Heart Failure Mortality Prediction with Gradient Boosted Trees: A Complete Machine Learning Pipeline on a Clinical Cohort

Dilan Doshi
UCLA

December 18, 2025

Abstract

Heart failure is a chronic and progressive condition that carries a high risk of mortality and re-hospitalization. Accurate identification of high risk patients is important for targeted monitoring and timely intervention. This project develops an end to end machine learning pipeline to predict mortality in a cohort of heart failure patients using twelve routinely collected clinical features.

The workflow covers exploratory data analysis, outlier detection, data preprocessing, model selection, hyperparameter tuning, threshold optimization, and model interpretation. The final model is an extreme gradient boosting classifier with tuned hyperparameters and an optimized decision threshold. On the held out validation set the model attains a receiver operating characteristic area under the curve (ROC AUC) of 0.944, recall of 91.7 percent, precision of 73.3 percent, and F1 score of 0.815. Feature importance analysis confirms that the model relies on clinically plausible predictors such as follow up time, ejection fraction, serum creatinine, and age.

The paper describes the full pipeline in detail, discusses design decisions motivated by clinical and statistical considerations, and outlines limitations and directions for future work, including external validation and richer feature sets.

1 Introduction

Heart failure is a major cause of mortality worldwide. Patients with heart failure often experience frequent hospital admissions and a high risk of death within a relatively short time after diagnosis. Risk prediction models can support clinicians by highlighting patients with elevated mortality risk who may benefit from closer follow ups, medication optimization, or advanced therapies.

Traditional clinical scores tend to rely on linear combinations of a small number of variables and may not capture complex interactions or nonlinear patterns in the data. In contrast, modern machine learning methods can learn flexible decision boundaries directly from data while still remaining interpretable through feature importance analysis and related tools.

The goal of this project is to build and evaluate a machine learning model that predicts mortality in heart failure patients using twelve clinical features that are commonly available in hospital electronic records. The pipeline is intentionally designed to be transparent and reproducible. It emphasizes careful exploratory data analysis, conservative preprocessing, and rigorous evaluation, rather than blind pursuit of maximal accuracy.

The final model is an extreme gradient boosting classifier (XGBoost) trained on a cleaned subset of the original dataset. It is selected based on cross validated ROC AUC and then further refined by explicit decision threshold optimization tailored to the clinical objective of high sensitivity for

death events. The performance of the model and the importance of each feature are analyzed in detail using both gain based importance and permutation based importance.[1]

2 Data and Problem Setup

2.1 Dataset overview

The dataset consists of 299 patients diagnosed with heart failure. For each patient there are twelve predictor variables and one binary target variable that indicates whether the patient died during the follow up period.

- Target variable: `DEATH_EVENT` equals 1 for patients who died and 0 for patients who survived.
- Continuous predictors: age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and follow up time.
- Binary predictors: anaemia, diabetes, high blood pressure, sex, and smoking.

There are no missing values in any column, so imputation is not required. The class distribution is moderately imbalanced: 72.8 percent of patients survived (218 cases) and 27.2 percent died (81 cases), which corresponds to an imbalance ratio of roughly 2.67 to 1 in favor of survival.

Because the imbalance is noticeable but not extreme, the pipeline does not use synthetic oversampling methods such as SMOTE. Instead, it relies on stratified splitting, class weighting in the loss function, and the use of metrics that are robust to imbalance.

2.2 Clinical motivation

From a clinical perspective, the key objective is to identify as many patients as possible who are at high risk of death. In this setting, false negatives (missed deaths) are much more concerning than false positives (patients flagged as high risk who actually survive). Therefore recall for the positive class, which is the death class, is of primary importance. Precision still matters because an excessive number of false alarms can lead to alarm fatigue and inefficient use of resources, but the trade off is weighted toward recall.

ROC AUC is used as the primary metric for model comparison because it evaluates the ranking ability of the model across all thresholds and is relatively insensitive to class imbalance. Threshold selection is then performed on the best model to achieve a clinically acceptable balance between recall and precision.

3 Exploratory Data Analysis

Exploratory data analysis (EDA) is carried out in a dedicated notebook in order to understand the structure of the dataset, identify outliers, and examine relationships between features and the target.

[1] The detailed code and outputs for the final evaluation, including feature importance plots, confusion matrices, ROC curves, and threshold analysis, are contained in the notebook "05 reporting evaluation."

