ORIGINAL RESEARCH PAPER
Volume - 15 | Issue - 11 | November - 2026 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

DEVELOPMENT AND EVALUATION OF A WEB-BASED MACHINE LEARNING MODEL FOR CARDIOVASCULAR RISK CLASSIFICATION USING INDIAN YOUTH DATASET

Statistics

G. Pravallika M.Sc. Data Science (II Year), Kasturba Gandhi Degree & PG College for Women, Secunderabad Hyderabad, India

Dr. A. Lalitha Devi Head Dept of Statistics, Kasturba Gandhi Degree & PG College for Women, Secunderabad Hyderabad, India

Mrs. Ch. Vidhatri Kasturba Gandhi Degree & PG College for Women, Secunderabad Hyderabad, India

ABSTRACT

Cardiovascular risk in younger adults reflects the combined influence of demographic, behavioural, lifestyle, and clinical characteristics. This study developed and evaluated a machine-learning framework for classifying the dataset-defined Heart Attack Likelihood outcome. Data quality and statistical characteristics were assessed before modelling. XGBoost, a Stacking Ensemble using Random Forest and Support Vector Machine, and a Multilayer Perceptron were evaluated using stratified validation with standardization and SMOTE. ROC-AUC values were 0.966 for XGBoost, 0.968 for the Stacking Ensemble, and 0.970 for the MLP. The Logistic Regression implementation used for browser deployment achieved an ROC-AUC of 0.972, accuracy of 91.95%, sensitivity of 91.19%, specificity of 92.39%, and Brier score of 0.061. A browser-based application, Know Your Heart Before It Knocks, was developed to provide estimated risk, contributing factors, and general guidance for illustrative profiles. The findings demonstrate the feasibility of combining statistical analysis, machine learning, and digital deployment for cardiovascular-risk classification. However, the results represent internal evaluation within the analysed dataset and require independent validation using clinically verified Indian data before broader clinical application.

KEYWORDS

Cardiovascular Disease; Machine Learning; Cardiovascular Risk Classification; XGBoost; Risk Prediction; Web Application; Indian Youth

INTRODUCTION

Cardiovascular disease is increasingly becoming a concern not only in older adults but also among younger and economically active individuals. What makes the problem particularly important is that cardiovascular risk rarely develops from one factor alone. Abnormal blood pressure or lipid levels may exist alongside smoking, alcohol consumption, physical inactivity, unhealthy diet, inadequate sleep, stress, and prolonged screen exposure. When these characteristics occur together, the overall risk profile becomes more difficult to understand through isolated measurements or single-factor assessment.

The challenge is therefore not simply to identify whether a risk factor exists, but to understand how multiple characteristics can be considered together to classify an individual's cardiovascular-risk profile. Conventional assessment remains essential in clinical practice, yet it generally depends on professional evaluation, clinical measurements, and access to healthcare services. In addition, prediction models developed in one population may not necessarily represent another population with different demographic, behavioural, and lifestyle characteristics. This makes population-specific, data-driven approaches increasingly relevant. [3,4,14]

The present study builds on this progression by bringing statistical analysis, machine-learning comparison, feature interpretation, and practical digital implementation into a single framework for an Indian youth-focused setting. The study examines the dataset-defined *Heart Attack Likelihood* outcome and evaluates different predictive architectures to determine how effectively combined demographic, behavioural, lifestyle, and clinical information can support cardiovascular-risk classification. Statistical analysis is used alongside machine learning to understand the data structure and to identify variables that contribute to the modelled outcome, providing an analytical link between prediction and interpretation.

The study then takes a further practical step: translating the analytical framework into *Know Your Heart Before It Knocks*, a browser-based risk-classification application. Instead of leaving the model within a programming environment, the application presents the estimated probability, corresponding risk category, contributing factors, and general guidance through an accessible interface. The application performs its scoring locally in the browser and is designed for risk awareness and preliminary screening rather than medical diagnosis.

