International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Machine Learning Approaches for Accurate Heart Disease Classification

M. Vilas [1], B. Venkateshwara Rao [2], J. LVSR Rajkumar [3], K. Shiva Kumar [4], Mrs.N.Bhargavi [5]
[1,2,3,4] UG Scholars, [5] Assistant Professor

[1,2,3,4,5] Department of Computer Science and Engineering (Data Science),

[1,2,3,4,5] Siddhartha Institute of Technology and Sciences, Hyderabad, Telangana, India.

Abstract

Heart disease (HD), including heart attacks, is a leading cause of death worldwide, making accurate determination of a patient's risk a significant challenge in medical data analysis. Early detection and continuous monitoring by physicians can significantly reduce mortality rates, but heart disease is not always easily detectable, and physicians cannot monitor patients around the clock. Machine learning (ML) offers a promising solution to enhance diagnostics through more accurate predictions based on data from healthcare sectors globally. This study aims to employ various feature selection methods to develop an effective ML technique for early-stage heart disease prediction. The feature selection process utilized three distinct methods: chi-square, analysis of variance (ANOVA), and mutual information (MI), leading to three selected feature groups designated as SF-1, SF-2, and SF-3. We then evaluated ten different ML classifiers, including Naive Bayes, support vector machine (SVM), voting, XGBoost, AdaBoost, bagging, decision tree (DT), K-nearest neighbor (KNN), random forest (RF), and logistic regression (LR), to identify the best approach and feature subset. The proposed prediction method was validated using a private dataset, a publicly available dataset, and multiple cross-validation techniques.

Keywords: Machine Learning (ML), Heart Disease Classification, Predictive Modeling, Cardiovascular Disease (CVD), Classification Algorithms

1. Introduction

Heart disease (HD) is one of the leading causes of death worldwide, with conditions such as heart attacks accounting for a significant portion of global mortality. Early detection and continuous monitoring are key factors in improving patient outcomes, as they allow for timely intervention and management. However, diagnosing heart disease can be challenging, as symptoms may be subtle or non-specific, and physicians cannot provide constant monitoring for all patients. Traditional diagnostic methods often face limitations in terms of cost, time, and availability of healthcare professionals. To overcome these challenges, machine learning (ML) techniques have emerged as a promising solution for improving the accuracy and efficiency of heart disease prediction. This project aims to develop an effective ML-based prediction system for early-stage heart disease by utilizing various feature selection methods and evaluating a range of machine learning classifiers. By employing techniques such as chi-square, analysis of variance (ANOVA), and mutual information (MI), the project identifies the most relevant features that contribute to accurate heart disease diagnosis. The study compares the performance of several classifiers, including Naive Bayes, support vector machine (SVM), XGBoost, AdaBoost, and Random Forest, to determine the best approach for prediction. The project also addresses the challenge of imbalanced data by applying the Synthetic Minority Oversampling Technique (SMOTE) to enhance model performance. Through extensive evaluation using multiple datasets and cross-validation techniques, the project identifies AdaBoost as the optimal model, achieving high accuracy and sensitivity.

2. Related work

The first groundwork related to “Classification and Prediction of Heart Diseases using Machine Learning Algorithms” and the details are presented as follows. Heart disease is a serious worldwide health issue because it claims the lives of many people who might have been treated if the disease had been identified earlier. The leading cause of death in the world is cardiovascular disease, usually referred to as heart disease. Creating reliable, effective, and precise predictions for these diseases is one of the biggest issues facing the medical world today. This experiment examined a range of machine learning approaches, including Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Artificial Neural Networks, to determine which machine learning algorithm was most effective at predicting heart diseases. One of the most often utilized data sets for this purpose, the UCI heart disease repository provided the data set for this study. The K-Nearest Neighbor technique was shown to be the most effective machine learning algorithm for determining whether a patient has heart disease. It will be beneficial to conduct further studies on the application of additional machine learning algorithms for heart disease prediction.

