Predicting Lethal Outcomes After Myocardial Infarctions Using 24-hour Post-Admission Patient Data

Kathryn Lee
Brown University
GitHub Repository: https://github.com/nowyouleeme/data1030-final-project

1. Introduction

1.1 Significance of Myocardial Infarctions

Myocardial infarction (MI) presents as one of the most complex challenges in modern medicine. MI can occur with or without complications, and these complications can significantly impact long-term prognosis. While some complications have minimal effects, others can severely exacerbate the condition and lead to death. This variability complicates the ability of experienced specialists to predict complications reliably and promptly [1]. Therefore, accurately predicting whether a patient will experience lethal MI complications or not is essential for enabling early intervention and implementing effective preventive measures.

1.2 MI Complications Dataset

The MI complications dataset from the UC Irvine Machine Learning Repository includes data on patients admitted to the Krasnoyarsk Interdistrict Clinical Hospital in Russia between 1992 and 1995. It includes data from 1700 separate patients who were admitted for MI, with detailed medical records and information about complications following the event [2].

1.3 Current Research

A 2020 study by Golovenkin et al. leveraged this dataset to model MI complications using unsupervised elastic principal trees (EPT) [3]. The study aimed to organize the data into a multi-class series of clinical trajectories, representing different patterns of MI complication progression. While the predictive power of the resulting model is not directly evaluated using traditional metrics such as accuracy, precision, and recall, it is demonstrated through the model's ability to capture 52.4% of the variance in the dataset [3]. This indicates that the model effectively identifies key patterns and relationships, although it does not provide explicit classification predictions. Instead, the model’s strength lies in its ability to assist clinical decision-making by organizing and interpreting complex clinical data.

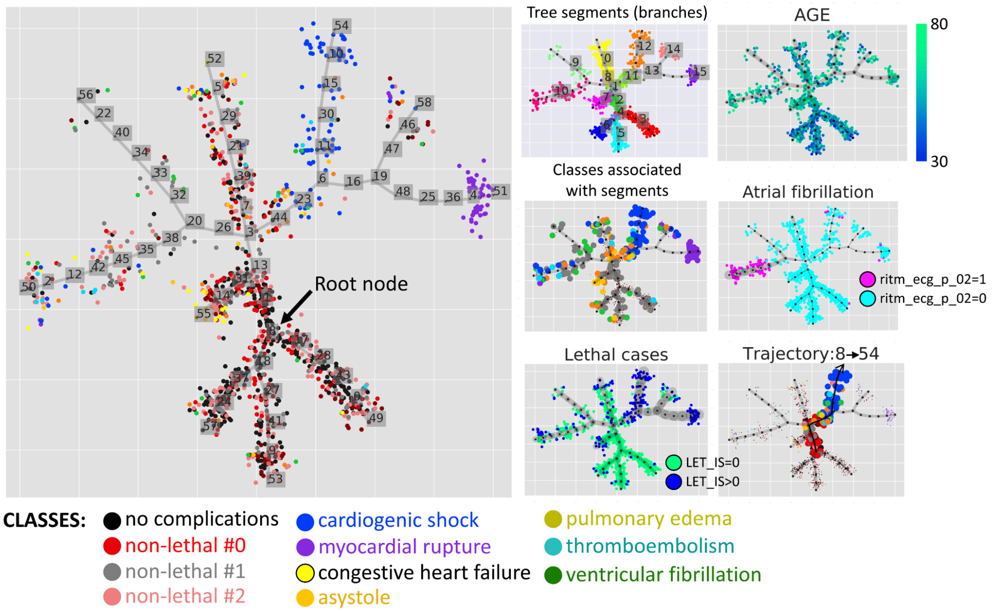


Figure 1. Summary of the EPT models used in the study, with the large panel showing how data points are distributed along the tree and the smaller panels highlight different types of data visualizations [3].

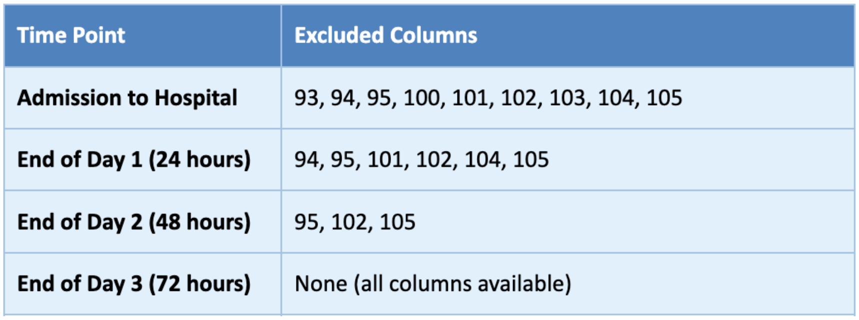
1.4 Project Focus

The focus of this project is to develop a binary classification supervised learning model that can predict whether or not a patient will experience a lethal complication 24 hours after their admission to the hospital for an MI. Focusing on the 24-hour period is ideal because it is early enough to detect critical complications that may arise soon after admission, but it also provides enough time to evaluate whether initial treatments are effective in stabilizing the patient. These are factors critical to any medical diagnostic tool aimed at improving patient outcomes and guiding timely interventions. By identifying patients at risk for lethal complications within this window, specialists can make informed decisions about treatment adjustments, potentially lowering mortality rates and increasing the likelihood of successful recovery.

