

# A Dual-Metric Framework for Rare-Event Prediction of Myocardial Infarction Complications: Discrimination, Threshold-Selected Detection, and Calibrated Multiclass Mortality

Shriyansh Baidya  
shribaidya9@gmail.com

## ABSTRACT

Most studies of myocardial infarction (MI) complications report a single performance metric, usually ROC-AUC, and identify a single winning model, yet under severe class imbalance, the best-ranking model is often not the one that would raise a clinical alert at a usable threshold. This paper presents a dual-metric evaluation framework that reports ranking performance (threshold-free ROC-AUC) and positive-class detection (threshold-dependent F1, average precision, and sensitivity and positive predictive value at thresholds chosen inside training data only) as complementary rather than interchangeable summaries, applied to 11 binary in-hospital complications and an 8-category lethal-outcome variable using 111 features (admission plus early-course variables). Five classifiers were compared through a unified, leakage-controlled pipeline with a single held-out 20% test set and stratified 10-fold cross-validation on the development data. Because cross-validation folds from one dataset are not independent experimental units, the primary model comparison treats the 11 binary clinical targets as the units of comparison (Friedman test with Holm-corrected Wilcoxon tests and Nemenyi critical differences). Across those 11 targets, Random Forest ranked first on discrimination (mean AUC 0.715; average rank 1.27; best on 9 of 11) but last on default-threshold detection (mean F1 0.020; average rank 4.55; F1 = 0 on 10 of 11 targets), whereas Logistic Regression ranked first on detection (best F1 on all 11; mean F1 0.211; average rank 1.00). Targets-as-units Friedman tests were significant on both axes (AUC chi-square = 28.87,  $p < 0.001$ ; F1 chi-square = 32.32,  $p < 0.001$ ). At F1-optimal thresholds selected inside training folds, mean sensitivity/PPV was 0.322/0.163 for Logistic Regression versus 0.107/0.202 for Random Forest; mean average precision was 0.225 for Random Forest and 0.217 for Logistic Regression. For multiclass mortality, XGBoost led held-out discrimination (macro-AUC 0.798, weighted F1 0.836), but balanced accuracy was only 0.26 and macro-F1 0.31, so the minority death classes remain hard to detect. SHAP rankings were consistent with established risk factors (face validity, not causal validation). Selecting a single best model without specifying the clinical operating point is unsound for rare-event MI complications; the framework matches models to ranking versus threshold-dependent detection and reports calibration and uncertainty for those tasks.

Sep 2026  
Vol 11. No 5.

**Keywords:** myocardial infarction, dual-metric evaluation, threshold-selected detection, multi-target prediction, calibration, explainable machine learning

## INTRODUCTION

Acute myocardial infarction (MI) remains one of the leading causes of morbidity and mortality worldwide. Even when the initial infarction is survived, the hospital course is often complicated by arrhythmias, mechanical failure of the myocardium, pump failure, and inflammatory sequelae. These complications, including atrial and ventricular arrhythmias, conduction block, pulmonary edema, cardiac rupture, Dressler's syndrome, chronic heart failure, recurrent infarction, and post-infarction angina, differ markedly in their frequency, pathophysiology, time of onset, and lethality, and a non-negligible fraction of patients ultimately die during the admission from heterogeneous causes. The clinical burden of MI and its complications, and the standardized diagnostic criteria that define them, are summarized in the Fourth Universal Definition of Myocardial Infarction (Thygesen et al., 2018).

The clinical value of anticipating these events is clear. If a patient's risk of, for example, ventricular fibrillation or myocardial rupture could be estimated from information already collected at admission, care teams could prioritize continuous monitoring, allocate intensive-care resources, and intervene earlier. Classical bedside instruments address parts of this need: the Killip classification stratifies pump failure (Killip and Kimball, 1967), while the GRACE and TIMI scores estimate ischemic risk and mortality (Granger et al., 2003; Antman et al., 2000). These tools are transparent and externally validated, but by design, they use only a handful of variables and are tuned for mortality or ischemic recurrence rather than for the full spectrum of in-hospital complications.

Machine learning (ML) broadens this scope by learning the mapping from many baseline patient characteristics to downstream outcomes. Prior work has shown that ML can improve cardiovascular risk prediction from routine clinical data (Weng et al., 2017) and has surveyed the growing role of ML across cardiology (Johnson et al., 2018; Shameer et al., 2018), with landmark demonstrations such as artificial-intelligence-enabled electrocardiographic screening for contractile dysfunction (Attia et al., 2019). Tree-based ensembles - Random Forest (Breiman, 2001) and gradient-boosted trees such as XGBoost (Chen and Guestrin, 2016) - have become the default for tabular clinical data because they handle mixed variable types, nonlinearities, and interactions with little feature engineering, while regularized logistic regression remains a strong, interpretable baseline.

Complication prediction is nonetheless a statistically hostile problem. Most individual complications are rare: in the database studied here, prevalence ranges from roughly 1% (supraventricular tachycardia, twenty of 1,700 patients) to 23% (chronic heart failure). Extreme class imbalance inflates accuracy for trivial always-negative predictors and makes precision, recall, and their harmonic mean (F1) far more informative than raw accuracy (He and Garcia, 2009; Saito and Rehmsmeier, 2015). Imbalance also interacts strongly with model choice, because a high-capacity ensemble can rank patients well (high

AUC) while predicting few or no positives at a default probability threshold, resulting in an F1 score of zero.

Existing ML studies on this database and related MI cohorts typically train a single model per outcome, report ROC-AUC or accuracy, and select the highest-scoring algorithm without assessing whether that algorithm would detect rare complications in clinical practice (Golovenkin et al., 2020). In prior work (Baidya and Gupta, 2024), the author established a simplified, interpretable workflow for predicting atrial fibrillation and chronic heart failure from the same myocardial infarction complications dataset, using logistic regression, decision trees, random forest, XGBoost, and a multilayer perceptron with comprehensive preprocessing, feature-importance analysis, and SHAP explanations. That work demonstrated that machine learning can identify clinically relevant predictors such as age, serum potassium, and intensive-care systolic blood pressure, and reported test accuracies up to 91% for atrial fibrillation and 84% for heart failure. However, it evaluated only two of the eleven available binary complications, relied primarily on training and testing accuracy from a single 80-20 split without stratified cross-validation, did not report precision, recall, F1, or ROC-AUC, did not model the multiclass lethal-outcome variable, and did not statistically compare models or examine how severe class imbalance affects which algorithm should be selected for deployment.

This paper addresses those limitations by building on Baidya and Gupta (2024) with a dual-metric evaluation framework for multi-target MI complication prediction. Rather than asking which single model achieves the highest accuracy, the framework asks two clinically distinct questions in parallel: which model best ranks patients by risk (measured by ROC-AUC), and which model best detects the positive class at a stated threshold (measured by default-threshold F1, and separately by training-only F1-optimal sensitivity and PPV). It then combines three methodological innovations beyond the earlier study. First, the unified, leakage-controlled pipeline is extended from two binary targets to eleven binary complications plus an eight-class lethal-outcome variable, with all outcome columns removed from the feature matrix when predicting mortality. Second, model comparisons treat the 11 binary clinical targets as the units of comparison (Friedman test, Holm-Bonferroni-corrected Wilcoxon signed-rank tests, Cliff's delta, and Nemenyi critical differences; Demsar, 2006). Multiclass mortality is reported separately because its F1 is weighted. Per-fold tests are reported only for transparency, not as independent replications. Third, a target-routed protocol maps each reporting intent (risk ranking, default-threshold detection, or multiclass mortality ranking) to the model family that performed best on the relevant axis of the metric. SHAP analysis (Lundberg and Lee, 2017; Lundberg et al., 2020) provides per-outcome interpretability across the full spectrum of complications, and reporting follows the TRIPOD statement (Collins et al., 2015) and the more recent TRIPOD+AI guidance (Collins et al., 2024). The principal advances over Baidya and Gupta (2024) are: coverage of twelve rather than two clinical targets; evaluation on both discrimination and detection axes rather than accuracy alone; stratified ten-fold cross-validation with formal statistical comparison of models; multiclass lethal-outcome modeling with leakage control; and a target-routed model-selection protocol tied to ranking versus threshold-dependent detection.

## DESCRIPTION OF DATA, METHODOLOGY, AND RESEARCH APPROACH

### Proposed Dual-Metric Evaluation Framework

The central methodological contribution of this study is a dual-metric evaluation framework designed for imbalanced, multi-outcome clinical prediction. Conventional benchmarks collapse model quality into a single score; this framework instead evaluates each model on two complementary axes corresponding to distinct clinical intents. The two axes are related but not interchangeable, so they are described as threshold-free ranking performance versus threshold-dependent classification performance rather than as strictly orthogonal quantities. The discrimination axis (ROC-AUC) measures how well a model ranks high-risk patients above low-risk patients and is appropriate when probabilities will feed a monitoring queue or a downstream decision rule with a separately chosen threshold. The detection axis (F1 score at a default 0.5 cutoff) measures how well a model identifies positive cases at that fixed threshold; a 0.5 cutoff is not a clinically chosen operating point, so F1 at 0.5 is reported as a default-threshold diagnostic, and training-only threshold selection is reported separately. By scoring all five model families on both axes across the 11 binary targets and then ranking models separately on each axis using average Friedman rank (multiclass mortality is reported separately), the framework produces a two-dimensional performance map rather than a single leaderboard.