The second groundwork related to “Heart disease prediction using machine learning techniques” and the details are presented as follows. Machine Learning (ML), which is one of the most prominent applications of Artificial Intelligence, is doing wonders in the research field of study. In this paper machine learning is used in detecting if a person has a heart

disease or not. A lot of people suffer from cardiovascular diseases (CVDs), which even cost people their lives all around the world. Machine learning can be used to detect whether a person is suffering from a cardiovascular disease by considering certain attributes like chest pain, cholesterol level, age of the person and some other attributes. Classification algorithms based on supervised learning which is a type of machine learning can make diagnoses of cardiovascular diseases easy. Algorithms like K-Nearest Neighbor (KNN), Random Forest are used to classify people who have a heart disease from people who do not. Two supervised machine learning algorithms are used in this paper which are, K-Nearest Neighbor (K-NN) and Random Forest. The prediction accuracy obtained by K-Nearest Neighbor (K-NN) is 86.885% and the prediction accuracy obtained by Random Forest algorithm is 81.967%.

3. Research methodology

AdaBoost (Adaptive Boosting), an ensemble learning technique that aims to improve the performance of weak classifiers by combining them into a stronger model. AdaBoost works by iteratively training a series of weak learners, typically decision trees, where each new classifier in the sequence focuses on the misclassifications made by the previous classifiers. This adaptive approach enables AdaBoost to reduce bias and variance, resulting in a more robust predictive model. One of the main advantages of AdaBoost is its ability to effectively handle noisy data and outliers by giving more weight to misclassified instances, allowing the model to correct errors progressively. Additionally, AdaBoost is computationally efficient, requiring fewer resources compared to other ensemble methods like random forests, and is less prone to overfitting, particularly when used with simple base learners. The proposed system employs AdaBoost to predict heart disease by focusing on the most important features from the dataset, ensuring that the model delivers high accuracy and generalizability. Its simplicity and interpretability make it a suitable choice for real-time medical applications, where quick, reliable predictions are essential for decision-making. The adaptability of AdaBoost makes it an ideal candidate for this project, providing an effective and efficient solution for early-stage heart disease detection.

MODULES:

Data Collection:

The first step in the project is to gather relevant datasets containing medical and clinical data for heart disease prediction. The data is sourced from both publicly available healthcare datasets and private datasets, providing a comprehensive set of features such as age, blood pressure, cholesterol levels, heart rate, and more. These datasets are essential for training the machine learning models and ensuring they can accurately predict the likelihood of heart disease based on a variety of patient characteristics.

Data Preprocessing:

Once the data is collected, it undergoes preprocessing to ensure it is clean, consistent, and ready for analysis. This step involves handling missing values, normalizing numerical features to a standard scale, and encoding categorical variables such as gender and smoker status. Additionally, to address class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) is applied. This technique generates synthetic samples of the minority class (e.g., patients with heart disease) to ensure the model does not become biased toward the majority class.

Feature Extraction:

Feature extraction involves selecting the most relevant and informative features for the prediction task. Various techniques such as chi-square, analysis of variance (ANOVA), and mutual information are employed to identify the key attributes that contribute most to predicting heart disease. These methods help in reducing the dimensionality of the dataset and improving the performance of machine learning models by focusing on the most influential features, while removing redundant or irrelevant ones.

Model Application:

With the preprocessed data and selected features, several machine learning classifiers are applied to the dataset to predict heart disease risk. The classifiers tested include Naive Bayes, Support Vector Machine (SVM), XGBoost, AdaBoost, Random Forest, K-Nearest Neighbours (KNN), and Decision Trees. Each model is trained using the training data, and its performance is assessed using evaluation metrics such as accuracy, sensitivity, specificity, and precision.

Accuracy and Prediction:

The accuracy and prediction performance of each model are evaluated based on various metrics. Accuracy measures the overall percentage of correct predictions made by the model, while precision, sensitivity, and specificity provide insights into how well the model distinguishes between patients with and without heart disease. Additionally, the F1 score is calculated to measure the balance between precision and recall. The final chosen model is validated using cross-validation to ensure robustness, and its prediction results are incorporated into a mobile application for real-time predictions.