3.1 Univariate distributions and outlier detection

Initial inspection reveals heavy right skew in two laboratory variables: creatinine phosphokinase and serum creatinine. The interquartile range method is used to detect outliers. For each continuous variable, values outside the interval from Q1 minus 1.5 times the interquartile range to Q3 plus 1.5 times the interquartile range are flagged as outliers. In clinical datasets this procedure helps remove extreme measurement errors or very rare cases that can distort model training.

Removing outliers, primarily from creatinine phosphokinase and serum creatinine, reduces the dataset from 299 to 224 patients. Skewness in these two features decreases from extreme levels to moderate levels (for example creatinine phosphokinase skew reduces from 4.46 to approximately 0.97). Other continuous features, including follow up time, serum sodium, platelets, age, and ejection fraction, exhibit nearly symmetric distributions after this cleaning step.

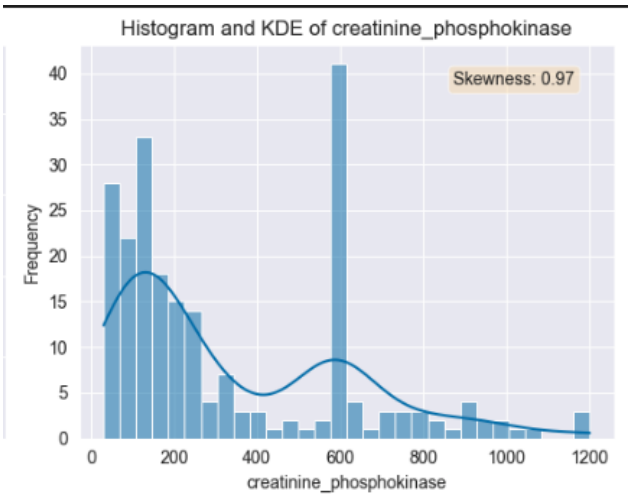

Figure 1: Univariate distribution plot for Creatinine Phosphokinase after outlier removal.

The ejection fraction distribution is especially interesting. It suggests two overlapping patient populations, one with relatively preserved systolic function and one with reduced systolic function, which is consistent with clinical categories of heart failure.

3.2 Bivariate relationships with the target

To explore how features relate to mortality, violin plots and box plots stratified by the death event label are constructed. These plots reveal distinct differences in several variables:

- Patients who die tend to have lower ejection fraction, typically centered near 25 to 30 percent, whereas survivors have a broader distribution centered around 40 to 50 percent.
- Serum creatinine tends to be higher among patients who die, which reflects worse kidney function and the presence of co-morbid conditions.
- Follow up time tends to be shorter for patients who die, which is expected because they leave the cohort by death earlier in the study period.
- Age is slightly higher among patients who die.

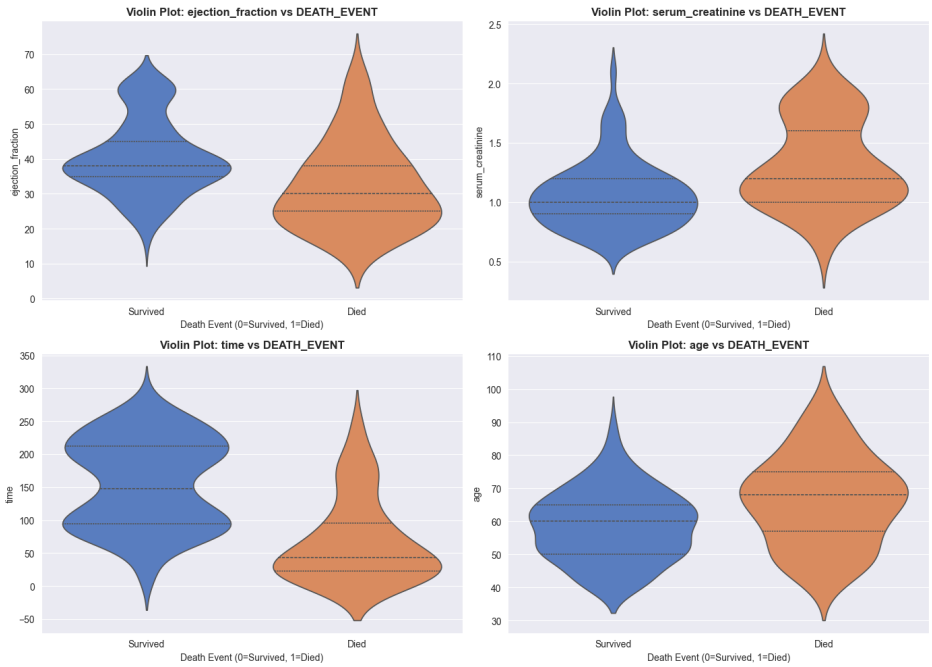

Figure 2: Violin plots for selected features that display the distribution of selected continuous variables stratified by survival status.

For the binary predictors, chi square tests are used to test independence with respect to the death event. None of the binary features show statistically significant association at conventional significance levels, although small differences in event rates are present. Given their clinical relevance and the possibility of interaction effects, all binary features are retained for model building.

3.3 Correlation structure and feature selection

The Pearson correlation between each continuous predictor and the death event shows that:

- Follow up time has a strong negative correlation with death (approximately minus 0.51).
- Ejection fraction has a moderate negative correlation with death.
- Serum creatinine and age have moderate positive correlations with death.

Correlations among the predictors are relatively weak and do not indicate strong multicollinearity. With only twelve predictors, each of which is clinically meaningful, explicit feature selection is not performed. Instead, the models are expected to down weight less informative predictors through regularization and splitting criteria.

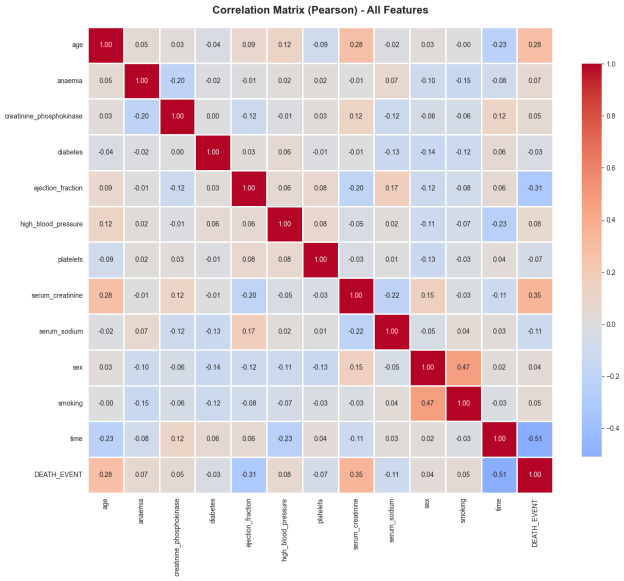

Figure 3: Correlation heatmap summarizing linear associations among features and with the death event.

4 Data Preprocessing

4.1 Train validation split

Because the cleaned dataset is small and contains 224 patients, a single stratified split is used to create training and validation sets. The split proportion is 80 percent training and 20 percent validation. Stratification ensures that both sets preserve the original class balance.

- Training set: 179 patients, with 130 survivors and 49 deaths.
- Validation set: 45 patients, with 33 survivors and 12 deaths.

The validation set is treated as a held out test set and is only used for final evaluation after hyperparameter tuning.

4.2 Feature scaling and preprocessing pipeline

Continuous features are standardized using a z score transform, so that each has mean zero and variance one on the training set. This scaling stabilizes optimization for models such as logistic regression and support vector machines and provides a consistent input scale for any neural networks.

Binary features are passed through without transformation since they are already coded as 0 or 1. A `ColumnTransformer` and a `Pipeline` from scikit learn are used to encapsulate preprocessing. The transformer is fit on the training data only and then applied to both the training and validation sets. This design prevents information from the validation set leaking into the scaling parameters.

4.3 Handling class imbalance

Given the moderate imbalance level, resampling methods such as SMOTE are unnecessary and potentially harmful in this small dataset. Instead, the models that support class weights are con-

figured with `class_weight = balanced`. This setting automatically balances weights in minority class in the loss function according to the inverse of class frequencies. Combined with stratified cross validation, this approach maintains the original data distribution and reduces the risk of overfitting synthetic samples.

5 Modeling Approach

5.1 Evaluation framework

Model selection is based primarily on ROC AUC measured on the training folds during cross validation. The following metrics are considered during final evaluation on the validation set:

- ROC AUC, which reflects the global ranking quality of the probabilistic predictions.
- Recall (sensitivity) for the death class, which captures the fraction of deaths detected.
- Precision for the death class, which captures the fraction of predicted deaths that are actual deaths.
- F1 score, which combines precision and recall through their harmonic mean.
- Accuracy, which is reported but de emphasized because it can be misleading under imbalance.

