Ms. Voliviya Veronica Sam. N, 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

An Intelligent System For Cardiovascular Disease Risk Prediction Using Xgboost

Ms. Voliviya Veronica Sam. N
[1]II M.Sc Data Analytics, Sri Ramakrishna College Of Arts & Science For Women
Mail Id: 242cf010@Srcw.Ac.In

ABSTRACT-The increasing incidence of cardiovascular disease (CVD) highlights the need for accurate and early risk prediction methods. Traditional approaches often fail to capture complex relationships among multiple health factors, leading to lower prediction efficiency. This paper proposes a machine learning–based framework for early prediction of cardiovascular disease risk using patient health and lifestyle data.
The system incorporates data preprocessing, feature selection using Recursive Feature Elimination (RFE), and class balancing using SMOTE to improve model performance. The optimized dataset is trained using the XGBoost classifier, which effectively handles nonlinear relationships and enhances prediction accuracy.
Experimental results demonstrate that the proposed model achieves high accuracy with improved precision, recall, and F1-score. The system provides a reliable and efficient solution for early detection of cardiovascular disease, supporting timely medical intervention and preventive healthcare strategies.

Keywords:- Cardiovascular Disease, Machine Learning, Early Risk Prediction, XGBoost Classifier, Recursive Feature Elimination (RFE), SMOTE, Feature Selection, Data Preprocessing, Class Balancing, Health Data Analysis, Lifestyle Factors, Disease Prediction, Predictive Modeling, Precision, Recall, F1-Score, Preventive Healthcare, Medical Diagnos
---

I. INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death worldwide, affecting millions of people each year. It includes various conditions such as coronary artery disease, heart attack, stroke, and hypertension. The increasing prevalence of CVD is largely influenced by lifestyle factors such as poor diet, lack of physical activity, stress, smoking, and other medical conditions like diabetes and obesity. Early prediction of cardiovascular risk is essential to reduce mortality and improve patient outcomes.

Traditional diagnostic methods mainly rely on clinical evaluation and statistical analysis, which may not effectively capture complex relationships among multiple health parameters. With the rapid growth of healthcare data, machine learning techniques have emerged as powerful tools for analyzing large datasets and identifying hidden patterns. These techniques enable faster and more accurate prediction compared to conventional methods.

In this study, a machine learning–based framework is proposed for early prediction of cardiovascular disease risk using patient health data. The system incorporates data preprocessing, feature selection using Recursive Feature Elimination (RFE), and class balancing using SMOTE to improve prediction performance. The XGBoost classifier is used to model complex nonlinear relationships and achieve high prediction accuracy.

The objective of this work is to develop an efficient and reliable system that can classify individuals into different risk levels and support early diagnosis. This approach aims to assist healthcare professionals in making informed decisions and contribute to preventive healthcare strategies.

II. INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death worldwide, affecting millions of people every year. It includes various conditions such as coronary artery disease, heart attack, stroke, and hypertension. The increasing prevalence of CVD is mainly attributed to unhealthy lifestyle habits, lack of physical activity, poor diet, stress, smoking, and other medical conditions such as diabetes and obesity. Early detection of cardiovascular risk is therefore essential to

© 2026 Ms. Voliviya Veronica Sam. N, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

reduce mortality rates and improve patient outcomes [1].

With the rapid growth of healthcare data, large volumes of patient information are now available, including clinical measurements and lifestyle-related attributes. However, traditional diagnostic methods often rely on manual evaluation and basic statistical techniques, which may not effectively capture complex relationships among multiple health parameters. This limitation makes it difficult to accurately predict cardiovascular disease risk at an early stage [2].

Given these challenges, there is a growing need for intelligent, scalable, and efficient systems that can analyze high-dimensional medical datasets and provide accurate predictions. Machine learning techniques have emerged as powerful tools in the healthcare domain, enabling automatic pattern recognition and data-driven decision-making. These techniques can process multiple features simultaneously and identify hidden patterns that are not easily detectable through conventional methods [3].

To address these limitations, this study proposes a machine learning–based framework for early prediction of cardiovascular disease risk. The system incorporates data preprocessing techniques such as handling missing values, encoding categorical variables, and feature scaling to ensure data consistency. Feature selection is performed using Recursive Feature Elimination (RFE) to identify the most relevant health attributes and reduce dataset dimensionality. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) is applied to handle class imbalance and improve prediction performance.

The optimized dataset is then trained using the XGBoost classifier, which is capable of capturing complex nonlinear relationships among clinical and lifestyle factors. This approach improves prediction accuracy, reduces misclassification, and enhances the reliability of the system. By integrating feature optimization with advanced machine learning techniques, the proposed model provides an efficient solution for early cardiovascular disease risk prediction.

Overall, this system aims to support healthcare professionals in timely diagnosis and preventive decision-making. It contributes to the development of intelligent healthcare systems that improve patient care and enable early intervention to reduce the impact of cardiovascular diseases.

III. LITERATURE REVIEW

Several research studies have explored the application of machine learning techniques for predicting cardiovascular disease risk and improving early diagnosis. These studies focus on analyzing patient health data and identifying key risk factors associated with heart-related conditions.

Detrano et al. (1989) introduced one of the earliest structured datasets for heart disease prediction and applied statistical and machine learning techniques to identify important clinical attributes such as age, cholesterol, and blood pressure. This work laid the foundation for future research in cardiovascular disease prediction.

Palaniappan and Awang (2008) compared various machine learning algorithms including Naïve Bayes, Decision Trees, and Neural Networks for predicting heart disease. Their results showed that machine learning models can achieve higher accuracy compared to traditional statistical methods, highlighting their importance in healthcare applications.

Jabbar et al. (2013) applied data mining techniques to analyze patient medical records and identify patterns related to cardiovascular disease. Their study emphasized the role of feature selection in improving prediction accuracy and reducing model complexity.

Kourou et al. (2015) discussed the use of machine learning algorithms in healthcare systems for disease prediction and decision support. Their work demonstrated that advanced algorithms can effectively handle large medical datasets and improve diagnostic performance.

Chen and Guestrin (2016) introduced the XGBoost algorithm, an optimized gradient boosting framework known for its high performance and efficiency. XGBoost has been widely used in healthcare prediction tasks due to its ability to handle complex nonlinear relationships and large datasets.

Chawla et al. (2002) proposed the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in datasets. This method improves the model's ability to detect minority classes, which is

particularly important in medical datasets where disease cases are often fewer than non-disease cases.

Guyon et al. (2002) introduced Recursive Feature Elimination (RFE), a feature selection technique that iteratively removes less important features to improve model performance. This method helps in identifying the most significant predictors of cardiovascular disease risk.

Recent studies have also explored ensemble and hybrid models for cardiovascular disease prediction, combining multiple algorithms to improve accuracy and reliability. These approaches demonstrate that integrating feature selection, data balancing, and advanced machine learning models can significantly enhance prediction performance.

Despite these advancements, challenges such as class imbalance, feature redundancy, and the need for highly accurate prediction models still exist. These limitations motivate the development of the proposed system, which integrates preprocessing, feature optimization, and the XGBoost classifier to improve cardiovascular disease risk prediction.

IV. PROPOSED WORK

The proposed work aims to develop an efficient and accurate machine learning–based system for early prediction of cardiovascular disease (CVD) risk using patient health data. The system is designed to analyze multiple clinical and lifestyle attributes, identify significant patterns, and classify individuals into different risk levels.

The overall framework consists of several stages, including data preprocessing, feature selection, class balancing, and model training. Initially, the collected dataset is processed to handle missing values, remove inconsistencies, and ensure data quality

Numerical features are standardized to maintain uniform distribution, and categorical variables such as gender, smoking status, and alcohol consumption are converted into numerical form using encoding techniques.

To improve model performance and reduce computational complexity, feature selection is performed using Recursive Feature Elimination (RFE). This method identifies the most relevant health attributes that significantly influence cardiovascular disease risk and removes redundant or less important features. By focusing only on important predictors, the system achieves better efficiency and accuracy.

Since medical datasets often suffer from class imbalance, where the number of non-disease cases is higher than disease cases, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE generates synthetic samples for the minority class, ensuring a balanced dataset and improving the model's ability to correctly identify high-risk patients.

The core component of the proposed system is the machine learning model built using the XGBoost classifier. XGBoost is chosen due to its high performance, scalability, and ability to handle complex nonlinear relationships between multiple features. It works by constructing an ensemble of decision trees in a sequential manner, optimizing prediction accuracy through gradient boosting techniques.

Once trained, the model is used to predict cardiovascular disease risk levels based on input health parameters. The output classifies individuals into categories such as low risk, medium risk, and high risk. These predictions help in early identification of potential cardiovascular conditions and support timely medical intervention.

By integrating data preprocessing, feature optimization, class balancing, and advanced machine learning techniques, the proposed system overcomes the limitations of traditional methods. It improves prediction accuracy, reduces misclassification, and provides a reliable solution for early cardiovascular disease risk prediction in healthcare applications.

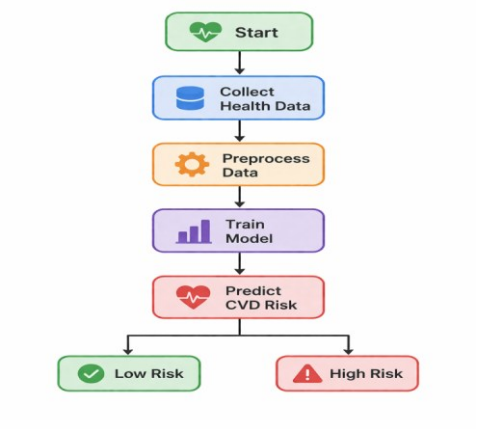


Figure 1: Proposed System Architecture for Cardiovascular Disease Prediction

V. METHODOLOGY

The methodology of the proposed system focuses on developing an efficient machine learning model for predicting cardiovascular disease (CVD) risk using patient health data. The process consists of several stages, including data collection, preprocessing, feature selection, class balancing, model training, and prediction.

