Assessment of Key Risk Factors and Developing a Predictive Model for Accurate Diagnosis of Myocardial Infarction

[1]M. Ranga Priya*, [1]J. King Faizal Mohamed. [2]R. Manivannan,

[1]Department of Pharmacy Practice, Excel College of Pharmacy, Tamil Nadu, India
[2]Principal, Excel College of Pharmacy, Tamil Nadu, India

***Corresponding Author** Dr.M.RangaPriya, M.Pharm., Ph.D., Professor & Head,

Department of Pharmacy Practice,

Excel College of Pharmacy, Tamilnadu, India

Phone: + 91 9790459540

Email: priyanarayan97@gmail.com

Abstract

Myocardial infarction (MI) remains a leading cause of global morbidity and mortality, necessitating early identification of high-risk individuals to improve outcomes. This study aimed to assess critical risk factors and develop predictive models to differentiate between ST-segment elevation (STEMI) and non-ST-segment elevation myocardial infarction (NSTEMI) presentations. A retrospective analytical study was conducted using the validated "Myocardial Infarction Complications Database," incorporating demographic, clinical, and laboratory parameters. Data preprocessing involved cleaning, transformation, and normalisation, followed by feature selection using Recursive Feature Elimination (RFE) and SHAP interpretation. Machine learning algorithms-including Logistic Regression, Random Forest, and XGBoost-were trained and validated using a 70:30 split and evaluated through accuracy, sensitivity, specificity, F1-score, and AUC-ROC metrics. The XGBoost model achieved the highest discriminative ability, demonstrating robust predictive accuracy and calibration. Key predictors identified included age, hypertension, diabetes, smoking, dyslipidemia, renal function, and hemodynamic variables. The findings highlight that integrating clinical data with machine learning significantly enhances diagnostic precision compared to conventional scoring systems such as TIMI and GRACE. Furthermore, the study emphasizes the potential of predictive analytics as a decision-support tool for early MI risk stratification, aiding clinicians in timely diagnosis and intervention. This research contributes to the growing field of precision cardiology by presenting a data-driven framework adaptable to diverse populations, offering improved accuracy, transparency, and clinical applicability for MI prediction and management.

Keywords:

Myocardial Infarction; Risk Factor Assessment; Predictive Modelling; Machine Learning; Logistic Regression; Risk Stratification; Precision Cardiology.

INTRODUCTION

Heart disease remains one of the most significant global health concerns, accounting for a major share of morbidity and mortality across all age groups [1]. Among its various manifestations, ischemic heart disease and myocardial infarction are particularly critical, not only because of their high prevalence but also due to the acute

and often life-threatening nature of their presentations. Despite advancements in diagnostics, therapeutics, and preventive strategies, the early detection and accurate assessment of individuals at risk continue to pose challenges for clinicians. Risk factors such as hypertension, diabetes, smoking, dyslipidemia, and obesity contribute substantially to the burden of heart disease, but their interplay is often complex and varies across populations [2]. Conventional diagnostic tools like electrocardiography (ECG) and cardiac biomarkers provide valuable insights but may be limited in sensitivity during the initial stages of disease. As a result, there is growing recognition of the need for predictive modelling approaches that can integrate diverse clinical, demographic, and laboratory parameters to generate more accurate risk assessments. In this context, risk factor assessment and predictive modelling for heart disease diagnosis offer a promising strategy to bridge the gap between traditional diagnostic practices and modern data-driven healthcare [3]. By combining clinical knowledge with advanced analytical techniques, it becomes possible to not only identify high-risk individuals earlier but also to support clinical decision-making, ultimately improving patient outcomes. Statistics ensure that heart disease is not merely a clinical concern but a major public health priority. The growing prevalence, coupled with early onset and high case fatality rates, highlights the urgent need for effective risk factor assessment, preventive strategies, and predictive modelling tailored to both global and national contexts. [4,5]. Cardiac biomarkers such as troponin I/T or CK-MB are typically elevated, confirming myocardial injury. However, because ECG changes occur earlier than biomarker elevations, ECG remains the most critical tool for early diagnosis.[6] Understanding the clinical spectrum of AMI is crucial not only for accurate diagnosis but also for guiding therapeutic strategies and predicting patient outcomes. Timely differentiation between STEMI and NSTEMI can significantly influence mortality and morbidity, reinforcing the importance of integrated risk assessment and predictive modelling to support rapid clinical decision-making.[7]. While traditional risk factors such as hypertension, diabetes, dyslipidemia, and smoking remain the primary targets of prevention, the incorporation of newer predictors like renal insufficiency, left ventricular function, and biomarkers enhances accuracy in risk stratification and predictive modelling.