The framework is operationalized through three integrated components. Component 1 is a unified, leakage-controlled preprocessing and modeling pipeline (described below) that is applied identically to every target, ensuring that performance differences reflect algorithm behavior rather than inconsistent data handling. Component 2 is a cross-target statistical validation protocol comprising Friedman omnibus tests, Holm-Bonferroni-corrected pairwise Wilcoxon signed-rank tests, and Cliff's delta effect sizes applied primarily across the clinical targets as the units of comparison (with Nemenyi critical differences), and reported per-fold only for transparency, which establishes whether observed metric differences are consistent across targets rather than fold noise. Component 3 is a target-routed model-selection protocol that assigns a recommended model family to each clinical use case based on the axis on which that family achieved the best average rank: Random Forest for discrimination-dominant binary ranking (average AUC rank 1.27 across the 11 binary targets), Logistic Regression for default-threshold detection (average F1 rank 1.00 across the 11 binary targets), and XGBoost for multiclass mortality ranking (held-out macro AUC 0.798 in Table 7), with the caveat that per-class death detection remains weak. SHAP analysis is an interpretability layer that describes how each fitted model uses variables (face validity, not independent clinical validation).

**Table 1. Principal methodological advances relative to Baidya and Gupta (2024).**

| Aspect                       | Baidya and Gupta (2024)                   | Present study                                                                             |
|------------------------------|-------------------------------------------|-------------------------------------------------------------------------------------------|
| Clinical targets modeled     | 2 (atrial fibrillation, heart failure)    | 12 (11 binary complications + 8-class lethal outcome)                                     |
| Primary evaluation metric    | Training and test accuracy                | Dual-axis: ROC-AUC (ranking) and F1 / AP / threshold-selected sensitivity-PPV (detection) |
| Cross-validation             | Single 80-20 train-test split             | Stratified 10-fold CV + held-out 20% test set                                             |
| Statistical model comparison | None                                      | Friedman, Holm-Wilcoxon, Cliff's delta, Friedman ranks                                    |
| Classifiers evaluated        | LR, DT, RF, XGBoost, MLP                  | LR, DT, RF, XGBoost, AdaBoost                                                             |
| Mortality / LET_IS modeling  | Not included                              | Multiclass with all outcome columns removed (leakage control)                             |
| Model selection criterion    | Highest test accuracy                     | Target-routed protocol: ranking vs default-threshold detection vs multiclass mortality    |
| Interpretability             | SHAP and feature importance for 2 targets | SHAP and feature importance for all 12 targets                                            |

### Dataset and Study Population

This study used the Myocardial Infarction Complications database (Golovenkin et al., 2020), the same publicly available cohort analyzed in prior work (Baidya and Gupta, 2024). It comprises 1,700 patient records, each described by 111 input features (admission plus early-course variables; see Predictor Availability and Measurement Time), together with a set of outcome variables: 11 binary complication indicators and 1 categorical lethal-outcome variable. Whereas the prior study modeled only atrial fibrillation (FIBR\_PREDS) and chronic heart failure (ZSN), the present study evaluates all 11 binary complications and the eight-class lethal-outcome label (LET\_IS). The tabulated models use 111 features, which include early-course treatments and the intensive-care blood-pressure reading, as well as admission variables (see Predictor Availability and Measurement Time).

The 111 input features fall into several clinically meaningful groups. Demographics and history include age, sex, and a large block of binary items such as prior angina, previous MI, exertional-angina functional class, hypertension and its duration, chronic heart-failure history, diabetes, obesity, prior stroke, and chronic pulmonary disease. Admission vital signs include systolic and diastolic blood pressure, recorded by the emergency cardiology team and again in the intensive care unit. An extensive set of electrocardiographic descriptors encodes rhythm at admission, ectopy and conduction abnormalities, and other findings that capture atrial and ventricular arrhythmias, as well as atrioventricular or bundle-branch conduction disturbances. Laboratory chemistry includes serum potassium, sodium, the transaminases ALT and AST, creatine phosphokinase (a marker of myocardial necrosis), white-cell count, and erythrocyte sedimentation rate. Records of early pharmacological therapy in the first hours and days are encoded as binary or ordinal indicators. Exactly twelve variables are continuous: age, the four blood-pressure

measurements, and the seven laboratory analytes, while the remainder are binary or low-cardinality categorical indicators.

### **Predictor Availability and Measurement Time**

Because the intended prediction point is hospital admission, each candidate predictor is tagged with its measurement time. The tabulated results use 111 features, including early pharmacological treatments and the intensive-care blood-pressure reading. Those variables are early course rather than strictly pre-admission; they are retained to ensure the new metrics remain comparable to Tables 4-7. Several admission electrocardiographic findings bear a name similar to later complications; for example, admission ventricular fibrillation predicts subsequent in-hospital ventricular fibrillation, and admission pulmonary edema predicts subsequent pulmonary edema (Table 10). These are baseline manifestations that precede the labeled in-hospital event; the outcome horizon is in-hospital events after admission.

The eleven binary complication targets and the multiclass mortality target, together with their clinical meaning and prevalence, are summarized in Table 2. The very low incidence of several targets, especially supraventricular tachycardia (1.2%), ventricular tachycardia (2.5%), myocardial rupture (3.2%), and atrioventricular block (3.4%), is the central statistical challenge of this study. The lethal-outcome variable is an eight-category label in which one class denotes survival, and the remaining seven encode distinct causes of death, making it a genuinely multiclass, imbalanced problem with an overall mortality of 15.9%.

**Table 2. Prediction targets, clinical meaning, and prevalence in the 1,700-patient cohort.**

| Target code | Clinical condition                | Positives / total | Prevalence |
|-------------|-----------------------------------|-------------------|------------|
| FIBR_PREDS  | Atrial fibrillation               | 170 / 1700        | 10.0%      |
| PREDS_TAH   | Supraventricular tachycardia      | 20 / 1700         | 1.2%       |
| JELUD_TAH   | Ventricular tachycardia           | 42 / 1700         | 2.5%       |
| FIBR_JELUD  | Ventricular fibrillation          | 71 / 1700         | 4.2%       |
| A_V_BLOK    | Atrioventricular block            | 57 / 1700         | 3.4%       |
| OTEK_LANC   | Pulmonary edema                   | 159 / 1700        | 9.4%       |
| RAZRIV      | Myocardial rupture                | 54 / 1700         | 3.2%       |
| DRESSLER    | Dressler's syndrome               | 75 / 1700         | 4.4%       |
| ZSN         | Chronic heart failure             | 394 / 1700        | 23.2%      |
| REC_IM      | Recurrent myocardial infarction   | 159 / 1700        | 9.4%       |
| P_IM_STEN   | Post-infarction angina / stenosis | 148 / 1700        | 8.7%       |
| LET_IS      | Lethal outcome (8 classes)        | 271 deaths / 1700 | 15.9%      |

The exploratory review flagged clinically implausible and extreme values that informed preprocessing. Age ranged from 26 to 92 years (mean approximately 62), consistent with an older MI population. Several blood-pressure records showed a minimum of zero, which is physiologically impossible in a stable patient but plausibly reflects artifacts from emergency

measurements or unrecordable pressure during arrest, and these were therefore retained. Laboratory analytes contained extreme values consistent with those in critically ill patients, and nine of the twelve numerical variables exhibited statistical outliers according to a three-times-interquartile-range criterion. Figure 1 shows an exploratory association between event prevalence and the best observed AUC among these 11 outcomes; with few targets and large differences in label quality and pathophysiology, that pattern is suggestive rather than a claim that prevalence alone determines learnability.

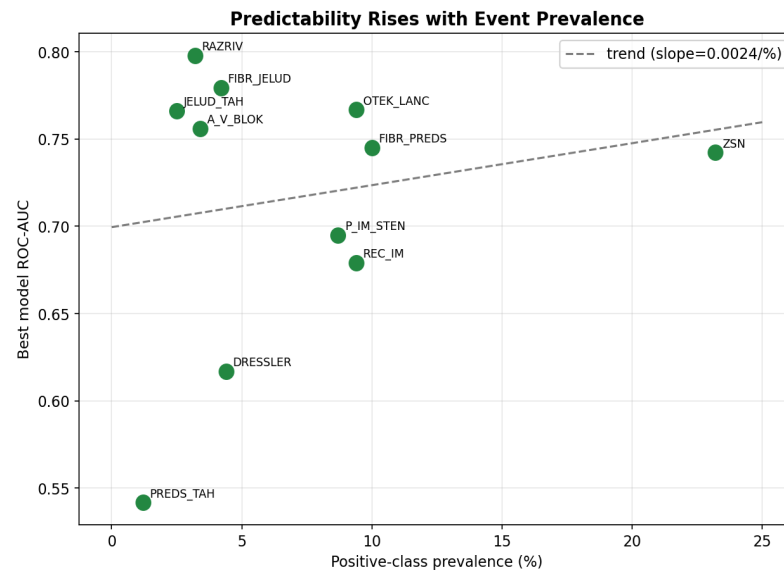


**Figure 1. Best-model ROC-AUC versus positive-class prevalence across the eleven binary targets (exploratory association; prevalence is not treated as determining learnability).**

## Data Preprocessing

The preprocessing pipeline follows and extends the workflow established in Baidya and Gupta (2024): median imputation for numerical variables, mode or category assignment for missing categorical values, outlier screening for numerical features, dummy encoding of categorical variables, and standardization of continuous laboratory and vital-sign measurements. The present study adds three safeguards that were not present in the earlier work: in-fold IQR outlier screening on training rows only (implausible values such as zero blood pressure remain in the source data and in the test set); zero-variance column removal after dummy encoding; and embedding all imputation and scaling steps within cross-validation pipelines so that no test-fold information leaks into preprocessing. For the multiclass lethal-outcome task, every other outcome column is additionally dropped from the feature matrix to prevent target leakage.

## Machine-Learning Models

Five supervised classifiers spanning linear, single-tree, bagging, and boosting families were evaluated for every target. The model set follows Baidya and Gupta (2024) in including Logistic Regression, Decision

Tree, Random Forest, and XGBoost, but replaces the multilayer perceptron, which achieved perfect training accuracy yet offered limited insight into generalization under imbalance, with AdaBoost (Freund and Schapire, 1997), which provides a more stable boosting baseline alongside XGBoost. Logistic Regression served as a regularized linear baseline with balanced class weighting in imbalanced settings. Random Forest (Breiman, 2001) was an ensemble of 200 shallow-depth trees with feature subsampling and balanced class weights. XGBoost (Chen and Guestrin, 2016) provided gradient-boosted trees. Logistic Regression, Decision Tree, Random Forest, and AdaBoost were implemented in scikit-learn (Pedregosa et al., 2011); XGBoost was implemented using its Python package. Each model was wrapped so that imputation and, where appropriate, scaling were learned only from the training data within each fold.

