Using Machine Learning for Heart Disease Prediction

Mrs T.Seeniselvi[1], Ms V.HariniPriya[2], Mr P.Gopinath[3]
[1]*Associate Professor, Department of Computer Science, Hindusthan College of Arts& Science, Coimbatore*
[2]*IIPG Student, Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore*
[3]*IIPG Student, Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore*

Abstract: Heart disease remains a leading cause of mortality worldwide, necessitating effective predictive tools to aid in early diagnosis and prevention. This study explores the application of machine learning (ML) techniques for predicting heart disease using a dataset of 1,200 patient records from the Mohand Amokrane EHS Hospital in Algiers, Algeria. The dataset includes 20 attributes related to patient health, such as age, sex, cholesterol levels, and blood pressure. We employed three prominent ML algorithms Neural Networks, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN)—to develop predictive models. The study involved several key phases: data collection, manual exploration, pre-processing, and modeling. Feature selection was performed using a Pearson correlation matrix, which identified 13 significant attributes for model training. The dataset was then split into training and testing subsets, with 80% allocated for training and 20% for testing. The performance of each algorithm was evaluated across different dataset sizes (600, 800, 1000, and 1200 records). Our results demonstrate that Neural Networks achieved the highest accuracy, consistently outperforming SVM and KNN, with an overall accuracy of 93%. In contrast, SVM and KNN achieved accuracies of 90% and 85.5%, respectively. The stability and superior performance of the Neural Network model make it the most effective choice for heart disease prediction in this study. These findings underscore the potential of machine learning in enhancing early heart disease diagnosis and suggest directions for future research, including the integration of additional data types and advanced algorithms to further improve predictive accuracy.

Keywords: Ml, KNN, NN (Neural Networks), SVM

I. INTRODUCTION

Cardiovascular disease (CVD) was responsible for 30% of the 58 million global deaths in 2015. This figure is equivalent to the combined toll of infectious diseases, nutritional deficiencies, and maternal and perinatal conditions. Notably, nearly 46% of these CVD- related deaths occurred in individuals under 70 years old, a demographic in their prime productive years. Additionally, 79% of the overall disease burden from CVD affects this age group. Between 2015 and 2021, non- communicable disease deaths (with cardiovascular disease accounting for half) were projected to rise by 17%, while deaths from infectious diseases, nutritional deficiencies, and maternal and perinatal conditions were expected to decrease by 3%. In low- and middle-income countries, non-communicable diseases already account for almost half of the total disease burden.

Heart disease remains the leading cause of death worldwide, claiming over 1.6 million lives annually. The term "heart disease" encompasses various conditions affecting the heart. The most prevalent is coronary artery disease, which can result in heart attacks. Other forms include issues with the heart valves or heart failure, where the heart is unable to pump blood effectively. Some individuals are born with heart disease, and it can develop in anyone, including children. Heart disease generally occurs when plaque builds up in the arteries. Risk factors include smoking, unhealthy diets, and a sedentary lifestyle, as well as conditions like high cholesterol, high blood pressure, and diabetes.

Preventative measures can be classified into natural and medical methods. Lifestyle changes such as quitting smoking, maintaining a healthy weight, eating a balanced diet, and engaging in regular physical activity arekey natural strategies for reducing heart disease risk. Scientific approaches, including medication and surgery, are also vital. Predicting the likelihood of heart disease before its onset is an important preventative measure, and machine learning algorithms have become widely used in this field.

This study focuses on the application of data science to detect heart disease by analyzing specific attributes such as age, gender, cholesterol levels, and blood pressure. The data used to train machine learning algorithms was collected from hospitals in Algeria, consisting of both healthy and affected individuals. The study begins with a review of recent research in this domain, followed by data preprocessing to select the most relevant attributes using correlation matrices. Finally, machine learning algorithms are applied to datasets of varying sizes (600, 800, 1000, and 1200 records) to develop an effective and stable predictive model for heart disease.

II. LITERATURE STUDY

The application of machine learning (ML) in healthcare, particularly for heart disease prediction, has grown significantly in recent years. Studies have focused on how different ML algorithms can enhance the accuracy and reliability of predicting heart disease. Early efforts in this field relied on traditional statistical methods for analyzing patient data. However, with advancements in computational power and the availability of larger datasets, machine learning emerged as a more efficient alternative. Researchers such as Babu et al. (2017) applied data mining techniques, demonstrating how structured data and effective algorithms can improve prediction outcomes. Comparative studies have evaluated various ML algorithms, such as Decision Trees, Naive Bayes, and Support Vector Machines (SVM), with SVM and Neural Networks often delivering high accuracy. Effective feature selection, as shown by Cai et al. (2018), is crucial for building robust models. Methods like the Pearson correlation matrix help identify key variables related to heart disease.