2. Exploratory Data Analysis

2.1 Initial Processing

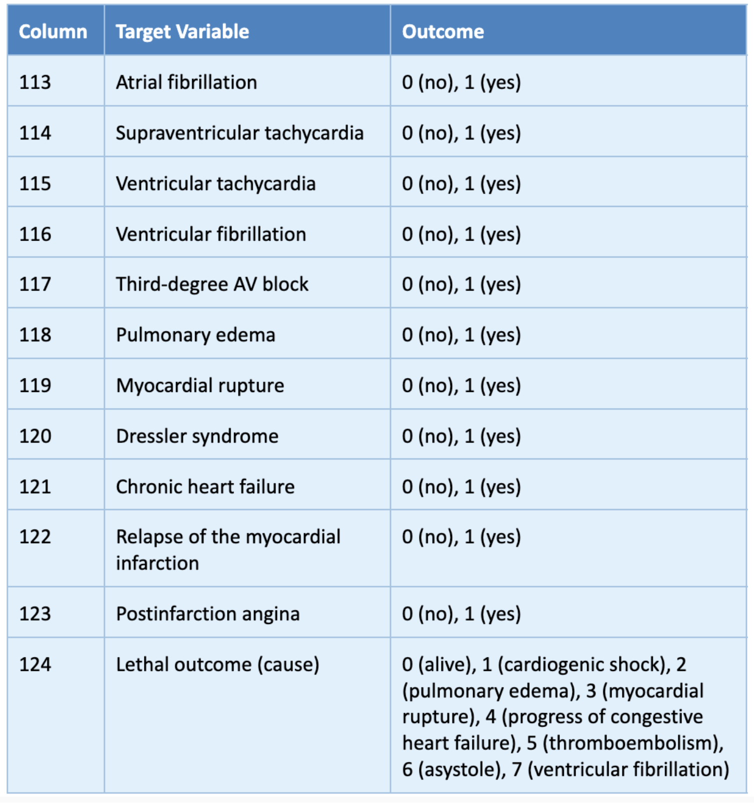
All values of the dataset are floats or integers. The dataset has 2-112 input columns, each representing features for predicting complications at different time points. Patient IDs (in column 1) were excluded, as they provide no predictive value, and some columns must also be excluded due to inaccessible data at the chosen time points:



Time Point	Excluded Columns
Admission to Hospital	93, 94, 95, 100, 101, 102, 103, 104, 105
End of Day 1 (24 hours)	94, 95, 101, 102, 104, 105
End of Day 2 (48 hours)	95, 102, 105
End of Day 3 (72 hours)	None (all columns available)

Table 1. Patient data exclusion by time point for analysis, advised by dataset creators.

This project predicts patient outcomes 24 hours after hospital admission, so columns 94, 95, 101, 102, 104, and 105 were excluded. The dataset also includes output columns (113-124) representing potential complications after an MI:



Column	Target Variable	Outcome
113	Atrial fibrillation	0 (no), 1 (yes)
114	Supraventricular tachycardia	0 (no), 1 (yes)
115	Ventricular tachycardia	0 (no), 1 (yes)
116	Ventricular fibrillation	0 (no), 1 (yes)
117	Third-degree AV block	0 (no), 1 (yes)
118	Pulmonary edema	0 (no), 1 (yes)
119	Myocardial rupture	0 (no), 1 (yes)
120	Dressler syndrome	0 (no), 1 (yes)
121	Chronic heart failure	0 (no), 1 (yes)
122	Relapse of the myocardial infarction	0 (no), 1 (yes)
123	Postinfarction angina	0 (no), 1 (yes)
124	Lethal outcome (cause)	0 (alive), 1 (cardiogenic shock), 2 (pulmonary edema), 3 (myocardial rupture), 4 (progress of congestive heart failure), 5 (thromboembolism), 6 (asystole), 7 (ventricular fibrillation)

Table 2. Target variables used for prediction of complications post-MI.

Due to the complexity of handling multivariate outcomes, columns 113-123 were dropped, leaving column 124 (lethal outcomes) as the target variable. This decision was made because column 124 captures a broad range of lethal outcomes, providing a more comprehensive risk assessment of patient mortality. After processing, the dataset contains 1700 data points and 106 columns (105 features, 1 target).

For EDA, the target variable was converted from multi-class to binary (alive vs. death due to MI complications) to simplify analysis and focus on general critical outcomes. The class balance was calculated and visualized, revealing a highly imbalanced distribution. This imbalance must be considered during data splitting for training, cross-validation, and testing.

