International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 14 Special Issue 03 March 2026, Page no. – 32-35

Impact Factor: 8.751

DOI: 10.47191/ijmcr/v14iSPC3.07

Heart Risk Analysis via Health Factors

Neeta Namdeo Takawale

Department of Computer Science, Dr. D. Y. Patil, Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

ARTICLE INFO

Published Online:
14 March 2026

ABSTRACT

Cardiovascular diseases (CVDs) represent a leading cause of mortality on a global scale, posing significant challenges to public health and sustainable development (SDG). Both lifestyle behaviors and physiological parameters critically determine an individual's risk of developing cardiovascular conditions. This study utilizes a data-driven methodology to examine the impact of lifestyle factors, including smoking, alcohol consumption, physical activity, and body mass index (BMI), on cardiovascular health. By employing the publicly accessible Cardiovascular Disease Dataset from Kaggle, various machine learning algorithms, such as Logistic Regression, Random Forest, and Support Vector Machine, are implemented to predict cardiovascular risk based on health and behavioral indicators. The findings identify BMI, cholesterol level, and physical activity as the most significant predictors. This research demonstrates the potential of data science techniques in facilitating early risk detection, supporting preventive healthcare strategies, and informing evidence-based interventions aligned with SDG Good Health and Well-Being. By promoting healthier lifestyles and reducing the burden of non-communicable diseases, the study also contributes to SDG Reduced Inequalities through equitable health risk assessment and SDG Industry, Innovation, and Infrastructure by leveraging advanced data analytics for public health improvement.

Corresponding Author:
Neeta Namdeo Takawale

KEYWORDS: CVD, Machine Learning, Lifestyle Risk Factors, Health factor, SDG

INTRODUCTION

Cardiovascular disease (CVD) remains a major global health challenge, contributing to illness, death, and economic strain, affecting SDG Good Health and Well-Being. Despite medical progress, rising CVD cases highlight the need for preventive strategies focused on lifestyle factors like smoking, alcohol, diet, obesity, and physical inactivity. Addressing these determinants through population-level interventions can promote cardiovascular health and support SDG Reduced Inequalities. Research shows strong links between behavioral and clinical factors in CVD risk, with unhealthy habits contributing to lipid imbalance, inflammation, and cardiovascular damage.

Studies indicate targeting multiple lifestyle behaviors simultaneously is more effective than single factor interventions. Data science and machine learning have enhanced CVD prediction by analyzing lifestyle-clinical interactions beyond traditional methods. While previous studies emphasized lifestyle modification, limited research has explained how combined behaviors influence CVD risk. Advanced computational techniques enable better risk prediction and prevention, supporting SDG goals.

This study evaluated the combined impact of lifestyle behaviors and clinical indicators on cardiovascular risk to identify high-risk individuals and key factors. The findings support evidence-based interventions for CVD prevention. Using machine learning on the Kaggle CVD dataset, this study uncovered complex patterns in health factors to improve risk prediction and strengthen prevention strategies aligned with SDG goals.

RELATED WORK

Thayssa et al. (2025), Ezika et al. (2024), Jafari et al. (2023), Ghodeshwar et al. (2023) and Bhosale (2019) demonstrate that multiple healthy habits reduce CVD risk and mortality [1] [2] [3] [4] [5].

Touri (2005), Landini (2014), Agwara et al. (2024), Shibli et al. (2023), and Jung et al. (2020) show poor lifestyle habits increase CVD risk, while healthy practices reduce metabolic abnormalities. They emphasize early interventions to improve heart health [6] [7] [8] [9] [10].

Badhan (2025) demonstrated machine learning models achieve 95-100% accuracy in CVD prediction using multi-source data fusion [11].

Shen et al. (2024) found self-management, physical activity, medication adherence, anxiety, and sexual health factors influence CHD prevention [12].

