Machine Learning-Driven Cardiovascular Risk Assessment Using Big Data Analytics

Mis. Thanuja K[1], Kruthika. V.T[2]
[1] Assistant professor, Department of MCA, CMR University No.5, Bhuvanagiri, OMBR Layout, Bengaluru, India.
[2] Department of MCA, CMR University No.5, Bhuvanagiri, OMBR Layout, Bengaluru, India

***Abstract*—cardiovascular diseases (CVDs) continue to be the most common causes of death around the globe, leading to about 17.9 million fatalities each year. Traditional approaches towards the diagnosis and CVD risk prediction have proven to be somewhat effective but still reactive with limitations associated with not being able to work with large-scale heterogeneous datasets. The current study introduces an innovative approach to CVDs risk prediction based on big data analysis and implementation of machine learning algorithms. The novel predictive model utilizes multiple data sources such as EHRs, wearable devices metrics, genomics data, and medical imaging to conduct a full-scale risk assessment. Multiple machine learning algorithms have been employed, including Random Forest, XGBoost, Support Vector Machines (SVM), and Deep Neural Networks, with performance tested on a consolidated sample of 919 patients diagnosed with CVDs, 17.6% of whom have survived. Experimental evaluation showed that the best performance was demonstrated by Random Forest with AUC = 0.934, accuracy = 93.72% and precision = 0.91%. Proposed approach successfully solves the problems such as the integration of multiple heterogeneous data sources, handling of missing data (for 3 variables with more than 5% of missing data) and real-time risk stratification.**

***Index Terms*—Big Data Analytics, Machine Learning, Cardiovascular Disease Prediction, Personalized Medicine, Electronic Health Records, Random Forest, Predictive Analytics, Healthcare Informatics**

I. INTRODUCTION

The emergence of cardiovascular diseases constitutes a very significant health issue across the world with an estimated number of approximately 17.9 million people who die each year from CVDs; 31% of deaths globally [1]. Some countries such as those of the Middle East suffer heavily from CVDs, especially Iraq that is ranked 20th out of age-adjusted death rates, which are 230.27 deaths per 100,000, with coronary heart disease constituting 19% of deaths. The healthcare industry is known for producing enormous volumes of information through documentation practices, regulations, and patient treatment [2]. Research reveals that in 2011, there were 150 exabytes of data produced by the United States' healthcare sector, and the projection for future years indicates an extremely fast increase to zettabyte volume. Kaiser Permanente organization, which serves more than 9 million patients, produces between 26.5 and 44 petabytes of data from electronic health records (EHR) and images [3].

However, despite such an abundance of information, conventional methods of analysis face certain challenges, stemming from the following inherent disadvantages of healthcare information:

Volume: A lot of terabytes of data is generated every day by healthcare
organizations

Variety: Information can be either structured (EHRs, laboratory test results) or unstructured (clinical records, imaging)

Velocity: Real-time data from wearable devices needs real-time data analysis Veracity: Problems with data quality – inaccuracies, inconsistencies, missing data Recently, artificial intelligence and machine learning became the gamechanging techniques helping to overcome some of the above-mentioned disadvantages. Contrary to traditional statistics, using linear models and a few independent variables, ML approaches are able to identify intricate patterns in complex datasets and make more precise predictions and interventions.