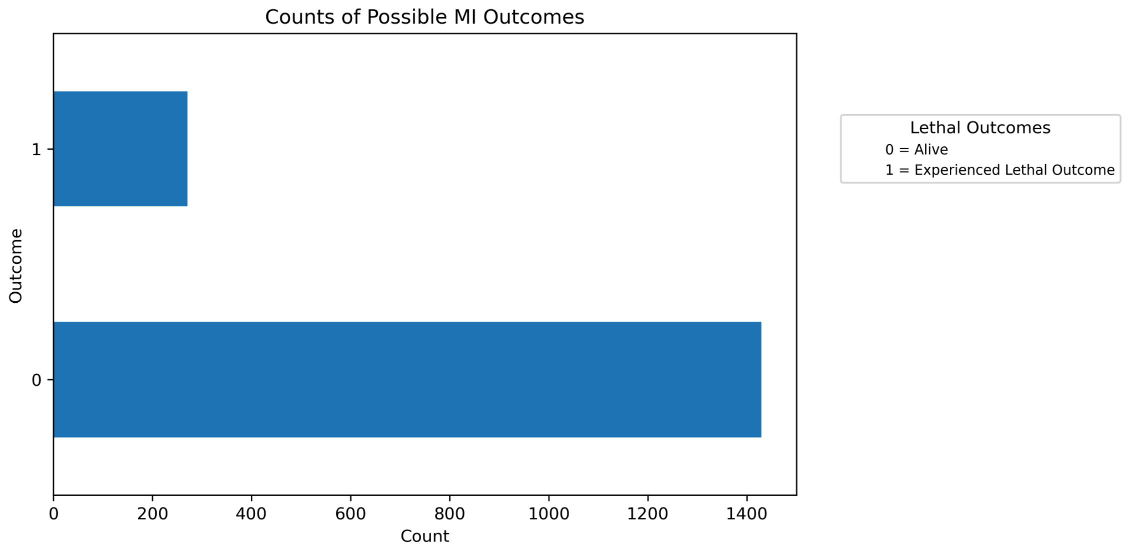


Figure 1. Imbalanced class distribution of the modified target variable. ~85% of patients are alive at 24 hours and ~15% experiencing lethal complications.

Each feature was then visualized against the target variable. Because of the high-dimensionality of the dataset, only one visualization for each type (continuous, ordinal, categorical) is included below:

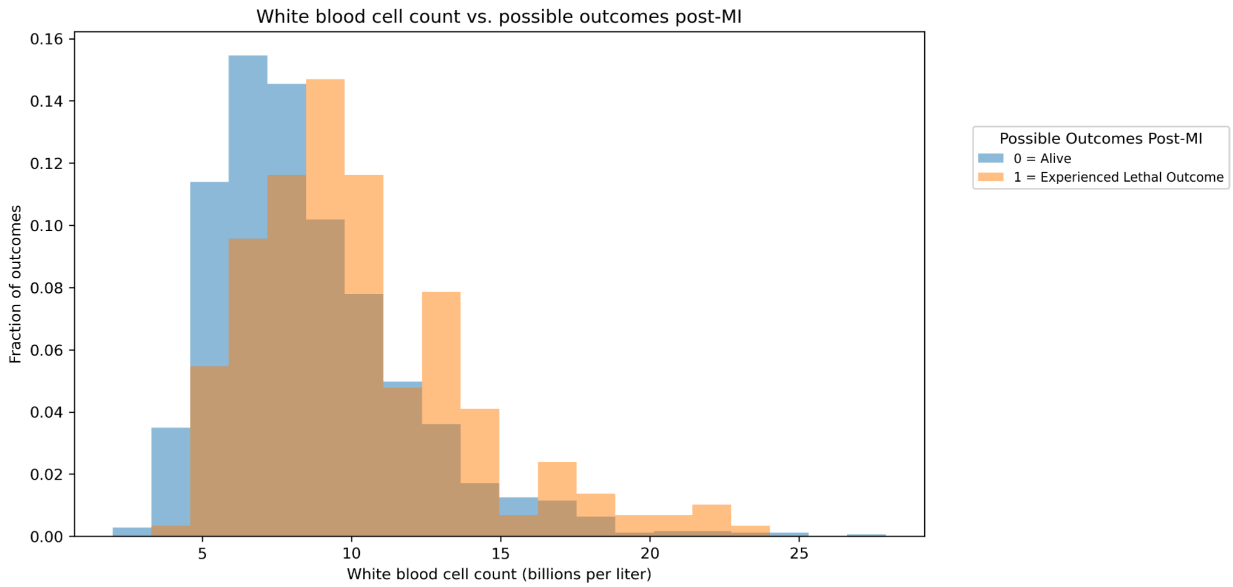


Figure 2. Category-specific histogram showing the relationship between white blood cell count and post-MI outcomes, with higher counts slightly associated with lethal complications.

