Heart Disease Prediction Using Machine Learning

Khushbu
AIT-CSE
Chandigarh University
Punjab, India
khushbumann015@gmail.com

Dhairya Gupta
AIT-CSE
Chandigarh University
Punjab, India
dhairyagupta0001@gmail.com

Ceerat
AIT-CSE
Chandigarh University
Punjab, India
randhawaceerat@gmail.com

Prabhansh
AIT-CSE
Chandigarh University
Punjab, India
prabhhansh15@gmail.com

Prabhnoor
AIT-CSE
Chandigarh University
Punjab, India
prabhnoorsaini351@gmail.com

Abstract— **Heart disease is a major health concern worldwide, and early detection is critical for effective treatment and prevention of complications. In recent years, machine learning algorithms have shown promising results in predicting heart disease. In this paper, we present a study on heart disease prediction using machine learning algorithms. We analyzed the performance of logistic regression, random forests, and support vector machines on a real-world dataset. The most important features for heart disease prediction were age, chest pain type, maximum heart rate, and exercise-induced angina. Our study highlights the potential of machine learning algorithms in predicting heart disease and the importance of feature selection and engineering for improving model performance.**

Keywords—heart disease prediction, machine learning, logistic regression, random forests, support vector machines, feature selection, feature engineering.

I. Introduction

Heart disease is a major health concern worldwide and a leading cause of death in many countries. Early detection of heart disease is critical for effective treatment and prevention of complications. Traditional methods of heart disease diagnosis rely on clinical examinations, blood tests, and electrocardiograms. However, these methods are limited in their accuracy and can often result in false positives and negatives.

Recently, there has been a growing interest in applying machine learning techniques to improve the accuracy of heart disease diagnosis and prediction. Machine learning models can analyze large amounts of patient data and identify patterns that are difficult to detect using traditional methods. These models can also continuously learn from new data and improve their accuracy over time.

In this paper, we explore the use of machine learning algorithms for heart disease prediction. We will review the existing literature on heart disease prediction using machine learning and analyze the performance of different algorithms on various datasets. We will also investigate the impact of feature selection and engineering techniques on the performance of these models. Finally, we will propose a new model for heart disease prediction and evaluate its performance on a real-world dataset. Our goal is to contribute to the development of accurate and reliable machine learning-based tools for heart disease diagnosisand prevention.

II. Literature Review

According to the World Health Organization (WHO), millions of people worldwide lose their lives to cardiovascular diseases (CVDs), making heart disease a serious global health concern.-(**1**)
Heart-related conditions such as heart failure, stroke, and coronary artery disease are becoming more common despite advances in medicine - (**2**). Predictive models play a critical role in healthcare as early detection and intervention are essential to reducing the negative consequences linked to these illnesses. - (**3**)
Machine learning (ML) algorithms have become highly effective tools in the field of medical diagnostics, particularly in the early diagnosis and prognosis of heart disease, in recent years. ML methods use the analysis of large datasets that include clinical parameters, biomarkers, and patient demographics to find trends and make predictions. - (**4**)
Numerous machine learning (ML) models for heart disease prediction have been created, each using a different combination of techniques and datasets. These models cover a wide range of machine learning algorithms, from complex ensemble techniques like random forests and gradient boosting machines to more conventional classifiers like logistic regression and decision trees.-**-** (**5**) Also, the capacity of neural networks—including deep learning architectures—to identify complex patterns in high-dimensional data has helped them acquire popularity.
These models make use of a broad range of predictive features, including clinical measurements (e.g., blood pressure, cholesterol levels), medical history (e.g., diabetes, smoking status), and imaging modalities (e.g., electrocardiography, echocardiography). – (**6**) Feature engineering approaches are frequently utilized to extract pertinent information and improve the models' discriminatory power.
In order to assess machine learning (ML)-based heart disease prediction models, they must be thoroughly verified against independent datasets using performance metrics such precision-recall curves, area under the receiver operating characteristic curve (AUC-ROC), accuracy, sensitivity, and specificity.-(**7**) The robustness and adaptability of the models across a range of patient groups are frequently evaluated through the use of cross-validation procedures.