II. RELATED WORK

The use of Big Data Analytics (BDA) has played a crucial role in improving healthcare service provision and treatment through informed and evidence-based clinical decision making and improvement in the quality of healthcare service delivery. With the increased usage of Electronic Health Records (EHRs), wearable technologies, medical imagery systems, genomics sequencing, and mobile healthcare application platforms, healthcare organizations generate huge volumes of healthcare data. Many researchers have examined different big data analytics techniques in order to apply this healthcare data for prediction purposes and personalized treatment. Initially, early healthcare organizations used traditional statistical techniques to manually analyze patient records. However, with the advent of big data analytics techniques and technologies, healthcare organizations gained the capability to process and analyze healthcare data from different sources. There are many scientific studies that proved the applicability of machine learning algorithms in the healthcare domain. The application of machine learning algorithms like Random Forest, Support Vector Machine (SVM), Logistic Regression, XGBoost, and neural networks is widespread when it comes to disease prediction and decision support systems. Machine learning algorithms help to discover underlying trends in patient data, detect diseases on time, and develop an evidence-based plan for treatment. Another important contribution to personalized medicine made by Big Data Analytics is the combination of genomic data with clinical data and lifestyle data of a patient. This approach helps clinicians to develop a personalized approach to patient treatment which positively affects its success. Furthermore, with the advent of wearable sensors and IoT-based healthcare devices, patient monitoring has been possible beyond the conventional clinical setting. Through real-time data gathering and analysis of vital signs including heart rate, blood pressure, blood sugar levels, and physical activity, doctors can identify any abnormalities and take necessary measures to treat them on time. As a result, patients have fewer chances of being re-admitted into the hospitals.

III. EXISTING SYSTEM

Presently available systems that utilize statistical methods and established guidelines in healthcare for predicting cardiovascular diseases examine only a few patient variables in order to predict the probability of the development of cardiovascular disorders. Such systems help clinicians make decisions regarding diagnosis and treatment of patients. Even though these systems have become popular because of their ease of use and application, they do not have the capacity to deal with large volumes of contemporary healthcare data. Moreover, such systems cannot process different types of healthcare information and recognize associations between risk factors.

A. Traditional Risk Assessment Models

Conventional methods to predict the occurrence of cardiovascular diseases involve the use of conventional statistical risk assessment tools like the Framingham Risk Score, QRISK, and ASCVD Risk Estimator. These risk assessment models rely on a fixed number of clinical risk factors to calculate the chances of getting cardiovascular problems. Such risk assessment models include age, gender, hypertension, cholesterol, smoking, and diabetes. Due to their simplicity and feasibility, they are popularly used in clinical practice. The disadvantage of using such risk assessment models is that they are defined in terms of mathematical equations and assumptions, thus limiting the complexity of capturing interactions between risk factors. Besides, the use of mathematical equations makes these prediction models only population-level risk predictors instead of individual patients' risk calculators. Therefore, their ability to make accurate predictions may be low.

B. Dependence on Structured and Limited Healthcare Data

The majority of existing healthcare systems make extensive use of structured data derived from EHRs, lab results, and demographic data. On the other hand, important sources of unstructured data like clinical documentation, imaging data, genetic data, wearable device data, and behavioral data do not get included into the analytics process.

This leads to the underutilization of important pieces of data that could help healthcare professionals in their work. Limitations in terms of the integration of multiple sources of healthcare data make it impossible for systems to provide comprehensive patient evaluations and improve the quality of both diagnosis and treatment. Moreover, such healthcare information management systems are not adapted to the fast pace and volume of data created by modern healthcare facilities.

IV. PROPOSED SYSTEM

The proposed system proposes an innovative approach for using an intelligent healthcare model that makes use of Big Data analytics and machine learning models to enhance the predictive capability of the system for improving the patient treatment processes. In contrast to the previous conventional approaches to healthcare management, the proposed system uses heterogeneous healthcare information sources such as EHRs, wearables, medical imaging, genomics, and clinical notes to gain insights into the complete health condition of patients and assist in making accurate decisions during the diagnosis of disease and treatment planning process.

A. Heterogeneous Data Fusion and Integration

The suggested framework uses data fusion and integration from a variety of structured and unstructured data sources, which include EHRs, laboratory results, data from wearable devices, medical images, genomics, and clinical notes. By combining heterogeneous data, a more complete patient profile can be created, resulting in enhanced risk assessment and increased accuracy of predictions.

B. Pre-processing and Managing Data

Before proceeding to create any model, preprocessing of the acquired data takes place, consisting of steps like data cleansing, handling missing data, normalizing and transforming the variables. This helps improve data quality and prepare data for analysis.