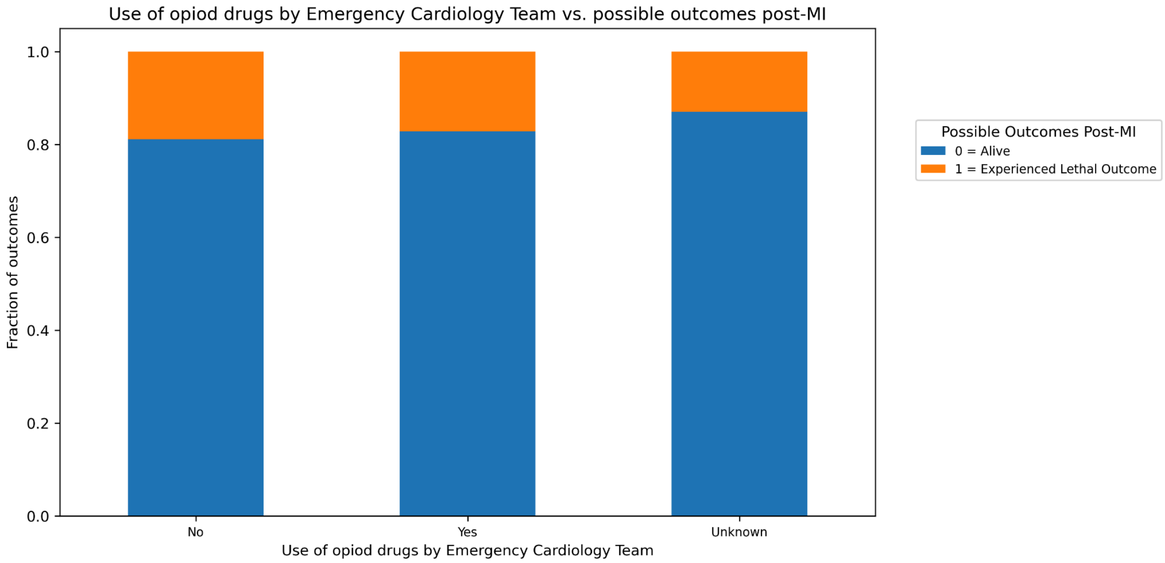


Figure 3. Stacked bar plot showing opioid use to treat MI relative to the possible post-MI outcomes. The proportion of lethal complications increases as the severity of chronic heart failure rises.

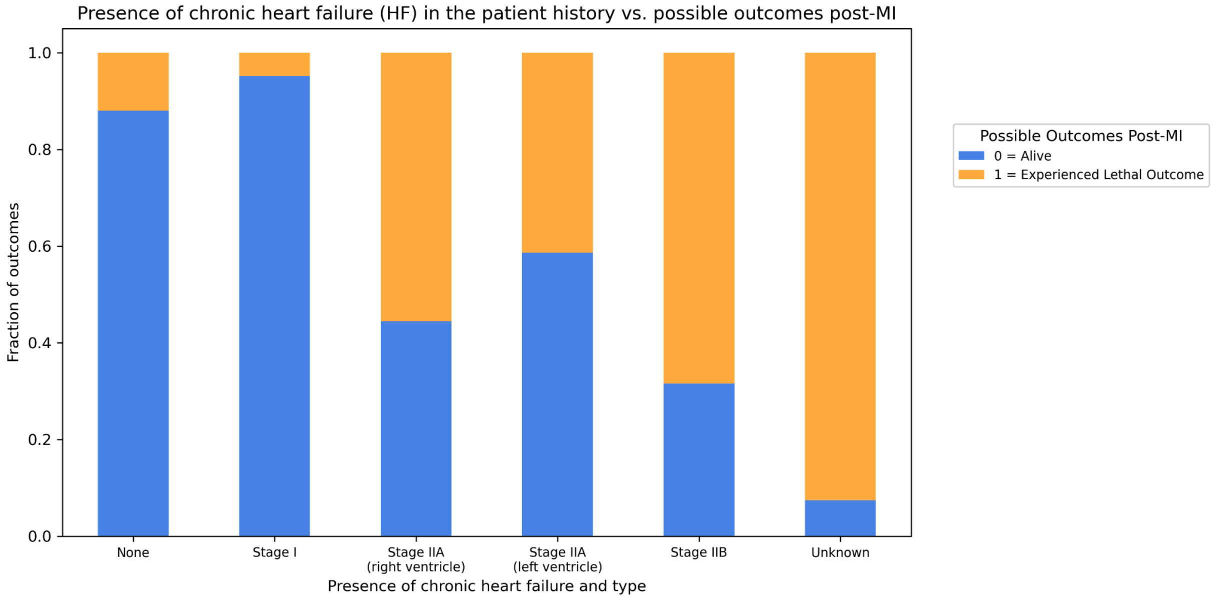


Figure 4. Stacked bar plot illustrating breakdown of the post-MI outcomes by patient history of chronic heart failure. Fractions of lethal complications appear to increase with severity.

2.2 Approach to Dropping Features

2.2a Missing values

An assessment of missing values revealed that all 1700 data points contained at least one missing feature value.Two columns, one representing a categorical feature and the target variable, had no missing values. However, the remaining 103 columns all contained at least one missing value, two of which had missing value percentages above 90%:

Variable	Missing Values Percentage
LET_IS (target variable)	0%
SEX	0%
IBS_NASL	95.76%
KFK_BLOOD	99.76%

Table 3. Sorted table of feature missing value percentages. IBS_NASL represents heredity of chronic heart disease, and KFK_BLOOD represents serum CPK content in IU/L.

Based on this evaluation, the two features with missing values percentage over 80% were removed, as they lack sufficient information to contribute to any predictions. Additionally, new categories were introduced to represent missing values in categorical and ordinal features, leaving only the continuous features with missing values to be handled later in the ML pipeline.

2.2b Pearson correlation

The dataset was also evaluated using a Pearson correlation matrix for continuous feature selection, and one pair of continuous variables was found to have a strong correlation. Since a high Pearson correlation indicates feature redundancy and can lead to multicollinearity (which affects model performance), one feature (highlighted in red) was removed from the dataset before training:

Features	Correlation
D_AD_ORIT and S_AD_ORIT	86%

Table 4. PPair of continuous variables with Pearson correlation coefficients above 85%. D_AD_ORIT and S_AD_ORIT represent diastolic and systolic blood pressure (in mmHg) from the ICU, with the highlighted feature removed.