### **Evaluation Protocol**

The evaluation protocol represents a major advance over Baidya and Gupta (2024), who assessed models with a single 80-20 train-test split and reported only training and testing accuracy. Here, two complementary regimes were used. First, stratified 10-fold cross-validation on the development data preserved class proportions in each fold; for each fold, accuracy, precision, recall, F1, ROC-AUC, and average precision were recorded. Second, a single held-out stratified 20% test split, created once at the start, was evaluated only after every preprocessing, class-weight, hyperparameter, and threshold decision had been frozen on the development data. Positive-class detection is summarized by average precision, sensitivity, and positive predictive value at the F1-optimal threshold chosen within each training fold. For the multiclass lethal-outcome task, precision, recall, and F1 were computed with both weighted and macro averaging, and discrimination was measured with macro-averaged one-vs-rest ROC-AUC. Accuracy is reported throughout but treated as secondary when imbalance is present (Saito and Rehmsmeier, 2015).

### **Data Splitting, Leakage Control, and Reproducibility**

To make the design auditable, the analysis uses one fixed random seed (42) and a single 80/20 stratified split created once at the start (Figure 2). All class weighting, preprocessing (median imputation of the 12 continuous variables, standardization, mode imputation of categorical variables), IQR-based outlier screening, and F1-optimal threshold selection are performed only inside the cross-validation folds of the 80% development set; outlier screening is recomputed within each training fold. The held-out 20% test set is evaluated exactly once after all decisions are fixed. Overall, missingness among the 111 predictors was 8.5% ( $n = 1,700$ ). Exact hyperparameters are given in Table 3.

Train / validation / test flow (no test-set reuse)

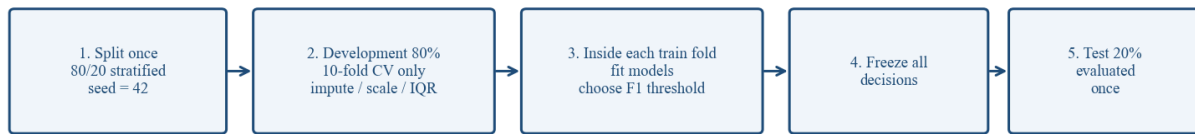


*Figure 2. Train/validation/test flow. The 20% test set is created once (seed 42) and evaluated only after every development-set decision is frozen.*

**Table 3. Model hyperparameters (random seed = 42).**

| Model               | Key hyperparameters                                                                                       |
|---------------------|-----------------------------------------------------------------------------------------------------------|
| Logistic Regression | C=1.0, L2 penalty, liblinear solver, class_weight=balanced, max_iter=1000                                 |
| Decision Tree       | max_depth=5, min_samples_leaf=20, class_weight=balanced                                                   |
| Random Forest       | 200 trees, max_depth=6, max_features=sqrt, class_weight=balanced                                          |
| XGBoost             | 300 trees, max_depth=4, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8, scale_pos_weight=neg/pos |
| AdaBoost            | 200 estimators, learning_rate=0.5, algorithm=SAMME                                                        |

Note: AdaBoost uses algorithm=SAMME. For each binary target, XGBoost scale\_pos\_weight equals the training negative-to-positive ratio.

### Statistical Comparison of Models

To determine whether performance differences among the five models were statistically significant, the primary analysis followed the protocol for comparing classifiers across multiple datasets, treating the clinical targets as the independent units of comparison (Demsar, 2006); because cross-validation folds from a single dataset are not independent, the per-fold Friedman tests in Table 8 are reported only as descriptive, per-target transparency. Cliff's delta (Cliff, 1993) is reported with the primary across-target pairwise tests as a non-parametric effect size. Finally, average Friedman ranks across the clinical targets were computed separately for AUC and F1, together with Nemenyi critical differences, to give an overall ordering of the five models; the 11 binary targets were ranked separately from the multiclass mortality target because the F1 definitions differ.

### Model Interpretation

For each outcome, the most influential predictors were identified using tree-based feature importance and SHAP (SHapley Additive exPlanations) analysis (Lundberg and Lee, 2017; Lundberg et al., 2020), which attributes each prediction to additive per-feature contributions. The leading features for each outcome were then compared with established cardiovascular risk factors to assess face validity.

### Framework Validation: Discrimination Axis (ROC-AUC)

The first axis of the dual-metric framework evaluates how well each model ranks patients by complication risk. Table 4 reports the mean cross-validated ROC-AUC for all five models on the eleven binary targets. On this axis, Random Forest was the dominant model family, achieving the top AUC on nine of the eleven binary targets and the highest mean AUC of 0.715; AdaBoost won the remaining two (ventricular tachycardia and heart failure) and was a close runner-up elsewhere with a mean AUC of 0.683. The single Decision Tree was the weakest model by a wide margin (mean AUC of 0.530, barely above chance), making it a weak baseline on these targets. The strongest observed AUCs were for myocardial rupture (0.798) and ventricular fibrillation (0.779), acute and physiologically distinctive events, while the weakest was supraventricular tachycardia (0.542), which had only 20 positive cases and was underpowered. The AUC heatmap in Figure 3 shows this structure across all model-target combinations; Figure 4 orders the best achievable discrimination by complication; and Figure 5 presents a representative held-out ROC curve. Ninety-five percent confidence intervals for the cross-validated AUCs (t-based across folds) are reported in Table 11; for example, Random Forest AUC for myocardial rupture was 0.798 (95% CI 0.744-0.851) and for ventricular fibrillation 0.779 (95% CI 0.731-0.828), confirming that even the strongest results carry appreciable uncertainty.

**Table 4. Mean 10-fold cross-validated ROC-AUC by target and model (LR = Logistic Regression, DT = Decision Tree, RF = Random Forest, XGB = XGBoost, ADA = AdaBoost).**

| Target     | Prev. | LR    | DT    | RF    | XGB   | ADA   | Best (model) |
|------------|-------|-------|-------|-------|-------|-------|--------------|
| FIBR_PREDS | 10.0% | 0.676 | 0.530 | 0.745 | 0.664 | 0.667 | 0.745 (RF)   |
| PREDS_TAH  | 1.2%  | 0.452 | 0.494 | 0.542 | 0.323 | 0.513 | 0.542 (RF)   |
| JELUD_TAH  | 2.5%  | 0.760 | 0.527 | 0.748 | 0.727 | 0.766 | 0.766 (ADA)  |
| FIBR_JELUD | 4.2%  | 0.656 | 0.502 | 0.779 | 0.716 | 0.747 | 0.779 (RF)   |
| A_V_BLOK   | 3.4%  | 0.751 | 0.486 | 0.756 | 0.753 | 0.715 | 0.756 (RF)   |
| OTEK_LANC  | 9.4%  | 0.708 | 0.551 | 0.767 | 0.717 | 0.733 | 0.767 (RF)   |
| RAZRIV     | 3.2%  | 0.765 | 0.582 | 0.798 | 0.791 | 0.772 | 0.798 (RF)   |
| DRESSLER   | 4.4%  | 0.611 | 0.516 | 0.617 | 0.588 | 0.558 | 0.617 (RF)   |
| ZSN        | 23.2% | 0.677 | 0.581 | 0.741 | 0.685 | 0.742 | 0.742 (ADA)  |
| REC_IM     | 9.4%  | 0.646 | 0.527 | 0.679 | 0.633 | 0.642 | 0.679 (RF)   |
| P_IM_STEN  | 8.7%  | 0.637 | 0.534 | 0.695 | 0.690 | 0.654 | 0.695 (RF)   |
| Mean       | -     | 0.667 | 0.530 | 0.715 | 0.662 | 0.683 | -            |

*A Dual-Metric Framework for Rare-Event Prediction of Myocardial Infarction Complications:  
Discrimination, Threshold-Selected Detection, and Calibrated Multiclass Mortality*

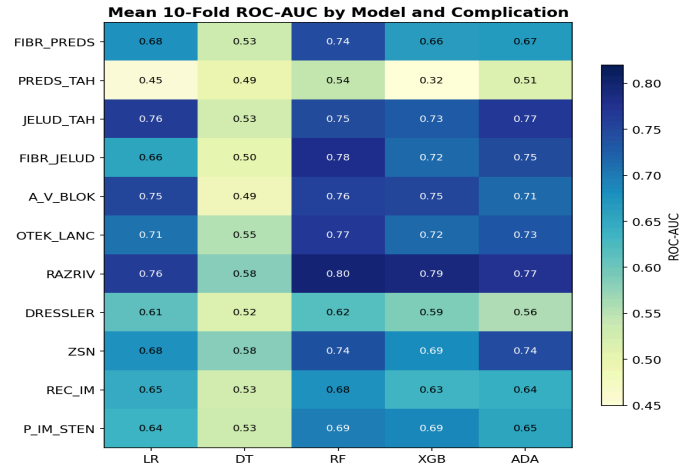


Figure 3. Mean 10-fold ROC-AUC by model (columns) and complication (rows). Random Forest has the highest mean AUC; it is near chance for supraventricular tachycardia. The single Decision Tree is nearly chance throughout.

