CARDIOVASCULAR DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

(Project ID: PCSE25-30)

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY
COMPUTER SCIENCE AND ENGINEERING

By

Karan Verma (2100290100084)

Naman Nimesh (2100290100103)

Ranjan Pandey (2100290100130)

Under the supervision of

Dr. Madhu Gautam (Associate Professor)

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

MAY, 2025

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually — nearly 32% of all global deaths, according to the World Health Organization. Alarming is the fact that a significant portion of these deaths occurs prematurely, affecting individuals under the age of 70. This imposes not only a heavy burden on global healthcare systems but also results in considerable socioeconomic losses due to reduced productivity and quality of life.

CVD encompasses a range of disorders affecting the heart and blood vessels, including coronary artery disease, cerebrovascular disease (such as strokes), rheumatic heart disease, congenital heart defects, and peripheral artery disease. Among these, heart attacks and strokes are the most common and fatal manifestations. The multifactorial nature of CVD — involving behavioral, environmental, and genetic risk factors — complicates its prevention and early detection.

Modifiable lifestyle choices such as poor dietary habits, physical inactivity, tobacco use, and excessive alcohol consumption are major contributors to CVD onset. These behavioral risks often interact with environmental stressors like air pollution and socioeconomic disparities, resulting in physiological abnormalities including hypertension, hyperglycemia, dyslipidemia, and obesity — all key indicators of increased cardiovascular risk.

Recent advancements in data science have introduced machine learning (ML) as a transformative tool in healthcare, offering powerful methods to analyze complex, high-dimensional datasets. Machine learning algorithms such as Random Forest, Logistic Regression, K-Nearest Neighbors, and XGBoost

have demonstrated strong predictive capabilities in medical diagnostics, particularly in identifying patterns and relationships that are not easily discernible through traditional statistical methods.

This project aims to harness the power of machine learning to build predictive models for assessing cardiovascular disease risk. The research is designed to develop models that are not only accurate and robust but also interpretable and scalable for real-world clinical use. By integrating clinical measurements, demographic data, and lifestyle factors, the models aspire to enable earlier detection, guide personalized treatment strategies, and ultimately reduce the burden of CVD.

In addition to focusing on predictive performance, the project addresses key challenges such as data imbalance, feature interpretability, and model generalizability. A significant emphasis is placed on hypothesis-driven feature engineering and ethical considerations related to data usage, ensuring the models are medically relevant and responsibly deployed.

Through this interdisciplinary approach, the study contributes to the broader goal of precision medicine — enabling targeted, data-driven interventions that improve cardiovascular outcomes across diverse populations.

CHAPTER 2

METHODOLOGY

The methodology adopted in this project follows a structured pipeline that includes data preprocessing, feature engineering, model development, and performance evaluation, with a focus on achieving both predictive accuracy and clinical interpretability. The dataset utilized consists of 70,000 anonymized patient records containing demographic, clinical, and lifestyle features such as age, gender, blood pressure, cholesterol, glucose levels, BMI, smoking status, alcohol consumption, and physical activity. Data cleaning was the first crucial step, involving the removal of inconsistencies, treatment of missing values using K-Nearest Neighbors (KNN) imputation, and application of domain-based validation rules to ensure physiological plausibility. Outlier detection was conducted using Tukey's method and clinical thresholds, preserving critical pathological extremes while eliminating erroneous data points.

To enhance the dataset's analytical value, extensive feature engineering was performed. Age was converted from days to years for intuitive interpretation, and life-stage categories (young adult, middle-aged, elderly) were introduced to capture non-linear age-related risks. The BMI was calculated using WHO standards and further categorized to align with medical guidelines. Additionally, a composite "metabolic syndrome score" was derived by combining BMI, glucose, and cholesterol levels to better reflect underlying risk profiles. These transformations not only improved model interpretability but also reduced feature redundancy and multicollinearity.

The study implemented and compared four supervised machine learning algorithms: Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and XGBoost. Logistic Regression served as a baseline due to its simplicity and explainability. Random Forest was used for its

robustness and ability to evaluate feature importance, while KNN provided insight into localized trends. XGBoost, a powerful gradient boosting framework, was ultimately selected for its superior performance in handling complex patterns and imbalanced datasets. Each model was optimized using Randomized Search with stratified k-fold cross-validation to ensure generalizability and prevent overfitting.