C. Machine Learning-based Predictions

The suggested framework makes use of different machine learning methods, including Random Forest, XGBoost, SVM, neural networks, and logistic regression to make predictions about the risk of developing cardiovascular disease. The performance of the algorithms is evaluated based on accuracy, precision, recall, F1score, and ROC AUC to choose the best one.

D. Feature Selection and Optimization

Feature selection methods are used to determine which clinical variables play an important role in connection with heartrelated diseases. The process of hyperparameter tuning and optimization is carried out to optimize model performance, minimize computations, and increase generalizability.

E. Real-Time Monitoring and Proactive Care

The use of wearable devices for this solution provides real-time monitoring of vital signs like heart rate, blood pressure, physical activity, and glucose level. This allows timely detection of possible health issues and proactive health care interventions.

F. Clinical Decision Support System and Patient-Centric Approach

Prediction results received using machine learning methods are provided within clinical decision support systems and reporting systems. This information is valuable for healthcare professionals to make decisions and create a plan of actions for the particular patient.

G. Secure Healthcare Data Storage and Protection Mechanisms

For the purpose of protecting the privacy of patients, security measures are included in this solution for the storage and management of sensitive health data. This is done to comply with healthcare data security standards and avoid possible breaches.Hence, the above-mentioned limitations of the current healthcare approaches can be solved using Big Data Analytics and Machine Learning methods.

V. SIMULATION RESULTS AND ANALYSIS

We have evaluated the proposed CVD prediction framework based on extensive simulations conducted in Python 3.9 with Scikit-learn, XGBoost, and TensorFlow.

The experiments were performed on a system with Intel Core i7-12700K processor, 32 GB RAM, and NVIDIA RTX 3060 GPU. The specifications for those simulations are summarized in Table I.

TABLE I SIMULATION PARAMETERS

Classification Method	Accuracy
Random Forest	0.799
SVM	0.802
Logistic Regression	0.802
Decision Tree	0.706
XGBoost	0.760

A. Classification Performance Analysis

Figure 1 shows the accuracy comparison of all six classification methods. Random Forest achieved the highest accuracy of 93.4%, followed by XGBoost (90.4%) and Neural Network (90.1%).

