

Heart Disease Prediction using Machine Learning Algorithms

Vinay Kumar Sriperambuduri¹, Bingi Srikanth² and Kakumanu Balasubbarao³

¹⁻³Vasavi College of Engineering, Hyderabad, India

Email: s.vinaykumar@staff.vce.ac.in, srikanthbingi028@gmail.com, rkakumanu09@gmail.com

Abstract— Heart disease is one of the main causes of death globally, and successful treatment depends on precise diagnosis and prognosis. However, predicting heart disease is challenging and requires expertise due to the potential consequences of errors. Large datasets from data repositories are available for real-world applications. The proposed system utilizes a dataset of patient medical records, which includes several features such as age, sex, blood pressure, cholesterol levels, and smoking status, among others. These characteristics are employed in the training and testing of various machine learning algorithms, such as support vector machines, random forests, and decision trees, to forecast the probability that a patient would have cardiac disease. We have used K-Nearest Neighbour, Random Forest, Logistic Regression, Support Vector Machine (SVM), and XG-boost. Our goal is to find hidden patterns and choose an appropriate machine learning method for heart disease prediction that is accurate and computationally efficient.

Index Terms—heart disease, prediction, machine learning, medical records, healthcare.

I. INTRODUCTION

The World Health Organization estimates that heart disease is responsible for the deaths of 12 million people worldwide each year. Heart disease is a major cause of illness and mortality in humans. Prediction of cardiovascular disease has become an important part using data analysis, as the occurrence of this disease has been increasing rapidly across the globe in recent years. Extensive research has been conducted by authors to identify the major risk factors that contribute to the heart disease, and measure the overall risk accurately [1]. Heart disease referred to as a "silent assassin," can lead to sudden fatal outcomes without showing any clear warning signs. Detecting it at early stages is very crucial, especially for high-risk individuals. This allows the individuals to make better lifestyle choices that can reduce the severity of the condition.

Machine learning has emerged as a valuable tool in this context, helping analyze the vast amount of healthcare data to assist in predicting heart disease and improving decision-making for prevention and treatment. This study focuses on analyzing patient data to predict the future likelihood of cardiac disease, providing timely insights for intervention [2-5].

II. RELATED WORK

Heart disease is one of the major killers around the world and has been killing millions of people annually. Early detection is critically essential because the disease may grow until the treatment options run out.

Machine learning algorithms are currently increasingly becoming important in the development of predictive models for earlier detection of heart disease in healthcare facilities. This literature review discusses the last critical reviews of the relevant works in the application of ML algorithms for predicting heart disease, offering insight into the effectiveness of various models [5,6].

Many researches have attempted to develop systems which predict the prevalence of heart disease based on ML algorithms. For example, Siddiquei et al. (2020) propose a cardiac disease prediction system using logistic regression and ANN.

The system was trained on a data set containing 303 patient records with an accuracy of 85.15%, which speaks to its usage as a useful prediction tool for heart disease [9,10]. Another prediction system was designed by Kaur et al. using the random forest (RF) algorithm that also analyzed 303 patient records. The accuracy of this above RF-based system is found to be 93.70%, hence the authors have concluded that RF will be useful in the early detection of heart disease [11, 12].

These discoveries hold the promise of utilizing machine learning in early diagnosis and further critical insight into the timely intervention and management system by healthcare providers for better patient outcomes. The increasing trend of the application of ML in heart disease prediction is expected to further provide more personalized and proactive care.

III. PROPOSED SYSTEM

The proposed system encompasses Machine Learning Algorithms such as K-Nearest Neighbour, Regression using Logistic Regression, Random Forest Classification, XG-Boost (Extreme Gradient Boosting), Support Vector Machine as depicted in Figure 1. The proposed system runs every algorithm one by one to find the model that produces the best prediction score and accuracy.

Benefits of the Proposed System:

A. Increased Algorithm Diversity

More than the established algorithms, such as KNN, and SVM, a set larger than the afore-mentioned models will be used within the proposed system, including XGBoost, logistic regression, and random forest classification. The above set of larger differences of different model types allows for a more expansive review of several types of models and may help in improving performance as well as predictability. The system is designed to improve predictions through various approaches, considering the fact that cardiac disease is dynamic, which makes the results more reliable towards early diagnosis and decision-making. This flexibility is key to optimizing the prediction of heart disease since algorithms can capture other different patterns within data and, therefore lead to more accurate results.

B. Advanced Techniques

The techniques of advanced machine learning to be used will include, for example, gradient boosting (XGBoost), among others. The ensemble methods like random forest classification are part of the system and help improve the performance of the model by bringing together the strengths of multiple models which effectively integrate their powers of predictions. This proves useful especially in complex datasets that contain nonlinear relations.