Neural Networks have emerged as one of the most promising approaches, consistently achieving higher accuracy compared to traditional algorithms. Hybrid models, combining multiple techniqueslike Decision Trees and Random Forests, have also shown improved results. However, challenges remain, includingdata quality, model interpretability, and generalizability across populations. Future research will likely focus on applying deep learning and expanding datasets to enhance accuracy and applicability.

Drawbacks

While machine learning shows promise for heart disease prediction, several challenges persist. High-quality, large datasets are essential for training accurate models, yet data from low- and middle-income regions may be incomplete or inconsistent, affecting prediction accuracy. Overfitting is another concern, where models may excel on training data but struggle with new data. Additionally, complex models like neural networks often lack interpretability, making it difficult to understand decision- making processes. Bias in training data can lead to skewed predictions for certain populations.

Machine learning models require continuous updates to remain relevant, demanding ongoing data collection and maintenance. Integrating these models into clinical workflows and addressing ethical and legal concerns, such as data privacy, are crucial for effective implementation.

III. DEVELOPMENT OF USING MACHINE LEARNING FOR HEART DISEASE PREDICTION

The development of machine learning (ML) models for heart disease prediction involves several critical stages, each aimed at enhancing the accuracy, efficiency, and applicability of predictive algorithms. This section outlines the key steps involved in developing robust ML models for predicting heart disease, from data acquisition to model deployment.

The main contributions in this project are: The development of machine learning techniques for heart disease prediction represents a significant advancement in health care analytics. This project has made several noteworthy contributions to the field, enhancing the accuracy, efficiency, and applicability of predictive models. The main contributions are outlined as follows:

Enhanced Data Collection and Integration: The project utilized a comprehensive data set of 1,200patientrecordsfromtheMohand AmokraneEHS Hospital in Algiers, Algeria. This dataset includes a wide range of attributes relevant to heart disease prediction, such as demographic information, medical history, and lifestyle factors. By integrating diverse data points, the project provides a robust foundation for developing accurate predictive models.

Model Training Module:
Objective: Train machine learning models using pre-processed data to predict the likelihood of heart disease.

Components:
Algorithm Selection: Implementation of various algorithms, such as Neural Networks, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN).
Training Process: Techniques for splitting the dataset into training and validation sets, and for tuning hyperparameters.

Model Evaluation: Metrics for assessing model performance, including accuracy, precision, recall, and F1 score

IV. RESULTANDDISCUSSION

This section outlines the process followed in the study, from data collection to the application of different machine learning techniques. We present the steps involved in achieving the final results, explained in the following phases: Data Collection
Data collection is the critical first step in any research, where information is gathered, measured, and analyzed to evaluate a hypothesis. In our study, we utilized a structured dataset of 1,200entries with 20 attributes, collected from patients at the Mohand Amokrane EHS Hospital in Algiers, Algeria. The dataset includes variables such as age, sex, chest pain type, blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, exercise- induced angina, ST depression, and other factors (e.g., smoking, alcohol use, obesity). The complete data set is presented.

Manual Exploration
Manual data exploration involves analyzing the dataset in an unstructured way to identify preliminary patterns, characteristics, and points of interest. The goal is not to extract every bit of information but to create a broad understanding of key trends. In this study, we added a "Target" column to the dataset, labeling each individual as either sick (1) or not sick (0). This helps in distinguishing between healthy individuals and those with heart conditions, as shown.

V. CONCLUSIONANDFUTUREENHANCEMENT

Heart diseases are becoming increasingly common, including in Algeria, and early prediction can significantly reduce the risk of fatality. This area has garnered considerable research interest, and our study contributes to the ongoing work on heart disease detection and prediction. Specifically, we applied machine learning algorithms to a real dataset of Algerian individuals, focusing on three widely-used algorithms: Neural Networks, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The results of our study were promising, with Neural Networks achieving an accuracy rate of 93%. A key strength of our research was the evaluation of the stability of these algorithms on varying dataset sizes.

REFERENCE

[1] Babu, S., et al.: Heart disease diagnosis using data mining techniques. In: 2017 International Conference on Electronics, Communication, and Aerospace Technology (ICECA), pp. 750-753. IEEE (2017).
[2] Cai, J., et al.: Feature selection in machine learning: A fresh perspective. Neurocomputing, 300, 70-79 (2018).
[3] Dangare, C.S., Apte, S.S.: Enhanced heart disease prediction usingdatamining classification methods. Int. J. of Computer Applications, 47(10), 44-48 (2012).
[4] Fang, X., et al.: Modeling wind powerspatial-temporal correlation using sparse correlation matrix in multi-interval power flow. Applied Energy, 230, 531-539 (2018).