IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning Techniques

[1]Yash Urkude, [2]Abhay Thakare, [3]Pratiksha Nimbokar, [4]Smita Rathod, [5]Gaurav Sawale,

[1,2,3,4] Student, [5]Assistant Professor
Dept of Computer Science and Engineering
Prof. Ram Meghe Institute of Technology & Research, Badnera-Amravati,India

Abstract : In today's fast-paced technological era, health is often neglected, leading to a rise in lifestyle and hereditary diseases, particularly heart disease. According to the WHO, over 11 million deaths occur annually due to heart-related complications, highlighting the urgent need for early diagnosis and preventive measures. However, the healthcare system faces scalability challenges due to limited workforce availability. Our project, "Heart Health Classification and Early Diagnosis of Heart Disease," leverages machine learning to predict heart disease risk based on historical medical data. The system processes input data using five machine learning models to provide accurate risk assessments. Designed as a user-friendly, remote-access platform, it facilitates early screening, reducing the burden on healthcare infrastructure while promoting timely medical intervention.

Keywords **- Decision Tree, KNN, SVM, Logistic Regression, Random Forest, Heart Disease Prediction**

INTRODUCTION

Heart disease is a one of the major public health concern worldwide, causing for millions of deaths each year. This paper discusses different data mining methods used in heart disease prediction. The human heart is responsible for controlling the flow of blood all over the body, and any impairment in its functioning may cause serious health issues. With the fast life people are living nowadays, heart disease has emerged as a major cause of death due to unhealthy living habits, smoking, consumption of alcohol, and high-fat diets leading to hypertension. More than 10 million people die every year from heart conditions, as cited by the World Health Organization. Early diagnosis and a healthy diet are important factors in avoiding diseases.

In healthcare, the major challenge involves ensuring efficient and accurate diagnosis. Although heart diseases are a prominent cause of deaths, they can be treated very well if identified early. The focus of this research is on early heart disease detection through machine learning (ML) methods, preventing serious outcomes. Medical databases are filled with many patient records, which, through data mining algorithms, can yield important information. Decision-making in the case of discrete medical data is usually challenging. ML, being a subset of data mining, processes massive datasets efficiently in order to help in disease diagnosis, detection, and prediction. The aim of this study is to create a tool that helps physicians identify heart disease at an early stage, enabling timely treatment and minimizing complications. ML methods are important in revealing underlying patterns in medical data, enhancing prediction accuracy. This paper compares the performance of different ML algorithms, such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Random Forest, in predicting heart disease at an early stage.

I.LITERATURE SURVEY

Machine learning has significantly contributed to the early detection of heart disease by analyzing medical data and improving diagnostic accuracy. Tithi et al. (2019) explored ECG data analysis using machine learning algorithms, demonstrating its effectiveness in detecting heart disease at an early stage. Their study emphasized the importance of ML in medical diagnostics, particularly for analyzing ECG signals to predict potential cardiac issues **[1]**. Similarly, Jin et al. (2018) proposed a predictive model using Electronic Health Record (EHR) sequential data, highlighting how deep learning models can enhance risk assessment for heart failure based on temporal health patterns **[2]**.

Several researchers have worked on optimizing machine learning models for better classification accuracy. Javeed et al. (2017) introduced an optimized Random Forest model integrated with a Random Search algorithm to enhance heart disease prediction. Their study showed that feature selection and hyperparameter tuning play a crucial role in improving ML performance for medical applications **[3]**. Muhammad et al. (2020) further contributed by developing an intelligent computational model that integrates multiple ML techniques, proving effective in early and accurate heart disease detection **[4]**.

These studies highlight the potential of machine learning in transforming heart disease diagnosis by leveraging patient data for early risk assessment. While ML-based models have shown promising results, challenges such as data quality, feature selection, and real-world implementation remain key areas for further research and improvement.

II.PROPOSED MODEL

The study under consideration is based on the prediction of heart disease using five classification algorithms and comparing their performance. The major task is to correctly specify whether a patient is prone to heart disease or not. In this process, medical practitioners provide patient health parameters to the system, which analyses the data through machine learning models. The model predicts the input and gives an output score, which is the probability of having heart disease. Figure 1 shows the overall process of the prediction.

