Intelligent Cardiovascular Disease Risk Prediction Using Machine Learning

Abhay Shivaji Dahe[1], Rutwick Santosh Pawar[2], Yogesh Rajendra Borse[3]
[1,2,3] *Department of Computer Science and Engineering, Sandip University, Nashik, India*

Abstract— cardiovascular diseases (CVDs) are among the leading causes of mortality worldwide and represent a serious challenge to modern healthcare systems. These diseases often develop silently and are diagnosed only after severe complications occur. Early prediction of cardiovascular risk can significantly reduce mortality by enabling timely medical intervention and lifestyle modification. With the rapid growth of electronic health records and medical datasets, machine learning techniques have become effective tools for analyzing complex clinical data and identifying hidden risk patterns. This paper proposes an intelligent machine learning–based system for predicting cardiovascular disease risk using patient health parameters such as age, gender, blood pressure, cholesterol level, blood sugar, heart rate, and other related indicators. The dataset is preprocessed through data cleaning, normalization, and feature selection to improve data quality and model performance. Multiple machine learning algorithms including Logistic Regression, Decision Tree, and Random Forest are implemented and evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. Experimental results demonstrate that the Random Forest algorithm achieves superior performance compared to other classifiers. The proposed system provides a reliable, cost-effective, and scalable decision-support tool that can assist healthcare professionals in early diagnosis and preventive cardiovascular care.

Index Terms — Cardiovascular Disease, Machine Learning, Risk Prediction, Healthcare Analytics, Random Forest

I. INTRODUCTION

Cardiovascular diseases are a major cause of death globally, including conditions such as coronary artery disease, heart attacks, strokes, and hypertension. One of the key challenges in managing cardiovascular diseases is their gradual development and lack of noticeable symptoms during the early stages. As a result, many patients are diagnosed only after the disease has progressed to a critical stage, leading to increased treatment costs and higher mortality rates. Traditional diagnostic approaches rely on clinical tests and expert interpretation, which can be time-consuming and may vary depending on practitioner experience. With the increasing availability of digital medical records, there is a growing need for intelligent systems that can analyze patient data efficiently and provide early risk predictions. Machine learning offers powerful techniques for extracting meaningful insights from large healthcare datasets. By learning from historical patient records, machine learning models can identify complex relationships among multiple risk factors and predict the likelihood of cardiovascular disease. This research focuses on developing a machine learning–based risk prediction system to support early diagnosis and preventive healthcare.

II. LITERATURE REVIEW

Numerous studies have explored the application of machine learning techniques for cardiovascular disease prediction. Logistic Regression is commonly used due to its simplicity and effectiveness in binary classification tasks. However, its performance is limited when handling non-linear relationships among medical attributes. Decision Tree algorithms provide interpretability and can identify important risk factors such as blood pressure and cholesterol levels, but they are prone to overfitting, especially with noisy datasets. Ensemble learning methods such as Random Forest have gained popularity due to their improved accuracy and robustness. By combining multiple decision trees, Random Forest reduces variance and handles complex feature interactions effectively. Several researchers have reported that Random Forest outperforms individual classifiers in heart disease prediction tasks.

Support Vector Machines and neural networks have also been applied, achieving good accuracy but often requiring complex parameter tuning and higher computational resources. Based on the literature, Random Forest is selected as the primary model in the proposed system.

III. METHODOLOGY

The proposed system follows a structured methodology consisting of data collection, preprocessing, feature selection, model training, and evaluation. A cardiovascular disease dataset containing clinical and lifestyle attributes is used for experimentation. Data preprocessing includes handling missing values, removing duplicate records, normalizing numerical features, and encoding categorical variables to ensure data consistency. Feature selection techniques are applied to identify the most relevant attributes influencing cardiovascular risk, thereby reducing model complexity and improving performance. The preprocessed dataset is divided into training and testing sets. Machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest are trained using the training data and evaluated on the test data using standard performance metrics.

IV. RESULTS AND DISCUSSION

The performance of the implemented machine learning models is evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. Logistic Regression provides reasonable accuracy but struggles with complex non-linear relationships. Decision Tree models are easy to interpret but tend to overfit the training data. The Random Forest algorithm achieves the highest prediction accuracy and demonstrates better generalization due to its ensemble learning approach. These results indicate that Random Forest is a reliable and effective model for cardiovascular disease risk prediction. The system can assist healthcare professionals by providing early warnings and supporting preventive decision-making.

V. CONCLUSION

This paper presented an intelligent machine learning–based system for predicting cardiovascular disease risk using patient health data. A comparative evaluation of multiple algorithms showed that the Random Forest model outperforms other classifiers in terms of accuracy and robustness. The proposed system can serve as a valuable decision-support tool for early diagnosis and preventive healthcare. Future work may include the integration of larger datasets, real-time health monitoring data, and advanced deep learning models to further enhance prediction accuracy.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the faculty members of the Department of Computer Science and Engineering, Sandip University, Nashik, for their guidance, encouragement, and support throughout the completion of this research work.

REFERENCES

[1] M. Akhil Jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm,” CIMTA, 2013.

[2] C. S. Dangare, “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques,” International Journal of Computer Applications, 2012.

[3] N. G. B. Amma, “Cardiovascular Disease Prediction System Using Genetic Algorithm,” IEEE International Conference, 2012.

[4] I. S. Siva Rao and T. Srinivasa Rao, “Performance Identification of Different Heart Diseases Based on Neural Network Classification,” 2016.