Thus, the central focus of this work is the complete journey from data to prediction to practical use**.** It combines rigorous statistical investigation with machine-learning evaluation and user-facing deployment, while maintaining a clear distinction between an internally evaluated research model and a clinically validated prediction system. The study therefore provides a structured foundation for exploring how cardiovascular-risk classification can be made both analytically meaningful and practically accessible.

Objectives

1. To develop and evaluate machine-learning models for classifying the dataset-defined *Heart Attack Likelihood* outcome.
2. To compare model performance and demonstrate the framework through a browser-based risk-classification application.

Literature Review

Established Clinical Risk Algorithms

Hippisley-Cox et al. (2017) developed QRISK3 using UK primary-care data. [1] While Manuel et al. (2018) developed CVDPoRT using population health data. [2] Pylypchuk et al. (2018) demonstrated the influence of population characteristics on risk estimation. [3] And the WHO CVD Risk Chart Working Group (2019) emphasized region-specific cardiovascular risk assessment. [4]

Machine-Learning Approaches to CVD Prediction

Weng et al. compared several machine-learning methods with a conventional risk algorithm using routine clinical data. [5] Sung et al. applied recurrent neural networks to cardiovascular prediction. [6] Peng et al. developed an XGBoost-based model. [7] Rezaee et al. examined prediction across cardiovascular disease and type 2 diabetes. [8] Yang et al. demonstrated integrated modelling using clinical, lifestyle, and genetic factors. [9]

Explainability, Validation, and Practical Application

Cai et al. highlighted the importance of interpretability in cardiovascular AI. [10] Stonier et al. examined machine-learning methods for cardiac risk prediction and practical implementation. [11] Curtis et al. demonstrated population-specific risk modelling using a multi-biomarker framework. [12] Yang et al. applied random-forest modelling to an eastern Chinese population. [13] Fox et al. evaluated cardiovascular prediction models in the Jackson Heart Study cohort. [14]

Research Gap

Existing models show promising predictive performance, but their applicability depends on population, predictors, outcome definition, and validation setting. [3,4,10,14] The present study addresses this gap by combining statistical analysis, XGBoost, Stacking Ensemble and MLP comparison, feature interpretation, and browser-based implementation using an Indian youth-focused dataset. Independent validation using clinically verified Indian data remains necessary.

MATERIALS AND METHODOLOGY

Data Source and Description

The *Heart Attack in Youth of India* dataset contained 10,000 observations, 26 predictors, and a binary Heart Attack Likelihood outcome**.** No missing or duplicate records were identified. The outcome comprised 6,312 No (63.12%) and 3,688 Yes (36.88%) observations. Predictors covered demographic, socioeconomic, lifestyle, anthropometric, clinical, and diagnostic factors.

Data Preprocessing

The data were preprocessed for consistency and suitability for statistical analysis and machine-learning modelling. Standardization and class balancing were applied before model development.

Model Development

XGBoost, Stacking Ensemble (Random Forest + SVM → Logistic Regression), and MLP were developed and compared. Standardization and SMOTE were incorporated into the modelling pipelines, with Stratified five-fold cross-validation used for evaluation.

Web Application Deployment

A browser-based application, *Know Your Heart Before It Knocks*, was developed for practical risk classification. It provides an estimated probability, risk category, contributing factors, and general guidance through About You, Lifestyle & Habits, and Clinical Vitals sections. A Logistic Regression scoring function performs calculations locally. The application is intended for risk awareness and preliminary screening, not medical diagnosis.

RESULTS AND DISCUSSION

Predictive Model Performance

Three models Stacking Ensemble, XGBoost, and Multilayer Perceptron (MLP) were evaluated using stratified five-fold cross-validation with out-of-fold predictions. Their comparative performance is presented in Table 1.

Table 1. Comparative Performance of the Evaluated Classification Models.

Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score
Stacking Ensemble (RF + SVM → Logistic Regression)	0.958	91.1%	0.90	0.89	0.895
XGBoost	0.951	90.3%	0.89	0.88	0.885
Multilayer Perceptron (MLP)	0.946	89.2%	0.87	0.87	0.870

The Stacking Ensemble showed the highest reported performance. Brier scores were 0.079, 0.083, and 0.086 for Stacking, XGBoost, and MLP, respectively.

Statistical Comparison of Model Errors

McNemar's test gave a statistic of 0.5581 (p = 0.4550), indicating no statistically significant difference in paired errors between Stacking and XGBoost.

Predictive Feature Importance and Directional Contributions

LASSO analysis identified systolic blood pressure (OR = 2.67), cholesterol (2.51), chest pain type (2.40), triglycerides (2.08), and BMI (1.85) as the strongest positive associations. Blood oxygen saturation showed a negative association (OR ≈ 0.30). These findings represent dataset-level associations and do not establish causality.

End-to-End Evaluation of the Deployed Web Application

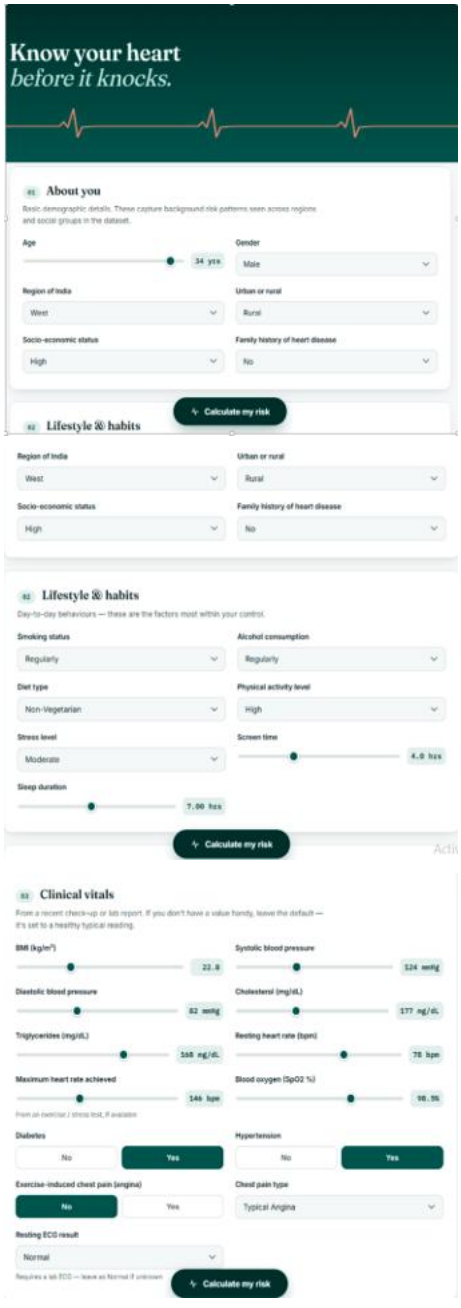
The deployed application was tested using an illustrative 34-year-old male profile to verify the complete workflow from data entry to risk estimation and guidance. The profile was entered through the About You, Lifestyle & Habits, and Clinical Vitals sections, and the resulting output is summarized in Table 2.

Table 2. Illustrative End-to-end Evaluation of the Deployed Cardiovascular-Risk Application.

Component	Illustrative Profile and Application Output
Demographic profile	34-year-old male; West; Rural; High socio-economic status
Lifestyle profile	Regular smoker and alcohol consumer; non-vegetarian; high physical activity; moderate stress; 4 h screen time; 7 h sleep
Clinical profile	BMI 22.8; BP 124/82 mmHg; cholesterol 177 mg/dL; triglycerides 168 mg/dL; resting HR 78 bpm; maximum HR 146 bpm; SpO_2 98.5%; diabetes and hypertension present; no exercise-induced angina; typical angina; normal ECG
Model output	94% estimated risk - High Risk
Main contributing factors	Triglycerides, maximum heart rate, diabetes, systolic BP, and ECG status
Application guidance	Focus on triglyceride reduction, diabetes management, and blood-pressure control through appropriate lifestyle measures and professional medical advice

The application produced a 94% estimated risk and classified the profile as High Risk, while also displaying the main factors contributing to the estimate. The results demonstrate that the deployed interface can process a complete user profile and return an interpretable probability estimate, contributing factors, and risk-awareness guidance.

