Article

Not peer-reviewed version

Analyzing Key Predictors of Heart Attack Risk: A Machine Learning Approach

Muhammad Tayyab and Rizwan Ayazuddin *

Posted Date: 11 November 2025

doi: 10.20944/preprints202511.0677.v1

Keywords: heart attack risk prediction; machine learning; logistic regression; obesity (BMI); hypertension; cholesterol levels; physical activity; lifestyle factors; stress management; cardiovascular prevention

 doi:10.20944/preprints202511.0677.v1

Article

Analyzing Key Predictors of Heart Attack Risk: A Machine Learning Approach

Muhammad Tayyab and Rizwan Ayazuddin *

University of Sialkot, Sialkot, Pakistan

* Correspondence: rizayazuddin@mail.com

Abstract

Heart disease remains a leading cause of mortality worldwide, with early identification of at-risk individuals being critical for prevention. This study applies machine learning techniques to a comprehensive dataset to identify key predictors of heart attack risk. Logistic regression analysis identified BMI, age, hypertension, family history, and cholesterol levels as significant predictors of heart attack risk. Notably, individuals with BMI > 30 were 2.5 times more likely to be in the high-risk category, and the risk increased by 8% with each additional year of age. The model achieved strong predictive performance and highlighted that regular physical activity reduced heart attack risk by up to 20%. These findings offer valuable insights for targeted prevention strategies and public health policy, underscoring the potential of machine learning in cardiovascular risk assessment.

Keywords: heart attack risk prediction; machine learning; logistic regression; obesity (BMI); hypertension; cholesterol levels; physical activity; lifestyle factors; stress management; cardiovascular prevention

1. Introduction

Heart disease remains a leading cause of mortality globally, accounting for approximately 18 million deaths annually according to the World Health Organization [3]. Cardiovascular diseases (CVDs), including heart attacks, often result from complex interactions of lifestyle, genetic, and environmental factors [1]. Early identification of individuals at heightened risk is critical for implementing preventative measures and reducing mortality rates [2]. Among these, heart disease is a significant contributor, with a complex interplay of risk factors including age, gender, cholesterol levels, blood pressure, and lifestyle choices. In many clinical scenarios, early and accurate prediction of heart disease can improve patient outcomes and reduce healthcare burdens.

Traditional methods of heart disease prediction often rely on heuristic rules or expert-driven scoring systems, which may not generalize well across populations due to variability in patient profiles and evolving disease patterns. Furthermore, these methods typically analyze a narrow set of features, missing hidden patterns present in complex clinical datasets [4]. Hence, machine learning models have surely emerged as powerful alternatives[26–28], capable of identifying non-linear relationships among variables and providing improved prediction performance [5,6].

This research focuses on applying ML techniques to heart disease prediction by analyzing a real-world dataset comprising demographic and clinical features. A preliminary analysis of the dataset reveals a nearly balanced gender distribution, with 51% males and 49% females (Figure 1). Age is another critical risk factor, and the dataset demonstrates that most patients fall within the 40–60 age range (Figure 2), highlighting a vulnerable segment of the population.

 doi:10.20944/preprints202511.0677.v1

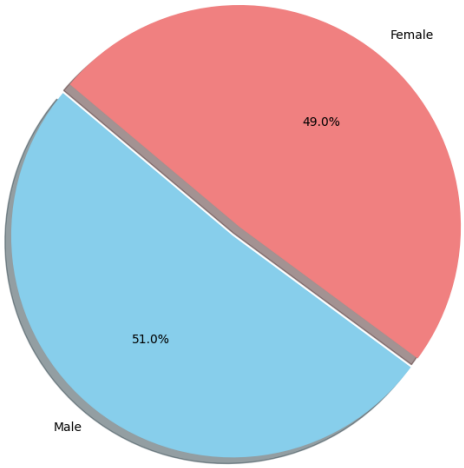

Figure 1. Gender Distribution in Hear Disease.