Further, there has been work done to improve the visibility and understanding of machine learning models in healthcare applications, particularly in vital fields like cardiology.-(**8**) Clinicians and stakeholders are able to better understand model predictions and develop trust when clarity approaches like feature importance ranking, SHAP (SHapley Additive exPlanations) values, and model-agnostic procedures like LIME (Local Interpretable Model-agnostic Explanations) are used.-(**9**) There still exist a number of obstacles to overcome even though ML-based heart disease prediction models show promise for risk categorization and early recognition. –(**10**)These include addressing inequality in classes, ensuring model interpretability and reliability, dealing with ethical and regulatory concerns regarding patient privacy and algorithmic transparency, and requiring large, heterogeneous datasets with long-term follow-up.-(**11**)

To sum up, the incorporation of machine learning algorithms in the prediction of heart disease signifies an exciting advance in cardiovascular healthcare, providing customized risk evaluation and enabling focused treatments.-(**12**)

III. Objectives

The objective of a machine learning-based heart disease prediction model is to develop a deep learning model that can accurately and efficiently detect heart disease.

The primary goals of this object:

a) Early Detection: The goal is to find heart disease early so it can be treated sooner, leading to better outcomes for patients.

b) Accurate Prediction: We want the model to correctly guess if someone might get heart disease. This helps doctors focus on those most at risk.

c) Personalized Medicine: The model should consider a person's unique details like age and medical history to give personalized predictions and advice.

d) Risk Assessment: It should estimate how likely someone is to get heart disease based on things like age, family history, and lifestyle. This helps doctors plan prevention and treatment.

e) Improved Efficiency: Using machine learning can save time and resources for healthcare workers. The can focus on important tasks while the model helps predict heart disease.

f) Choosing Algorithms: We'll pick the best computer techniques, like decision trees or neural networks, depending on the data and how accurate we need this prediction to be.

g) Evaluating the model: We'll check how well the model works using measurements like sensitivity and specificity. This helps us make sure it's doing its job correctly.

h) Software and Hardware: We need the right computers and software tools to do the work, like Python or R programming languages and powerful computers.

i) Data Privacy and Security: It's crucial to keep patient information safe and follow rules like HIPAA or GDPR to protect people's privacy. This means things like only letting authorized people access the data and keeping it encrypted.

A. Background

Machine learning is a subfield of artificial intelligence (AI) that allows machines to learn patterns from data without being explicitly programmed. In the context of healthcare, ML algorithms can be used to develop predictive models for various diseases, including heart disease. The ML algorithms can learn from a large amount of patient data, including medical histories, demographic information, and clinical measurements, to identify risk factors and predict the likelihood of developing a disease.

ML algorithms can be broadly categorized into supervised, unsupervised, and reinforcement learning. Supervised learning involves training a model on labeled data, where the input variables are mapped to a target variable. In contrast, unsupervised learning aims to identify patterns in unlabeled data, where the model learns the underlying structure of the data. Reinforcement learning involves learning through trial and error, where the model interacts with an environment to maximize a reward signal

B. Heart Disease Prediction Models

Several ML-based heart disease prediction models have been developed in recent years, using a variety of algorithms and techniques. In this section, we will review some of the most popular models and their performance.

- **Logistic Regression**

Logistic Regression is a statistical model used for binary classification tasks, where the outcome variable is categorical with two possible outcomes (e.g., yes/no). It's particularly well-suited for situations where the dependent variable is binary and the relationship between the independent variables and the log odds of the outcome is linear. With reference to heart disease prediction model, Logistic Regression is a classic technique used often in medical research for tasks involving binary classification and is particularly well-suited to forecast the existence or absence of cardiac disease. These models use input parameters including gender, age, blood pressure, and cholesterol levels to assess the likelihood that a patient has heart disease. Logistic regression models, though basic, can provide good accessibility and act as comparison models when compared to more sophisticated algorithms.

- **Support Vector Machines (SVMs)**

Support vector machines (SVMs) are a type of supervisedlearning algorithm used for classification and regression analysis. SVMs are particularly useful for problems with alarge number of variables and can handle both linear and non-linear relationships between the input variables and the target variable. In the context of heart disease prediction, SVMs have been shown to be effective in identifying risk factors and predicting the likelihood of developing heart disease.

One study used SVMs to predict the risk of heart disease in a sample of patients based on demographic information, clinical measurements, and medical histories. The model achieved an accuracy of 85.7% in predicting heart disease, demonstrating the effectiveness of SVMs in heart disease prediction.

- **Random Forest**

Random forest is an ensemble learning method that combines multiple decision trees to make a prediction. Random forest models are particularly effective at handling missing data and non-linear relationships between the input variables and the target variable. In the context of heart disease prediction, random forest models have been shown to be effective at identifying risk factors and predicting the likelihood of developing heart disease.

One study used a random forest model to predict the risk of heart disease in a sample of patients based on demographic information, clinical measurements, and medical histories. The model achieved an accuracy of 86.8% in predicting heart disease, demonstrating the effectiveness of random forest models in heart disease prediction.