After addressing missing values and feature correlation, the dataset consists of 1700 data points and 102 columns (101 features and 1 target variable).

3. Methods

3.1 Splitting

Given the independent and identically distributed (IID) nature of the data and imbalanced classes, Stratified KFold was used to preserve the class distribution across training, cross-validation, and test sets. Simplifying the target variable to binary classification during EDA addressed challenges from the splitting strategy, as individual lethal outcomes had small, imbalanced sample sizes. This approach helped generate more balanced splits, leading to more robust model learning and better handling of the class imbalance.

First, a stratified split divided the data into an 80% training-validation set and a 20% holdout test set. TThe training-validation set was further split using StratifiedKFold with 3 folds, resulting in approximately 53% of the data used for training and 27% for validation in each fold. This 3-fold split was then applied for cross-validation to evaluate model performance:

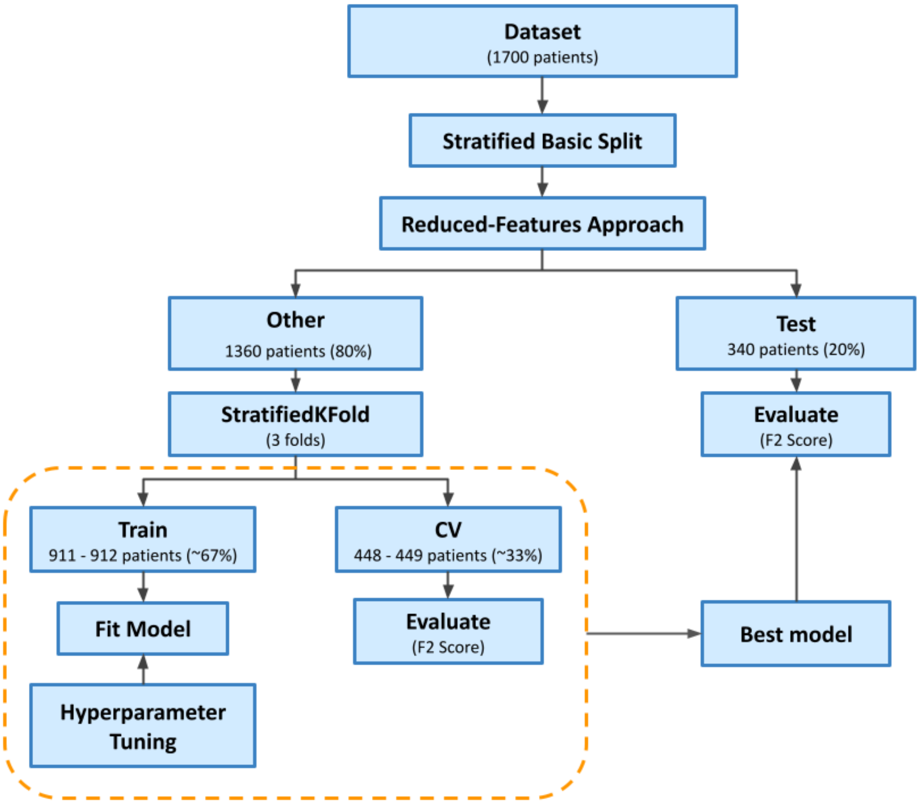


Figure 5. Schematic for model implementation and cross-validation pipeline.

3.2 Preprocessing

For data preprocessing, OneHotEncoder was used to handle categorical features to avoid misleading interpretations of nominal data. Continuous features were scaled using either MinMaxScaler or StandardScaler, based on EDA visualizations. StandardScaler was applied to features with significant outliers, while MinMaxScaler was used for features with defined ranges. Ordinal features, which are already numeric with an intrinsic order, did not require additional encoding. Finally, StandardScaler was applied to all features to normalize them with a mean of 0 and a standard deviation of 1, ensuring consistency for model training and feature importance analysis.

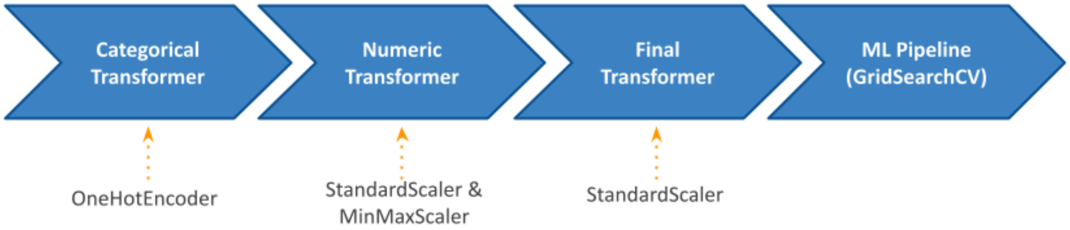


Figure 6. Workflow of preprocessing (encoding and scaling) data.

3.3 Cross-Validation Pipeline and Algorithms

Five reduced-feature supervised ML models were implemented with GridSearchCV for hyperparameter tuning with cross-validation. The reduced-feature approach was applied to address continuous features with missing values, as the models used do not natively handle missing data.

