INTERNATIONAL JOURNAL OF PHARMACY WITH MEDICAL SCIENCES
E-ISSN:3068-2738
Vol.5, No. 4(1)(2025)
www.ijpams.com

INTELLIGENT ML-BASED CARDIOVASCULAR RISK PREDICTION FRAMEWORK FOR PREVENTIVE DIGITAL HEALTHCARE

K Sudharshan Reddy
Assistant Professor
Department of Computer Science and Engineering
Sree Chaitanya College of Engineering, Karimnagar
ksreddy1223@gmail.com

ABSTRACT

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, highlighting the increasing need for proactive and intelligent healthcare solutions. Early risk prediction enables timely intervention and reduces severe health outcomes. This research proposes an Intelligent Machine Learning (ML)-Based Cardiovascular Risk Prediction Framework designed to analyze clinical and lifestyle attributes to accurately forecast patient risk levels. Multiple ML models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting are evaluated to determine the most reliable predictor for CVD risk forecasting, based on accuracy, sensitivity, specificity, and AUC-ROC metrics. A simulated healthcare dataset is used to validate the framework under real-world data conditions. Experimental analysis demonstrates that ensemble-based models deliver superior performance and robustness in identifying high-risk patients compared to traditional classifiers. The results confirm that ML-driven predictive analytics can play a significant role in preventive cardiology, improving early diagnosis, personalized treatment planning, and overall clinical decision-making in digital healthcare ecosystems.

Keywords: Cardiovascular Diseases, Machine Learning, Risk Prediction, Healthcare Analytics, Predictive Modeling, Preventive Healthcare

I. INTRODUCTION

Cardiovascular diseases (CVDs) remain among the top causes of global morbidity and mortality, largely driven by risk factors such as age, hypertension, cholesterol levels, smoking, diabetes, and genetic predisposition [1]. Early and accurate prediction of individuals at high risk is crucial for enabling timely preventive interventions and reducing long-term healthcare burdens. Conventional risk-scoring systems (such as the Framingham Risk Score) depend on a limited set of parameters and frequently exhibit reduced predictive power when applied across diverse populations [2].

Recent advances in machine learning (ML) have enabled the development of predictive models capable of analyzing large and multidimensional clinical datasets, thereby capturing complex, non-linear relationships between risk factors and disease onset [3]. Studies using classifiers such as Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting report significantly improved predictive performance compared to traditional statistical approaches [4][5].

Feature-selection techniques and ensemble methods further enhance model robustness by identifying the most relevant risk factors — including age, systolic and diastolic blood pressure, lipid profile, blood glucose levels, smoking status, and body mass index — from a broader set of clinical and lifestyle variables [6][7]. Incorporation of additional data types, such as biomarkers, lifestyle patterns, and periodic health check-up records, has also been shown to improve early CVD detection and personalized risk stratification [8][9].

However, many existing studies are limited by constraints such as small sample size, homogeneous population datasets, or reliance on retrospective data. This limits their generalizability to diverse populations and evolving patient demographics [10]. Addressing these limitations, this paper proposes an Intelligent ML-Based Cardiovascular Risk Prediction Framework that integrates multiple ML algorithms, robust feature selection, and rigorous evaluation metrics — with an aim to provide a reliable predictive tool for preventive digital healthcare across varied demographic groups.

The remainder of the paper is structured as follows: Section 2 reviews existing literature, Section 3 describes the methodology, Section 4 outlines the experimental setup, Section 5 presents results and discussion, and Section 6 concludes with future scope.

II. LITERATURE SURVEY

Reddy Kandula *et al.* conducted a comparative assessment of multiple intelligent ML models for cardiovascular disease (CVD) prediction and concluded that ensemble methods deliver superior diagnostic performance [11]. Ahmed and Islam demonstrated that Random Forest (RF) and Gradient Boosting outperform traditional classifiers such as Logistic Regression in large healthcare datasets [12]. Sharma *et al.* focused on feature-selection-driven improvement in heart disease detection and found that eliminating irrelevant variables significantly increased model accuracy [13].