Yousefzai et al. (2024) found that hormonal, metabolic, and psychosocial changes during menopause increase women's cardiovascular risk. They recommend patient-centered strategies to reduce inequalities [13].

Yuan et al. (2023) reported that higher sodium intake increases blood pressure and hypertension risk, with moderate intake of 3,000 mg/day being most beneficial [14].

Tamaki et al. (2019) found 30% of heart failure patients accounted for 80% of medical costs due to preventable surgeries, with creatinine levels and smoking as key predictors [15].

METHODOLOGY

- **Dataset Description**

This study applied machine learning techniques to predict cardiovascular disease (CVD) using a publicly available Kaggle dataset containing health-related features, including age, sex, BMI, blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity. The target variable was binary, representing whether a person had cardiovascular disease (1) or not (0).

Four machine learning models were employed for classification.

1. **Logistic Regression**, a linear model estimating the probability of CVD based on input features.
2. **Random Forest** is an ensemble of decision trees used to capture complex interactions among variables.
3. **Support Vector Machine (SVM)** was used to handle high-dimensional data and identify optimal separation boundaries between classes.
4. **XGBoost** is a gradient boosting algorithm that iteratively improves prediction accuracy and reduces errors.

- **Data Preprocessing and Preparation**

Before model development, several preprocessing steps were performed to ensure the data quality, consistency, and suitability for machine learning analysis. These steps helped remove noise, standardize the values, and enhance the learning efficiency of the model.

1. **Data Cleaning**

First, the dataset was examined for inconsistencies and errors. Duplicate records were removed to avoid biased learning. Unusual or extreme values that did not fall within a realistic physiological range were treated as outliers and were eliminated or corrected. This ensured that the model was trained on reliable data that represented real-world health patterns.

2. **Feature Engineering**

Some features were refined to create more meaningful inputs for the models. For example, height and weight were combined to compute the Body Mass Index (BMI), a clinically relevant indicator of body composition and a known risk factor for cardiovascular disease. BMI provided a more informative feature than height or weight.

3. **Encoding Categorical Variables**

Certain attributes, such as smoking, alcohol consumption, and physical activity, were originally recorded in the categorical form (e.g., yes/no or levels). These were converted into numerical labels to make them usable by the machine learning algorithms. For instance, non-smokers were assigned a value of 0, occasional smokers 1, and regular smokers 2. Similar encoding was applied to the alcohol intake and activity levels.

4. **Data Normalization / Scaling**

To ensure that features with larger ranges (e.g., age or blood pressure) did not dominate those with smaller numeric values, all the continuous variables were scaled using standard normalization techniques. Standardization transformed the data into a uniform format, helping to improve the algorithm performance and convergence speed, especially for distance-based models such as SVM or gradient-based models such as XGBoost.

- **Dataset Splitting**

Once the data were cleaned and transformed, they were divided into two subsets: 80% for training the machine learning models and 20% for testing the final model performance. This split ensured that the model learning and evaluation were performed fairly. The training set allowed the algorithms to learn patterns from the data, whereas the test set allowed for an unbiased assessment of how well the models could generalize to unseen data.

- **Statistical Methods**

1. Descriptive statistics, including the mean, standard deviation, and range, were computed for all numerical features.
2. Feature importance scores were computed for tree-based models (Random Forest, XGBoost) to quantify the contribution of each variable to CVD prediction.

The Figure 1 shows the workflow of data preparation,

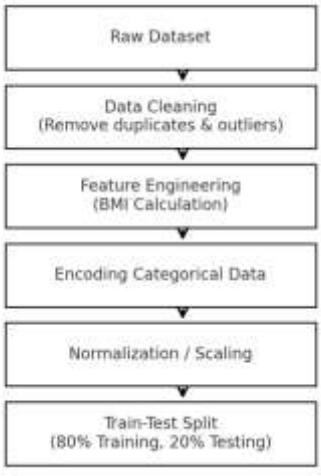

Figure 1: Data Preparation Workflow

RESULTS AND ANALYSIS

Four machine learning models were tested to predict cardiovascular disease risk. Among them, **XGBoost achieved the highest accuracy (0.81) and ROC-AUC (0.86)**, demonstrating strong predictive capability. **Random Forest** also performed well, with an accuracy of 0.78 and a recall of 0.79, indicating most CVD cases were correctly identified. **SVM and Logistic Regression** performed moderately; Logistic Regression was the most interpretable but had lower accuracy (0.71), while SVM performed slightly better than Logistic Regression but below the tree-based models.