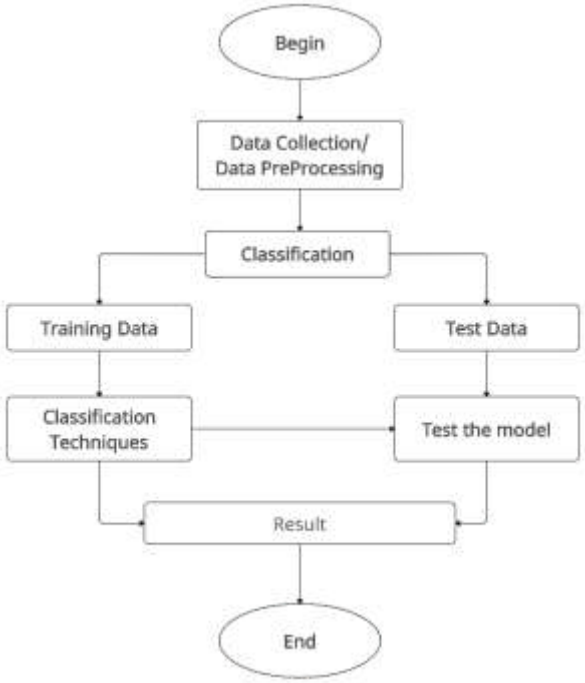

Fig. 1: Framework for Heart Disease Prediction Using Machine Learning

2.1 Data Collection and Preprocessing

The data is present in the Kaggle, our analysis website. There are 10 independent variables. As the dataset contains 1025 instances, observations were carried out for data preparation. This gives us lesser number of the observations giving useless training to our variables which makes logistic regression suitable for classification model. Therefore, we moved ahead with data imputation with the mean value of the observations and scaling them using SimpleImputer and StandardScaler modules of Sklearn.

TABLE I. FEATURES SELECTED FROM DATASET

Sr. No.	Attribute Description	Distinct Values of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0Female, 1-Male)	0,1
3.	CP- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	Resting BP-It represents the patient's BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2

8.	MaxHRt- shows the max heart beat of patient	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	OldPeak- describes patient's depression level.	Multiple values between 0 to 6.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3.

2.2 Classification

The features detailed in Table 1 serve as an input to some machine learning algorithms for classification, namely Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The data are split into 70% training data and 30% test data to effectively evaluate the model. The training data are employed in creating and fine-tuning the model, and the test data evaluate its predictive accuracy. The performance of each algorithm is evaluated based on important evaluation measures like accuracy, precision, recall, and F-measure, which are explained in detail in the subsequent sections. The classification methods investigated in this research are as follows.

i. Random Forest

Random Forest is an ensemble method algorithm employed in both classification and regression. It works by creating numerous decision trees while training and aggregating their predictions to improve accuracy and prevent overfitting. The approach performs well with handling large data and is stable even if some data values are missing. The algorithm entails constructing several decision trees on random subsets of data and combining their predictions to come up with the final result. Random Forest finds extensive application in medical data analysis because it can effectively deal with complex, nonlinear relationships between features.

ii. Decision Tree

Decision Tree is another popular classification algorithm that depicts decisions in a flowchart-like format. Every internal node is a decision on a feature, every branch is an outcome, and every leaf node is a classification result. Decision Trees are favored due to their simplicity, interpretability, and efficiency. The classification process begins from the root node, where there is a decision based on the value of a feature, and then moves along with that corresponding branch until it hits a leaf node. This process is efficient for predicting heart disease since it offers a direct and organized method for reading medical information so that it can be of assistance to healthcare professionals.

iii. Logistic Regression

Logistic Regression is a classification model mostly applied for binary classification problems, e.g., predicting the occurrence or absence of heart disease. It does not fit a straight line but fits a logistic (sigmoid) function, which maps values between 0 and 1. The algorithm returns the probability that the patient has heart disease based on input features, and a threshold value is applied to decide the classification.

iv. Support Vector Machine (SVM)

Support Vector Machine (SVM) is another supervised learning algorithm that determines the best decision boundary, or hyperplane, to distinguish between various classes. SVM is especially helpful in high-dimensional space and uses kernel tricks to deal with non-linearly separable data. It maximizes the margin between the two classes, which results in improved generalization when making predictions on new data.

v. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric classifier that classifies new instances based on how close they are to already labeled data. KNN believes that similar instances lie close to each other and uses the majority vote of the nearest instances to determine the classification. Although simple and efficient, KNN's accuracy can be affected by the number of instances and the distance measure used. This approach is intuitive and suits the case when decision boundaries are not overly complex.

III.RESULT AND ANALYSIS

The outcomes achieved by implementing Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression are compared by the use of important performance measures including Accuracy, Precision, Recall, and F-measure. These measures assist in assessing how well each algorithm makes heart disease predictions. Precision, as equation (1) displays, indicates the percentage of accurately classified positive instances out of all the predicted positive instances. Recall, as defined in equation (2), measures the fraction of true positive cases that have been correctly identified. The F-measure, as outlined in equation (3), weighs Precision and Recall equally, giving an overall measure of accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad \ldots.(1)$$

$$\text{Recall} = (TP) / (TP + FN) \quad \ldots.(2)$$

$$\text{F– Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \ldots.(3)$$

• TP True positive: the patient has the disease and the test is positive.
• FP False positive: the patient does not have the disease but the test is positive.
• TN True negative: the patient does not have the disease and the test is negative.
• FN False negative: the patient has the disease but the test is negative.