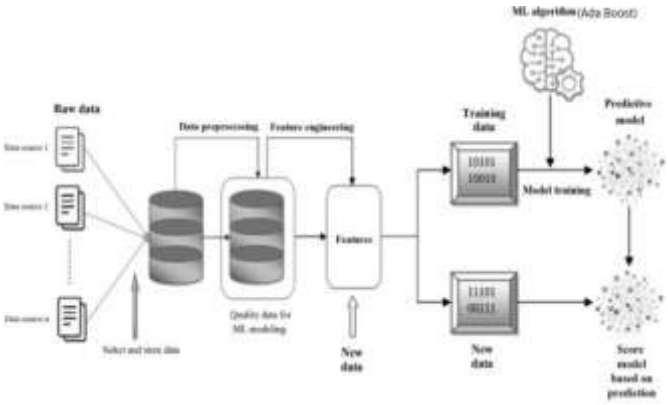

Fig: System Architecture of Research Methodology

4. Conclusion and Future Scope

In conclusion, this project has demonstrated the significant potential of machine learning, specifically the AdaBoost classifier, in enhancing early-stage heart disease prediction. By leveraging diverse feature selection methods (chi-square, ANOVA, and mutual information) and applying cross-validation techniques to multiple datasets, we developed a robust model capable of achieving high accuracy, sensitivity, specificity, precision, and F1 score. The incorporation of the Synthetic Minority Oversampling Technique (SMOTE) addressed data imbalance, further improving the model's performance, while the explainable AI approach using SHAP provided critical insights into the prediction process, ensuring transparency and trust in the system's decisions. The integration of this model into a mobile app offers a practical solution for users to receive rapid heart disease risk assessments, promoting early detection and timely interventions.

The future enhancement of the heart disease prediction project, utilizing the AdaBoost classifier, can focus on several key areas

to further improve its performance, scalability, and impact. First, expanding the data sources by incorporating real-time patient data from wearable devices, fitness trackers, and continuous monitoring tools could enable more dynamic and timely predictions. Integrating additional factors such as genomic data, family history, and lifestyle choices (diet, smoking, physical activity) would provide a more comprehensive view of a patient's risk profile. To enhance the model's feature engineering, exploring advanced techniques like Recursive Feature Elimination (RFE), deep learning-based feature extraction, and time-series analysis for capturing patient data trends over time could improve prediction accuracy. Furthermore, the use of hybrid ensemble methods, such as combining AdaBoost with other classifiers (e.g., XGBoost, Random Forest) or employing stacking and blending techniques, could optimize performance. Addressing data imbalance with more sophisticated techniques, such as cost-sensitive learning algorithms, could improve performance on underrepresented classes.

5. REFERENCES

[1] (2023). World Health Organization. Cardiovascular Diseases (CVDs). Accessed: May 5, 2023. [Online]. Available: https://www.afro.who.int/ health-topics/cardiovascular-diseases.
[2] Z. Alom, M. A. Azim, Z. Aung, M. Khushi, J. Car, and M. A. Moni, Early stage detection of heart failure using machine learning techniques,"in Proc. Int. Conf. Big Data, IoT, Mach. Learn., Cox's Bazar, Bangladesh,2021, pp. 23–25.
[3] S. Gour, P. Panwar, D. Dwivedi, and C. A. Mali, "Machine learning approach for heart attack prediction," in Intelligent Sustainable Systems.Singapore: Springer, 2022, pp. 741–747.
[4] C. Gupta, A. Saha, N. S. Reddy, and U. D. Acharya, "Cardiac disease prediction using supervised machine learning techniques," J. Phys., Conf. Ser., vol. 2161, no. 1, 2022, Art. no. 012013.
[5] K. Shameer, "Machine learning predictions of cardiovascular disease risk in a multi-ethnic population using electronic health record data," Int. J.Med. Inform., vol. 146, Feb. 2021, Art. no. 104335.
[6] M. Liu, X. Sun, Y. Liu, X. Yang, Y. Xu, and X. Sun, "Deep learningbased prediction of coronary artery disease with CT angiography," Jpn.J. Radiol., vol. 38, no. 4, pp. 366–374, 2020.
[7] N. Zakria, A. Raza, F. Liaquat, and S. G. Khawa ja, "Machine learning based analysis of cardiovascular disease prediction," J. Med. Syst., vol. 41, no. 12, p. 207, 2017.
[8] M. Yang, X. Wang, F. Li, and J. Wu, "A machine learning approach to identify risk factors for coronary heart disease: A big data analysis," Comput. Methods Programs Biomed., vol. 127, pp. 262–270, Apr. 2016.
[9] C. Ngufor, A. Hossain, S. Ali, and A. Alqudah, "Machine learning algorithms for heart disease prediction: A survey," Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 2, pp. 7–29, 2016.
[10] A. Shoukat, S. Arshad, N. Ali, and G. Murtaza, "Prediction of cardiovascular diseases using machine learning: A systematic review," J. Med. Syst., vol. 44, no. 8, p. 162, Aug. 2020.