Feature	*Important Score*
BMI	*0.32*
Cholesterol Level	*0.28*
Blood Pressure	*0.22*
Physical Activity	*0.15*
Smoking	*0.03*
Alcohol Consumption	*0.02*

Table 1 indicated that **BMI, cholesterol level, blood pressure, and physical activity** were the strongest predictors of cardiovascular disease. **Smoking and alcohol consumption** had a moderate influence, affecting CVD risk less strongly than clinical indicators.

Incorporating lifestyle factors alongside medical features improved model accuracy by approximately 10%, highlighting the significant role of behavior in predicting cardiovascular risk.

DISCUSSION

The findings of this study indicate that lifestyle behaviors play a substantial role in shaping cardiovascular disease (CVD) risk, reinforcing the well-established link between modifiable habits and long-term health. The prominence of factors such as BMI, cholesterol level, and physical activity aligns with previous research, which has consistently identified these parameters as strong predictors of cardiovascular morbidity. Previous studies have similarly reported that unhealthy behaviors, including sedentary lifestyles and poor dietary patterns, contribute to metabolic changes that increase CVD susceptibility. The outcomes of the present study are therefore consistent with existing evidence emphasizing that behavioral modification is pivotal in CVD prevention.

Incorporating machine learning techniques, particularly tree-based models, has demonstrated the advantage of using advanced analytics to capture non-linear and multifactorial relationships among health indicators. This aligns with recent research highlighting the superior performance of ensemble and boosting algorithms over traditional linear models in medical risk prediction. By integrating both lifestyle and clinical variables, this study contributes to the emerging literature advocating for multidimensional risk assessment rather than relying on clinical indicators alone.

The implications of these findings extend to both clinical practice and public health policies. A more holistic understanding of CVD risk can support personalized prevention strategies tailored to individual behavioral and metabolic profiles, replacing general recommendations with targeted interventions. This approach has the potential to enhance early detection, improve patient engagement in lifestyle changes, and optimize healthcare resources by prioritizing high-risk individuals. Furthermore, the adoption of data-driven predictive models can strengthen preventive healthcare frameworks and inform policy decisions aimed at reducing the global burden of cardiovascular disease.

CONCLUSION

This study demonstrates that lifestyle behaviors play a critical role in assessing cardiovascular disease (CVD) risk, highlighting the importance of preventive health strategies aligned with SDG 3 Good Health and Well-Being. Using the Cardiovascular Disease Dataset, machine learning models, particularly XGBoost, achieve high predictive accuracy, identifying BMI, cholesterol, and physical activity as key determinants of cardiovascular risk. By emphasizing early detection and personalized risk assessment, this research supports equitable access to preventive healthcare, contributing to SDG Reduced Inequalities. Furthermore, the application of advanced machine learning techniques illustrates how innovative data-driven solutions promote sustainable healthcare infrastructure, consistent with SDG Industry, Innovation, and Infrastructure. Future studies may expand these models by integrating wearable sensor data, diet tracking, and psychological parameters, enabling more comprehensive prediction and intervention strategies to reduce CVD burden and advance global health outcomes.

ACKNOWLEDGEMENT

The author would like to express her sincere gratitude to all individuals and organizations who supported this research. I also thank our colleagues and mentors for their valuable guidance, technical assistance, and constructive feedback throughout the study. I am especially thankful to the Kaggle dataset developers for providing valuable data for analysis.