The pre-processed data is utilized to perform the experiments in the experiment and the following mentioned algorithms are investigated and employed. The aforementioned performance measures are derived using confusion matrix. Confusion Matrix refers to the model's performance. The confusion matrix derived by the proposed model based on various algorithms is presented here in Table 2. Accuracy score derived on Random Forest, Decision Tree, Logistic Regression, SVM and KNN classification methodologies is presented below in Table

TABLE II. VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	125	40	24	119
SVM	120	48	43	97
Random Forest	152	0	9	9
Decision Tree	168	0	0	140
KNN	123	37	40	108

TABLE III. ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	Precision	Recall	F1	Accuracy
Decision Tree	0.975	0.943	0.976	97.88%
Logistic Regression	0.793	0.778	0.795	79.35%
Random Forest	0.987	0.952	0.989	98.76%
KNN	0.727	0.691	0.707	70.77%
SVM	0.708	0.688	0.704	70.45%

CONCLUSION

Heart Disease Prediction Using Machine learning demonstrates the power of machine learning in addressing critical healthcare challenges. By leveraging a range of machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest, can develop a robust system for predicting heart disease risk. the approach utilizes a dataset with attributes, including age, sex, resting blood pressure, and serum cholesterol, among others, to provide accurate predictions. In this analysis, effectively prepared and evaluated the dataset through comprehensive preprocessing, including feature extraction, outlier treatment, categorical encoding, and feature scaling. In this analysis, the dataset is effectively prepared and evaluated through comprehensive preprocessing steps, including feature extraction, outlier treatment, categorical encoding, and feature scaling. Additionally, model performance is enhanced through techniques like cross-validation ensuring optimal results. By interpreting model outputs with techniques like feature importance analysis, insights can be gained into the most significant risk factors, ultimately aiding healthcare professionals in making informed decisions. Table 1 Table Type Styles

REFERENCES

[1] Tithi, S. R., Aktar, A., Aleem, F., & Chakrabarty, A. (2019). *ECG data analysis and heart disease prediction using machine learning algorithms*. Proceedings of 2019 IEEE Region 10 Symposium.

[2] Jin, B., Che, C., Liu, Z., Zhang, S., Yin, X., & Wei, X. (2018). Predicting the Risk of Heart Failure with EHR Sequential Data Modeling. IEEE Access, Volume 6.

[3] Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., Nour, R., Wali, S., & Basit, A. (2017). *An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection*. IEEE Access, Volume 4

[4] Muhammad, Y., Tahir, M., Hayat, M., et al. (2020). *Early and accurate detection and diagnosis of heart disease using intelligent computational model*. Sci Rep 10, 19747

[5] Sushmita Roy Tithi ,AfifaAktar , Fahimul Aleem , Amitabha Chakrabarty ‘’ECG data analysis and heart disease prediction using machine learning algorithms’’. Proceedings of 2019 IEEE Region 10 Symposium

[6] Bo Jin ,Chao Che, Zhen Liu, Shulong Zhang, XiaomengYin,AndXiaopeng Wei, “Predicting the Risk of Heart Failure WithEHR Sequential Data Modeling” ,IEEE Access Volume 6 2018.

[7] Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim,Adeeb Noor, Redhwan Nour4, Samad Wali And Abdul Basit ,“An Intelligent Learning System based on Random SearchAlgorithm and Optimized Random Forest Model forImproved Heart Disease Detection” , IEEE Access 2017.This work is licensed under a Creative Commons Attribution 4.0 License Volume 4 2016.

[8] Hai Wang et. al.,”Medical Knowledge Acquisition through Data Mining”, Proceedings of 2008 IEEEInternational Symposium on IT in Medicine and Education 978-1-4244- 2511-2/08©2008 Crown.

[9] LathaParthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3,Page No. 3, 2008.

[10] Chaitrali S. Dangare, Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 888)Volume 47No.10, June 2012.

[11] S. Vijiyarani et. al., “An Efficient Classification Tree Technique for Heart Disease Prediction”,International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications (IJCA) (0975 – 8887), 2013 (pp 6-9).

[12] Harsh Vazirani et. al.," Use of Modular Neural Network for Heart Disease", Special Issue of IJCCT Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010], 3-5 August 2010 (pp 88-93)