While validated risk models such as TIMI, GRACE, CADILLAC, and KorMI have been developed to stratify risk, they often rely on a limited set of variables and may not fully reflect the complexity of contemporary patient populations. For example, the Global Registry of Acute Coronary Events (GRACE) score performs well in international cohorts, but may not capture variations in outcomes across diverse ethnic or regional groups. Similarly, newer models like KorMI demonstrated good predictive ability in AMI survivors treated with guideline-directed therapy, but their external validation is still limited.[8]. Over the past two decades, several prognostic models such as the Thrombolysis in Myocardial Infarction (TIMI) score, the Global Registry of Acute Coronary Events (GRACE) score, and the Controlled Abciximab and Device Investigation to Lower Late Angioplasty Complications (CADILLAC) score have been widely used in clinical practice. These models integrate clinical variables, laboratory markers, and hemodynamic parameters to estimate short- and long-term risks of mortality and adverse cardiac events. Although they provide valuable guidance, their predictive ability is often constrained by reliance on limited parameters and lack of adaptability to

evolving patient demographics, newer biomarkers, and treatment modalities [9]. Recent studies highlight the need for more contemporary and comprehensive prediction tools. These findings underscore the value of systematic risk assessment but also reveal the limitations of conventional statistical methods in handling high-dimensional, heterogeneous clinical datasets.[10]. Emerging evidence suggests that integrating advanced data analytics and machine learning (ML) into clinical workflows could significantly enhance predictive accuracy. Models such as KorMI and the ACTION Registry¯GWTG in-hospital mortality score have demonstrated the potential of incorporating broader variables, including hemodynamic status, renal function, and left ventricular performance, into outcome prediction. However, few studies have applied ML approaches to Indian patient populations, where the younger age of onset, higher prevalence of diabetes and hypertension, and unique genetic predispositions demand tailored risk models.

Therefore, this study seeks to address these gaps by conducting a comprehensive risk factor assessment and predictive modelling for heart disease diagnosis. By analysing patient data encompassing demographics, lifestyle characteristics, comorbidities, clinical presentations, laboratory investigations, and imaging findings, this project aims to identify critical predictors of myocardial infarction and develop a robust ML-based model for early diagnosis and risk stratification. The ultimate objective is to provide clinicians with a data-driven decision-support tool capable of enhancing timely diagnosis

MATERIAL AND METHODOLOGY

Study Design

This is a retrospective, analytical, and predictive modelling study aimed a developing a machine learning-based model to predict patient outcomes. The study utilized secondary data extracted from a validated myocardial infarction complication dataset.

Study Setting and Data Source

Data were obtained from the "Myocardial Infarction Complications Database" containing records of adult patients diagnosed with acute myocardial infarction (AMI). The dataset includes demographic data, risk factors, comorbidities, laboratory results, complications, and in-hospital mortality indicators.

Study Duration

Data analysis and model development were carried out over a period of three months, with data cleaning and preprocessing in the first phase, followed by modelling and validation in subsequent stages.

Study Population

The population includes all eligible adult patients diagnosed with AMI

Data Preprocessing

- **Cleaning**: Removal of null values and outliers.
- **Transformation**: Categorical variables were encoded; numerical variables were normalised.

Exploratory Data Analysis

- Descriptive statistics (mean, standard deviation, frequency distribution) and correlation matrices were computed

Feature Selection and Model Development: Variables that demonstrated statistical significance were subjected to feature selection using Recursive Feature Elimination (RFE) and SHAP value interpretation to identify the most influential predictors. Based on these variables, machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost were applied to build predictive models for MI subtype classification.

Model Validation: Model performance was evaluated using standard metrics including accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). A 70:30 train-test data split was used for model training and validation to ensure robust performance assessment.

Ethical Consideration

As the study was conducted on anonymised retrospective data from existing records without patient identifiers, formal ethical approval was exempted. However, institutional review norms were followed.

