Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning and Explainable Artificial Intelligence for Predicting 10-Year Mortality after Cardiac Surgery: A Retrospective Cohort Study

Ruonan Jiao[1]

The Affiliated Guangdong Second Provincial General Hospital of Jinan University

Chenghui Zhou

chenghuizhou@vip.163.com

Capital Medical University

Research Article

Keywords: Neutrophil-to-lymphocyte ratio, Cardiac surgery, All-cause mortality, Machine learning, Explainable artificial intelligence, SHAP

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11213235/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Preoperative neutrophil-to-lymphocyte ratio (NLR) is associated with short-term outcomes after cardiac surgery, but its value for long-term mortality prediction remains unclear. We developed a machine learning model incorporating NLR to predict 10-year all-cause mortality after cardiac surgery, and used explainable artificial intelligence to interpret the role of NLR.

Methods

We included adult patients undergoing elective cardiac surgery from the MIMIC-IV database. The primary outcome was 10-year all-cause mortality. Six machine learning models were compared using grid search, stratified 5-fold cross-validation, and bootstrap internal validation. The best model was interpreted using SHAP on the test set. The NLR cutoff was identified by piecewise regression on the training set. Calibration curves and decision curve analysis were used to evaluate model performance.

Results

Among 4,891 patients (median age 68 years), 222 (4.54%) died within 10 years. After tuning, XGBoost performed best (optimism-corrected AUC = 0.815; 95% CI: 0.758–0.861), outperforming LASSO regression (AUC = 0.780). NLR alone had an AUC of 0.709, similar to APACHE III (0.732; P = 0.630). Combining NLR with APACHE III improved the AUC to 0.761 (P = 0.002). SHAP analysis identified APACHE III (0.371) and NLR (0.291) as the top two predictors. Piecewise regression identified an NLR cutoff of 6.5. At a 15% risk threshold, the positive predictive value was 23.3% (5.1-fold enrichment). Decision curve analysis showed net benefit across thresholds of 5%-30%.

Conclusions

Preoperative NLR is an independent predictor of 10-year mortality after cardiac surgery, with performance comparable to APACHE III. The tuned XGBoost model showed good overall performance. An NLR threshold of 6.5 was identified and supported by SHAP analysis. NLR may complement existing risk scores for long-term risk stratification, but should not be used as a standalone tool.

Background

Cardiac surgery works well for a lot of complex heart disease, but patients still die years down the line. Figuring out who is at high risk early on matters if we want to intervene. The standard tools people use are EuroSCORE, STS score, APACHE III, and SOFA. They have their place, but they were built before blood-

based inflammation markers became routine, and most of them were validated for short-term outcomes like 30-day or in-hospital death rather than long-term survival.

NLR is readily available from routine blood work[28], calculated as neutrophils divided by lymphocytes. [29]. A fair amount of work now links high NLR to worse outcomes in cardiovascular disease [1–3] and other clinical conditions [14]. But most of that work looks at short-term results, and the NLR cutoff people report jumps around a lot (2.5 to 10, depending on the paper). Classic regression probably cannot capture the non-linear stuff going on between NLR and long-term death.Machine learning is good at non-linear relationships and has been picking up in clinical prediction [4–6]. The catch is that it often feels like a black box. SHAP uses game theory to show how each feature affects the model's prediction. [7, 8]. Beyond cardiac surgery, elevated NLR has also been linked to poor prognosis in several other conditions, including hepatocellular carcinoma[15, 16], suggesting that NLR may reflect a general state of inflammatory imbalance rather than being disease-specific.

We did this in two steps. First, we tuned and compared six machine learning models to find the best one. We then applied SHAP analysis to the test set to understand the role of NLR. We compared NLR with APACHE III, SOFA, and a traditional logistic regression model.We picked 10-year all-cause death as the outcome because most studies stop at 30 days or 1 year.

Methods

Data Source

We used MIMIC-IV, a public database covering ICU patients at Beth Israel Deaconess Medical Center from 2008 to 2019 [9–11]. The data are de-identified. Our use was approved by the relevant IRBs, and the need for informed consent was waived because this is retrospective.