ML model	Hyperparameters	Best Parameter Grid
Reduced-Feature Logistic Regression (l1)	C (alpha) solver max_iter tol	[1e2, 3.59381366e1, 1.29154967e1, **4.64e0**, **1.67e0**, 5.99e-1, 2.15e-1, 7.74e-2, 2.78e-2, 1e-2] ['saga'] [10000] [1e-4]
Reduced-Feature Logistic Regression (l2)	C (alpha) solver max_iter tol	[1e2, 3.59381366e1, 1.29154967e1, 4.64e0, **1.67e0**, **5.99e-1**, 2.15e-1, 7.74e-2, 2.78e-2, 1e-2] ['saga'] [10000] [1e-4]
Reduced-Feature Random Forest Classification	max_depth max_features	[1, 3, 10, **30**, 100], [0.5, **0.75**, 1.0]
Reduced-Feature KNN Classification	n_neighbors weights	[**1**, 3, 10, 30, 100], [**'uniform'**, 'distance']
Reduced-Feature SVM Classification	gamma C	[**1e-3**, 1e-1, 1e1, 1e3, 1e5], [1e-2, 1e-1, 1e0, 1e1, **1e2**]

Table 5. Supervised ML Models Evaluated and their Corresponding Hyperparameter Values.

Uncertainties from splitting and non-deterministic models were quantified by calculating the mean and standard deviation of the evaluation metric across 5 random states. The F2 score was chosen as the evaluation metric since it prioritizes recall over precision to minimize false negatives, which is crucial in medical diagnostics. The imbalanced target classes make the F2 score particularly suitable, as it prevents overemphasis on the majority class.

4. Results

Based on mean F2 scores of the models on the test set, the Random Forest Classifier performed the best with a score of 0.78 as shown below. While all models performed better than the baseline F2 score of 0.49, logistic regression performed the worst out of the 4 models.

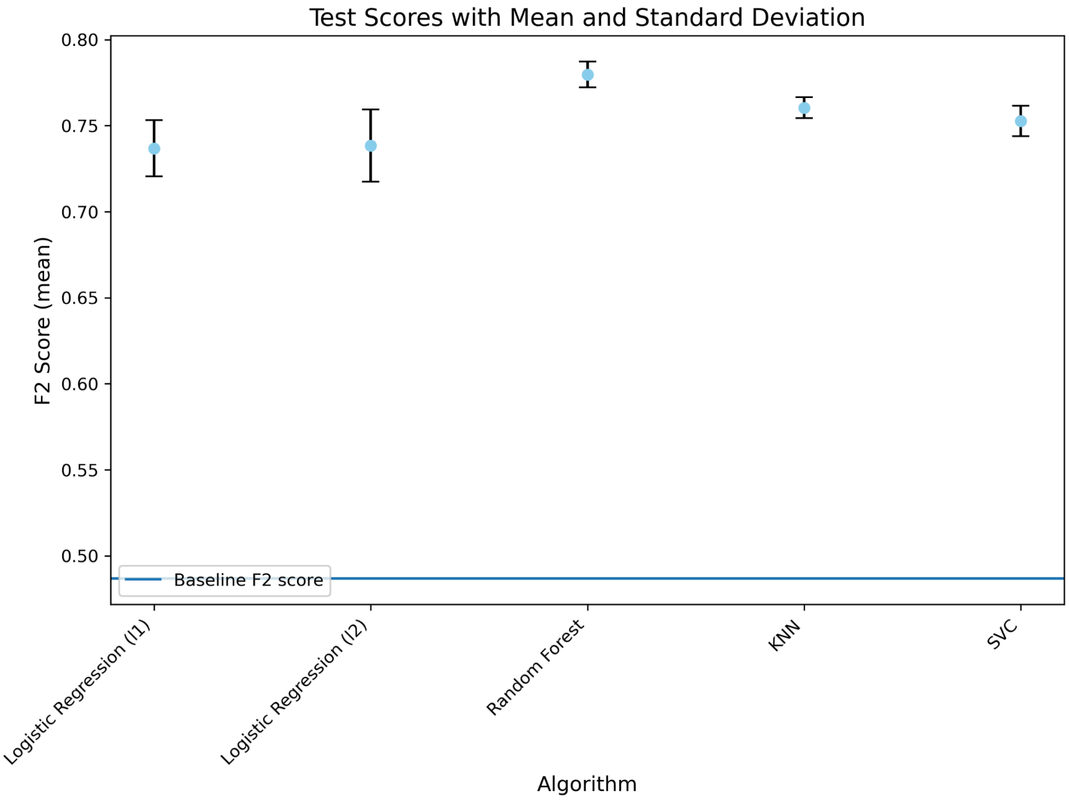


Figure 7. Comparison of mean and standard deviation of F2 scores for each model.

ML model	# of Standard Deviations Above Baseline
Reduced-Feature Logistic Regression (l1)	15.29
Reduced-Feature Logistic Regression (l2)	12.02
Reduced-Feature Random Forest Classification	39.09
Reduced-Feature KNN Classification	45.64
Reduced-Feature SVM Classification	30.38

Table 6. Number of standard deviations above baseline performance for all models.