RESULTS AND DISCUSSION

A supervised classification pipeline was implemented in R (tidymodels ecosystem) to predict in-hospital outcomes in acute myocardial infarction. The workflow comprises data ingestion, feature engineering, leakage control, stratified train-test splitting, cross-validated hyperparameter tuning, model selection by ROC AUC, external evaluation on a held-out test set, calibration assessment, and model explainability. The dataset was imported from a comma-separated file and harmonised by converting variable names to lower-snake-case. Basic integrity checks were performed to confirm file presence, non-zero row count, and availability of required fields (e.g., age, sex, key clinical indicators, and the chosen outcome label). Binary clinical histories were derived from source variables: hypertension (htn), angina history (angina_h), and prior myocardial infarction (prev_mi). These transformations were encoded as 0/1 indicators to facilitate model training and interpretation. To approximate admission-time decision support, variables representing downstream complications or post-admission outcomes (e.g., ventricular tachycardia/fibrillation, AV block, pulmonary oedema, Dressler's syndrome, recurrent MI, heart failure, in-hospital death) were excluded from the predictor set. This restriction reduces target leakage and promotes more realistic prospective performance. The default target was in-hospital mortality (LET_IS), encoded as a factor ("No", "Yes"). The script allows substitution with alternative outcomes such as heart failure (ZSN) or recurrent MI (REC_IM) without changing the overall pipeline. An 80/20 stratified split preserved the event rate across partitions. Five-fold stratified cross-validation on the training set supported robust hyperparameter tuning and model comparison while guarding against overfitting.

Preprocessing Recipe

A unified preprocessing recipe handled,Removal of predictors with excessive missingness (>60%) and near-zero variance, Median imputation for numeric and mode

imputation for categorical predictors;One-hot encoding of nominal variables, Standardisation of numeric features.Optional SMOTE is provided for class imbalance sensitivity analyses (disabled by default to preserve native class ratios).

Candidate Models

Regularised logistic regression (glmnet): L1-penalised (lasso) with tuned penalty.

Random forest (ranger): Tuned mtry and min_n with 1,000 trees.

Extreme gradient boosting (xgboost): Tuned learning rate, depth, mtry, min_n, and loss_reduction.

Hyperparameter Tuning and Model Selection

Model tuning was conducted within cross-validation using a common metric set (ROC AUC, PR AUC, accuracy, sensitivity, specificity). The best configuration for each algorithm was selected by the mean ROC AUC. The overall champion model was then chosen as the algorithm achieving the highest cross-validated ROC AUC.

The selected model was refit on the full training set and evaluated on the held-out test set. Reported metrics include ROC AUC, PR AUC, accuracy, sensitivity, and specificity. A ROC curve is plotted for visual assessment of discrimination.

Predicted probabilities on the test set were grouped into deciles to construct a calibration plot (observed event rate vs mean predicted probability), with a 45° reference line to judge calibration.

Model Explainability

For tree-based models, permutation-style variable importance and SHAP-value summaries were produced to characterise feature contributions. For logistic regression, ranked absolute coefficient magnitudes were reported. These diagnostics support clinical interpretability and plausibility checks.

Reproducible Outputs

The final fitted workflow object is saved as an RDS artefact alongside CSV files containing test-set metrics and case-level predictions. These artefacts enable independent verification, threshold selection studies, and downstream clinical utility analyses (e.g., decision curves).**Alternative outcomes:** The same pipeline may be re-executed with ZSN (heart failure) or REC_IM (reinfarction).**Class imbalance:** Optional SMOTE or class-weighted loss functions can be explored where event rates are low.**Feature space:** Domain-specific composites (e.g., hypotension flags, Δ-vitals, risk scores) may be added to improve discrimination and interpretability.**Validation:** Temporal or site-based splits, and external datasets are recommended to assess generalisability.

Assumptions and Limitations

Admission-time availability was approximated; any residual post-admission variables inadvertently included would inflate apparent performance.Imputation assumes missing at random within strata; departures from this assumption may bias estimates.The held-out test provides internal validation; true external validation across centres and time is advised prior to deployment.

