121.

14. Wikipedia contributors. (2022, June 20). Random forest. In Wikipedia, The Free Encyclopedia. Retrieved 06:41, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Random_forest&ol did=1094130824.

15. Wikipedia contributors. (2022, June 15). Decision tree learning. In Wikipedia, The Free Encyclopedia. Retrieved 06:42, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Decision_tree_lear ning&oldid=1093316444.

16. Wikipedia contributors. (2022, June 24). Gradient boosting. In Wikipedia, the Free Encyclopedia. Retrieved 06:43, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Gradient_boosting &oldid=1094845596.

17. nikki2398. (02 Sep, 2020). ML–Gradient Boosting. https://www.geeksforgeeks.org/ml-gradient-boosting/.

18. Wikipedia contributors. (2022, June 20). Support-vector machine. In Wikipedia, The Free Encyclopedia. Retrieved 06:51, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Supportvector_machine&oldid=1094109362.