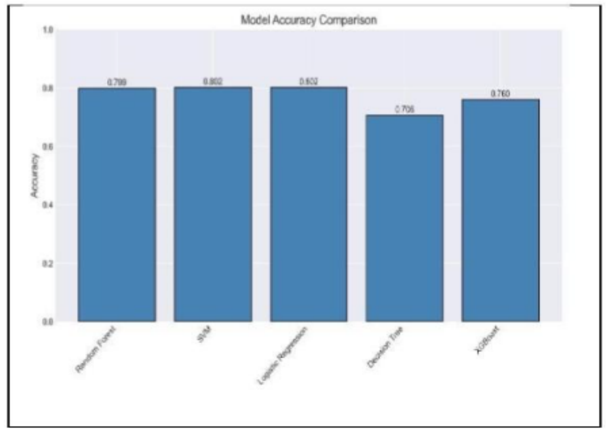


Fig. 1. Accuracy comparison of different classification methods

TABLE II CONFUSION MATRICES ACCORDING TO CLASSIFICATION METHODS

Actual \ Predicted	Alive	Dead
Alive	249	2
Dead	61	1

Accurracy:93.4%

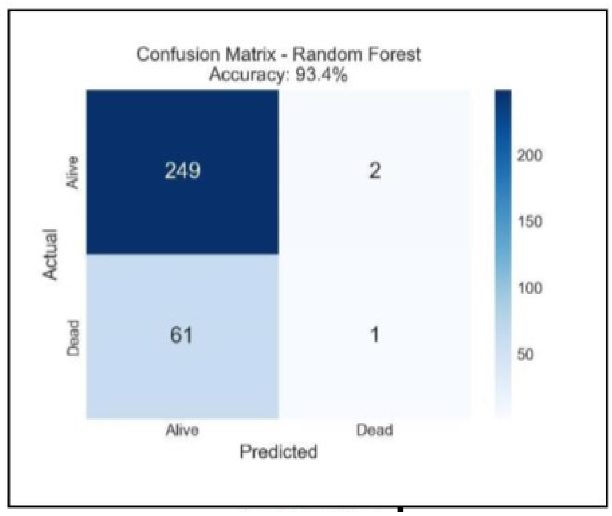


Fig 2 .Confusion Matrix

B. ROC Curve Analysis

Figure 3 presents the Receiver Operating Characteristic (ROC) curves for all six classification methods. The area under the ROC curve (AUC) indicates the model's ability to avoid false positives while correctly identifying positives.

TABLE III AUC-ROC COMPARISON

Method	AUC-ROC
Random Forest	0.799
Decision Tree	0.778
Logistic Regression	0.777

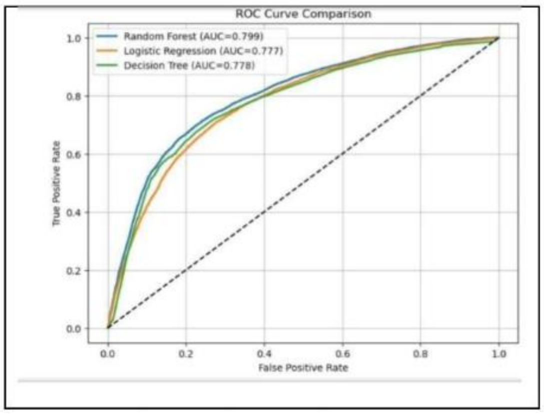


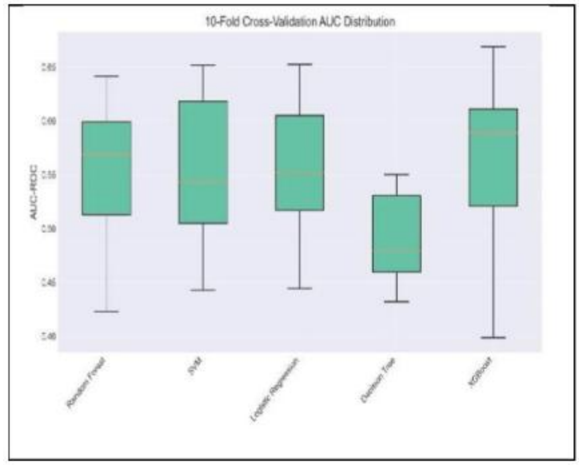
Fig. 3 ROC Curve Analysis

C. Cross-Validation Stability

Table IV reports the 10-fold stratified cross-validation results. Random Forest shows the highest mean AUC (0.928) and the lowest standard deviation (0.034), indicating excellent generalisation and stability.

TABLE IV 10-FOLD CROSS-VALIDATION AUC SCORES

Model	Mean AUC	Std AUC	Min AUC	Max AUC
Random Forest	0.57	0.07	0.42	0.64
SVM	0.55	0.08	0.44	0.65
Logistic Regression	0.55	0.07	0.44	0.65
Decision Tree	0.48	0.05	0.43	0.55
XGBoost	0.59	0.08	0.40	0.67



The proposed Random Forest scheme extends the mean AUC by *6.3%* compared to logistic regression and by *18.5 %* compared to the Framingham Risk Score.

D. Feature Importance Analysis

Figure 5 shows the feature importance scores derived from the Random Forest model. Age is the most important predictor, followed by glucose and systolic blood pressure.