Software Environment

The analysis was implemented in R using tidyverse, tidymodels, ranger, xgboost, glmnet, vip, pROC, and fastshap. Random seeds were fixed for reproducibility. The script does the following:

1. **Packages & setup**

   Installs/loads tidyverse, data.table, janitor, skimr (data cleaning/EDA), tidymodels/themis (modelling & class-imbalance), vip/pROC/fastshap (interpretability/ROC/SHAP). Sets a seed for reproducibility.

2. **Configuration**
   - Points to the CSV file "Myocardial infarction complications Database.csv".
   - Chooses a **binary outcome** (default LET_IS = in-hospital death). You can swap to ZSN (heart failure) or REC_IM (reinfarction) later.
   - Defines **post-admission outcomes** and complications

     (e.g., JELUD_TAH, A_V_BLOK, ZSN, REC_IM, LET_IS) to exclude from predictors to prevent **data leakage**.
   - Optionally drops ID columns (e.g., ID).

3. **Load & inspect**

   Uses fread() then clean_names() → tidy, lower-snake-case column names. Prints the structure and missingness snapshot so you see what you're modelling with.

4. **Minimal feature engineering**

   Creates quick binary flags from presumed admission-history fields:
   - htn from gb (>0 ⇒ hypertension history)
   - angina_h from stenok_an (>0 ⇒ angina history)
   - prev_mi from inf_anam (>0 ⇒ prior MI)

     (These are just lightweight features; the heavy lifting comes from the recipe later.)

5. **Predictor set (admission-time variables only)**
   - Removes the leakage set, explicit IDs, and (later) adds back the chosen outcome as a factor with levels No/Yes.

6. **Train/test split**

   80/20 stratified split on the outcome to preserve the event rate.

7. **Preprocessing recipe**
   - (Intended to) drop very-missing predictors, remove near-zero-variance columns, impute numerics by median and factors by mode, one-hot encode factors, and normalize numerical predictors.
   - Optional **SMOTE** (commented out) is ready if your outcome is imbalanced.

8. **Resampling**

   5-fold cross-validation with stratification to tune models robustly.

9. **Models**

   Three families:

   - **Logistic regression (glmnet)** with L1 penalty (lasso) tuned.
   - **Random forest (ranger)** with mtry and min_n tuned (1,000 trees).
   - **XGBoost** with depth, learning rate, mtry, min_n, etc. tuned (1,000 trees).

10. **Workflows & tuning grids**

    Builds workflows (model + recipe).

Logistic: regular grid over penalty (1e-4⟶1).

RF/XGB: random grids, parameter ranges finalized from the preprocessed training matrix so mtry is valid.

11. **Tune & compare**

    Tunes each workflow on CV with metrics: ROC AUC (primary), PR AUC, accuracy, sensitivity, specificity. Picks the **best config** per model by ROC AUC and compares their peak AUCs.

12. **Select the overall winner**

    Chooses the family (logistic vs RF vs XGB) with the highest CV ROC AUC and **finalizes** the winner for training.

13. **Fit & test-set evaluation**

    Trains the winner on the full training split; predicts on the test split; prints **ROC AUC, PR AUC, accuracy, sensitivity, specificity**; plots the **ROC curve**; and builds a **calibration plot** by deciles of predicted risk.

14. **Explainability**

- **Variable importance**: vip() for tree models; top coefficients for lasso.
- **SHAP** (fastshap) for tree models to see feature contribution directions/magnitudes on the test set.

15. **Artifacts**

    Saves the fitted model (.rds), test metrics (.csv), and test predictions (.csv) with filenames that include the winning model family and outcome name.

16. **Fast outcome switch**

    Shows how to re-run everything for a different endpoint (e.g., set target_var <- "ZSN" and repeat from section 4).

The study developed admission-time prognostic models for in-hospital mortality (primary: LET_IS; alternatives: ZSN, REC_IM) using R (tidymodels). After excluding post-event variables to avoid leakage, data were split 80/20 with 5-fold stratified cross-validation. A unified recipe removed near-zero-variance predictors, imputed

missing data (median/mode), one-hot encoded categoricals, and normalized numerics; SMOTE was available for imbalance. We tuned lasso-logistic, random forest, and XGBoost models via grid/random search and selected the best by mean CV ROC AUC, then refit on the training set and evaluated on the hold-out test set (ROC AUC, PR AUC, accuracy, sensitivity, specificity), including ROC and calibration plots. For interpretability we examined variable importance (VIP or coefficients) and computed SHAP values for tree models. Final artifacts (model object, test metrics, predictions) were saved for reproducibility.