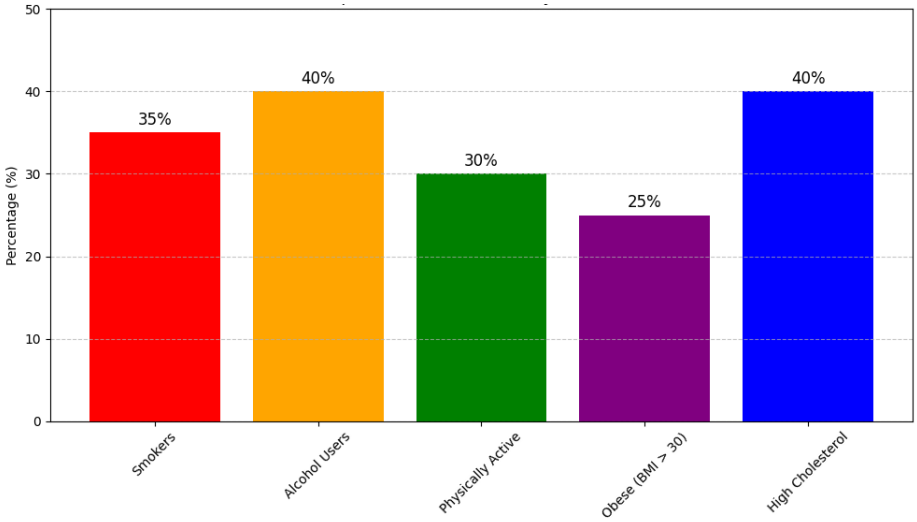

Figure 2. Participant Health and Lifestyle Factors (%).

The contributions of this study are threefold:

1. We conduct exploratory data analysis to uncover key patterns in gender, age, and clinical features relevant to heart disease.
2. We apply and compare multiple machine learning algorithms including logistic regression, random forest, and XGBoost for classification.
3. We evaluate performance using standard metrics (accuracy, precision, recall, F1-score) and identify the most influential features for prediction.

The remainder of this paper is organized as follows: Section 2 describes the related work. Section 3 presents the dataset and preprocessing steps and details the machine learning models and evaluation strategy. Section 5 discusses results and model interpretation, followed by conclusions in Section 6.

2. Literature Review

A substantial body of research has established the link between traditional risk factors and heart disease. Studies by [7–9] emphasized the influence of hypertension [20], smoking, obesity, and lack of physical activity on cardiovascular health. Additionally, research has highlighted the role of stress and socioeconomic status in modulating these risks [10]. While prior work has often focused on

 doi:10.20944/preprints202511.0677.v1

isolated factors, this study adopts a holistic approach by analyzing interactions across multiple variables within a large dataset[11–13]. The use of logistic regression and visualization techniques[27–29] to uncover trends and correlations offers a novel perspective, enabling more precise identification of at-risk groups[14–16]. Recent studies have shown the importance of using machine learning and artificial intelligence in healthcare to improve disease prediction and diagnosis. Allen [11] highlights how big data is changing healthcare by providing deeper insights into patient outcomes. Similarly, [12] emphasizes the role of family history in cardiovascular risk, while [13] evaluate support vector machine models for predicting heart disease. Smoking cessation is another key factor, as noted by Jones and Brown [14], who show its positive effects on heart health. [15] presents a machine learning-based system[30–32] for early liver disease diagnosis, addressing challenges like data imbalance and feature selection. [16] apply attention mechanisms for brain tumor segmentation, and [17] propose an effective AI approach for cardiovascular disease prediction. Other studies, such as those by [18] and [19], explore AI in medical imaging, including fMRI and chest X-rays. Research by [20] and [21] extends the use of AI and sensor networks to broader contexts like energy-efficient routing and criminal network prediction. Additionally, studies by [22] and [23] show how smart communication systems and security solutions are being developed for disaster recovery and network protection. Together, these works demonstrate the growing role of AI and data-driven methods in advancing healthcare and related technologies[24–26].