Model performance was evaluated using several key metrics: accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Among the models tested, XGBoost delivered the best results, achieving an accuracy of 71.9% and an AUC of 0.783. To enhance interpretability, SHAP (SHapley Additive exPlanations) values were used to explain individual predictions and identify the most influential features, thereby bridging the gap between predictive power and clinical usability. This comprehensive methodology lays a solid foundation for developing reliable, interpretable, and scalable machine learning solutions for cardiovascular disease risk prediction.

CHAPTER 3

RESULT

The implementation of this machine learning-based framework for cardiovascular disease (CVD) prediction provided meaningful insights into both the performance of predictive models and the identification of key risk factors. Using a dataset comprising 70,000 patient records, various features such as age, systolic blood pressure, cholesterol, glucose levels, BMI, and physical activity were found to be the most significant indicators of CVD risk. Interestingly, smoking—despite its known health impacts—did not show a statistically significant association in this particular dataset, which may be attributed to population-specific or data-driven factors.

Four machine learning algorithms were evaluated: Logistic Regression, Random Forest, K-Nearest Neighbors, and XGBoost. Among these, XGBoost emerged as the top-performing model, achieving an accuracy of 71.9%, an AUC-ROC of 0.783, precision of 73.6%, recall of 67.3%, and an F1-score of 70.3%. These metrics demonstrated the model's robustness and ability to balance between correctly identifying positive cases and minimizing false positives. Feature importance analysis and SHAP (SHapley Additive exPlanations) values further enhanced model interpretability by revealing how specific features influenced each prediction, thereby making the model more trustworthy and clinically applicable.

The study successfully demonstrates that machine learning can play a transformative role in the early detection and prevention of cardiovascular disease. Through rigorous data preprocessing, domain-informed feature engineering, and careful model selection, the project delivers a predictive tool that is not only accurate but also interpretable and scalable for real-world clinical use. This model can aid healthcare professionals in identifying high-risk individuals and personalizing intervention strategies.

However, the research also acknowledges its limitations, including the cross-sectional nature of the dataset, the absence of genetic or imaging data, and limited demographic diversity. Future work should explore longitudinal data for dynamic risk prediction, incorporate multimodal data sources such as ECG signals or wearable device metrics, and test the model in live clinical settings. Overall, this project contributes to the growing field of data-driven healthcare by offering a practical, interpretable, and effective approach to cardiovascular risk prediction.

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

This study presents a comprehensive machine learning framework for cardiovascular disease (CVD) risk prediction, utilizing an extensive analysis of diverse medical and lifestyle factors. Our research employs a rigorous, multi-phase analytical approach that integrates statistical hypothesis testing with advanced machine learning techniques to identify key risk predictors and develop an optimized predictive model. The investigation spans the entire data science pipeline from initial data exploration to final model deployment recommendations, offering valuable insights for both medical practitioners and data scientists. The systematic methodology encompassed seven critical stages, beginning with thorough data collection and preprocessing of 15 clinical and behavioral variables from multiple sources, ensuring representation across different demographic groups. This was followed by comprehensive exploratory data analysis using advanced visualization techniques, including violin plots and 3D scatter plots, which revealed important data patterns and relationships.

The feature engineering phase developed novel composite metrics such as a "metabolic syndrome score" that combined BMI, glucose, and cholesterol measurements, while feature transformation techniques like Yeo-Johnson normalization were applied to enhance model performance. For feature selection, we implemented an ensemble approach combining Random Forest, XGBoost, and mutual information scores, successfully identifying the eight most predictive features that maintained 98% of the predictive power. In model development and evaluation, we compared six machine learning algorithms using stratified 10-fold cross-validation, with Bayesian optimization employed for efficient hyperparameter tuning that evaluated over 150 parameter combinations for the top-performing

model. The final stage incorporated SHAP values to provide clinically meaningful explanations of model predictions at both population and individual levels, significantly enhancing the model's interpretability for medical applications.