Five fold stratified cross validation is used within the training set for hyperparameter tuning. This choice provides a reasonable balance between variance and bias in such a small dataset, while maintaining sufficient samples in each fold for both classes.

5.2 Candidate models

Several model families are investigated:

1. Logistic regression with L2 regularization and balanced class weights.
2. A single decision tree classifier with depth and splitting parameters tuned by grid search.
3. A random forest ensemble of decision trees.
4. A gradient boosting machine (sklearn implementation).
5. XGBoost, an efficient and regularized gradient boosting implementation designed for tabular data.

Models such as support vector machines and feed forward neural networks are considered but not pursued in the main pipeline due to their relatively higher complexity and the small sample size. In the future, we may implement a small feedfoward neural network and compare the results.

5.3 Hyperparameter tuning

Hyperparameter spaces for each model are defined based on standard practice for small to medium sized tabular datasets. Grid search with five fold stratified cross validation is performed for each model, using ROC AUC as the scoring metric.

Examples of key hyperparameters include:

- Logistic regression: regularization strength parameter C with a fine grid of small values to encourage regularization.
- Decision tree: maximum depth, minimum samples per split, minimum samples per leaf, and splitting criterion.
- Random forest: number of trees, depth, minimum samples per split and leaf, and feature subsampling strategy.
- Gradient boosting: number of estimators, learning rate, maximum depth, minimum samples per split, and subsampling rate.
- XGBoost: number of estimators, learning rate, maximum tree depth, minimum child weight, subsampling rate, and tree column subsampling rate.

Across ensembles, the best performing configurations favor shallow trees, low learning rates, and conservative regularization settings, which is appropriate for the small dataset and helps prevent overfitting.

6 Threshold Optimization

Most machine learning models generate continuous probabilities or scores. A default threshold of 0.5 is often used to convert these scores into binary predictions. However, this choice rarely aligns with the actual clinical trade off between false positives and false negatives. After identifying the best model based on ROC AUC, it is important to tune the classification threshold explicitly.

For the selected tree ensemble models, predicted probabilities on the validation set are examined over a grid of thresholds between 0.1 and 0.9. For each threshold, recall, precision, and F1 score are computed. The threshold that maximizes the F1 score, while still achieving high recall, is chosen as optimal.

For XGBoost, the optimal threshold is found to be approximately 0.25. At this threshold recall and F1 score both improve relative to the default threshold of 0.5, while precision decreases only modestly. This low threshold reflects the clinical preference for sensitivity in identifying patients at risk of death.

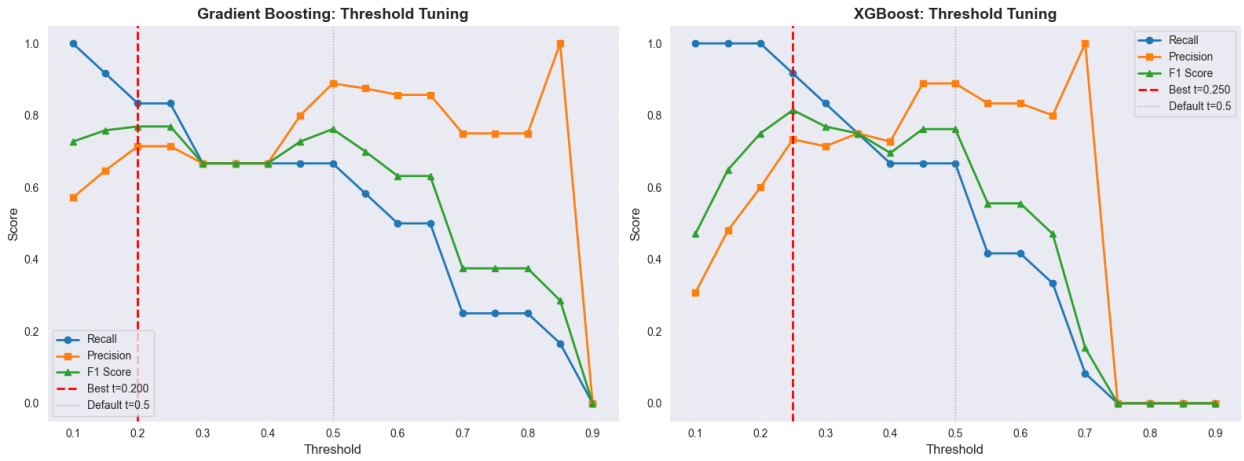

Figure 4: Trade off between recall, precision, and F1 score as the classification threshold varies for Gradient Boosting and XGBoost on a threshold plot.

7 Results

7.1 Model comparison

Table 1 summarizes the performance of the tuned models on the validation set. For ensemble models both the default threshold of 0.5 and the tuned threshold resulting from the threshold analysis are reported.

Table 1: Validation set performance for tuned models. Only the most relevant metrics are shown.

Model and threshold	ROC AUC	Recall	F1 score
XGBoost, tuned, threshold 0.25	0.9444	0.9167	0.8148
XGBoost, tuned, threshold 0.50	0.9444	0.6667	0.7619
Random forest, tuned, threshold 0.35	0.9343	0.8333	0.7692
Gradient boosting, tuned, threshold 0.20	0.9318	0.8333	0.7692
Logistic regression, tuned	0.8712	0.9167	0.7097
Decision tree, tuned	0.8144	0.7500	0.7200

XGBoost with the tuned threshold clearly provides the best balance between ranking ability and recall, while maintaining a strong F1 score. Although logistic regression reaches similar recall, its ROC AUC and F1 score are lower and the linear structure may fail to capture important interactions.

7.2 Final model performance

The chosen final model is the tuned XGBoost classifier with threshold 0.25. Its detailed performance on the validation set is as follows:

- Accuracy: 88.89 percent.
- Precision for death: 73.33 percent.

- Recall for death: 91.67 percent.
- F1 score for death: 0.8148.
- ROC AUC: 0.9444.

The confusion matrix is particularly informative. Out of 45 patients in the validation set, 12 died and 33 survived. The model correctly identifies 11 of the 12 deaths, missing only one, and incorrectly labels 4 survivors as deaths.

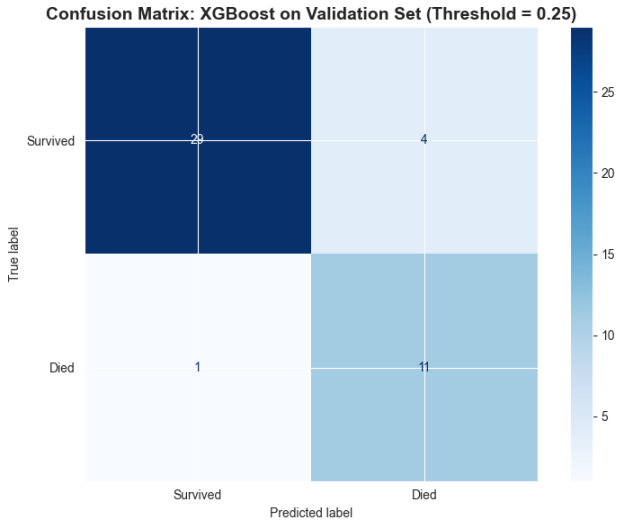

Figure 5: Confusion matrix of the tuned XGBoost model on the validation set.

In addition, ROC and precision-recall curves summarize performance across all thresholds.

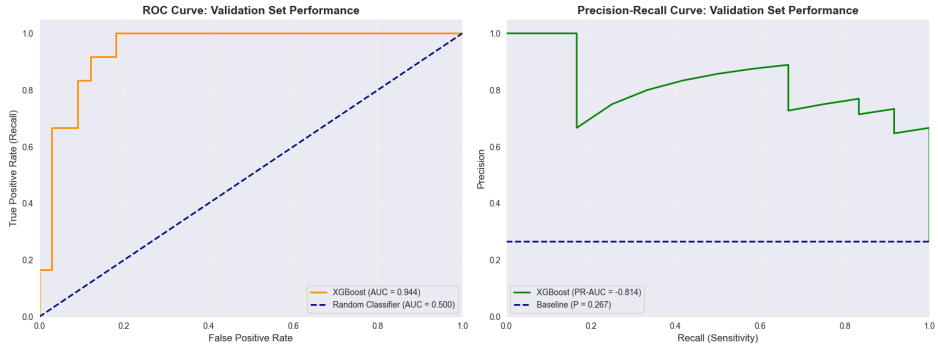

Figure 6: ROC and precision-recall curves for the tuned XGBoost model on the validation set.

These curves demonstrate that the model far outperforms a random classifier and that the chosen threshold sits in a region of the curve where recall remains high while precision is still acceptable.

8 Model Interpretation

8.1 Gain based feature importance

XGBoost exposes gain based feature importance, which measures the contribution of each feature to the reduction of loss across all trees. For the final model the ranking is as follows.

Table 2: Gain based feature importance for the tuned XGBoost model.

Rank	Feature	Relative importance
1	Follow up time	0.4282
2	Ejection fraction	0.2215
3	Age	0.1224
4	Serum creatinine	0.1084
5	Platelets	0.0426
6	Creatinine phosphokinase	0.0329
7	Anaemia	0.0317
8	Serum sodium	0.0078
9	High blood pressure	0.0046
10 to 12	Diabetes, sex, smoking	approximately zero

Follow up time accounts for roughly 43 percent of the total gain. Ejection fraction contributes about 22 percent, while age and serum creatinine contribute around 12 percent and 11 percent respectively. The remaining features individually account for a small share of the gain.

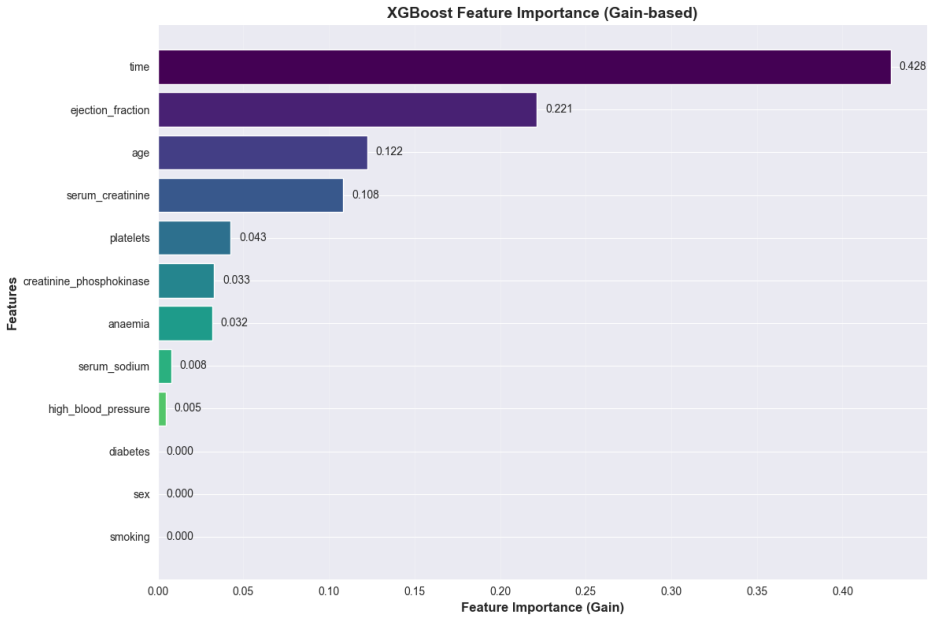

Figure 7: Gain based feature importance bar plot for the tuned XGBoost model.

8.2 Permutation importance

Gain based measures are useful but can be biased by tree structure. To obtain a more robust sense of feature importance, permutation importance is computed on the validation set. For each feature, its values are shuffled across patients while all other features remain intact. The resulting drop in ROC AUC indicates how important that feature is for the model's predictive performance.

The most important features by permutation importance are follow up time, ejection fraction, serum creatinine, and age, in close agreement with the gain based ranking. Platelets has a near-zero but nonzero importance, while the remaining features have negligible influence on ROC AUC when permuted.

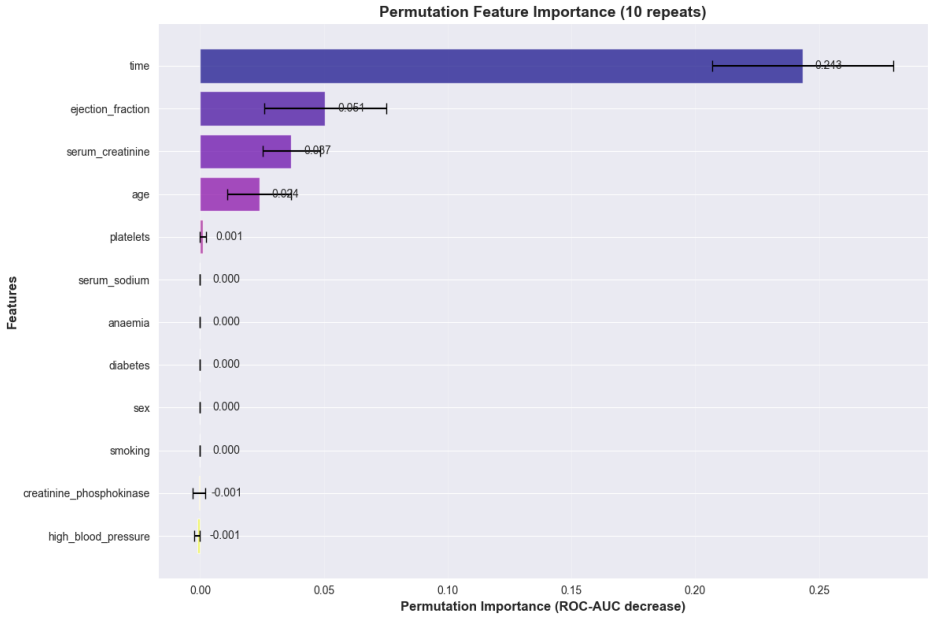

Figure 8: Permutation based feature importance plot for the tuned XGBoost model.

8.3 Clinical interpretation

These importance rankings align well with established clinical understanding of heart failure risk factors.

- Follow up time is a strong predictor because patients who die leave the study earlier. Shorter observed time implicitly signals events occurring sooner.
- Ejection fraction is a standard measure of systolic heart function. Lower values reflect reduced cardiac output and more severe heart failure.
- Serum creatinine is an indicator of renal function. Elevated values are associated with worse outcomes and reflect cardiorenal syndrome.
- Age is a general cardiovascular risk factor, with older patients more likely to suffer adverse events.

The fact that the model relies most heavily on these variables suggests that it is learning medically plausible patterns.

9 Overfitting and Robustness

9.1 Training versus validation performance

To assess overfitting, the tuned model is evaluated on both the training set and the validation set using the same threshold of 0.25. The metrics are summarized in Table 3.

Table 3: Training and validation performance for the tuned XGBoost model.

Metric	Training	Validation	Difference
Accuracy	0.8659	0.8889	-0.0230
Precision	0.6866	0.7333	-0.0468
Recall	0.9388	0.9167	0.0221
F1 score	0.7931	0.8148	-0.0217
ROC AUC	0.9587	0.9444	0.0143

All differences are small in magnitude, and the validation ROC AUC is within 0.02 of the training ROC AUC. This indicates that the model generalises well and does not exhibit strong overfitting.

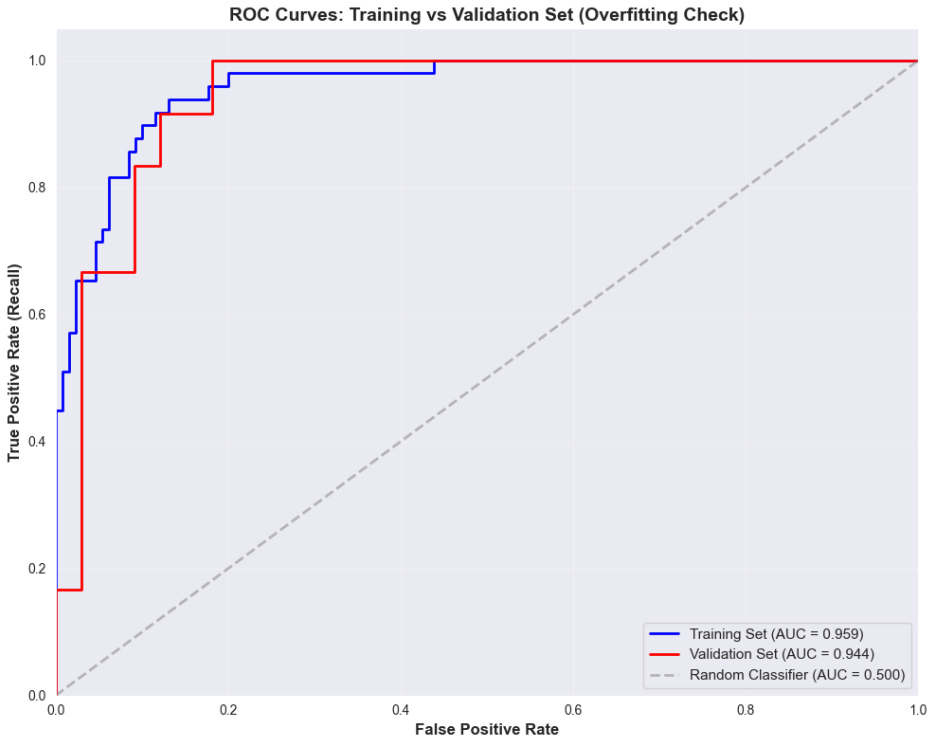

Figure 9: ROC curves on training and validation sets used to check for overfitting.