This would aggregate results from different models, thus reducing the chances of overfitting and generalization to unseen data in cases of ensemble methods, such as random forests. XGBoost refines the learning process through iterative correction of errors emanating from previous models, making it highly nuanced in prediction. All these place the system well-suited for the complexity offered in medical data, and all these pieces culminate towards better early detection and risk assessment of heart disease.

C. Model Selection for Best Performance

The proposed system tests several algorithms and chooses the best performing one according to the prediction score and accuracy of quality responses. Thus, the final model is optimized for the data present, enhancing the predictive result. The system will compare the output from several algorithms and be able to draw upon the model that best captures the complexity of its dataset, whether it is more accurate at disease diagnosis for heart diseases or precise regarding risk assessment.

This adaptive nature is critical for improving the reliability of prediction since varying algorithms may respond differently based on conditions. The system continuously evaluates and selects the best performing model to ensure outcomes that are consistent and trustworthy in delivering appropriate medical diagnoses and proper treatment planning.

D. Adaptability for Future Improvements

The algorithmic system designed here is intended to be extremely flexible and adaptive; thus, it integrates a wide range of algorithms and advanced methodologies. Consequently, it would be extremely easy to modify the model whenever new algorithms or techniques appeared that could easily be integrated into it.

Algorithm: Random Forest Algorithm

RandomForest(D, T, F):

Inputs:

D = Training dataset

T = Number of trees to build

F = Number of features to randomly select at each split

Initialize:

Forest = [] # To store the decision trees

For i = 1 to T:

- Create a bootstrap sample D_i by randomly selecting data points from D (with replacement).
- Build a decision tree T_i using D_i :
 - For each node in the tree:
 - Randomly select F features from the full feature set.
 - Find the best feature and split point among the F features using a criterion like Gini index (for classification) or mean squared error (for regression).
 - Split the node based on this feature.
 - Repeat until the stopping condition (like max depth or minimum samples per leaf) is met.
- Add tree T_i to the Forest.

Return Forest

Predict(X, Forest):

Inputs:

X = Test input data

Forest = Trained random forest model

Initialize:

Predictions = [] # To store predictions from each tree

For each tree T_i in Forest:

- Use T_i to predict the output for X.
- Store the prediction in Predictions.

For classification:

Return the majority vote (most common class) from Predictions.

For regression:

Return the average (mean) of all predictions in Predictions.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Several machine learning techniques were tested and trained, and after performing so, results indicated the XGBoost algorithm produces better accuracy than other algorithms. The performance was measured by using evaluation metrics- confusion matrices and correlation matrices, which gave it an accurate view of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It helped in determining the strength of the whole algorithm. Using the accuracy formula, it was also demonstrated that two of the classifiers, namely Random Forest Classifiers and XGBoost method, achieved an 86% accuracy with both algorithms, which simply means that one performed slightly better than the other in terms of correctness due to the ability to take care of complex data patterns. This thus determines that it is a highly effective choice when predicting the heart disease.

IV. CONCLUSION AND FUTURE WORK

Heart-related conditions rank among the world's leading causes of death, including in India. The potential social benefit of applying machine learning techniques for early cardiac disease prediction is substantial. Timely detection of heart disease enables individuals at high risk to make necessary lifestyle changes, leading to prevention of complications and improved medical outcomes. In this study, a dataset was utilized to evaluate the performance of five distinct machine learning algorithms, namely SVM, XG-Boost, Random Forest, KNN,

Logistic Regression. The system was assessed based on a selection of 14 essential features, with the potential for further improvement by incorporating additional features. Feature selection was carried out to enhance efficiency, eliminating highly correlated attributes. The accuracy levels of all seven machine learning techniques were compared, and the top-performing algorithm was utilized to construct a prediction model. Subsequent studies may explore the utilization of more advanced algorithms, larger datasets, and the integration of additional features or data sources to enhance prediction accuracy. Furthermore, validating the developed model in real- world clinical settings can provide insights into its applicability and generalizability.



Figure 1. System Design

The comparison of accuracy scores across different algorithms is presented in Table I and the results in Figure 2. The correlation matrix is represented in Figure 3.

TABLE I. ACCURACY AND SCORE MODELS

Model Name	Accuracy Score
Random Forest Tuned	86.89
Extreme Gradient Boost tuned	85.25
K-Nearest Neighbour Tuned	83.61
Logistic Regression Tuned	81.97
Logistic Regression	81.97
Support Vecor Machine Tuned	78.69
K-Nearest Neighbour	78.69
Support Vecor Machine	78.69
Random Forest	77.05
Extreme Gradient Boost	77.05