Study Population

We pulled adults (18+) who had elective cardiac surgery between 2008 and 2019. The procedures were CABG, aortic or mitral valve repair/replacement, or combined operations. Patients who had neutrophil and lymphocyte counts within 24 hours before surgery and complete follow-up was include. We excluded patients with missing NLR, abnormal NLR (< = 0 or > 100), or missing death data. We did not impute missing NLR because it was our main variable. Among 8,218 eligible patients, 4,891 were analyzed.

The main outcome was 10-year mortality.

Features that were reported as prognostic in the literature, had over 80% missingness in MIMIC-IV, and covered different clinical areas. We included 12 variables: NLR, age, sex, BMI, prior MI, heart failure,

diabetes, renal disease, isolated CABG, valve surgery, APACHE III on ICU day 1, and SOFA. LVEF, smoking, and medications were excluded because more than half of the data were missing.

Statistical Analysis

Multi-model Comparison and Hyperparameter Tuning

The data 70/30 was splited into training and test sets, stratified by death. Six models were compared: LASSO regression, random forest, XGBoost, LightGBM, SVM, and KNN. We tuned parameters using grid search and 5-fold cross-validation to maximize AUC. The parameter details are in Additional file 1: Table S1.

The event rate was only 4.54%, giving an events-per-variable ratio of about 18:1. To prevent overfitting, we used shallow trees (max_depth = 3) and 1,000 bootstrap resamples for internal validation. Each bootstrap iteration repeated the full training process including parameter tuning. We used class weighting for the imbalanced data. SVM, KNN, and LASSO were run on standardized features.

NLR Threshold Identification and Validation

Piecewise linear regression was used on the training set to find the best cutoff (based on AIC), then applied it to the test set. Both threshold selection and model training used the training data. We tested the cutoff using the maxstat method with 1,000 permutations, and estimated the 95% CI with 1,000 bootstrap resamples. The SHAP plot on the test set was for visual support only, not a formal test.

SHAP Explainability Analysis

TreeExplainer was used to compute SHAP values on the test set for both the feature importance ranking and the dependence plot. Doing it on the test set means we are looking at how the trained model handles new patients, which is what we actually care about clinically. All analysis was in Python 3.13 with scikit-learn, XGBoost, LightGBM, and SHAP.

Comparison with Traditional Scoring Systems

We computed AUC for NLR, APACHE III, and SOFA separately, and compared them with DeLong tests. Also built a logistic regression model with NLR plus APACHE III to assess the incremental benefit. The LASSO model with the same 12 features was our traditional benchmark. Then performed pairwise DeLong tests among the top three models.

Calibration and Clinical Utility

We assessed calibration using calibration curves (10 bins) and the Hosmer-Lemeshow test, and calculated the Brier score. Sensitivity, specificity, PPV, and NPV were calculated at both the Youden-optimal threshold and a clinical threshold of 15% predicted probability. Decision curve analysis covered threshold probabilities from 0% to 50%.

Sensitivity Analysis for Missing Data

MICE was used to impute missing NLR values. The imputation model included age, sex, APACHE III, SOFA, diabetes, heart failure, renal disease, and BMI. We imputed neutrophil and lymphocyte counts separately (predictive mean matching, 5 nearest neighbors), then recalculated NLR as the ratio. Five imputed datasets were generated. Results were pooled using Rubin's rules. The pooled XGBoost AUC was 0.818 (95% CI: 0.762–0.867), similar to the complete-case result of 0.822.

Results

Baseline Characteristics

Of the 4,891 patients, 222 (4.54%) died within 10 years. Median age was 68 (IQR 59-76), 30.1% were women, and median NLR was 3.01 (IQR 2.09-4.75). The training set had 3,423 patients (155 deaths, 4.53%) and the test set 1,468 (67 deaths, 4.56%).

3,327 patients was left out because NLR was missing. > Excluded patients were slightly younger (66.2 vs. 67.8, P < 0.001) and had lower APACHE III (49.8 vs. 52.3, P < 0.001), but 10-year mortality was similar (4.31% vs. 4.54%, P = 0.45). Feature completeness ranged from 82.3% (BMI) to 99.8% (age, sex). We median-imputed continuous variables and mode-imputed categorical ones when missingness was under 5%. BMI was missing in 17.7%, so we used a regression-based imputation with age, sex, and comorbidities. The MICE sensitivity analysis came back nearly identical (pooled AUC 0.818 vs. 0.822).

Table 1 Baseline characteristics of included vs. excluded patients

Characteristic	Included (n=4891)	Excluded (n=3327)	P-value
Age, mean (SD)	67.8 (12.1)	66.2 (13.0)	<0.001
Female, n (%)	1648 (30.1%)	1128 (33.9%)	0.85
NLR, median (IQR)	3.01 (2.09-4.75)	N/A	-
APACHE III, mean (SD)	52.3 (20.1)	49.8 (21.5)	<0.001
Diabetes, n (%)	1562 (31.9%)	1028 (30.9%)	0.34
Heart failure, n (%)	1023 (20.9%)	689 (20.7%)	0.82
10-year mortality, n (%)	222 (4.54%)	143 (4.31%)	0.45

Table 1. Continuous variables are presented as median (interquartile range) or mean ± standard deviation, and categorical variables as number (percentage). P values were calculated using the Mann-Whitney U test for continuous variables and the chi-square test for categorical variables. NLR, neutrophil-to-lymphocyte ratio; BMI, body mass index; CABG, coronary artery bypass grafting; APACHE III, Acute Physiology and Chronic Health Evaluation III; SOFA, Sequential Organ Failure Assessment.

Multi-model Performance Comparison

After tuning, XGBoost came out on top[26]: apparent AUC 0.822, optimism-corrected 0.815 (95% CI: 0.758-0.861). LightGBM (0.816) and random forest (0.811) were close behind. None of the top three were significantly different from each other by pairwise DeLong (XGBoost vs. LightGBM: z = 0.82, P = 0.412; XGBoost vs. RF: z = 1.15, P = 0.250; LightGBM vs. RF: z =0.31, P = 0.756). We picked XGBoost because it had the best AUC and the SHAP tools for it are mature.LASSO regression with the same 12 features had an AUC of 0.780. XGBoost improved the AUC by about 0.042 compared with this traditional model.

Table 2 Performance comparison of machine learning models (Youden-optimal threshold)

Model	Apparent AUC	Corrected AUC	Sensitivity	Specificity	PPV	NPV
XGBoost	**0.822**	**0.815**	**0.761**	**0.752**	**0.153**	**0.968**
LightGBM	0.816	0.809	0.746	0.745	0.148	0.966
Random Forest	0.811	0.804	0.731	0.740	0.143	0.964
LASSO Regression	0.780	0.772	0.687	0.718	0.131	0.960
SVM	0.732	0.724	0.657	0.685	0.115	0.955
KNN	0.710	0.701	0.627	0.673	0.108	0.952

Table 2. AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LASSO, least absolute shrinkage and selection operator; SVM, support vector machine; KNN, K-nearest neighbors. The Youden-optimal threshold was determined separately for each model based on the maximum of (sensitivity + specificity - 1).

Figure 1 shows the ROC curves of all six models. XGBoost had the best AUC LightGBM (0.816) and random forest (0.811) were very close. These three tree-based models did better than LASSO regression (AUC = 0.767). This means that non-linear relationships may be important for predicting long-term death. We also found that NLR alone (AUC = 0.709) was similar to APACHE III (AUC = 0.732). This is interesting because NLR is just a simple blood test, while APACHE III is a complex scoring system.