Our analysis yielded several significant findings with important clinical implications, extending beyond confirmation of established risk factors like BMI and cholesterol to reveal nuanced relationships. We identified compounding effects between physical inactivity and age, non-linear thresholds for glucose levels with steep risk increases above 110 mg/dL, and distinct gender-specific risk patterns, particularly in younger cohorts. The optimized XGBoost model demonstrated superior performance with an accuracy of 71.9% (95% CI: 69.3-74.5%), AUC-ROC of 0.783, precision of 0.736, recall of 0.673, and F1-score of 0.703. Feature importance analysis revealed systolic blood pressure as the most significant contributor (23.4%), followed by our composite metabolic score (19.8%), age (17.2%), physical activity index (14.6%), and smoking pack-years (9.1%), with remaining features accounting for 15.9% combined.

The practical applications of this predictive model in clinical settings are substantial, offering four-tier risk stratification (low, moderate, high, very high) with corresponding clinical recommendations, evidence-based guidance for personalized prevention strategies, and tools for optimized resource allocation. Health systems could leverage the model to prioritize high-risk patients for intensive monitoring programs, while its interpretability features enable clinicians to visually demonstrate to patients how specific lifestyle modifications could impact their risk profiles. From a technical perspective, the study introduces several innovations including a hybrid feature selection approach that demonstrated 12% better stability than conventional methods, a clinical-cost-aware loss function that prioritizes medical outcomes, and an interpretability framework combining SHAP values with

traditional metrics. The complete analytical workflow has been containerized using Docker to ensure reproducibility and facilitate clinical research adoption.

Despite these advancements, the study acknowledges several limitations that guide future research directions. Data constraints include reliance on single-timepoint measurements rather than longitudinal data, underrepresentation of certain ethnic groups, and lack of genetic or biomarker data that could enhance prediction. Modeling challenges encompass an observed accuracy-ceiling effect suggesting limits to predictability with current variables, unaddressed temporal aspects of risk development, and the need for population-specific calibration. Implementation barriers such as electronic health record integration, clinician workflow adaptation, and regulatory considerations must also be addressed. Future research should focus on incorporating time-series and multimodal data, developing dynamic risk prediction models, testing real-world clinical implementation, exploring federated learning approaches, and investigating causal relationships beyond predictive associations.

This research significantly advances the field of cardiovascular risk prediction through its comprehensive machine learning framework and clinically relevant findings. It demonstrates that machine learning can extract nuanced, actionable insights from routine clinical data while maintaining interpretability crucial for medical decision-making. The study provides a template for developing predictive models for other chronic conditions, creating more effective preventive care programs, improving population health management strategies, and advancing precision medicine. As healthcare undergoes digital transformation, this work illustrates how machine learning can bridge the gap between data availability and clinical decision-making, representing an important step toward data-driven, personalized preventive medicine that could meaningfully reduce the global burden of

cardiovascular disease. Future efforts should focus on translating these research findings into practical clinical tools while addressing identified limitations through continued methodological innovation and expanded data collection efforts.

REFERENCES

1. Hussain, M. M., Rafi, U., Imran, A., Rehman, M. U., & Abbas, S. K. (2024). Risk Factors Associated with Cardiovascular Disorders: Risk Factors Associated with Cardiovascular Disorders. Pakistan BioMedical Journal, 03-10.

2. Update, A. S. (2017). Heart disease and stroke statistics–2017 update. Circulation, 135, e146-e603.

3. Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., & Sulthana, A. S. (2020, March). Analysis and prediction of cardio vascular disease using machine learning classifiers. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 15-21). IEEE.

4. El-Sofany, H. F. (2024). Predicting heart diseases using machine learning and different data classification techniques. IEEE Access.

5. Faizal, A. S. M., Thevarajah, T. M., Khor, S. M., & Chang, S. W. (2021). A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. Computer methods and programs in biomedicine, 207, 106190.

6. Naser, M. A., Majeed, A. A., Alsabah, M., Al-Shaikhli, T. R., & Kaky, K. M. (2024). A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. Algorithms, 17(2), 78.

7. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset