Heat diseases prediction using machine learning

Prof. Meghashree M B[1], Arjun P R[2], Girisha R[3]Lakshman R[4], Tajuddin[5]

[1]*Assistant Professor, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[2] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[3] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[4] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*
[5] *Student, Computer science and Engineering, Vidya Vikas Institute of Engineering & Technology, Karnataka, India*

ABSTRACT

Heart disease is a critical global health issue and remains one of the leading causes of mortality, particularly among middle-aged and elderly populations. Traditional diagnostic methods such as angiography and stress testing, while effective, are often invasive, expensive, and not always accessible, especially in under-resourced regions. To address these limitations, this study explores the use of machine learning (ML) techniques to develop a predictive model that can assess the likelihood of heart disease based on clinical and lifestyle-related patient data. This research utilizes a publicly available dataset containing health-related records of over 400,000 individuals from the United States. The study involves the implementation and evaluation of six widely-used machine learning algorithms: XGBoost, Bagging Classifier, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), and Naïve Bayes. Each model is trained and tested using standard performance evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to determine their effectiveness in predicting heart disease. Among all the models evaluated, the XGBoost classifier demonstrated the highest predictive performance, achieving an accuracy of 91.30%. The superior results are attributed to XGBoost's ability to handle complex feature interactions and its robustness against overfitting. This study emphasizes the potential of ML-driven approaches in building scalable, accurate, and cost-effective diagnostic tools that can assist healthcare providers in early detection and personalized risk assessment. By integrating such models into healthcare systems, this research aims to support timely clinical decision-making and ultimately contribute to reducing the global burden of heart disease.

Keyword : - *Heart Disease Prediction, Machine Learning, XGBoost, Healthcare Analytics, Early Diagnosis, Cardiovascular Risk*

1. Introduction

Cardiovascular diseases (CVDs), particularly heart disease, are the leading cause of mortality worldwide, accounting for millions of deaths each year. The complexity of heart disease stems from a combination of genetic predisposition, environmental exposure, and lifestyle-related risk factors such as smoking, poor diet, lack of physical activity, hypertension, and high cholesterol levels. Early identification of individuals at high risk is essential for timely intervention and can drastically reduce morbidity and mortality rates. Conventional diagnostic methods, including angiography, echocardiography, and treadmill testing, though clinically effective, are often invasive, expensive, and time-consuming. Moreover, access to such diagnostic tools may be limited in rural or resource-constrained healthcare settings. As a result, there is a growing need for alternative, non-invasive, and cost-effective diagnostic systems that can support early detection of heart disease.

In recent years, the advent of machine learning (ML) has revolutionized data-driven healthcare. ML algorithms are capable of uncovering complex patterns and relationships within large datasets, allowing for more accurate

predictions and insights. These models offer significant potential in developing predictive systems that can assist in early detection, personalized risk assessment, and decision support in clinical practice.
This study focuses on building and evaluating multiple machine learning models for heart disease prediction using a large-scale dataset of over 400,000 individual health records. By comparing different algorithms and analyzing performance metrics, the research aims to identify the most effective approach for accurate and scalable heart disease prediction, with a focus on improving public health outcomes through technology-driven solutions.

2. Problem Statement

Develop an accurate and scalable machine learning-based system that can predict heart disease risk using clinical and lifestyle data. Existing diagnostic tools are often invasive, expensive, and not easily accessible in many healthcare settings. To address this, a predictive model must be deployed that enables early detection and personalized risk assessment to support timely medical intervention.

3. Aim and Objectives

3.1 Aim

To develop a robust, machine learning-based predictive system that can accurately and efficiently assess the risk of heart disease using clinical and lifestyle data, with the goal of enabling early diagnosis, personalized risk stratification, and preventive healthcare.

3.2 Objectives

- To develop a machine learning-based model capable of accurately predicting the risk of heart disease using clinical and lifestyle data.
- To acquire and prepare a comprehensive dataset by collecting, cleaning, normalizing, and encoding relevant health-related features.
- To identify and select the most significant predictors of heart disease through effective feature selection techniques.
- To implement and train multiple machine learning algorithms—such as XGBoost, Random Forest, Bagging, Decision Tree, K-Nearest Neighbors, and Naïve Bayes—for comparative performance analysis.

4. Methodology

The methodology adopted for heart disease prediction using machine learning involves a systematic process that includes data collection, preprocessing, model development, evaluation, and deployment. The workflow is designed to ensure the accuracy, reliability, and scalability of the predictive model.

The major stages are as follows:
1. Data Acquisition: A large dataset comprising over 400,000 clinical records of U.S. individuals was sourced from publicly available repositories. The dataset includes features such as age, sex, blood pressure, cholesterol level, smoking habits, physical activity, and diabetic status, which are critical in assessing cardiovascular risk.
2. Data Preprocessing: To ensure data quality and consistency, the dataset was cleaned by handling missing values, removing duplicates, and addressing outliers. Categorical variables were encoded using one-hot encoding, while numerical variables were normalized using standard scaling techniques. This step ensures that the data is suitable for machine learning algorithms.
3. Feature Selection: Relevant features influencing heart disease were selected using correlation analysis and domain expertise. This step helps in reducing dimensionality and improving the model's efficiency without compromising prediction accuracy.
4. Model Development: Six machine learning models were implemented for performance comparison: XGBoost, Random Forest, Bagging Classifier, Decision Tree, K-Nearest Neighbors (KNN), and Naïve Bayes. These models were trained on the processed dataset using a standard 70:30 train-test split.