At the Youden cutoff, the PPV was only 15.3%. This is not surprising because the event rate was low (4.54%). When few patients die, it is hard for any model to have a high PPV. But when we used a 15%

predicted risk threshold, things looked better. The model found 176 high-risk patients, and 41 of them died. This gave a PPV of 23.3% and an NPV of 95.8%. Compared with the baseline mortality of 4.54%, this is about 5.1 times higher. In simple words, a patient in this high-risk group has about a 1 in 4 chance of dying in 10 years, while the average patient has less than a 1 in 20 chance. This could help doctors find patients who need closer follow-up and more care after surgery.

Comparison with Traditional Scoring Systems

We first compared NLR with traditional scoring systems. NLR had an AUC of 0.709 (95% CI: 0.643-0.773). APACHE III had an AUC of 0.732 (95% CI: 0.673-0.783), and SOFA had 0.637 (95% CI: 0.557-0.707). NLR and APACHE III were not significantly different (DeLong z = -0.482, P = 0.630). The fact that they performed similarly suggests that inflammation may be a key driver of long-term outcomes after cardiac surgery[13].

When we put NLR and APACHE III together in a logistic regression model, the AUC went up to 0.761 (95% CI: 0.703-0.808). This was significantly better than APACHE III alone (P = 0.002). Adding NLR improved the AUC by 0.029. This number is small, but it was statistically significant.

We also checked how well the model was calibrated. Figure 2 shows the calibration curves for XGBoost, LightGBM, and random forest. All three models were close to the perfect calibration line, especially in the lower risk range. The Brier scores were similar across models: 0.041 for XGBoost, 0.042 for LightGBM, and 0.041 for random forest. The Hosmer-Lemeshow test gave a chi-square of 22.02 (P = 0.005), which suggests some miscalibration, but the Brier score was still low overall. This means the model's predicted probabilities were generally close to what actually happened.

Finally, we used decision curve analysis to see if the model would be useful in clinical practice. The DCA showed positive net benefit across threshold probabilities from 5% to 30%. This was better than either treating all patients or treating none.

Figure 3 presents the decision curve analysis. This analysis evaluates whether the model provides clinical utility by comparing its net benefit against two reference strategies: treating all patients and treating none.The XGBoost model demonstrated positive net benefit across threshold probabilities ranging from approximately 5% to 30%. The treat-all strategy only worked at very low thresholds, below 5%. After that, it dropped quickly.

These findings suggest that the XGBoost model has practical clinical value. By applying a threshold within the 5% to 30% range, clinicians can identify a high-risk subgroup for closer follow-up while avoiding excessive surveillance of low-risk patients. This range covers most clinically relevant decision thresholds, supporting the model's potential utility in routine practice.

SHAP Explainability Analysis

Figure 4 shows the SHAP feature importance ranking for the XGBoost model. [27] The bars represent the mean absolute SHAP value for each feature.APACHE III was the most important predictor, with a mean SHAP value of 0.371. NLR came in second at 0.291. Renal disease, BMI and SOFA followed. Age and isolated CABG had moderate importance.The less important features included sex, diabetes, heart failure, prior MI, and valve surgery.

Overall, the ranking shows that APACHE III, SOFA and NLR are the main drivers of long-term mortality prediction after cardiac surgery. NLR ranked second, supports its potential as a simple and useful prognostic marker.

Figure 5 shows the SHAP dependence plot for NLR. Shows how much NLR pushes that patient's predicted mortality risk up or down.We can see that when NLR is below 6.5, the SHAP value rises quickly. This means that each unit increase in NLR adds a lot to the predicted mortality risk in this range. But once NLR goes above 6.5, the curve flattens out. The risk does not increase much more after that point.

The threshold at NLR = 6.5 was identified by piecewise regression on the training set. It falls right at the point where the trend line starts to flatten.It is also worth noting that there are fewer patients with very high NLR values (> 10).A few patients with very high NLR actually had negative SHAP values This may be because other factors, such as active infection or immunosuppression, can make NLR harder to interpret in individual patients.

Subgroup Analysis

We evaluated the model's performance across different patient subgroups (Figure 6). The overall AUC in the test set was 0.822. The model performed consistently well in most subgroups, with AUC values ranging from 0.67 to 0.86.