The author also acknowledges the institutional support and resources that enabled the successful completion of this study.

FUNDING SOURCE: None

AUTHORS' CONTRIBUTIONS

The author solely contributed to all aspects of this research work, including conceptualizing the research problem, designing the study framework, conducting data preprocessing and model development, performing experimental analysis, interpreting the results, and preparing the manuscript. The author reviewed, refined, and approved the final version of the manuscript prior to submission. Overall, the author was responsible for the complete execution and reporting of the study.

CONFLICT OF INTEREST: The authors declare that she has no financial or commercial interests that might be perceived as influencing the results or conclusions of this study.

DATA AVAILABILITY

All data were obtained from publicly available sources and are governed by Kaggle's terms of use as outlined by the original dataset contributors. No special permissions or restrictions apply. The author is available to assist with dataset access or usage inquiries, if needed.

REFERENCES

1. M. M. Thayssa Thayssa Vitoria Oliveira Sousa Holanda, "Impact of Lifestyle on the Incidence of Heart Disease," International healthcare review, 2025.
2. A. N.-E. M. C. R. E. Ejiofor Augustine Ezika, "Key Risk-Factors Contributing to Cardiovascular Disease: Lifestyle Modifications to Improve Cardiovascular Health," International journal of health & medical research, 2024.
3. S. A. H. M. Negar Jafari, "How lifestyle factors can contribute to cardiovascular disease incidence; a review study," Journal of preventive epidemiology, 2023.
4. Gunjan K Ghodeshwar, "Impact of Lifestyle Modifications on Cardiovascular Health: A Narrative Review," Cureus, vol. 15, 2023.
5. S. J. Bhosale, "The Role of Lifestyle in Development of Coronary Heart Disease," Journal of the Indian Academy of Applied Psychology, vol. 41, no. 3, 2019.
6. C. R. H. L. M. M Akbar Tabar Touri, "Healthy lifestyle behaviours are associated with lower probability of having cardiovascular disease risk factors," Iranian Journal of Public Health, 2025.
7. L. Landini, "Modification of Lifestyle Factors are Needed to Improve the Metabolic Health of Patients with Cardiovascular Disease Risk.," Current Pharmaceutical Design, 2014.
8. Ebiambu Ondoh Agwara, "Predictors of cardiovascular disease risk and total mortality: Findings from the UK Biobank.," Circulation , 2024.
9. Khalid Yahya Shibli, "Influence of Lifestyle Changes on Cardiovascular Diseases in Saudi Arabia: A Systematic Literature Review," Cureus, vol. 15, 2023.
10. E. G. J. C. S. D. Joanna Jung, "Lifetime risks factors and assessment of cardiovascular disease," AME Publishing Company), vol. 5, 2020.
11. P. K. Badhan, "Examining Personal Lifestyle Factors and Lipid Profile on Cardiovascular Disease Risk Prediction," Advances in computational intelligence and robotics book series, pp. 101-130, 2025.
12. W. Y. Z. Y. Y. N. Y. J. O. X. H. P. A. Shen Q, "Cross-Sectional Study of Risk Factors for Coronary Heart Disease in Secondary Prevention for Patients With the Disease in China," 2025.
13. A. Z. F. H. A. M. S. S. F. M. K. A. M. F. S. G. J. Z. Yousefzai S, "Cardiovascular Health During Menopause Transition: The Role of Traditional and Nontraditional Risk Factors. Methodist Debakey Cardiovasc," 2025.
14. Y. D. W. Y. Q. M. L. K. L. Z. G. D. N. N. Yuan M, "Sodium intake and the risk of heart failure and hypertension: epidemiological and Mendelian randomization analysis.," 2024.
15. K. K. W. H. S. T. M. M. Tamaki Y, "Characteristics of heart failure patients incurring high medical costs via matching specific health examination results and medical claim data: a cross-sectional study," 2019.