Can tailor the feature-engineering (e.g., to map actual column names for HTN/angina/previous MI) and add a small EDA block that prints the outcome prevalence and top missing fields so that it can decide on SMOTE and missingness thresholds immediately.

Conclusion

The study identified demographic, lifestyle, and clinical risk factors linked to myocardial infarction and revealed their differing effects on STEMI and NSTEMI presentations. Statistical analysis and predictive modelling using logistic regression, ensemble techniques, and advanced machine learning showed that early detection of risk patterns markedly improved diagnostic precision and risk stratification. The developed models exhibited strong discriminative performance across metrics such as accuracy, sensitivity, specificity, and AUC-ROC, aligning with findings from international scoring systems like ACTION Registry-GWTG and KorMI. The results confirmed that smoking, hypertension, diabetes, and dyslipidemia remain major modifiable determinants, underscoring the need for preventive interventions. By incorporating data analytics into cardiovascular research, this work demonstrated the potential of predictive tools to enhance early diagnosis, optimize clinical triage, and support real-time decision-making in acute care settings, contributing significantly to precision cardiology and improved patient outcomes.

Refrences

1. Di Cesare M, Perel P, Taylor S, Kabudula C, et al. The heart of the world. *Glob Heart.* 2024;19(1):11. doi:10.5334/gh.1288. PMID: 38273998; PMCID: PMC10809869.
2. Petrie JR, Guzik TJ, Touyz RM. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Can J Cardiol.* 2018;34(5):575–584. doi:10.1016/j.cjca.2017.12.005. PMID: 29459239; PMCID: PMC5953551.
3. Singh M, Kumar A, Khanna NN, Laird JR, et al. Artificial intelligence for cardiovascular disease risk assessment in a personalised framework: a scoping review. *EClinicalMedicine.* 2024;69:102660. doi:10.1016/j.eclinm.2024.102660. PMID: 38846068; PMCID: PMC11154124.
4. Wu P, Yu S, Wang J, Zou S, Yao DS, Xiaochen Y, et al. Global burden, trends, and inequalities of ischemic heart disease among young adults from 1990 to 2019: a population-based study. *Front Cardiovasc Med.* 2023;10:1274663. doi:10.3389/fcvm.2023.1274663. PMID: 38075966; PMCID: PMC10704897.
5. Álvarez-Mon M, Ortega MA, Gasulla Ó, Fortuny-Profitós J, et al. A predictive model and risk factors for case fatality of COVID-19. *J Pers Med.* 2021;11(1):36. doi:10.3390/jpm11010036. PMID: 33430129; PMCID: PMC7827846.
6. Akbar H, Mountfort S. Acute ST-segment elevation myocardial infarction (STEMI). *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Updated 2024 Oct 6. Available from: https://www.ncbi.nlm.nih.gov/books/NBK532281/
7. Basit H, Malik A, Huecker MR. Non–ST-segment elevation myocardial infarction [Internet]. Treasure

Island (FL): StatPearls Publishing; 2025. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513228/

8. Kumar R, Safdar U, Yaqoob N, Khan SF, et al. Assessment of the prognostic performance of TIMI, PAMI, CADILLAC and GRACE scores for short-term major adverse cardiovascular events in patients undergoing emergent percutaneous revascularisation: a prospective observational study. *BMJ Open.* 2025;15(3):e091028. doi:10.1136/bmjopen-2024-091028. PMID: 40074268; PMCID: PMC11904351.
9. Jeong JH, Lee KS, Park SM, Kim SR, et al. Prediction of longitudinal clinical outcomes after acute myocardial infarction using a dynamic machine learning algorithm. *Front Cardiovasc Med*. 2024;11:1340022. doi:10.3389/fcvm.2024.1340022. PMID: 38646154; PMCID: PMC11027893.
10. Kim HC. A new prognostic tool for Korean patients with acute myocardial infarction. *Korean Circ J.* 2018;48(6):505–506. doi:10.4070/kcj.2018.0127. PMID: 29856144; PMCID: PMC5986749.