In patients with NLR < 6.5, the AUC was 0.84 (n = 1,250), compared with 0.69 in patients with NLR ≥ 6.5 (n = 218). The model also performed better in patients without renal disease (AUC = 0.83, n = 1,216) than in those with renal disease (AUC = 0.67, n = 252). Similarly, patients without heart failure had a higher AUC (0.86, n = 1,046) compared with those with heart failure (0.70, n = 422). For diabetes status, the AUC was 0.85 in patients without diabetes (n = 944) and 0.77 in those with diabetes (n = 524).

Regarding surgery type, the AUC was 0.85 for valve surgery (n = 504) and 0.80 for isolated CABG (n = 704). By sex, the AUC was similar between men (0.81, n = 1,057) and women (0.82, n = 411). Patients younger than 65 years had a higher AUC (0.86, n = 580) compared with those aged 65 years or older (0.79, n = 888).

Despite these descriptive differences, none of the interaction tests reached statistical significance. The interaction P values were 0.588 for age, 0.208 for sex, 0.168 for diabetes, 0.386 for heart failure, 0.946 for renal disease, and 0.790 for surgery type. Therefore, the observed differences across subgroups should be considered hypothesis-generating rather than definitive.

SHAP Beeswarm Analysis

To further understand how each feature affects individual predictions, we generated a SHAP beeswarm plot (Figure 7). Each dot represents one patient, and the color indicates the feature value (red for high, blue for low). The x-axis shows the SHAP value, where positive values increase the predicted mortality risk and negative values decrease it.

APACHE III was the most influential feature. Patients with higher APACHE III scores (red dots) tended to have positive SHAP values, meaning they were predicted to have higher mortality risk. Conversely, lower APACHE III scores (blue dots) were associated with negative SHAP values, reducing the predicted risk.

NLR showed a similar pattern. Higher NLR values (red dots) generally pushed the predicted risk upward, while lower NLR values (blue dots) pushed it downward. The spread of NLR dots was wider than most other features, confirming its importance as the second strongest predictor.

Renal disease also had a clear effect. Patients with renal disease had positive SHAP values. This means they had higher mortality risk.BMI was more complicated. Both high and low BMI were linked to higher risk in some patients. This suggests a possible U-shaped relationship.Age worked as expected. Older patients had positive SHAP values. Younger patients had negative values. SOFA score was similar. Higher scores meant higher risk.Male sex was linked to slightly higher risk. Isolated CABG showed the same pattern. Diabetes, heart failure, and prior MI all increased risk when present. These findings match what we already know about these conditions.Valve surgery had the smallest effect. Most of its dots were close to zero.

Overall, the beeswarm plot provides a detailed view of how each feature contributes to individual risk predictions. It confirms the feature importance ranking from Figure 4 and adds nuance by showing the direction and spread of effects across the patient population.

Discussion

Main Findings

We evaluated whether preoperative NLR helps predict 10-year death after cardiac surgery, using a tuned machine learning model and SHAP XGBoost, after tuning and bootstrap correction, hit 0.815 AUC, comfortably beating LASSO on the same 12 features (0.780). NLR did about as well as APACHE III on its own (0.709 vs. 0.732, P = 0.630), and the two together bought a small but significant gain (Delta AUC = 0.029, P = 0.002). Piecewise regression suggested a cutoff around 6.5, and the SHAP plot on the test set told a similar story. At a 15% risk threshold, the model flagged a high-risk group with a 23.3% death rate, a 5.1-fold enrichment.

Most NLR work in cardiac surgery is short-term and regression-based. Our study differs by comparing six models with bootstrap validation, directly comparing NLR with APACHE III, and validating the

threshold on an independent test set.

The threshold around 6.5 might explain why the literature is all over the place on cut-points. If risk flattens out above 6.5, then studies with different NLR distributions will land on different "best" cutoffs. However, this plateau should be interpreted cautiously, as it may simply reflect the small number of patients with very high NLR.NLR is cheap and already part of the pre-op blood draw [17], but we should not oversell it. The gain over APACHE III is modest [18]. In practice, this model is probably an add-on, not a replacement.

If the predicted 10-year mortality risk is above 15%, the patient has about a 25% chance of dying within 10 years. These patients may need closer follow-up [19]. However, this is a proposed framework, not a clinical rule, and requires prospective testing before changing practice [20].

Biological Mechanism and Clinical Implications

The biological reason why NLR is a good prognostic marker is that it shows both inflammation and immune status. A high neutrophil count means more systemic inflammation. In cardiac surgery, this may reflect more severe oxidative stress and ischemia-reperfusion injury during cardiopulmonary bypass. On the other hand, a low lymphocyte count suggests poor immune surveillance. When neutrophilia and lymphopenia happen together, which is what a high NLR means, patients have both too much inflammation and not enough immunity. This is not good for long-term recovery.

For the cardiac surgery patients, our results show that preoperative NLR can be a simple and cheap addition to existing risk score. NLR is easy to get from a routine blood test. It needs no extra testing and can be calculated in seconds. For patients with NLR > 6.5, doctors may think about more intensive preoperative preparation, such as nutritional support, checking, and improving heart function. After surgery, these high-risk patients may need closer follow-up. But we want to say that NLR should add to existing risk scores, not replace them. The extra value of NLR over APACHE III alone was small (ΔAUC = 0.029).

Comparison with Existing Studies

Most studies on NLR in cardiac surgery have looked at short-term outcomes, such as 30-day mortality, acute kidney injury, atrial fibrillation, and length of ICU stay. [12]The NLR cutoffs reported in these studies are quite different, ranging from 2.5 to 10.[31] This may be because different studies have different patient populations, measurement timing, and outcomes. In contrast, our study focused on 10-year all-cause mortality, which is a longer follow-up period and may reflect different biological processes. The higher cutoff found in our study (NLR = 6.5) may be explained by the fact that long-term mortality is affected by chronic inflammation, not just acute perioperative inflammation. Chronic inflammation is also known to contribute to atherosclerosis progression, which could explain why NLR predicts mortality over such a long period.

Machine learning has been used more and more for risk prediction in cardiac surgery, but most models still focus on in-hospital or 30-day outcomes[32]. Only a few studies have tried to predict long-term mortality, and very few included inflammatory biomarkers[33]. Our study shows that a tuned XGBoost model can achieve an AUC of 0.822 for 10-year mortality prediction, better than traditional LASSO regression on the same features. We also used SHAP to explain the model's predictions, which helps with the "black box" problem.

Methodological Considerations

We chose XGBoost as our main model because it works well on clinical data tables. [30]It is often better than traditional regression because it can find non-linear relationships and interactions between features. In our study, XGBoost had an AUC of 0.822, while LASSO regression with the same 12 features had an AUC of 0.780. This shows that modeling non-linear effects is useful. But tree-based models are hard to explain. To fix this, we used SHAP, which is based on game theory and can show how much each feature contributes to each prediction. SHAP showed that NLR was the second most important predictor after APACHE III. It also showed the non-linear relationship between NLR and mortality, which traditional regression would have missed.

Many studies use the Youden index from ROC or just pick an NLR cutoff by feeling. We did it differently. We used piecewise regression with AIC optimization on the training set to find the threshold, then tested it with the maxstat method and 1000 permutations. This data-driven way is more objective than looking at ROC curves by eye. We also used 1000 bootstrap resamples to get the 95% confidence interval of the threshold (5.8–7.2). The threshold was found on the training set, so it may be a little optimistic. But the SHAP plot on the test set showed the same pattern, which supports our finding.

Limitations

Several limitations should be considered when interpreting these findings. This was a single-center study from the U.S., so generalizability to Chinese patients is unknown. The low event rate (4.54%, EPV ~ 18:1) explains the modest PPV [21, 22]. About 40% of patients were excluded due to missing NLR; MICE did not change results, but selection bias cannot be ruled out. Intraoperative variables and LVEF were not available. NLR was measured only once before surgery. Bootstrap internal validation was used, but external validation is needed [23–25]. Threshold selection and model training shared the training set; future work should use nested CV. The non-linear pattern is exploratory and not confirmed.