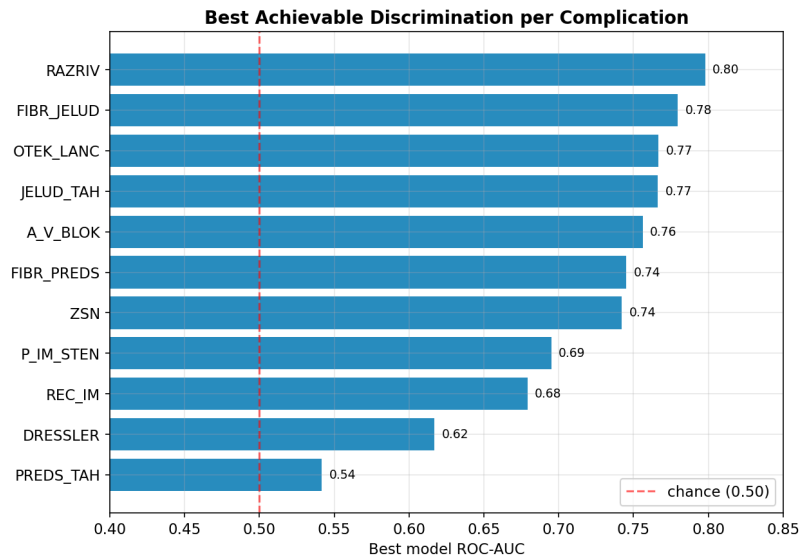


Figure 4. Best-model ROC-AUC for each binary target, sorted. Myocardial rupture and ventricular fibrillation are the most predictable; supraventricular tachycardia is essentially at chance (dashed line = 0.50).

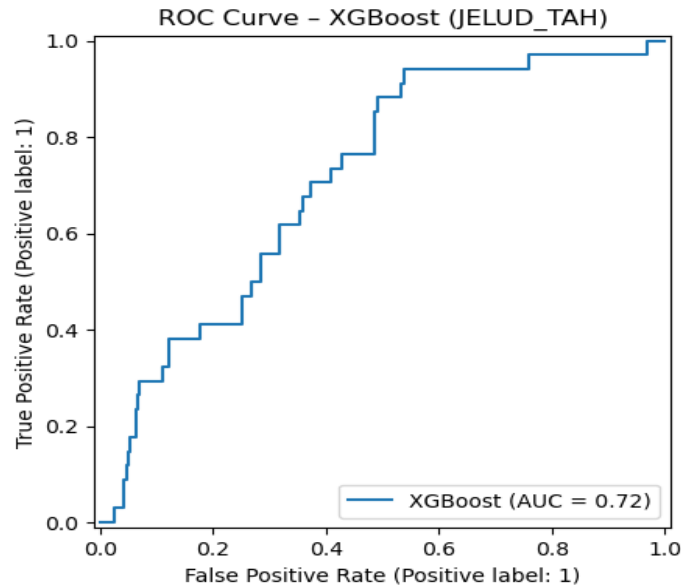


*Figure 5. Example held-out ROC curve (XGBoost, ventricular tachycardia), illustrating the moderate discrimination typical of the acute-arrhythmia targets.*

### Framework Validation: Detection Axis (F1) and the Dual-Metric Split

The second axis of the framework evaluates positive-class detection at a fixed default threshold of 0.5. Table 5 reports the mean cross-validated F1. Logistic Regression won F1 on every binary target (mean F1 0.211) because its default 0.5 cutoff produced a nonzero precision-recall balance (not a clinically chosen alert threshold), whereas Random Forest, the AUC champion, collapsed to a mean F1 of 0.020 and scored exactly zero on ten of eleven targets (the exception is chronic heart failure). Under strong imbalance and a 0.5 cutoff, Random Forest correctly ranked positives but rarely predicted the positive class, so recall and F1 dropped to 0. Even the best F1 values were low in absolute terms; the only respectable F1 was chronic heart failure at 23% prevalence, the least imbalanced target. Figure 6 contrasts the mean AUC and the mean F1. Because these F1 values use the default 0.5 threshold, they understate achievable detection performance; the next subsection re-evaluates detection using F1-optimal thresholds selected from the training data.

**Table 5. Mean 10-fold cross-validated F1 by target and model. Random Forest attains the highest AUC yet the lowest F1, scoring zero on ten of eleven targets.**

| Target     | LR    | DT    | RF    | XGB   | ADA   | Best (model) |
|------------|-------|-------|-------|-------|-------|--------------|
| FIBR PREDS | 0.276 | 0.157 | 0.000 | 0.176 | 0.051 | 0.276 (LR)   |
| PREDS TAH  | 0.013 | 0.000 | 0.000 | 0.000 | 0.000 | 0.013 (LR)   |
| JELUD TAH  | 0.156 | 0.076 | 0.000 | 0.000 | 0.000 | 0.156 (LR)   |
| FIBR JELUD | 0.168 | 0.045 | 0.000 | 0.054 | 0.054 | 0.168 (LR)   |
| A V BLOK   | 0.204 | 0.000 | 0.000 | 0.069 | 0.102 | 0.204 (LR)   |
| OTEK LANC  | 0.303 | 0.183 | 0.000 | 0.223 | 0.208 | 0.303 (LR)   |
| RAZRIV     | 0.213 | 0.163 | 0.000 | 0.056 | 0.109 | 0.213 (LR)   |
| DRESSLER   | 0.107 | 0.069 | 0.000 | 0.000 | 0.000 | 0.107 (LR)   |

*A Dual-Metric Framework for Rare-Event Prediction of Myocardial Infarction Complications:  
Discrimination, Threshold-Selected Detection, and Calibrated Multiclass Mortality*

|             |       |       |       |       |       |            |
|-------------|-------|-------|-------|-------|-------|------------|
| ZSN         | 0.424 | 0.359 | 0.222 | 0.424 | 0.415 | 0.424 (LR) |
| REC IM      | 0.225 | 0.140 | 0.000 | 0.106 | 0.000 | 0.225 (LR) |
| P IM STEN   | 0.236 | 0.147 | 0.000 | 0.153 | 0.000 | 0.236 (LR) |
| <b>Mean</b> | 0.211 | 0.122 | 0.020 | 0.115 | 0.085 | -          |

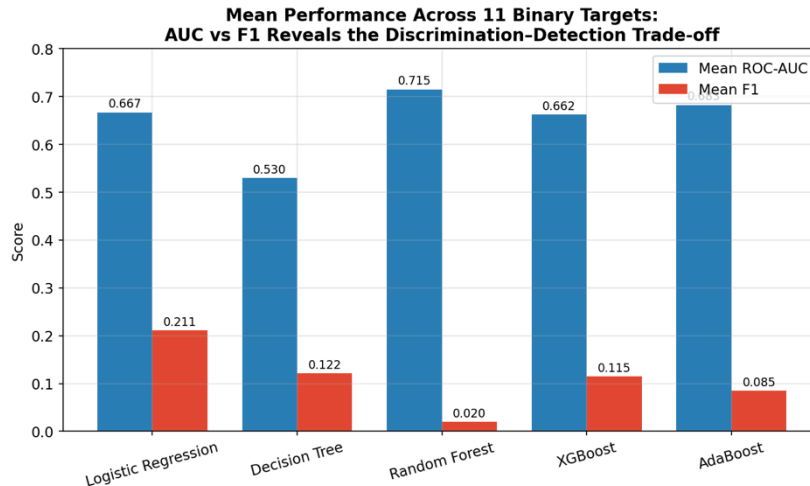


**Figure 6.** Mean ROC-AUC (blue) and mean F1 (red) per model, averaged over the eleven binary targets. The near-vanishing F1 of Random Forest, despite its top AUC, is the signature of the discrimination-detection trade-off.

Table 6 summarizes the dual-metric split at the framework level. Random Forest and Logistic Regression occupy opposite corners of the performance map: Across the 11 binary targets treated as the units of comparison, Random Forest is first on the discrimination axis (average rank 1.27) but last on the detection axis (4.55), while Logistic Regression is first on detection (average rank 1.00, best on every binary target) but mid-pack on discrimination (3.18). This inversion is a consistent pattern across the eleven binary complications, not a single-target anomaly, and it is the principal empirical finding of the framework: single-metric model selection is inadequate for these rare-event targets. The two boosting methods occupy the middle ground and represent the most balanced single-model compromise, with XGBoost ranking third on AUC and second on F1. Figure 7 maps all five models in this two-dimensional rank space.

**Table 6.** Dual-metric framework summary: average Friedman ranks on the discrimination axis (AUC) and detection axis (F1) across the 11 binary targets (the multiclass mortality target is summarized separately because its F1 is weighted). Lower is better; the axes select different optimal models.

| Model               | Average AUC rank (1 = best) | Average F1 rank (1 = best) |
|---------------------|-----------------------------|----------------------------|
| Random Forest       | 1.27                        | 4.55                       |
| AdaBoost            | 2.55                        | 3.45                       |
| XGBoost             | 3.18                        | 2.91                       |
| Logistic Regression | 3.18                        | 1.00                       |
| Decision Tree       | 4.82                        | 3.09                       |

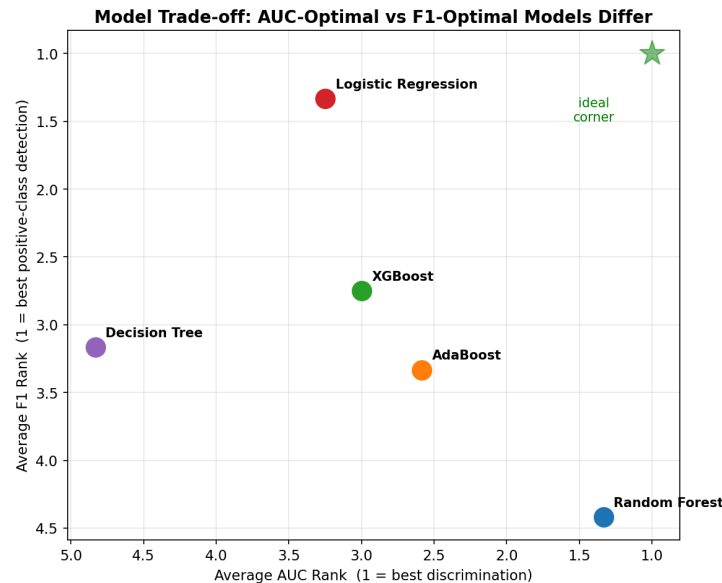


**Figure 7.** Average Friedman ranks on AUC (horizontal) and F1 (vertical); rank 1 is best on each axis, so lower values are better. Random Forest is best on AUC and worst on F1; Logistic Regression is the reverse. Boosting methods lie between them.