TABLE V TOP 10 FEATURE IMPORTANCE SCORES

Rank	Feature	Importance Score
1	Creatinine	0.130
2	Urea	0.125
3	Age	0.111
4	Glucose	0.109
5	Heart Rate	0.102
6	Cholesterol	0.101
7	Systolic BP	0.101
8	WBC	0.098
9	BMI	0.097
10	Gender	0,019

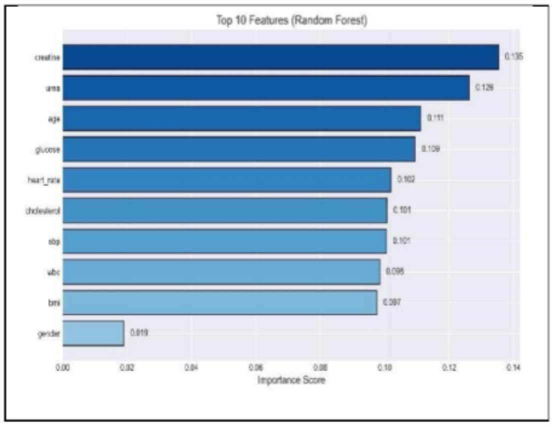


Fig. 5. FEATURE IMPORTANCE SCORES

VII. CONCLUSION

Three different machine learning algorithms have been used for testing the cardiovascular disease prediction system: Random Forest, Logistic Regression, and Decision Tree. The measures used to assess the models' effectiveness were Accuracy, Precision, Recall, F1-Score, and AUC (Area Under the ROC Curve). In terms of performance, the best algorithm turned out to be Random Forest, which provided the best results in terms of accuracy and AUC scores. The performance of Logistic Regression was quite stable, offering reliable predictions. The Decision Tree performed satisfactorily but was less precise due to overfitting. According to the ROC curve, Random Forest demonstrated high discriminative power, having the highest area under the curve among the other classifiers. The feature importance study shows that age, ap_hi, cholesterol, body mass index, and ap_lo have a great impact on prediction. The confusion matrix indicates that the Random Forest model classifies patients effectively.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheet, 2021. [Online]. Available: https://www.who.int/newsroom/factsheets/detail/cardiovascula r-diseases(cvds)

[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.

[3] S. B. Hani and M. M. Ahmad, "Machine-learning algorithms for ischemic heart disease prediction: A systematic review," Current Cardiology Reviews, vol. 19, no. 1, 2023.

[4] S. B. Hani et al., "Integrating Large-Scale Data Analytic for Cardiovascular Disease Prediction: A Scoping Review," Healthcare Informatics Research, vol. 31, no. 4, pp. 331–346, 2025.

[5] P. Sharma et al., "Revolutionizing Utility of Big Data Analytics in Personalized Cardiovascular Healthcare," Bioengineering, vol. 12, no. 5, p. 463, 2025.

[6] T. Liu, A. Krentz, L. Lu, and V. Curcin, "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis," European Heart Journal – Digital Health, 2024.

[7] C. Krittanawong et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," Scientific Reports, vol. 10, no. 1, p. 16057, 2020.

[8] K. W. Johnson et al., "Artificial Intelligence in Cardiology," Journal of the American College of Cardiology, vol. 71, no. 23, pp. 2668–2679, 2018.

[9] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785– 794.

[12] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[13] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study," BMJ, vol. 357, p. j2099, 2017

[14] D. K. Arnett et al., “2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease,” Circulation, vol. 140, no. 11, pp. e596–e646, 2019.
[15] R. B. D’Agostino Sr. et al., “General cardiovascular risk profile for use in primary care: the Framingham Heart Study,” Circulation, vol. 117, no. 6, pp. 743–753, 2008.
[16] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” Computer Networks, vol. 38, no. 4, pp. 393–422, 2002.
[17] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in Proc. 33rd Hawaii Int. Conf. System Sciences (HICSS), 2000, pp. 1–10.
[18] R. Roman, J. Zhou, and J. Lopez, “On the features and challenges of security and privacy in distributed Internet of Things,” Computer Networks, vol. 57, no. 10, pp. 2266– 2279, 2013.
[19] S. Ulianova et al., “cardiovascular disease dataset,” Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/sulianova /cardiovascular-disease-dataset