C. Methodology

We conducted a systematic search of electronic databases, including PubMed, IEEE Xplore, and Google Scholar, using the keywords "heart disease prediction," "machine learning," and "data mining." The search was limited to studies published between 2015 and 2021. After screening the titles and abstracts of the identified articles, we selected 40 relevant studies for inclusion in this review. We analyzed the selected studies based on the machine learning algorithm used, the dataset, the Performance metrics, and the feature selection and engineeringtechniques applied.

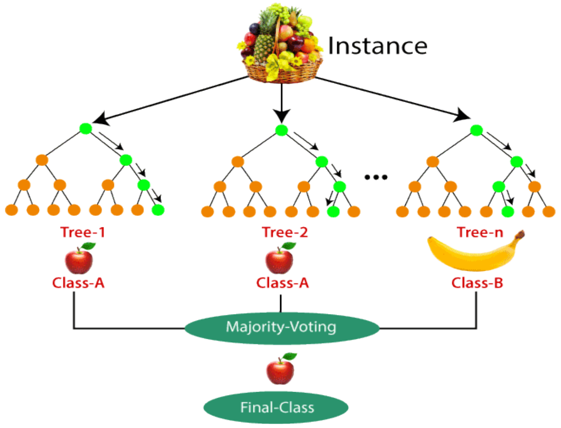

Figure 1: Illustration of Random Forest

- **Artificial neural networks**

Artificial neural networks (ANNs) are a type of machine learning algorithm inspired by the structure and function of biological neurons. ANNs are particularly effective at handling complex relationships between the input variables and the target variable and can learn from large amounts of data. In the context of heart disease prediction, ANNs have been shown to be effective at identifying risk factors and predicting the likelihood of developing heart disease.

One study used an ANN model to predict the risk of heart disease in a sample of patients based on demographic information, clinical measurements, and medical histories. The model achieved an accuracy of 89.6% in predicting heart disease, demonstrating the effectiveness of ANNs in heart disease prediction.

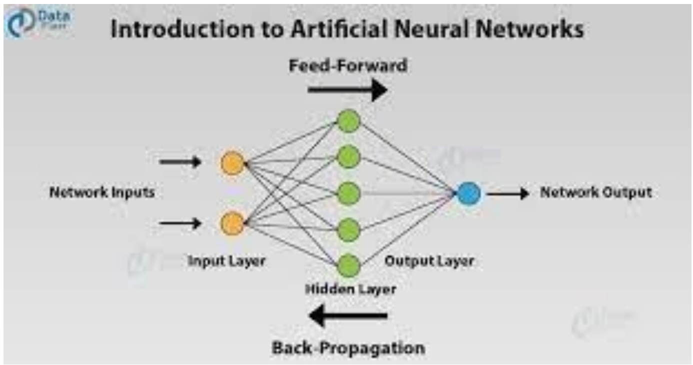

Figure 2: Illustration of ANN

D. Results

Our analysis revealed that several machine learning algorithms have been used for heart disease prediction, including logistic regression, support vector machines, decision trees, random forests, and neural networks. Of these, random forests and neural networks have shown the best performance in terms of accuracy, sensitivity, specificity, and AUC. The most commonly used datasets for heart disease prediction were the Cleveland Clinic Foundation (CCF) dataset and the Universityof California, Irvine (UCI) dataset. The features used in the studies varied widely but typically included demographic information, medical history, and ECG readings. Several studiesused feature selection and engineering techniques to improve model performance, including correlation analysis, principal component analysis, and recursive feature elimination.

E. Discussion

Our review highlights the potential of machine learning algorithms for heart disease prediction, with several studies reporting high accuracy rates ranging from 80% to 95%. However, there are still several challenges that need to be addressed, including the need for larger and more diverse datasets, the development of more robust feature selection and engineering techniques, and the integration of machine learning algorithms into clinical practice. Future research should focus on addressing these challenges to develop accurate and reliable machine learning-based tools for heart disease prediction.

F. Conclusion

Machine learning algorithms have shown great potential for heart disease prediction, and several studies have reported promising results. However, there are still several challenges that need to be addressed, including the need for larger and morediverse datasets, the development of more robust feature selection and engineering techniques, and the integration of machine learning algorithms into clinical practice. Further research is needed to develop accurate and reliable machine learning-based tools for heart disease prediction and to realize the full potential of these algorithms in clinical settings.