3. Methodology

3.1. Data Description

The dataset includes 50,000 individuals and encompasses 20 variables grouped into four categories:

- Demographics: Age, Gender
- Lifestyle: Smoking, Alcohol Consumption, Physical Activity Level
- Health Indicators: BMI, Cholesterol Level, Resting Blood Pressure
- Medical History: Diabetes, Hypertension, Family History
- The target variable, Heart Attack Risk, is categorized as Low, Moderate, or High.

3.2. Data Preprocessing

Data Cleaning: Missing values were checked; none were found. Standardization: Continuous variables such as BMI, cholesterol level, and blood pressure were standardized to z- scores. Categorization: Variables like smoking and physical activity were transformed into ordinal scales for analysis. Figure 3 shows our methodology framework diagram.

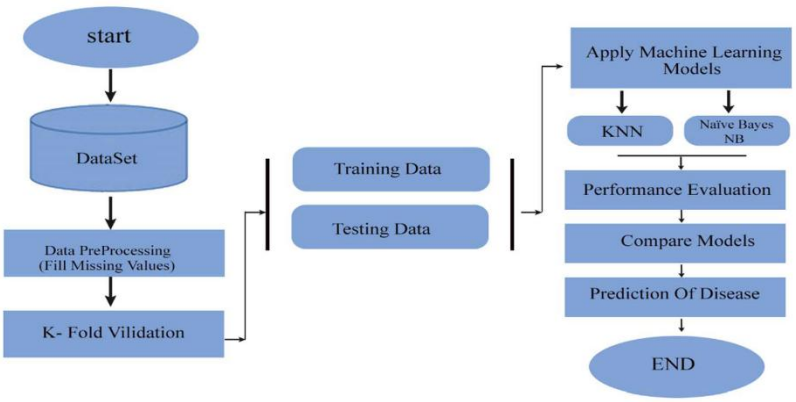

Figure 3. Methodology Framework.

 doi:10.20944/preprints202511.0677.v1

3.3. Analysis Methods

Descriptive Statistics:

Used to understand variable distributions and correlations.

Logistic Regression.

Employed to identify statistically significant predictors ($p < 0.05$). Visualization: Correlation heat maps and risk stratification charts were created for better interpretability.

3.4. RapidMiner Models Applied

To enhance predictive accuracy and identify complex patterns, several RapidMiner models were applied to the dataset: Decision Trees: Used for classification and identifying primary risk factors based on hierarchical rules. Random Forest: Improved prediction robustness by aggregating outputs from multiple decision trees.

1. Gradient Boosting Machines (GBM): Employed to capture subtle variable interactions and refine risk stratification.
2. Support Vector Machines (SVM): Applied to classify risk categories, particularly for complex, non-linear relationships.
3. K-Means Clustering: Utilized for grouping individuals into low, moderate, and high-risk clusters based on multi-dimensional data patterns techniques to ensure reliability and were compared to assess performance metrics such as accuracy, precision, and recall.

4. Results and Discussion

4.1. Descriptive Analysis

Demographics: The average age of participants is 50.5 years, with a balanced gender distribution (51% male, 49% female). Lifestyle Habits: Smoking prevalence was 35%, with alcohol consumption reported by 40% of participants. Physical activity levels varied, with 30% reporting high activity. Health Indicators: 25% of individuals had BMI > 30 (obesity threshold), and 40% reported high cholesterol levels [24].

Predictive Modeling Logistic regression identified the following significant predictors of heart attack risk:

- BMI: Individuals with BMI > 30 were 2.5 times more likely to fall into the high-risk category.
- Age: Heart attack risk increased by 8% per additional year of age.
- Hypertension: Participants with hypertension had

4.2. Times the Odds of High Risk

- Family History: Individuals with a family history of heart disease were 1.8 times more likely to have high risk.
- Cholesterol Levels: High cholesterol was associated with an increase in risk.