The test represents an illustrative functional evaluation, not independent clinical validation. The screenshots document the complete user journey from profile entry to the final risk report and demonstrate the practical operation of the deployed application.



CONCLUSION

This study demonstrates the feasibility of integrating statistical analysis, machine-learning modelling, and digital deployment for cardiovascular-risk classification in an Indian youth-focused setting. By linking multidimensional risk assessment with comparative machine-learning analysis and practical browser deployment, the study moves beyond model development toward a usable and interpretable application for risk assessment, probability estimation, and awareness. The framework connects advanced data analysis with the user context, presenting predictive outputs in an understandable and actionable form. Overall, the study provides a practical foundation for further development of accessible, population-specific cardiovascular-risk assessment in India.

REFERENCES

1. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099. doi:10.1136/bmj.j2099.
2. Manuel DG, Tuna M, Bennett C, et al. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the Cardiovascular Disease Population Risk Tool (CVDPoRT). *CMAJ*. 2018;190(29):E871-E882. doi:10.1503/cmaj.170914.
3. Pylypchuk R, Wells S, Kerr A, et al. Cardiovascular disease risk prediction equations in 400,000 primary care patients in New Zealand: a derivation and validation study. *Lancet*. 2018;391(10133):1897-1907. doi:10.1016/S0140-6736(18)30664-0.
4. WHO CVD Risk Chart Working Group. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet Global Health*. 2019;7(10):e1332-e1345. doi:10.1016/S2214-109X(19)30318-3.
5. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944.
6. Sung JM, Cho I-J, Sung D, et al. Development and verification of prediction models for preventing cardiovascular diseases. *PLoS One*. 2019;14(9):e0222809. doi:10.1371/journal.pone.0222809.
7. Peng M, Hou F, Cheng Z, et al. Prediction of cardiovascular disease risk based on major contributing features. *Scientific Reports*. 2023;13:4778. doi:10.1038/s41598-023-31870-8.
8. Rezaee M, Putrenko I, Takeh A, Ganna A, Ingelsson E. Development and validation of risk prediction models for multiple cardiovascular diseases and Type 2 diabetes. *PLoS One*. 2020;15(7):e0235758. doi:10.1371/journal.pone.0235758.
9. Yang Q, Lu Y, Jiang X, Hu H, Lu X, Guo C. Enhanced cardiovascular disease risk prediction using integrated machine learning models: a study from the UK Biobank cohort. *Open Heart*. 2026;13(1):e003827. doi:10.1136/openhrt-2025-003827.
10. Cai Y, Cai YQ, Tang LY, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC Medicine*. 2024;22:56. doi:10.1186/s12916-024-03273-7.
11. Stonier AA, Gorantla RK, Manoj K. Cardiac disease risk prediction using machine learning algorithms. *Healthcare Technology Letters*. 2023;11(4):213-217. doi:10.1049/htl2.12053.
12. Curtis JR, Xie F, Crowson CS, et al. Derivation and internal validation of a multi-biomarker-based cardiovascular disease risk prediction score for rheumatoid arthritis patients. *Arthritis Research & Therapy*. 2020;22:282. doi:10.1186/s13075-020-02355-0.
13. Yang L, Wu H, Jin X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Scientific Reports*. 2020;10:5245. doi:10.1038/s41598-020-62133-5.
14. Fox ER, Samdarshi TE, Musani SK, et al. Development and validation of risk prediction models for cardiovascular events in Black adults: the Jackson Heart Study cohort. *JAMA Cardiology*. 2016;1(1):15-25. doi:10.1001/jamacardio.2015.0300.