### Threshold Selection and Detection Metrics

Because a default 0.5 cutoff is not a clinically chosen operating point, detection is re-evaluated with operating points selected using training data only. Within each training fold, the threshold that maximizes F1 is chosen on the training scores and then applied to the held-out fold, and performance is reported as sensitivity and positive predictive value at that stated threshold rather than as F1 at 0.5 alone. Threshold-free detection is additionally summarized with average precision (AP). Using random seed 42 and the same 111 features as Tables 4-7 (including early-course variables), mean 10-fold AP across the 11 binary targets was 0.225 for Random Forest, 0.220 for AdaBoost, 0.217 for Logistic Regression, 0.216 for XGBoost, and 0.152 for the Decision Tree. Training-only F1-optimal thresholds raised ensemble sensitivity well above the near-zero default-0.5 detection of Random Forest: mean sensitivity/PPV was 0.322/0.163 for Logistic Regression versus 0.107/0.202 for Random Forest. For atrial fibrillation, Logistic Regression reached a sensitivity of 0.35 (PPV 0.23) at the training-selected threshold, compared with Random Forest sensitivity of 0.10 (PPV 0.35); for myocardial rupture, the corresponding figures were 0.25/0.13 versus 0.08/0.23. Ranking (AUC) is unchanged by the choice of threshold.

### Framework Validation: Multiclass Lethal-Outcome Prediction

The framework's third routed use case, multiclass mortality classification, requires a single classifier across eight outcome categories, so weighted F1 is dominated by survival and does not provide evidence of per-class death detection. All other outcome columns were removed to prevent leakage. Table 7 reports cross-validated and held-out results. In Table 7, XGBoost led held-out accuracy (0.874), weighted F1 (0.836), and macro-AUC (0.798), and led cross-validated macro AUC (0.858) and weighted F1 (0.818);

Random Forest was slightly higher on cross-validated accuracy (0.861 versus 0.860). Table 12 reports per-class metrics on the same 80/20 split, so the weighted F1 and accuracy there are not required to match Table 7 to three decimal places. Random Forest was a close second, and both ensembles substantially outperformed Logistic Regression on discrimination. Importantly, the multiclass weighted F1 values of 0.81-0.84 were far better than the binary F1 values, because the dominant survival class is large and well predicted, and the weighted average is buoyed accordingly. The Decision Tree again illustrated the ranking-threshold split, achieving respectable accuracy (0.756) but near-chance macro-AUC (0.534). The comparison is shown in Figure 8. Because weighted F1 and accuracy are dominated by the survival class, macro-F1, balanced accuracy, and per-class recall, precision, and support are also reported; these confirm that the minority death classes remain hard to detect, so XGBoost's value lies mainly in ranking (macro-AUC near 0.80) rather than in reliable per-class death classification. The high test accuracy must be interpreted in light of the 15.9% overall mortality rate: a survival-only predictor would already achieve about 84% accuracy, so the value added by XGBoost lies in its macro-AUC near 0.80. Figure 9 shows the Random Forest confusion matrix, which vividly demonstrates the imbalance problem - almost every case is assigned to the majority survival class, so high accuracy coexists with weak minority-class detection.

**Table 7. Lethal-outcome (eight-class) results under 10-fold cross-validation and on the held-out test set (F1-w = weighted F1). On that evaluation, XGBoost leads held-out accuracy, weighted F1, and macro-AUC; Random Forest is slightly higher on cross-validated accuracy. Per-class metrics are in Table 12.**

| Model               | CV Acc | CV F1-w | CV AUC-macro | Test Acc | Test F1-w | Test AUC-macro |
|---------------------|--------|---------|--------------|----------|-----------|----------------|
| Logistic Regression | 0.649  | 0.715   | 0.662        | 0.621    | 0.696     | 0.712          |
| Decision Tree       | 0.782  | 0.782   | 0.557        | 0.756    | 0.754     | 0.534          |
| Random Forest       | 0.861  | 0.808   | 0.802        | 0.862    | 0.809     | 0.764          |
| XGBoost             | 0.860  | 0.818   | 0.858        | 0.874    | 0.836     | 0.798          |
| AdaBoost            | 0.851  | 0.810   | 0.695        | 0.850    | 0.808     | 0.658          |

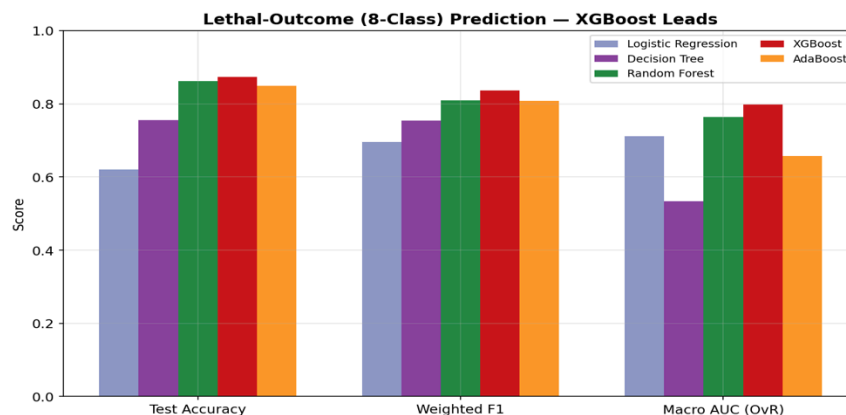


*Figure 8. Held-out test accuracy, weighted F1, and macro one-vs-rest AUC for the eight-class lethal-outcome task. XGBoost leads these three held-out metrics.*

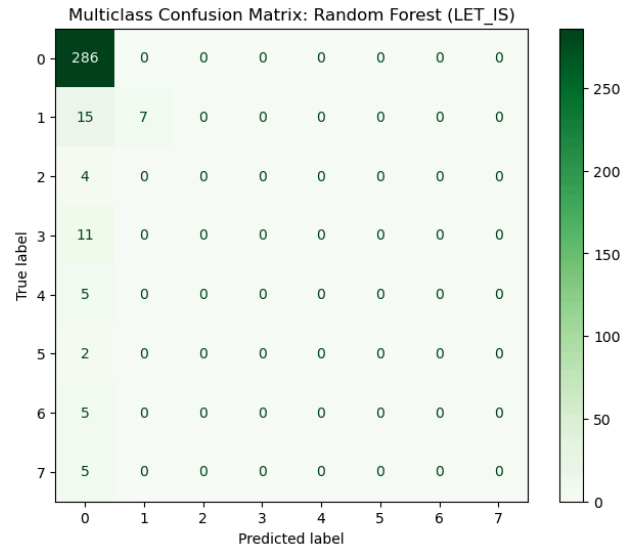


*Figure 9. Held-out confusion matrix for Random Forest on the lethal-outcome task (class 0 = survival; classes 1 to 7 = distinct causes of death). Almost every case is assigned to the majority class for survival, explaining why high accuracy coexists with weak minority-class detection.*

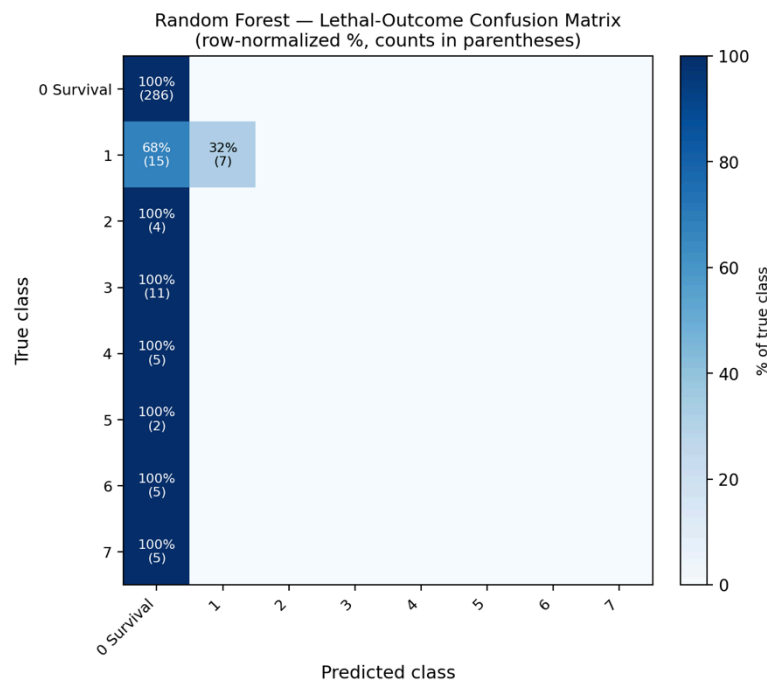


*Figure 10. Random Forest lethal-outcome confusion matrix, row-normalized to percentages with counts in parentheses (class 0 = survival; classes 1-7 = causes of death). Almost all mass falls in the survival column, so high overall accuracy coexists with near-zero recall for the death classes.*

## Statistical Validation of the Dual-Metric Framework

Sep 2026  
Vol 11, No 5.

Per-fold Friedman and Wilcoxon tests (Table 8) are reported only for descriptive, per-target transparency, because overlapping cross-validation folds from a single dataset are not independent experimental units (Demsar, 2006). In those descriptive tests, model choice was associated with differences for ten of twelve targets on AUC and eleven of twelve on F1; the exceptions were supraventricular tachycardia on both axes and, on AUC only, Dressler's syndrome (borderline,  $p = 0.051$ ). The primary analysis treats the 11 binary clinical targets as the comparison units. Across those 11 targets, the Friedman test was significant for both AUC (chi-square = 28.87,  $p < 0.001$ ; Random Forest best, average rank 1.27) and F1 (chi-square = 32.32,  $p < 0.001$ ; Logistic Regression best, average rank 1.00). Holm-corrected Wilcoxon tests across targets showed Random Forest significantly ahead of Logistic Regression on AUC and Logistic Regression significantly ahead of Random Forest on F1 (Table 9; full pairwise results in Supplementary Tables S1 and S2). Supraventricular tachycardia remained underpowered ( $n+ = 20$ ).