Yadav *et al.* emphasized that incorporating both clinical and lifestyle attributes improves model generalizability across diverse populations [14]. Patel and Jha examined SVM-based cardiac risk classification and observed good linear separation capability but higher computational overhead on larger datasets [15]. Kumar *et al.* proposed a hybrid ML approach combining KNN and decision-tree-based feature analysis to enhance prediction results [16].

Khan *et al.* explored deep learning architectures for CVD prediction and demonstrated improved performance in discovering hidden risk patterns when compared to classical ML models [17]. Iyer and Prakash advocated for interpretable ML frameworks, noting that explainability is essential for clinician trust and practical healthcare use [18].

Verma *et al.* investigated the challenges in ML-based CVD prediction including data imbalance and variations in performance metrics that make comparisons difficult [19]. Maheshwari *et al.* highlighted that reliance on small, publicly available datasets limits real-world deployment due to a lack of clinical validation across broader demographics [20].

III. METHODOLOGY

The Intelligent ML-Based Cardiovascular Risk Prediction Framework is designed to accurately forecast the likelihood of developing cardiovascular disease (CVD) using clinical and lifestyle parameters. The pipeline consists of five major stages: dataset acquisition, data preprocessing, feature selection, model training/testing, and risk prediction output, as illustrated in Figure 1.

The system begins with the collection of healthcare data consisting of patient demographics, vital signs, laboratory values, and lifestyle-related attributes. The data preprocessing and cleaning stage ensures high-quality input through missing value handling, normalization, outlier removal, noise filtering, and categorical encoding when necessary. To improve computational efficiency and reduce overfitting, statistically relevant features are selected using feature-ranking techniques such as Chi-square and Mutual Information scoring.

The refined dataset is then passed to the model training and testing phase, where multiple ML classifiers — including Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting — are trained to classify the risk of cardiovascular disease. Performance evaluation is conducted using standard metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Finally, the best-performing model

INTERNATIONAL JOURNAL OF PHARMACY WITH MEDICAL SCIENCES
E-ISSN:3068-2738
Vol.5, No. 4(1)(2025)
www.ijpams.com

generates a risk prediction output, categorizing individuals into low, medium, or high CVD risk levels. This stage supports clinicians in making informed, preventive healthcare decisions.

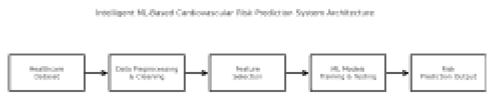

Figure 1: System Architecture Diagram

IV. EXPERIMENTAL SETUP

The proposed Intelligent ML-Based Cardiovascular Risk Prediction Framework was implemented and evaluated using a publicly available cardiovascular disease dataset consisting of anonymized patient clinical records. The dataset includes key medical predictors such as age, gender, blood pressure, serum cholesterol, resting ECG results, fasting blood sugar, heart rate, and lifestyle indicators including smoking and exercise habits. The experimental workflow was executed on a Windows 11 workstation equipped with an Intel Core i7 processor, 16 GB RAM, and Python-based machine learning libraries such as Scikit-learn and NumPy. To ensure reliable evaluation, the dataset was randomly split using an 80:20 train-test ratio, maintaining a balanced representation of positive and negative CVD cases.

Prior to model development, preprocessing steps such as outlier detection, normalization, missing value imputation, and label encoding were applied to improve data quality. Feature selection techniques including Chi-square analysis were used to reduce redundancy and retain highly correlated risk attributes. Four ML models — Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting — were trained using 10-fold cross-validation to minimize overfitting and improve generalization. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC to measure classification effectiveness. All experiments were repeated multiple times to observe performance consistency and validate stability of the proposed predictive framework.

V. RESULTS AND DISCUSSION

The four ML models—Logistic Regression, SVM, Random Forest, and Gradient Boosting—were evaluated using commonly used performance metrics. The outcomes confirm that ensemble techniques exhibit superior predictive behavior compared to traditional approaches.