The normalized confusion matrix for the best model (Random Forest Classifier) with random state 42, shown in Figure 8, reveals that the model performs well at predicting alive patient outcomes (class 0) with 93% accuracy. However, it struggles with predicting post-MI lethal outcomes (class 1), correctly classifying only 13% of such cases. Despite dataset stratification, k-fold cross-validation, and a focus on recall, the model's precision (65%) and recall (13%) for class 1 remain low, typical for models trained on imbalanced datasets. This suggests that the current feature set and model architecture may not effectively capture the signals needed to identify patients at risk of lethal outcomes.

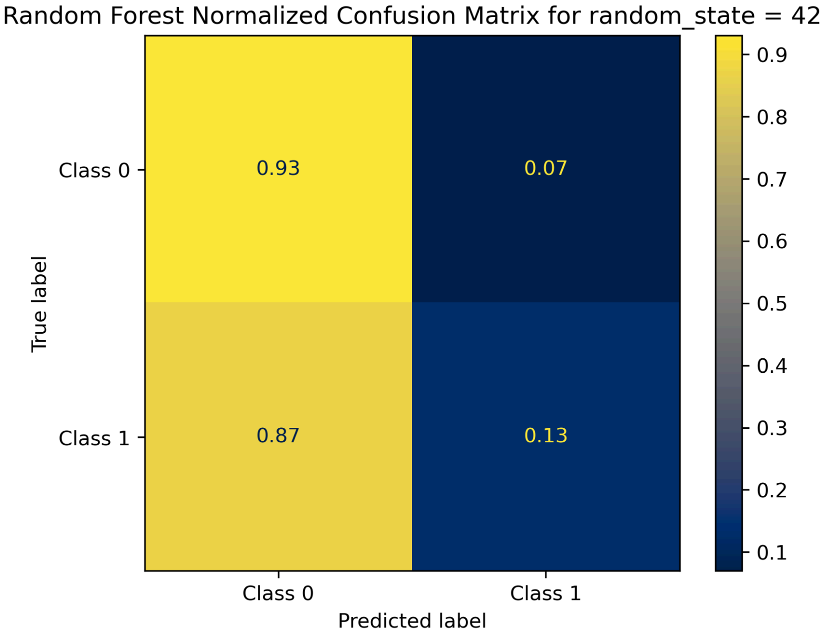


Figure 8. Normalized confusion matrix for the best classifier model: Random Forest (random_state = 42).

4.1 Feature Importance

To study which features are most important in the best performing model (Random Forest Classifier), global and local feature importance metrics were explored using SHAP values.

4.1a Global Feature Importance

Global feature importance highlights the features that have the greatest impact on the model's predictions. This understanding aids in evaluating and comparing models, as well as identifying key factors driving decision-making.