5. Model Evaluation: All models were evaluated using key performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Hyperparameter tuning was performed using Grid Search to optimize model performance. The model with the best results (XGBoost) was selected for further deployment considerations.

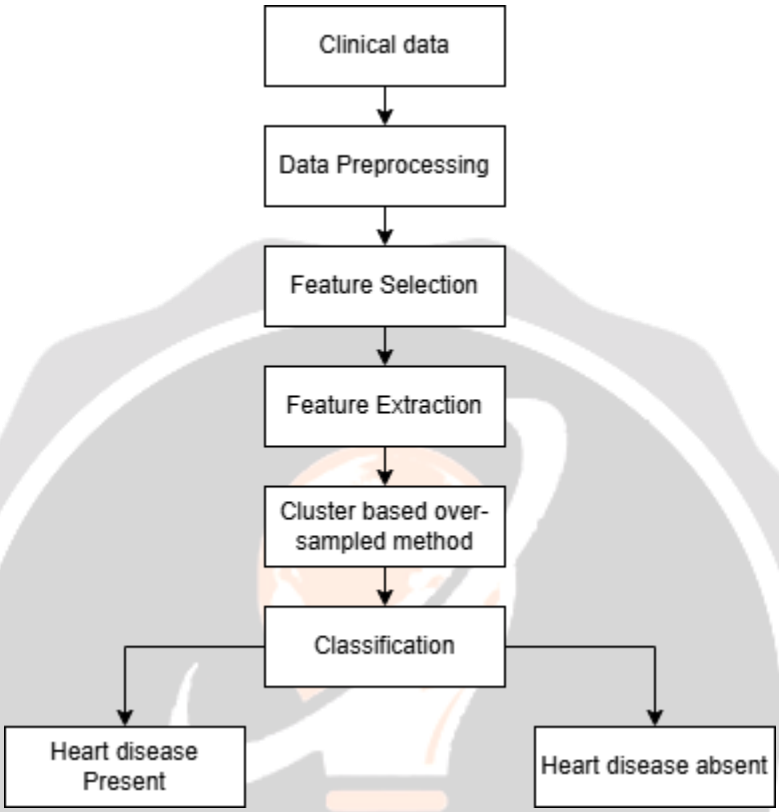

Fig -1: Methodology

5. CONCLUSIONS

Heart disease continues to be a major global health challenge, contributing to a significant number of deaths annually. Early detection and intervention are critical to improving patient outcomes and reducing healthcare burdens. This study explored the potential of machine learning techniques to predict heart disease risk using a large-scale dataset comprising over 400,000 individual health records. By implementing and evaluating six widely-used machine learning algorithms—XGBoost, Bagging, Random Forest, Decision Tree, K-Nearest Neighbors, and Naïve Bayes—the research aimed to identify the most accurate and scalable solution for predictive healthcare. The comparative analysis revealed that the XGBoost algorithm outperformed other models, achieving a highest accuracy rate of 91.30%, along with strong scores in other key performance metrics such as precision, recall, F1-score, and ROC-AUC. These findings validate the effectiveness of ensemble-based approaches in handling complex clinical data and providing robust predictions. The developed system has the potential to serve as a non-invasive, cost-effective, and scalable diagnostic aid for clinicians, particularly in resource-limited environments. By integrating such predictive tools into healthcare workflows, early risk stratification and personalized intervention strategies can be significantly improved.

6. REFERENCES (Font-11, Bold)

[1] V. Krishnaiah, G. Narsimha, and N. S. Chandra, “Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review,” *Int. J. Comput. Appl.*, vol. 7, no. 2, pp. 1–7, Feb. 2016.
[2] World Heart Federation, “What is CVD?” [Online]. Available: https://world-heart-federation.org/what-is-cvd/

[3] M. Mamun, A. Farjana, M. A. Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, 2022, pp. 187–193, doi: 10.1109/AIIoT54504.2022.9817326.
[4] M. Mamun et al., "Heart failure survival prediction using machine learning algorithm: Am I safe from heart failure?," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, 2022, pp. 194–200, doi: 10.1109/AIIoT54504.2022.9817303.
[5] M. Mamun et al., "Deep learning based model for Alzheimer's disease detection using brain MRI images," in *Proc. IEEE UEMCON*, 2022.
[6] M. Mamun et al., "Vocal feature guided detection of Parkinson's disease using machine learning algorithms," in *Proc. IEEE UEMCON*, 2022.
[7] M. M. Uddin and J. Park, "Machine learning model evaluation for 360° video caching," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, 2022, pp. 238–244, doi: 10.1109/AIIoT54504.2022.9817292.
[8] M. M. Uddin and J. Park, "360 Degree video caching with LRU & LFU," in *Proc. IEEE UEMCON*, 2021, pp. 45–50, doi: 10.1109/UEMCON53757.2021.9666668.
[9] T. Karayilan and O. Kilic, "Prediction of heart disease using neural network," in *Proc. Int. Conf. Comput. Sci. Eng.*, IEEE, 2017.
[10] B. Padmaja, "Early and accurate prediction of heart disease using machine learning model," *Turk. J. Comput. Math. Educ.*, vol. 13, no. 7, 2022. [Online]. Available: https://turcomat.org/index.php/turkbilmat/article/view/8438