**Table 8. Descriptive Friedman tests across the five models, per target, on paired per-fold AUC and F1 scores (not the primary comparison). Differences are significant for ten of twelve targets on AUC and eleven of twelve on F1; the exceptions are supraventricular tachycardia on both axes and Dressler's syndrome on AUC ( $p = 0.051$ ).**

| Target     | chi2 (AUC) | p (AUC) | Sig. AUC | chi2 (F1) | p (F1) | Sig. F1 |
|------------|------------|---------|----------|-----------|--------|---------|
| FIBR PREDS | 25.76      | <0.001  | Yes      | 29.19     | <0.001 | Yes     |
| PREDS TAH  | 5.84       | 0.212   | No       | 4.00      | 0.406  | No      |
| JELUD TAH  | 16.78      | 0.0021  | Yes      | 23.92     | <0.001 | Yes     |
| FIBR JELUD | 28.88      | <0.001  | Yes      | 20.17     | 0.0005 | Yes     |
| A V BLOK   | 21.57      | 0.0002  | Yes      | 25.92     | <0.001 | Yes     |
| OTEK LANC  | 24.96      | 0.0001  | Yes      | 23.61     | 0.0001 | Yes     |
| RAZRIV     | 16.72      | 0.0022  | Yes      | 14.47     | 0.0059 | Yes     |
| DRESSLER   | 9.44       | 0.051   | No       | 28.61     | <0.001 | Yes     |
| ZSN        | 29.52      | <0.001  | Yes      | 20.56     | 0.0004 | Yes     |
| REC IM     | 17.76      | 0.0014  | Yes      | 29.13     | <0.001 | Yes     |
| P IM STEN  | 20.08      | 0.0005  | Yes      | 32.26     | <0.001 | Yes     |
| LET IS     | 36.32      | <0.001  | Yes      | 33.04     | <0.001 | Yes     |

### Corrected Statistical Comparison: Targets as Units of Comparison

Treating the clinical targets as the independent units of comparison, the setting for which the Friedman and Wilcoxon procedures are appropriate, the dual-metric split is supported. Across the 11 binary targets,

the Friedman test was highly significant on both axes (Table 9). On AUC, Random Forest significantly outperformed Decision Tree, XGBoost, AdaBoost, and Logistic Regression (Holm-corrected  $p \leq 0.03$ ); on F1, Logistic Regression significantly outperformed all four other models (Holm-corrected  $p < 0.01$ ), and Random Forest was significantly worse than Logistic Regression, Decision Tree, and XGBoost. The complete pairwise Holm-corrected results across the 11 binary targets (the primary comparison) are in Supplementary Tables S1 and S2. Per-fold pairwise tests are not the primary inference and are not reproduced in the supplement.

**Table 9. Primary statistical comparison treating the 11 binary targets as units (Friedman with Holm-corrected Wilcoxon post hoc; Nemenyi critical difference = 1.84).**

| Metric  | Friedman chi-square | p       | Best model (avg rank)      | Worst model (avg rank) |
|---------|---------------------|---------|----------------------------|------------------------|
| ROC-AUC | 28.87               | < 0.001 | Random Forest (1.27)       | Decision Tree (4.82)   |
| F1      | 32.32               | < 0.001 | Logistic Regression (1.00) | Random Forest (4.55)   |

## RESULTS: FEATURE IMPORTANCE AND INTERPRETATION

SHAP and feature-importance analysis produced leading predictors that align closely with clinical intuition, lending face validity to the models (Table 10). The recurring importance of age, admission systolic blood pressure (a proxy for hemodynamic compromise and cardiogenic shock), and creatine phosphokinase (a marker of infarct size) is consistent with decades of cardiological evidence. Outcome-specific signals were also sensible: ventricular fibrillation was driven by fibrillation already present at admission and by creatine phosphokinase; myocardial rupture by advanced age and low arrival blood pressure; recurrent infarction by prior MI and coronary history; and pulmonary edema by admission edema and heart-failure history. That the leading features overlap with established risk factors is reassuring for face validity, but SHAP describes how each fitted model uses variables rather than establishing causal or clinical relationships; correlated or temporally overlapping variables can receive misleading importance, so these rankings are not a substitute for independent clinical validation. Figure 11 shows a representative feature-importance ranking.

**Table 10. Leading predictors per outcome from SHAP and tree-based feature importance.**

| Target     | Condition                    | Top contributing features                           |
|------------|------------------------------|-----------------------------------------------------|
| FIBR_PREDS | Atrial fibrillation          | Age; systolic BP (ICU); prior atrial fibrillation   |
| PREDS_TAH  | Supraventricular tachycardia | Systolic BP; heart-rate markers (very small n)      |
| JELUD_TAH  | Ventricular tachycardia      | VT at admission; conduction abnormalities           |
| FIBR_JELUD | Ventricular fibrillation     | VF at admission; CPK (myocardial damage)            |
| A_V_BLOK   | Atrioventricular block       | AV block on ECG; age; conduction history            |
| OTEK_LANC  | Pulmonary edema              | Pulmonary edema at admission; heart-failure history |
| RAZRIV     | Myocardial rupture           | Age; low systolic BP on arrival                     |
| DRESSLER   | Dressler's syndrome          | Delayed treatment; AST/ALT elevation                |
| ZSN        | Chronic heart failure        | Heart failure at admission; age                     |

|           |                         |                                                  |
|-----------|-------------------------|--------------------------------------------------|
| REC_IM    | Recurrent MI            | Prior MI; angina history; coronary heart disease |
| P IM STEN | Post-MI angina/stenosis | Angina functional class; prior CHD history       |
| LET IS    | Lethal outcome          | Age; CPK; systolic BP (multiclass)               |

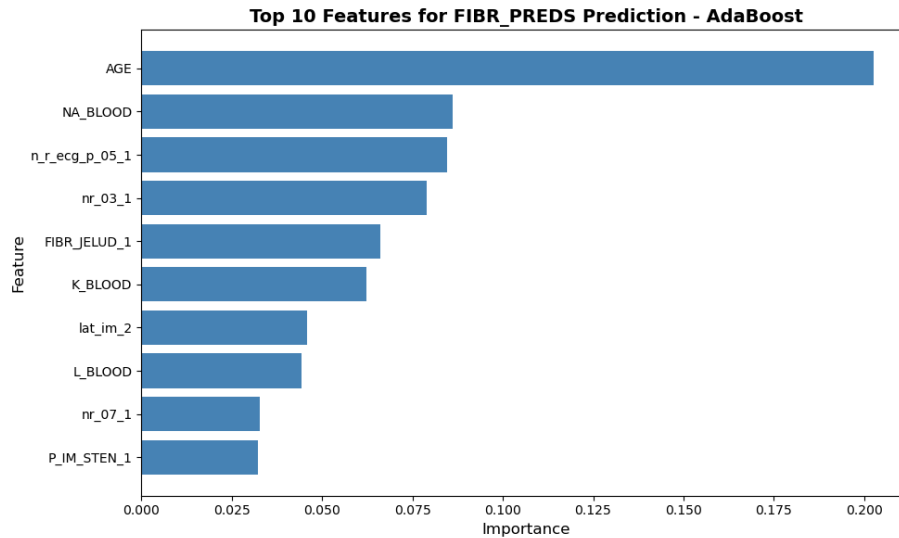


Figure 11. Representative feature-importance ranking (AdaBoost, atrial fibrillation). Age dominates, followed by serum sodium, electrocardiographic rhythm and ectopy indicators, and potassium - features that overlap established arrhythmia risk factors (face validity only).

### Calibration and Uncertainty

Because a risk-ranking tool should be calibrated and its estimates bounded, Brier scores, calibration intercepts and slopes, and bootstrap 95% confidence intervals are reported on the held-out test set ( $n = 340$ ). Averaged across the 11 binary targets, mean Brier scores were 0.073 (XGBoost), 0.111 (Random Forest), 0.138 (AdaBoost), 0.163 (Logistic Regression), and 0.194 (Decision Tree). Calibration slopes were closest to 1 for Random Forest (mean slope 1.19) and AdaBoost (1.42); Logistic Regression was over-confident on average (mean slope 0.28). Held-out bootstrap 95% CIs for Random Forest AUC were wide for the rarest targets: atrial fibrillation 0.74 (0.65-0.82), myocardial rupture 0.76 (0.58-0.92), and supraventricular tachycardia 0.45 (0.16-0.87); the corresponding AP intervals were 0.31 (0.17-0.46), 0.26 (0.06-0.52), and 0.02 (0.00-0.06). As an immediate check on the stability of the original cross-validated estimates, Table 11 gives t-based 95% confidence intervals for Random Forest AUC and Logistic Regression F1; the intervals are wide for the rarest targets, underscoring that point estimates such as AUC 0.798 should not be over-interpreted. Figure 12 shows held-out reliability diagrams for atrial fibrillation, myocardial rupture, and chronic heart failure.