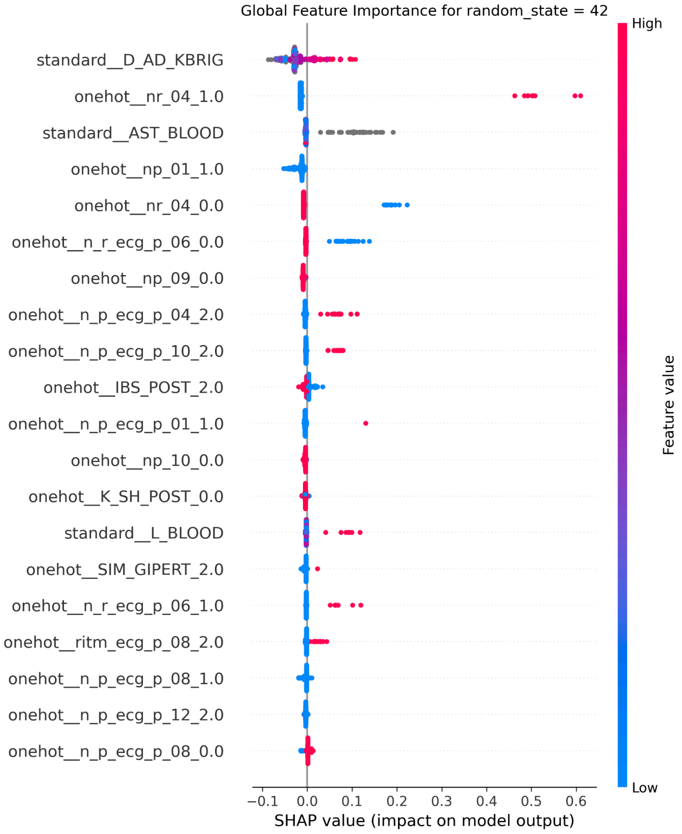
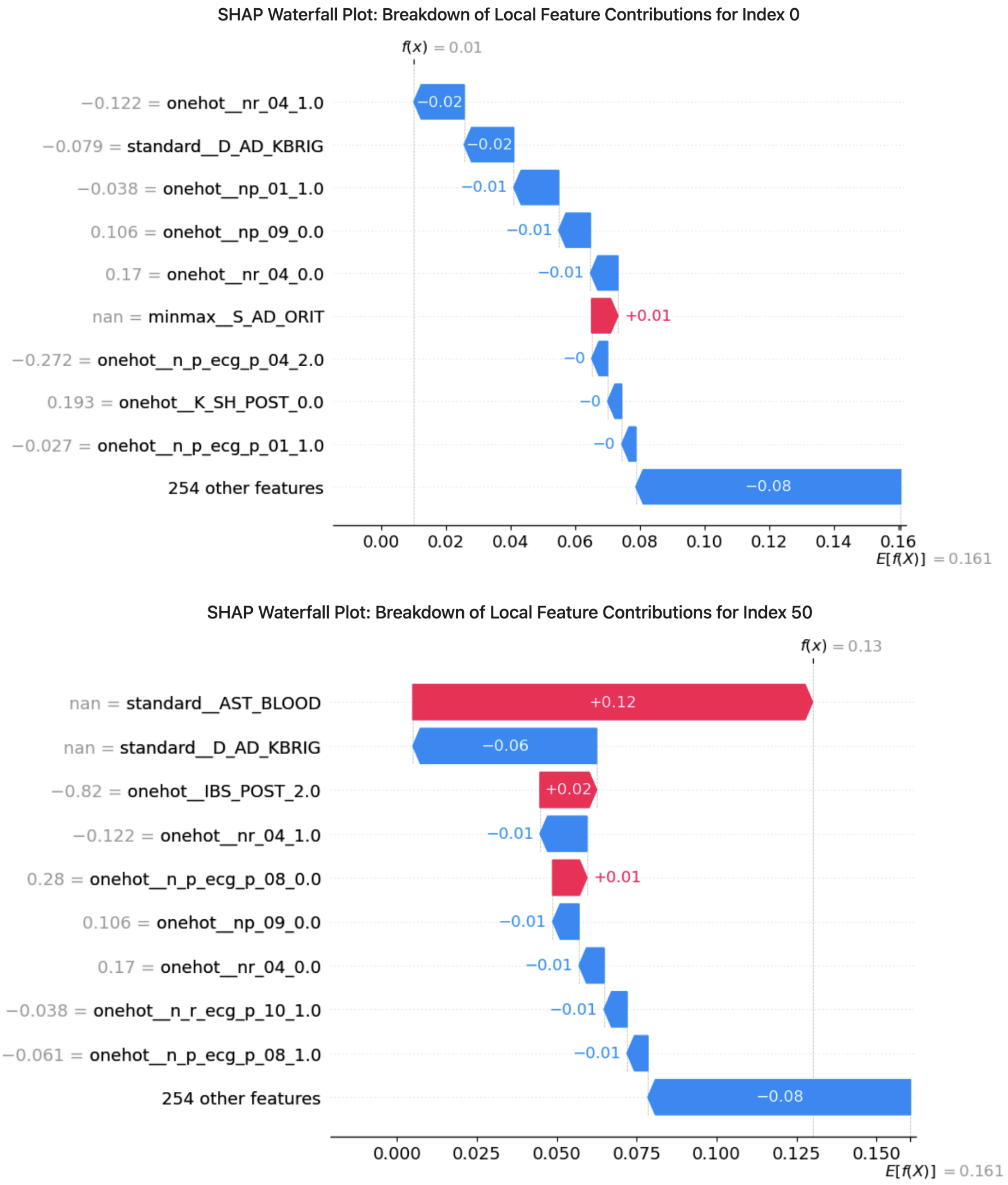


Figure 9. SHAP summary plot of the top 20 important features for the Random Forest Classifier (random_state = 42). It shows that the top five contributing features are diastolic blood pressure (according to the emergency cardiology team), patient history of persistent atrial fibrillation (with and without), serum AsAT content, and patient history of first-degree AV block. These features consistently appear in summary plots generated for different random states.

4.1b Local Feature Importance

Local feature importance focuses on how each feature impacts a specific prediction for an individual case. It offers a detailed understanding of why a model made a particular decision, which is crucial for interpreting medical diagnostic predictions. This is especially valuable for healthcare providers in making personalized recommendations or treatments for patients.



SHAP Waterfall Plot: Breakdown of Local Feature Contributions for Index 0

SHAP Waterfall Plot: Breakdown of Local Feature Contributions for Index 50

Figure 10. SHAP waterfall plots of the most important local features for patients 0 and 50 (Random Forest Classifier, random_state = 42) show different top contributing features: persistent patient history of atrial fibrillation for patient 0 and serum AsAT content for patient 50. Despite the different rankings, many of the other important features are similar for both patients.

5. Outlook

To improve the model's performance and interpretability, we could experiment with models like XGBoost, which may offer better predictive power. Expanding the hyperparameter search space could further enhance model performance. Additionally, additional interpretability techniques like LIME or perturbed feature analysis could provide new insights into model decisions, though challenges with the feature importance analysis and the reduced-features approach may need to be addressed. The current dataset is high-dimensional but suffers from missing data, which likely weakens the predictive power of the models. Collecting more complete data or improving missing data handling (e.g., imputation) would improve model quality.

6. References

[1] Griffin, B.P., Topol, E.J., Nair, D. and Ashley, K. eds., 2008. Manual of cardiovascular medicine. Lippincott Williams & Wilkins.

[2] Golovenkin, S., Shulman, V., Rossiev, D., Shesternya, P., Nikulina, S., Orlova, Y., & Voino-Yasenetsky, V. (2020). Myocardial infarction complications [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C53P5M.

[3] Golovenkin, S.E., Bac, J., Chervov, A.V., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N., & Zinovyev, A.Y. (2020). Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. GigaScience, 9. 10.1093/gigascience/giaa128.