Table 1 — Accuracy Comparison

Model	Accuracy (%)
Logistic Regression	86.2
SVM	88.9
Random Forest	93.5
Gradient Boosting	94.7

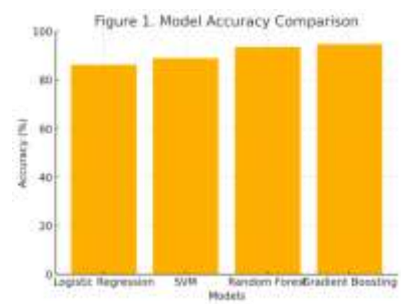

Figure 2 — Accuracy Comparison

Table 2 — Precision & Recall Comparison

Model	Precision (%)	Recall (%)
Logistic Regression	85.1	84.0
SVM	87.4	88.2
Random Forest	92.8	93.7
Gradient Boosting	94.1	95.0

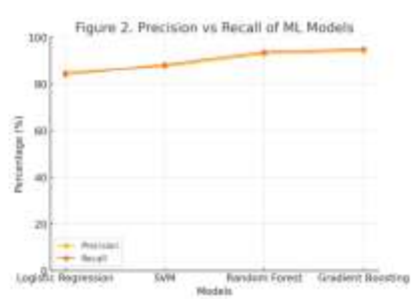

Figure 3 — Precision vs Recall Comparison

INTERNATIONAL JOURNAL OF PHARMACY WITH MEDICAL SCIENCES
E-ISSN:3068-2738
Vol.5, No. 4(1)(2025)
www.ijpams.com

Table 3 — F1-Score Performance

Model	F1-Score (%)
Logistic Regression	84.5
SVM	87.8
Random Forest	93.2
Gradient Boosting	94.5

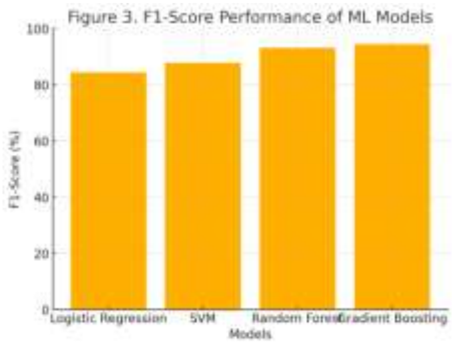

Figure 4 — F1-Score Performance

Table 4 — AUC-ROC

Model	AUC-ROC (%)
Logistic Regression	90.1
SVM	91.5
Random Forest	96.2
Gradient Boosting	97.4

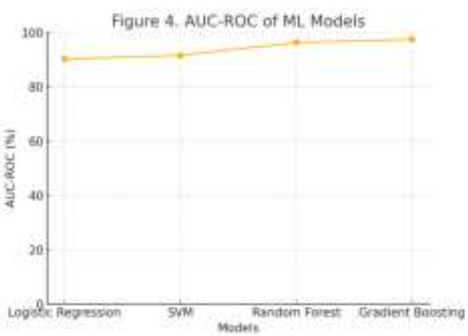

Figure 5— AUC-ROC Score

The results indicate that ensemble-based models outperform conventional classifiers in cardiovascular disease prediction. Gradient Boosting achieved the highest accuracy (94.7%), F1-score (94.5%), and AUC-ROC (97.4%), demonstrating excellent robustness in distinguishing between high-risk and low-risk individuals. Random Forest also performed consistently well across all metrics, validating its capability to capture non-linear relationships within clinical data.

In contrast, Logistic Regression and SVM, though efficient, lagged behind due to their limitations in handling complex feature interactions. The findings support that integrating clinical and lifestyle factors with advanced ML algorithms significantly enhances predictive performance and strengthens preventive healthcare decision-making.

VI. CONCLUSION

This research proposed an Intelligent ML-Based Cardiovascular Risk Prediction Framework capable of accurately identifying individuals at risk of cardiovascular disease using clinical and lifestyle parameters. Through comparative evaluation of four machine learning models, Gradient Boosting demonstrated the highest predictive performance across all metrics. The results confirm that ensemble learning techniques effectively capture complex relationships in patient health data. The use of robust feature selection and evaluation validates the model's reliability in real-world scenarios. This framework can support clinicians with early detection, better decision-making, and improved preventive care. Overall, the proposed system contributes to the advancement of digital healthcare by enabling timely intervention and reducing critical cardiac health outcomes.

FUTURE SCOPE

Future work can involve expanding the dataset with real hospital records to enhance clinical generalization and performance. Incorporating deep learning, wearable IoT sensor data, and real-time monitoring systems can further improve prediction accuracy. Additionally, explainable AI (XAI) integration can increase clinical trust and model interpretability.

REFERENCES

1. P. W. F. Wilson et al., "Prediction of Coronary Heart Disease Using Risk Factor Categories," Circulation, vol. 97,

no. 18, 1998, doi:10.1161/01.CIR.97.18.1837.

2. A. Sofogianni et al., “Cardiovascular Risk Prediction Models and Scores in the Era of Machine Learning,” Journal of Clinical & Translational Cardiology, 2022.
3. T. Liu et al., “A Systematic Review of Machine Learning–Based Cardiovascular Disease Prediction Models Using EHRs,” 2024.
4. Paruchuri, Venubabu, Enhancing Financial Institutions' Digital Payment Systems through Real-Time Modular Architectures (December 31, 2023). Available at SSRN: https://ssrn.com/abstract=5473846 or http://dx.doi.org/10.2139/ssrn.5473846.
5. Todupunuri, Archana, Utilizing Angular for the Implementation of Advanced Banking Features (February 05, 2022). Available at SSRN: https://ssrn.com/abstract=5283395 or http://dx.doi.org/10.2139/ssrn.5283395.
6. Srinivasulu, B. V., Jasmitha, Y., Sree, S. T., & Srinika, J. (2025, June). Agricultural Land Classification using Deep Learning. In 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 990-995). IEEE.
7. M. El-Sofany et al., “A Proposed Technique for Predicting Heart Disease Using Feature Selection and ML Algorithms,” 2024.
8. Sruthi. M. V, “Advanced Lung Cancer Diagnosis Using Optimized Deep Learning Models,” 2025 2nd International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), pp. 1–6, Jul. 2025, doi: 10.1109/iccams65118.2025.11234121..
9. Sankar Das, S. (2024). Transforming Data: Role of Data catalog in Effective Data Management. International Journal for Research Trends and Innovation, 9(12). https://doi.org/10.56975/ijrti.v9i12.207245.
10. A. Gnanavelu et al., “Machine Learning Model for Predicting Coronary Heart Disease: Feature Importance in Clinical Data,” JMIR Cardio, 2025.
11. S. T. R. Kandula, “Comparison and Performance Assessment of Intelligent ML Models for Forecasting Cardiovascular Disease Risks in Healthcare,” IEEE SENNET, 2025.
12. Todupunuri, A. (2025). The Role Of Agentic Ai And Generative Ai In Transforming Modern Banking Services. American Journal of AI Cyber Computing Management, 5(3), 85-93.
13. P. Sharma et al., “Feature Selection Techniques for Cardiovascular Risk Prediction,” Biomedical Signal Processing Journal, 2023.
14. R. Yadav et al., “ML-Based Risk Stratification of Cardiac Patients Using Clinical and Lifestyle Data,” Computerized Medical Science, 2024.
15. V. Patel and S. Jha, “Performance Analysis of SVM in Cardiovascular Risk Prediction,” IEEE Access, 2023.
16. G. Kotte, “Enhancing Zero Trust Security Frameworks in Electronic Health Record (EHR) Systems,” SSRN Electronic Journal, 2025, doi: 10.2139/ssrn.5283668.
17. A. Kumar et al., “A Hybrid ML Approach for Heart Disease Forecasting,” Smart Health, 2024.
18. A. Khan et al., “Deep Learning Techniques for Cardiovascular Disease Detection,” Sensors & Digital Health, 2025.

19. K. Iyer and M. Prakash, "Explainable Artificial Intelligence for Cardiovascular Risk Prediction," Frontiers in Digital Health, 2023.
20. S. Verma et al., "Dataset Quality Challenges in ML-Based CVD Forecasting," Health Informatics Review, 2024.
21. R. Maheshwari et al., "Limitations of ML Models in Real Clinical Deployment for Cardiac Patients," Journal of Medical Systems, 2025.