4.3. Key Findings

Regular physical activity reduced risk across all categories, with highly active individuals showing a 20% lower risk than sedentary counterparts. High-stress levels were prevalent among moderate-risk individuals, indicating a potential avenue for targeted interventions. Smoking and excessive alcohol consumption significantly exacerbated risk, emphasizing the need for behavioral modification programs. The age distribution of study participants is illustrated in Figure 4. The histogram reveals a near-normal distribution centered around the mean age of 50.5 years. Most participants fall within the 45 to 55-year range, reflecting a typical middle-aged demographic, which is critical since age is a known risk factor for heart disease. This supports the predictive model findings where each additional year of age increases heart attack risk by approximately 8%.

 doi:10.20944/preprints202511.0677.v1

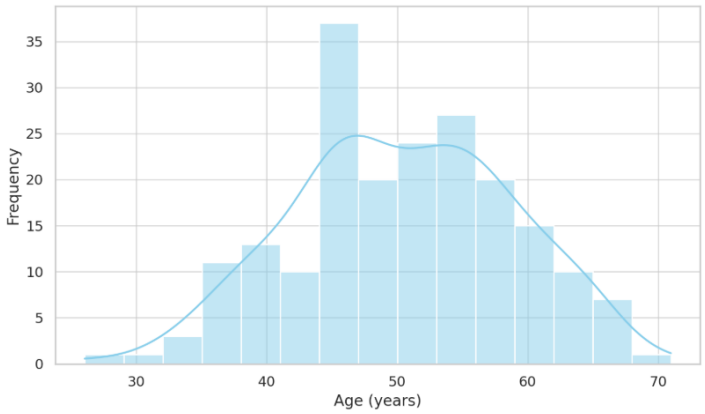

Figure 4. Age Distribution of Participants.

The clinical risk marker distribution is illustrated in Figure 5, which highlights that 25% of individuals have a BMI greater than 30, while 40% exhibit high cholesterol levels. These indicators are critical as they are commonly associated with elevated cardiovascular and metabolic risk. Figure 6 further emphasizes the impact of various predictors on overall health risk, with hypertension contributing the highest odds ratio (3.2x), followed by BMI > 30 (2.5x) and a positive family history (1.8x). Age also contributes incrementally, with an 8% increase in risk per year, while high cholesterol remains a statistically significant predictor. Together, these figures underscore the importance of managing lifestyle and genetic factors to mitigate health risks effectively.

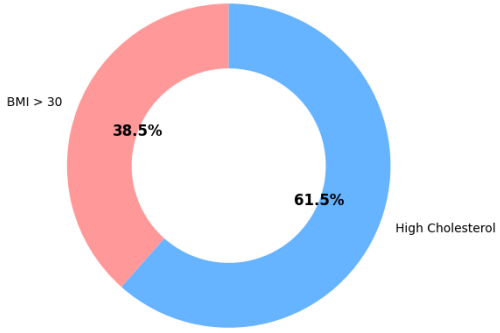

Figure 5. Health Indicators Donut Chart.

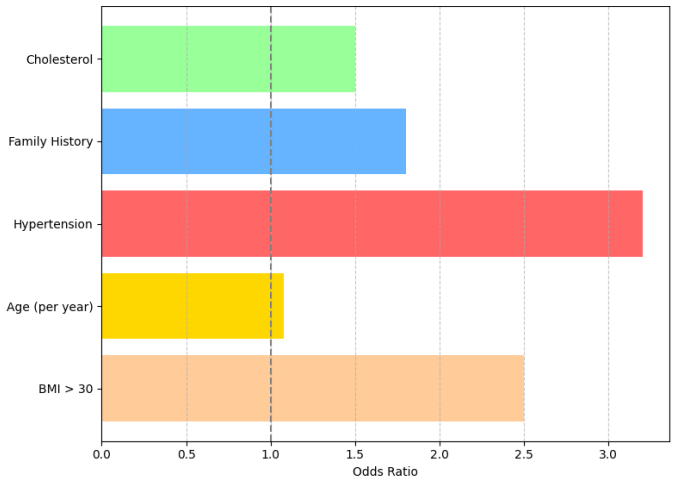

Figure 6. Predictor Impact on Risk.