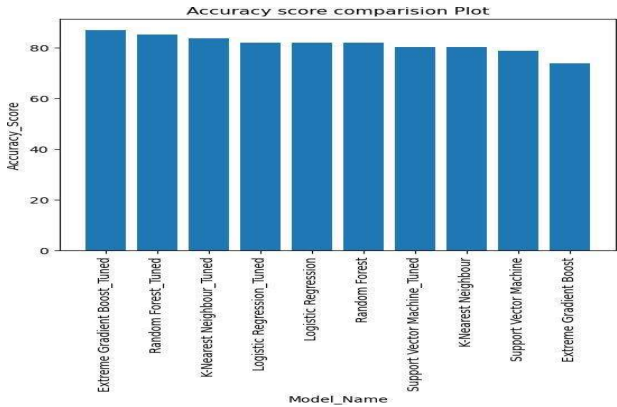


Figure 2. Accuracy Score Comparison

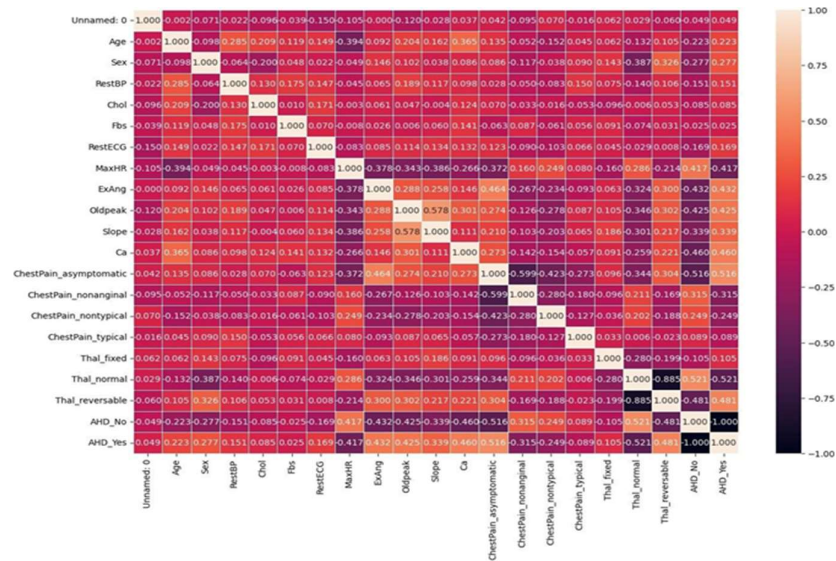


Figure 3. Correlation Matrix

ACKNOWLEDGMENT

We thank Vasavi College of Engineering (Autonomous), Hyderabad for the support extended towards this work.

REFERENCES

- [1] Soni J, Ansari U, Sharma D & Soni S, "Predictive data mining for medical diagnosis: an overview of heart disease prediction", International Journal of Computer Applications, 2020, 17(8), 43-8.
- [2] Narendra Mohan, Vinod Jain & Gauranshi Agrawal, "Heart disease prediction using supervised machine learning algorithms", 5th International Conference on Information Systems and Computer Networks (ISCON), 2021.
- [3] Ordonez C, "Association rule discovery with the train and test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, 2006, 10(2), 334-43.
- [4] Shinde R, Arjun S, Patil P & Waghmare J, "An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm", International Journal of Computer Science and Information Technologies, 2015, 6(1), 637-9.
- [5] Bashir S, Qamar U & Javed M Y, "An ensemble-based decision support framework for intelligent heart disease diagnosis", International Conference on Information Society (i-Society 2014), IEEE, 2014, pp. 259-64. ICCRDA
- [6] Jee S H, Jang Y, Oh D J, Lee S H, Park S W & Yun Y D, "A coronary heart disease prediction model: the Korean heart study", BMJ open, 2014, 4(5), e005025.
- [7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E, "Multilocus genetic risk scores for coronary heart disease prediction", Arteriosclerosis, thrombosis, and vascular biology, 2013, 33(9), 2267-72.
- [8] Jabbar M A, Deekshatulu B L & Chandra P, "Heart disease prediction using lazy associative classification", International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) 2013, IEEE, pp. 40- 6.
- [9] Brown N, Young T, Gray D, Skene A M & Hampton J R, "Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register", BMJ, 1997, 315(7101), 159-64.
- [10] Folsom A R, Prineas R J, Kaye S A & Soler J T, "Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women", International journal of epidemiology, 1989, 18(2), 361-7.
- [11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J, "HDPS: Heart disease prediction system", Computing in Cardiology, IEEE 2011, pp. 557- 60.
- [12] Parthiban, Latha and R Subramanian, "Intelligent heart disease prediction system using CANFIS and genetic algorithm", International Journal of Biological, Biomedical and Medical Sciences 3.3, 2008.
- [13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H, "Wireless body area network for heart attack detection", IEEE antennas and propagation magazine, 2016, 58(5), 84-92.
- [14] Patel S & Chauhan Y, "Heart attack detection and medical attention using motion sensing device-kinect", International Journal of Scientific and Research Publications, 2014, 4(1), 1-4.
- [15] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J, "Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone", Current controlled trials in cardiovascular medicine, 2002.