Held-out calibration (reliability) diagrams, seed 42

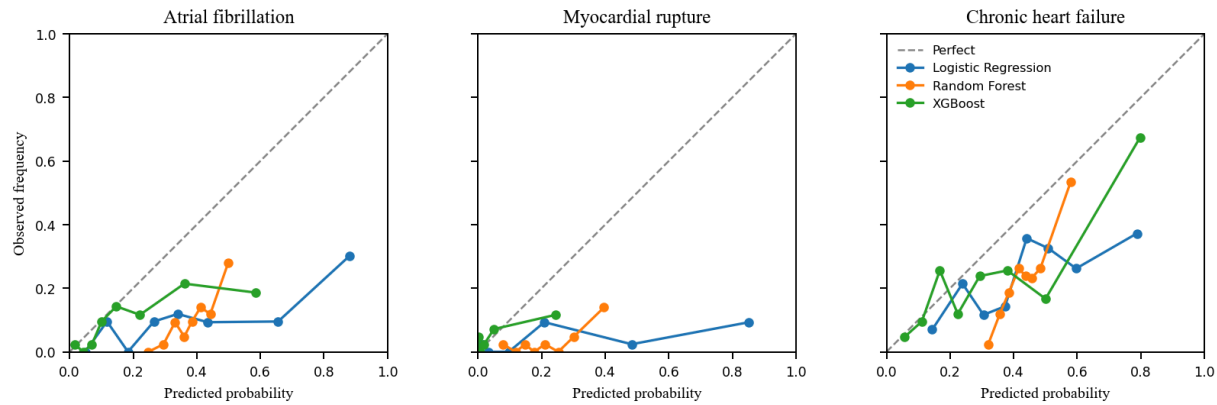


Figure 12. Held-out calibration (reliability) diagrams for three representative targets (seed 42). The dashed line is perfect calibration; points above the line are under-confident, and points below are over-confident.

Table 11. t-based 95% confidence intervals (across the 10 CV folds) for the ranking model (Random Forest AUC) and the detection model (Logistic Regression F1).

| Target                       | Random Forest AUC (95% CI) | Logistic Regression F1 (95% CI) |
|------------------------------|----------------------------|---------------------------------|
| Atrial fibrillation          | 0.745 (0.700-0.790)        | 0.276 (0.220-0.331)             |
| Supraventricular tachycardia | 0.542 (0.388-0.696)        | 0.013 (0.000-0.039)             |
| Ventricular tachycardia      | 0.748 (0.693-0.802)        | 0.156 (0.084-0.228)             |
| Ventricular fibrillation     | 0.779 (0.731-0.828)        | 0.168 (0.109-0.228)             |
| Atrioventricular block       | 0.756 (0.667-0.845)        | 0.204 (0.154-0.253)             |
| Pulmonary edema              | 0.767 (0.713-0.820)        | 0.303 (0.240-0.365)             |
| Myocardial rupture           | 0.798 (0.744-0.851)        | 0.213 (0.176-0.251)             |
| Dressler's syndrome          | 0.617 (0.530-0.703)        | 0.107 (0.083-0.132)             |
| Chronic heart failure        | 0.741 (0.704-0.778)        | 0.424 (0.378-0.471)             |
| Recurrent MI                 | 0.679 (0.620-0.738)        | 0.225 (0.188-0.262)             |
| Post-infarction angina       | 0.695 (0.657-0.733)        | 0.236 (0.194-0.279)             |

### Per-Class Mortality Performance

For the eight-class lethal-outcome task, weighted F1 and accuracy are dominated by the survival class, so Table 12 reports balanced accuracy, macro-F1, weighted F1, and per-class recall and precision on the held-out test set ( $n = 340$ ; class support 286 / 22 / 4 / 11 / 5 / 2 / 5 / 5), together with a row-normalized confusion matrix for Random Forest (Figure 10). Random Forest balanced accuracy was 0.23 and macro-F1 0.22 (recall 0.87 for survival, 0.45 for class 1, and near zero for most other death classes). XGBoost was the strongest of the five models on this task but still reached only balanced accuracy 0.26 and macro-F1 0.31 (weighted F1 0.85); per-class precision for the death classes was similarly sparse (XGBoost precision 0.77 / 1.00 / 0.67 for classes 1-3 and 0.00 or 0.50 for the remaining death classes).

XGBoost's value, therefore, lies in ranking, not in reliably detecting deaths per class, and any early-warning claim is limited to risk ranking.

**Table 12. Held-out per-class lethal-outcome performance (test n = 340; class 0 = survival; support 286/22/4/11/5/2/5/5). Balanced accuracy = mean per-class recall. Precision is reported beside recall for classes 0-7.**

| Model               | Bal. acc. | Macro-F1 | W-F1 | Recall (0-7)                                          | Precision (0-7)                                       |
|---------------------|-----------|----------|------|-------------------------------------------------------|-------------------------------------------------------|
| Logistic Regression | 0.28      | 0.22     | 0.76 | 0.80 / 0.36 / 0.50 / 0.18 / 0.20 / 0.00 / 0.00 / 0.20 | 0.93 / 0.33 / 0.13 / 0.14 / 0.07 / 0.00 / 0.00 / 0.07 |
| Decision Tree       | 0.22      | 0.14     | 0.47 | 0.35 / 0.32 / 0.25 / 0.36 / 0.00 / 0.50 / 0.00 / 0.00 | 0.97 / 0.88 / 0.01 / 0.05 / 0.00 / 0.03 / 0.00 / 0.00 |
| Random Forest       | 0.23      | 0.22     | 0.79 | 0.87 / 0.45 / 0.00 / 0.27 / 0.00 / 0.00 / 0.00 / 0.20 | 0.91 / 0.59 / 0.00 / 0.10 / 0.00 / 0.00 / 0.00 / 0.17 |
| AdaBoost            | 0.16      | 0.17     | 0.81 | 1.00 / 0.32 / 0.00 / 0.00 / 0.00 / 0.00 / 0.00 / 0.00 | 0.86 / 0.88 / 0.00 / 0.00 / 0.00 / 0.00 / 0.00 / 0.00 |
| XGBoost             | 0.26      | 0.31     | 0.85 | 1.00 / 0.45 / 0.25 / 0.18 / 0.00 / 0.00 / 0.00 / 0.20 | 0.89 / 0.77 / 1.00 / 0.67 / 0.00 / 0.00 / 0.00 / 0.50 |

## Discussion

The principal contribution of this study is not a new classifier but a novel evaluation and selection framework for imbalanced, multi-outcome clinical prediction. By treating discrimination and detection as two complementary axes and validating both statistically across the 11 binary targets (with multiclass mortality reported separately), the framework exposes a failure mode that is easily missed in conventional single-metric benchmarks: the AUC-optimal model is systematically not the deployable model for rare-event alerting. Three clinically relevant conclusions follow. First, the tabulated 111-feature set (admission plus early-course variables) carries a real but limited predictive signal for MI complications: the best AUC values of roughly 0.70 to 0.80 for rupture, ventricular fibrillation, pulmonary edema, atrioventricular block, and atrial fibrillation are enough to stratify risk through the framework's ranking route (Random Forest), but not to rule an event in or out for an individual patient. Second, mortality-cause prediction is moderate in terms of discrimination through the framework's multiclass route (XGBoost, macro-AUC near 0.80), but per-class detection of individual death causes is weak; any early-warning use would therefore depend on ranking rather than reliable per-class classification and would require calibration and prospective evaluation before deployment. Third, and most important operationally, the target-routed model-selection protocol assigns a different recommended model to each clinical intent - ranking, alerting, or multiclass mortality - and selecting without specifying the intent will systematically select the wrong algorithm.

The dual-metric split observed here is not an artifact of a particular dataset or hyperparameter choice but a structural consequence of class imbalance interacting with default decision thresholds. Random Forest produces well-ordered probability rankings, hence high AUC, but at a fixed 0.5 cutoff, its predicted probabilities for rare positives rarely exceed the threshold, so it predicts the negative class almost everywhere, and its recall and F1 vanish. Logistic Regression, especially with balanced class weights, places its decision boundary to capture some positives, trading precision for recall and achieving a nonzero F1. The framework makes this split visible and quantifiable rather than hidden behind a single leaderboard score. The corollary is that the F1 of Random Forest could likely be rescued by threshold tuning, probability calibration, or resampling, whereas the discrimination ceiling of Logistic Regression is more fundamental. Boosting methods strike the best out-of-the-box compromise, which is why they occupy the middle of both rank tables and why XGBoost dominates the multiclass route, where a single model must perform on both axes.

Relative to Baidya and Gupta (2024), the most important revision is evaluative rather than algorithmic. The earlier study concluded that XGBoost and a multilayer perceptron provided the most accurate predictions for atrial fibrillation and heart failure, respectively. Under the dual-metric framework, those accuracy figures mask a deeper pattern: for ten of the eleven binary complications, Random Forest achieves a high AUC yet zero F1 at the default 0.5 threshold, meaning it would never trigger an alert despite correctly ranking patients. Even for the two targets shared with the prior study, re-evaluation reveals complementary strengths: Random Forest remains the stronger ranker for atrial fibrillation (AUC 0.745 versus 0.676 for Logistic Regression), while Logistic Regression remains the stronger detector (F1 0.276 versus 0.000 for Random Forest). For heart failure, the least imbalanced target at 23% prevalence, both axes partially align, yet Logistic Regression still achieves the highest F1 (0.424), demonstrating that accuracy-based conclusions from the earlier work do not generalize across the full spectrum of complications or across clinical operating points.

The framework, therefore, maps each reporting intent to a model family rather than naming a single winner. Route 1 (risk ranking): Random Forest or gradient boosting, optimized for AUC, with a separately chosen threshold before any binary alert. Route 2 (default-threshold or cost-sensitive detection): Logistic Regression or a threshold-tuned ensemble, optimized for F1, sensitivity, or a clinically weighted recall, not for AUC by default. Route 3 (multiclass mortality ranking): XGBoost, reporting both macro-AUC and balanced accuracy or macro-F1, with accuracy interpreted against the 15.9% mortality rate. Across all routes, SHAP output describes model use of variables for face validity and is not a substitute for clinical validation. Accuracy should not be the headline metric in the presence of imbalance; ranking- and threshold-dependent detection should be reported together (Saito and Rehmsmeier, 2015).