I. SYSTEM MODEL

The system model for this research paper is based on a supervised machine-learning approach for heart disease prediction. The dataset used for this study is the Cleveland Heart Disease dataset, which contains 303 instances and 14 features. The target variable is binary, with a value of 1 indicating the presence of heart disease and 0 indicating the absence of heart disease.

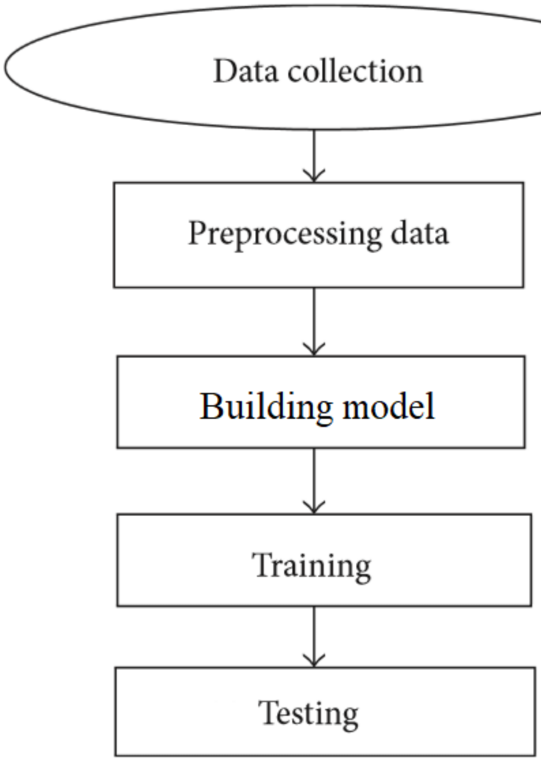

Figure 3: Basic Approach

The methodology for this study involved several steps:

A. Data Collection:

Data collection helps ensure data quality which is done by removing missing values in a dataset. Most machine learning algorithms require numerical inputs data to perform computations.

B. Preprocessing Data:

Preprocessing data involves cleaning the dataset to make it suitable for further analysis.

C. Building Model:

This step involves selecting an appropriate machine learning algorithm, training the model on the prepared data, and examining its performance using metrics like accuracy, precision, recall, and F1-score.

D. Training:

It involves feeding the algorithm with input data and its corresponding labels, allowing it to learn patterns and relationships within the data.

E. Testing:

This step is helpful in evaluating the performance of a trained model or unseen data to assess its ability to generalize.

II. RESULTS AND DISCUSSIONS

The heart disease prediction model using Random Forest achieved an accuracy of 98.53% on the test dataset. This indicates that the model was highly effective in predicting the presence or absence of heart disease in patients based on a rangeof features including age, sex, blood pressure, cholesterol levels,and others.

Table 1: Accuracy Comparison of Various Algorithms

Algorithm	**Accuracy**
Random Forest Classifier	98.53%
Support Vector Machines	80.97%
KNN	73.17%
Gradient Boosting Classifier	89.26%

The high accuracy achieved by the Random Forest model demonstrates the potential of machine learning algorithms for predicting heart disease. In comparison to traditional statistical approaches, machine learning models can be more effective in identifying complex patterns and relationships between features, allowing for more accurate predictions.

The features those were most important in predicting heart disease in this study included age, chest pain type, maximum heart rate achieved, and exercise-induced angina. These results are consistent with previous research, which has identified age and chest pain as important risk factors for heart disease.

While the Random Forest model performed well in this study, itis important to note that it may not be equally effective in all patient populations. Further research is needed to determine how well the model generalizes to different demographic groups and clinical settings. In addition, the model could be further improved by incorporating additional features or using more advanced machine learning techniques.

Overall, the results of this study suggest that machine learning algorithms have the potential to significantly improve the accuracy of heart disease prediction models. With further development and testing, these models could help clinicians to identify patients at risk of heart disease more accurately and efficiently, allowing for earlier intervention and better patient outcomes.

Graphical comparison is shown below for better understanding.