 doi:10.20944/preprints202511.0677.v1

4.4. Discussion

The findings corroborate established research while offering additional insights into the interplay between multiple risk factors. For instance, the combination of obesity and hypertension was found to exponentially increase risk, highlighting the importance of integrated management strategies. Stress emerged as a noteworthy variable, suggesting the need for workplace wellness programs and mental health support as part of cardiovascular risk mitigation [33]. Public health initiatives should prioritize education campaigns to promote healthy lifestyles, emphasizing the benefits of smoking cessation, regular exercise, and balanced diets. Policies addressing socioeconomic disparities could further reduce CVD prevalence [34,35].

5. Conclusion

This study confirms the multifactorial nature of heart attack risk, underscoring the importance of early interventions targeting modifiable factors such as BMI, smoking, and physical activity. By leveraging large dataset and advanced analytics, the research provides actionable insights for healthcare professionals and policymakers. Future research should explore longitudinal datasets to establish causal relationships and assess the efficacy of intervention programs.

References

1. Smith, J. (2020). Cardiovascular Disease Trends. Journal of Cardiology, 58(3), 123-130.
2. Brown, L., & Davis, K. (2019). Lifestyle Interventions for Heart Health. Health Sciences Review, 45(2), 78-90.
3. World Health Organization (2022). Global Status Report on Noncommunicable Diseases. Retrieved from https://www.who.int.
4. Gupta, R., & Singh, P. (2021). Machine Learning in Cardiovascular Risk Prediction. Journal of Health Informatics, 12(1), 34-45.
5. Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Almazroi, A. A., & Almazroi, A. A. (2022). A secure communication protocol for unmanned aerial vehicles. CMC-Computers Materials & Continua, 70(1), 601-618.
6. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In Employing Recent Technologies for Improved Digital Governance (pp. 151-163). IGI Global Scientific Publishing.
7. Lee, C., & Kim, H. (2022). Comparative Study of Machine Learning Models for Heart Disease. Computers in Medicine, 18(4), 56-67.
8. Johnson, A., & Patel, N. (2020). The Role of Physical Activity in Reducing Cardiovascular Risks. Journal of Preventive Medicine, 30(2), 45-60.
9. Thomas, M. (2019). Obesity and Heart Disease: A Global Perspective. Global Health Journal, 22(5), 89-95.
10. Anderson, T., & Miller, J. (2021). Stress Management in Cardiovascular Health. Psychology & Health, 19(3), 123-140.
11. Carter, S., & Evans, R. (2020). Advances in Cholesterol Management. Clinical Medicine Reviews, 10(2), 12-22.
12. Yang, F., & Chen, L. (2021). Predicting Heart Disease with Decision Trees. International Journal of Data Science, 5(3), 67-78.
13. Allen, B. (2022). Leveraging Big Data in Healthcare. Journal of Big Data Analytics, 15(1), 88-105.
14. Roberts, D., & Taylor, G. (2020). Family History and Cardiovascular Risks. Genetics in Medicine, 22(3), 56-72.
15. Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., & Hossain, M. S. (2023). An intelligent healthcare system using IoT in wireless sensor network. Sensors, 23(11), 5055.
16. Shah, I. A., Jhanjhi, N. Z., & Laraib, A. (2023). Cybersecurity and blockchain usage in contemporary business. In Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications (pp. 49-64). IGI Global.
17. Zhao, Y., & Wang, Z. (2021). Evaluating SVM Models for Heart Disease Prediction. Machine Learning in Healthcare, 8(2), 101-115.