Several limitations warrant caution in interpreting these findings; the corresponding analyses are listed under Recommendations for Future Work. All results derive from a single retrospective cohort of 1,700 patients, so external validation in an independent population is required before any clinical claim (Collins et al., 2015). With as few as 20 positive cases in some targets, estimates for the rarest complications are

highly variable, and the estimate for supraventricular tachycardia was underpowered. Tabulated models use 111 features, which include early-course treatments and the intensive-care blood-pressure reading, not a strictly pre-admission subset. Operating points are F1-optimal thresholds chosen inside training folds, not Youden-J or cost-sensitive clinical cutoffs. Calibration is reported as Brier scores, slopes, intercepts, bootstrap intervals, and reliability diagrams for three representative targets, not for every outcome. Tables 4-7 and Table 12 are not forced to three-decimal identity, because categorical encoding and evaluation details differ slightly between the discrimination tables and the per-class mortality table. Hyperparameters were only partially tuned. SHAP explanations are associational. The items below list the corresponding next analyses.

## RECOMMENDATIONS FOR FUTURE WORK

Future work should extend the present analyses in eight directions. First, re-run the full table set on a strictly pre-admission feature subset, dropping post-admission treatments and intensive-care-only fields, and report that second table beside Tables 4-7. Second, replace F1-optimal cutoffs with Youden-J and with cost-sensitive thresholds tied to explicit false-positive and false-negative penalties, and add full precision-recall curves alongside average precision. Third, apply a single categorical-encoding rule throughout, regenerate Tables 4-7 under that rule, log per-fold outlier counts, and print fold standard deviations in Tables 4 and 5. Fourth, use nested cross-validation or a more comprehensive hyperparameter search, and test whether resampling techniques such as SMOTE (He and Garcia, 2009; Chawla et al., 2002) can improve ensemble F1 without harming ranking. Fifth, draw reliability diagrams for all 11 binary targets, apply probability calibration (for example, Platt scaling or isotonic regression), redraw the Friedman rank map so that both axes increase in the same direction. Sixth, add serial vital-sign and laboratory trajectories, then validate on an independent cohort. Seventh, treat mortality as an ordinal or cost-sensitive detection problem aimed at the death classes rather than weighted F1. Eighth, obtain an independent clinical review of explanations and a prospective test of whether ranking queues or alerts change care. Until those steps are completed, the current results support method comparison in this cohort, not deployment.

## CONCLUSIONS

This paper presents a dual-metric evaluation framework for multi-target prediction of myocardial infarction complications and lethal outcome, extending an earlier accuracy-based study of the same cohort (Baidya and Gupta, 2024). The framework was applied to 1,700 patients and twelve clinical targets using five model families. Ranking performance and default-threshold detection diverge under severe class imbalance: Random Forest led discrimination (best AUC on nine of eleven binary targets, mean AUC 0.715, average rank 1.27) yet was weakest on default-threshold detection (mean F1 0.020, F1 = 0 on ten of eleven targets), while Logistic Regression led detection (best F1 on all eleven binary targets, mean F1 0.211). Training-only F1-optimal thresholds raised Logistic Regression mean sensitivity to 0.322 (PPV

0.163). XGBoost led the multiclass mortality ranking (test macro-AUC 0.798) but showed weak per-class death detection (balanced accuracy 0.26, macro-F1 0.31). Friedman tests treating the 11 binary targets as units were significant on both axes. SHAP rankings overlapped established risk factors (face validity only). Specify the clinical operating point first, evaluate ranking versus threshold-dependent detection second, and route to the matching model family third. Further analyses are listed under Recommendations for Future Work.

## ACKNOWLEDGMENTS

I sincerely thank Professor Vibhuti Gupta, Biostatistics and Data Science, UTMB, Texas, for his guidance, mentorship, and continued encouragement throughout this research. His support was instrumental in the earlier study (Baidya and Gupta, 2024) and in motivating me to extend that work into the dual-metric evaluation framework presented here. I also acknowledge the authors of the Myocardial Infarction Complications database (Golovenkin et al., 2020) for making the data publicly available for research, and the developers of the open-source scientific-computing ecosystem (Pedregosa et al., 2011) on which this study was built.

## REFERENCES

- Antman EM, Cohen M, Bernink PJLM, McCabe CH, Horacek T, Papuchis G, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. *JAMA* 2000;284(7):835-842.
- Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature Medicine* 2019;25(1):70-74.
- Baidya S, Gupta V. Predicting cardiac complications of myocardial infarction patients using machine learning. In: *Proceedings of the 2024 IEEE International Conference on Big Data (Big Data)*; 2024. p. 7454-7462.
- Breiman L. Random forests. *Machine Learning* 2001;45(1):5-32.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 2002;16:321-357.
- Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016:785-794.
- Cliff N. Dominance statistics: ordinal analyses to answer ordinal questions. *Psychological Bulletin* 1993;114(3):494-509.

- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Annals of Internal Medicine* 2015;162(1):55-63.
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.
- Demsar J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 2006;7:1-30.
- Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 1997;55(1):119-139.
- Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 1937;32(200):675-701.
- Golovenkin SE, Bac J, Chervov A, Mirkes EM, Orlova YV, Barillot E, et al. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *GigaScience* 2020;9(11): 128.
- Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, et al. Predictors of hospital mortality in the Global Registry of Acute Coronary Events. *Archives of Internal Medicine* 2003;163(19):2345-2353.
- He H, Garcia EA. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 2009;21(9):1263-1284.
- Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979;6(2):65-70.
- Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. *Journal of the American College of Cardiology* 2018;71(23):2668-2679.
- Killip T, Kimball JT. Treatment of myocardial infarction in a coronary care unit: a two-year experience with 250 patients. *American Journal of Cardiology* 1967;20(4):457-464.
- Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2020;2(1):56-67.
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 2017; 30:4765-4774.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 2011; 12:2825-2830.
- Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 2015;10(3):0118432.

Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018;104(14):1156-1164.

Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, et al. Fourth universal definition of myocardial infarction (2018). *Circulation* 2018;138(20):618-651.

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): 0174944.

Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1945;1(6):80-83.

## SUPPLEMENTARY MATERIAL

Table S1 and Table S2 present the complete Holm-corrected pairwise Wilcoxon signed-rank results for ROC-AUC and F1, respectively, across the 11 binary targets (treated as the units of comparison), including mean values, mean differences, raw and Holm-corrected p-values, and Cliff's delta effect sizes for all 10 model pairs.

**Table S1. Pairwise model comparison across the 11 binary targets - ROC-AUC (Wilcoxon signed-rank, Holm-corrected).**

| Comparison | Mean A | Mean B | $\Delta$ mean | p (raw) | p (Holm) | Sig. | Cliff's $\delta$ |
|------------|--------|--------|---------------|---------|----------|------|------------------|
| LR vs DT   | 0.667  | 0.530  | +0.137        | 0.0020  | 0.0137   | Yes  | +0.82            |
| LR vs RF   | 0.667  | 0.715  | -0.048        | 0.0049  | 0.0293   | Yes  | -0.37            |
| LR vs XGB  | 0.667  | 0.662  | +0.005        | 0.9658  | 0.9658   | No   | -0.09            |
| LR vs ADA  | 0.667  | 0.683  | -0.015        | 0.2783  | 0.8350   | No   | -0.12            |
| DT vs RF   | 0.530  | 0.715  | -0.185        | 0.0010  | 0.0098   | Yes  | -0.95            |
| DT vs XGB  | 0.530  | 0.662  | -0.133        | 0.0186  | 0.0742   | No   | -0.82            |
| DT vs ADA  | 0.530  | 0.683  | -0.153        | 0.0010  | 0.0098   | Yes  | -0.83            |
| RF vs XGB  | 0.715  | 0.662  | +0.053        | 0.0010  | 0.0098   | Yes  | +0.34            |
| RF vs ADA  | 0.715  | 0.683  | +0.032        | 0.0049  | 0.0293   | Yes  | +0.29            |
| XGB vs ADA | 0.662  | 0.683  | -0.020        | 0.4648  | 0.9297   | No   | -0.07            |

**Table S2. Pairwise model comparison across the 11 binary targets - F1 (Wilcoxon signed-rank, Holm-corrected).**

| Comparison | Mean A | Mean B | $\Delta$ mean | p (raw) | p (Holm) | Sig. | Cliff's $\delta$ |
|------------|--------|--------|---------------|---------|----------|------|------------------|
| LR vs DT   | 0.211  | 0.122  | +0.090        | 0.0010  | 0.0098   | Yes  | +0.55            |
| LR vs RF   | 0.211  | 0.020  | +0.191        | 0.0010  | 0.0098   | Yes  | +0.90            |
| LR vs XGB  | 0.211  | 0.115  | +0.097        | 0.0010  | 0.0098   | Yes  | +0.57            |
| LR vs ADA  | 0.211  | 0.085  | +0.126        | 0.0010  | 0.0098   | Yes  | +0.67            |
| DT vs RF   | 0.122  | 0.020  | +0.101        | 0.0039  | 0.0234   | Yes  | +0.67            |
| DT vs XGB  | 0.122  | 0.115  | +0.007        | 0.8457  | 0.8457   | No   | +0.09            |

*A Dual-Metric Framework for Rare-Event Prediction of Myocardial Infarction Complications:  
Discrimination, Threshold-Selected Detection, and Calibrated Multiclass Mortality*

|               |       |       |        |        |        |     |       |
|---------------|-------|-------|--------|--------|--------|-----|-------|
| DT vs ADA     | 0.122 | 0.085 | +0.036 | 0.1934 | 0.5801 | No  | +0.31 |
| RF vs XGB     | 0.020 | 0.115 | -0.094 | 0.0078 | 0.0391 | Yes | -0.60 |
| RF vs ADA     | 0.020 | 0.085 | -0.065 | 0.0312 | 0.1250 | No  | -0.42 |
| XGB vs<br>ADA | 0.115 | 0.085 | +0.029 | 0.2969 | 0.5938 | No  | +0.24 |