Initially, the dataset containing patient health records is collected. It includes various attributes such as age, gender, height, weight, blood pressure (ap_hi and ap_lo), cholesterol levels, glucose levels, smoking habits, alcohol consumption, and physical activity. These features play a significant role in determining cardiovascular risk.

In the preprocessing stage, data cleaning is performed to handle missing values and remove inconsistencies. Categorical variables are converted into numerical form using encoding techniques, and numerical features are standardized to ensure uniform scaling. This step improves data quality and enhances model performance.

Next, feature selection is carried out using Recursive Feature Elimination (RFE). This technique selects the most relevant features by iteratively removing less important attributes, thereby reducing dimensionality and improving computational efficiency. Selecting important features also helps in increasing prediction accuracy.

To address the issue of class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE generates synthetic samples for the minority class, ensuring a balanced dataset and improving the model's ability to correctly classify high-risk cases.

The processed data is then used to train the XGBoost classifier. XGBoost is an advanced ensemble learning algorithm that builds multiple decision trees and combines them to improve prediction accuracy. It is capable of handling complex nonlinear relationships among features and provides efficient performance for large datasets.

Finally, the trained model is used to predict cardiovascular disease risk levels. The output categorizes individuals into different risk groups, enabling early detection and supporting timely medical intervention. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure reliability.

VI. RESULTS

The performance of the proposed cardiovascular disease (CVD) prediction system is evaluated using standard classification metrics. The model is trained and tested on the processed dataset after applying preprocessing, feature selection, and class balancing techniques. The effectiveness of the system is measured using accuracy, precision, recall, and F1-score.

The XGBoost classifier demonstrates strong performance in predicting cardiovascular disease risk. The model achieves an overall accuracy of approximately **96%**, indicating its ability to correctly classify both high-risk and low-risk patients. The high precision value shows that the model produces fewer false positive predictions, while the high recall indicates its effectiveness in identifying actual disease cases. The F1-score further confirms a balanced performance between precision and recall.

A confusion matrix is used to analyze the classification results in detail. It shows that the number of true positive and true negative predictions is significantly higher compared to false classifications. This indicates that the model performs well in distinguishing between patients with and without cardiovascular disease.

The results also demonstrate the impact of feature selection and SMOTE in improving model performance. Feature selection helps in reducing noise and focusing

on the most relevant attributes, while SMOTE ensures balanced data distribution, leading to better prediction accuracy.

Overall, the experimental results confirm that the proposed system is efficient, reliable, and capable of accurately predicting cardiovascular disease risk. This model can be effectively used as a decision-support tool in healthcare applications for early diagnosis and prevention.

VII. CONCLUSION

This study presents a machine learning–based framework for the early prediction of cardiovascular disease (CVD) risk using patient health data. The proposed system integrates data preprocessing, feature selection using Recursive Feature Elimination (RFE), and class balancing using SMOTE to improve the quality and balance of the dataset. The XGBoost classifier is employed to model complex relationships among various clinical and lifestyle attributes, resulting in high prediction accuracy.

The experimental results demonstrate that the model achieves strong performance in terms of accuracy, precision, recall, and F1-score, making it a reliable tool for identifying individuals at risk of cardiovascular disease. The use of feature selection and data balancing techniques significantly enhances the model's efficiency and reduces misclassification.

Overall, the proposed system provides an effective solution for early detection of cardiovascular disease and supports healthcare professionals in making timely and informed decisions. Future work can focus on integrating deep learning techniques, real-time data analysis, and explainable AI methods to further improve prediction performance and interpretability.

REFERENCES

1. R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiology, 1989.
2. S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," IEEE/ACS Int. Conf., 2008.
3. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease prediction system using associative classification," IEEE Conf., 2013.
4. K. Kourou et al., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, 2015.
5. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD, 2016.
6. N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, 2002.
7. I. Guyon et al., "Gene selection for cancer classification using support vector machines," Machine Learning, 2002.
8. R. Deo, "Machine learning in medicine," Circulation, 2015.
9. World Health Organization, "Cardiovascular diseases (CVDs)," 2023.
10. UCI Machine Learning Repository, "Heart Disease Dataset," 2019.
11. H. Chen et al., "A hybrid model for heart disease prediction using machine learning," IEEE Access, 2020.
12. S. Sharmila and S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using feature selection," Int. J. Computer Applications, 2018.
13. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.
14. M. S. Uddin et al., "Comparative study of machine learning algorithms for heart disease prediction," IEEE Access, 2019.
15. J. Brownlee, "XGBoost with Python: Gradient Boosted Trees," Machine Learning Mastery, 2016.
16. A. K. Sen et al., "Heart disease classification using machine learning techniques," Int. J. Computer Science and Information Technologies, 2021.