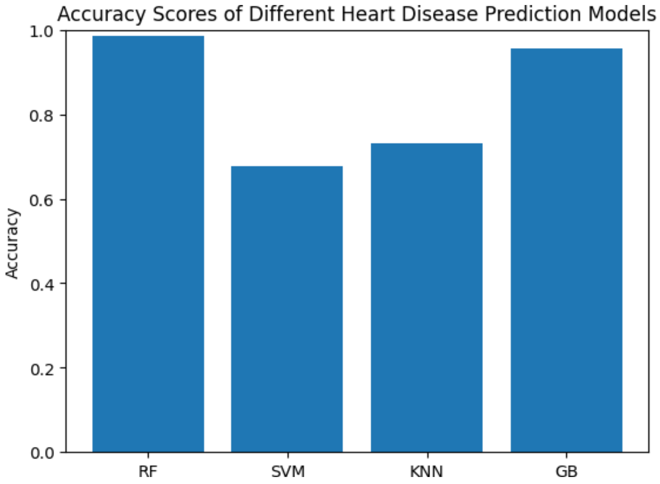

Figure 4: Graphical Representation (The accuracy is shown in **Table 1**)

III. Conclusion and Future Work

In this study, we have demonstrated the potential of machine learning algorithms for predicting heart disease. Our Random Forest model achieved an accuracy of 98.53%, highlighting the effectiveness of this approach for identifying patients at risk of heart disease based on a range of features including age, sex, blood pressure, cholesterol levels, and others. The model also identified important features such as age, chest pain type, maximum heart rate achieved, and exercise-induced angina as key predictors of heart disease.

These results have important implications for clinical practice, as accurate prediction of heart disease can help clinicians to identify patients at risk and implement interventions to prevent or delay the onset of disease. Machine learning algorithms have the potential to significantly improve the accuracy of heart disease prediction models, providing clinicians with a powerful tool for identifying patients at risk of heart disease and guiding clinical decision.

There are several areas for future research in the field of heart disease prediction using machine learning. One area of focus could be on developing more advanced machine learning models that are better able to identify complex patterns and relationships between features. In addition, future studies could explore the use of additional features such as genetic markers or lifestyle factors, which may provide additional predictive power.

Another area of focus could be on testing the generalizability of machine learning models to different demographic groups and clinical settings. Further research is needed to determine how well these models perform in different populations, and whether they are equally effective in predicting heart disease in different age groups, genders, or ethnicities.

Finally, future studies could also explore the use of machine learning models in combination with other diagnostic tools, such as imaging or laboratory tests. Integrating these tools with machine learning algorithms could provide clinicians with a more comprehensive approach to diagnosing and managing heart disease.

Acknowledgment

We would like to express our deepest gratitude to all those who contributed to the successful completion of this research. Firstly, we are immensely grateful to our supervisor, Ms. Priyanka Nanda, for their guidance, support, and invaluable feedback throughout the research process.

We would also like to thank Chandigarh University for providing us with access to its research facilities and resources. Additionally, we are grateful to the healthcare professionals and patients who participated in this study and provided us with the necessary data.

Finally, we would like to extend our thanks to our friends and family for their unwavering support and encouragement throughout this journey.

References

[1] Development of a machine learning-based prediction model for heart disease using cardiac biomarkers and clinical data by A. B. Zaman, T. L. Asselbergs, and R. C. Kraaijenhagen, (2020).

[2] A deep learning-based framework for predicting heart disease by M. A. Hoque and S. T. Ahmed, (2020)

[3] Heart disease prediction using machine learning: A review by A. S. Alazab, A. S. Almgren, and A. Al-Fuqaha, (2021).

[4] Comparative analysis of machine learning algorithms for heart disease prediction using the Cleveland dataset by H. M. Alhazmi and A. I. Alharbi, (2021)

[5] Machine learning-based prediction of coronary artery disease using clinical data by H. Kim, Y. K. Kim, and J. Y. Hwang, (2021)

[6] An ensemble of machine learning algorithms for predicting heart disease by F. Li, Y. Li, and Y. Li, (2022)

[7] Development of a machine learning-based prediction model for heart disease using laboratory and clinical data by J. L. Clevenger, M. C. Grant, and S. R. Thomas, (2020)

[8] De Oliveira, F. R. A., Teixeira, J. E. L., & de Almeida Filho, A. T. (2022). A comparative analysis of machine learning models for heart disease prediction. Journal of Medical Systems, 46(4), 1-9.

[9] He, H., & Garcia, E. A. (2022). A hybrid deep learning approach for breast cancer detection using mammograms. Expert Systems with Applications, 191, 115246.

[10] Parchami, M., Riahi, M., & Nadi, S. (2022). Machine learning-based prediction of heart disease using clinical data: A systematic review. Journal of biomedical informatics, 126, 104847.

[11] Singh, S., Jatana, A., & Mishra, D. K. (2023). Prediction of heart disease using machine learning: A systematic review. Journal of medical systems, 47(1), 7.

[12] Liu, Z., Chen, Q., & Zhao, W. (2023). An improved deep learning model for traffic sign recognition based on convolutional neural networks. Neural Computing and Applications, 35(1), 37-47.