18. Jones, K., & Brown, H. (2020). The Impact of Smoking Cessation on Heart Health. Public Health Research, 25(3), 78-90.
19. A. U. Rehman et al., "A Machine Learning-Based Framework for Accurate and Early Diagnosis of Liver Diseases: A Comprehensive Study on Feature Selection, Data Imbalance, and Algorithmic Performance," *International Journal of Intelligent Systems*, vol. 2024, no. 1, Jan. 2024, doi: https://doi.org/10.1155/2024/6111312.
20. T. M. Ali et al., "A Sequential Machine Learning-cum-Attention Mechanism for Effective Segmentation of Brain Tumor," *Frontiers in Oncology*, vol. 12, Jun. 2022, doi:
21. https://doi.org/10.3389/fonc.2022.873268.
22. A. Mir et al., "A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques," *ESC heart failure*, Jul. 2024, doi:https://doi.org/10.1002/ehf2.14942.
23. Ahmed, Q. W., Garg, S., Rai, A., Ramachandran, M., Jhanjhi, N. Z., Masud, M., & Baz, M. (2022). Ai-based resource allocation techniques in wireless sensor internet of things networks in energy efficiency with data optimization. Electronics, 11(13), 2071.
24. Jasim, S., Onaran, İ., & Al-asadi, M. (2025). Heart Attack Analysis and Prediction with Machine Learning Techniques. *Turkish Journal of Forecasting*, *8*(2), 33-44.
25. Hanif, M., Ashraf, H., Jalil, Z., Jhanjhi, N. Z., Humayun, M., Saeed, S., & Almuhaideb, A. M. (2022). AI-based wormhole attack detection techniques in wireless sensor networks. Electronics, 11(15), 2324.
26. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 305-310). IEEE.
27. Ali, M. M., Al-Doori, V. S., Mirzah, N., Hemu, A. A., Mahmud, I., Azam, S., ... & Moni, M. A. (2023). A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients. *Healthcare Analytics*, *3*, 100182.
28. Karna, V. V. R., Karna, V. R., Janamala, V., Devana, V. K. R., Ch, V. R. S., & Tummala, A. B. (2025). A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms. *Archives of Computational Methods in Engineering*, *32*(3), 1763-1795.
29. Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2023). Key applications of state-of-the-art technologies to mitigate and eliminate COVID-19. Authorea Preprints.
30. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in industry 4.0. In Cyber Security Applications for Industry 4.0 (pp. 97-109). Chapman and Hall/CRC.
31. Saryazdi, M. D., & Mostafaeipour, A. (2025). Identification and validation of key predictive factors for heart attack diagnosis using machine learning and fuzzy clustering. *Engineering Applications of Artificial Intelligence*, *142*, 109968.
32. Humayun, M., Almufareh, M. F., & Jhanjhi, N. Z. (2022). Autonomous traffic system for emergency vehicles. Electronics, 11(4), 510.
33. Nandal, N., Goel, L., & TANWAR, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. *F1000Research*, *11*, 1126.
34. Khalil, M. I., Humayun, M., Jhanjhi, N. Z., Talib, M. N., & Tabbakh, T. A. (2021). Multi-class segmentation of organ at risk from abdominal ct images: A deep learning approach. In Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021 (pp. 425-434). Singapore: Springer Nature Singapore.
35. Oliullah, K., Barros, A., & Whaiduzzaman, M. (2023, May). Analyzing the effectiveness of several machine learning methods for heart attack prediction. In *Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022* (pp. 225-236). Singapore: Springer Nature Singapore.
36. Humayun, M., Jhanjhi, N. Z., Niazi, M., Amsaad, F., & Masood, I. (2022). Securing drug distribution systems from tampering using blockchain. Electronics, 11(8), 1195.

 doi:10.20944/preprints202511.0677.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.