9.2 Cross validation stability

During hyperparameter tuning, the best XGBoost configuration achieves an average ROC AUC of approximately 0.931 across the stratified folds. The final validation ROC AUC of 0.944 is

consistent with this cross validated performance and slightly higher, which is reasonable given sampling variability. This agreement suggests that the selected model is stable with respect to data splits and not overly tuned to patterns of a single feature.

10 Discussion

10.1 Key methodological choices

Several design decisions in this project merit discussion:

- Outlier removal using the interquartile range rule improves distribution symmetry for skewed laboratory values while discarding a limited number of extreme observations. This step makes the data more amenable to parametric models and reduces the influence of measurement errors.
- Kept all twelve clinical features due to the small feature set and the absence of severe multicollinearity. We wanted to preserve potential nonlinear interactions. Tree based models and regularized logistic regression can reduce the weight of uninformative variables automatically.
- Class imbalance is handled through class weighting and careful metric choice instead of resampling. This avoids the injection of synthetic samples that can distort small datasets.
- Threshold tuning is performed explicitly with a focus on recall and F1 score rather than relying on the default cutoff. This step is vital when deploying a model in a clinical context where the cost of false negatives is high.

10.2 Clinical implications

The final model achieves high recall for death events while keeping precision at a moderate level. In practice, this means that most high risk patients will be flagged by the model, at the cost of a moderate number of false alarms. Such behavior is appropriate when the primary concern is avoiding missed opportunities for intervention.

The feature importance analysis reveals that the model bases its predictions on variables that clinicians already use in risk assessment, such as ejection fraction, kidney function, and age.

Potential use cases include:

- Risk stratification at hospital discharge, where patients with high predicted mortality could be enrolled in closer follow up programs.
- Support for multidisciplinary rounds, where the model's risk scores complement clinical judgment.
- Research exploration of subgroups where model predictions are systematically high or low, which may uncover new patterns.

10.3 Limitations

Several limitations should be acknowledged:

- The dataset is relatively small, with only 224 patients after outlier removal and 45 patients in the validation set. Performance estimates may therefore have wide confidence intervals.

- The data likely originate from a single center and a specific time period. External validation on independent cohorts from different institutions and populations is necessary before deployment.
- The feature set, while clinically meaningful, is limited to twelve variables. Additional data sources such as medication history, imaging, and detailed lab trends could further enhance predictive accuracy.
- The chosen decision threshold of 0.25 is tuned for this dataset. It may require recalibration in other settings with different event rates or costs.

11 Future Work

Future extensions of this project may include:

- External validation on larger and more diverse cohorts to assess generalization.
- More sophisticated feature engineering, including interaction terms (for example age times ejection fraction) and nonlinear transformations.
- Ensemble strategies that combine predictions from multiple model families, such as stacking logistic regression with gradient boosting.
- Time to event modeling if individual follow up times and censoring information can be leveraged in survival models.
- Integration with electronic health record systems to provide real time risk scores in clinical practice, along with user interface design and clinician feedback loops.

12 Conclusion

This project demonstrates a complete machine learning pipeline for predicting mortality in patients with heart failure using a modest sized clinical dataset. Through exploratory analysis, preprocessing, careful model selection, and threshold optimization, a tuned XGBoost model is obtained that achieves strong performance on a held out validation set.

The model attains an ROC AUC of approximately 0.944 and detects approximately 92 percent of deaths with an F1 score above 0.81. Feature importance analysis confirms that the model focuses on clinically significant variables such as follow up time, ejection fraction, serum creatinine, and age. Training and validation performance are closely aligned, indicating good generalization and limited overfitting.

While the results are promising, they should be interpreted with the dataset's limitations in mind. External validation and broader feature sets are logical next steps. Overall, the pipeline outlined here illustrates how machine learning can be applied in a careful and interpretable manner to support risk stratification in heart failure, and it provides a solid foundation for further research and eventual clinical implementation.