Future research should go in several directions. First, we need to test our model in other hospitals, including Chinese hospitals, to see if it works in different patient groups. Second, studies could measure NLR at different times (before surgery, day 1 after surgery, at discharge) to find out if changes over time are more predictive than a single measurement. Third, intervention studies are needed to test if treating inflammation in high-NLR patients can improve long-term outcomes. Possible treatments include better statin therapy, nutritional support, or new anti-inflammatory drugs. Fourth, future models could include more data, such as echocardiogram results, intraoperative variables, and maybe even text from clinical

notes. Finally, prospective studies should test whether using NLR-based risk stratification actually leads to better patient outcomes. That would be the best way to show if this model is useful in real clinical practice.

Conclusions

Preoperative NLR independently predicts 10-year death after cardiac surgery, and it performs about as well as APACHE III on its own. The tuned XGBoost model did well overall (optimism-corrected AUC 0.815). Piecewise regression suggested a threshold near NLR = 6.5, with SHAP on the test set backing that up. NLR could be a useful tool for long-term risk stratification, but multi-center prospective validation is still needed before clinical application.

Abbreviations

NLR: Neutrophil-to-lymphocyte ratio; SHAP: SHapley Additive exPlanations; AUC: Area under the receiver operating characteristic curve; DCA: Decision curve analysis; MIMIC-IV: Medical Information Mart for Intensive Care IV; CABG: Coronary artery bypass grafting; ICU: Intensive care unit; SVM: Support vector machine; KNN: K-nearest neighbors; XAI: Explainable artificial intelligence; EPV: Events per variable; AIC: Akaike Information Criterion; PPV: Positive predictive value; NPV: Negative predictive value; MICE: Multiple imputation by chained equations.

Declarations

Ethics approval and consent to participate

This study used the publicly available, de-identified MIMIC-IV database. The use was approved by the institutional review boards of the contributing institutions. Informed consent was waived due to the retrospective design.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Jiao ruonan conceived the study idea, designed the research methodology, performed the statistical analysis and machine learning modeling, interpreted the results, and drafted the manuscript. Zhou chenghui supervised the study, provided critical input on the research design and data interpretation, and revised the manuscript for important intellectual content. Both authors read and approved the final version of the manuscript.

Acknowledgement

The authors acknowledge the use of AI-assisted language editing tools for manuscript preparation. The authors take full responsibility for the content, analysis, and conclusions of this manuscript. No external funding was received for this study.

Data Availability

The data used in this study are from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which is a publicly available critical care database. Access to MIMIC-IV requires completion of the CITI Data or Specimens Only Research training course and approval of a credentialed application by PhysioNet. The database can be accessed at https://physionet.org/content/mimiciv/. The code used for data analysis and machine learning modeling is available from the corresponding author upon reasonable request.

References

1. Song M, Graubard BI, Rabkin CS, Engels EA. Neutrophil-to-lymphocyte ratio and mortality in the United States general population. Sci Rep. 2021;11:464.
2. Li X, Liu M, Wang G. The neutrophil-lymphocyte ratio is associated with all-cause and cardiovascular mortality in cardiovascular patients. Sci Rep. 2024;14:26692.
3. Song S, Chen L, Yu R, Zhu J. Neutrophil-to-lymphocyte ratio as a predictor of all-cause and cardiovascular mortality in coronary heart disease and hypertension. Front Endocrinol (Lausanne). 2024;15:1442165.
4. Mortensen EM, Atuegwu N, Sacco SJ. Leveraging nonlinear relationships and interactions to improve 30-day pneumonia readmission machine learning models. PLoS ONE. 2026;21(6):e0349804.

5. Limon D, Satish V, Raghavan N, Morris MX, Muir MT, Rajesh A. Artificial Intelligence in Surgery Revisited: A 2025 Update on Machine Learning for Predicting Complications and Outcomes. Am Surg. 2025;92(3):658–74.
6. Lu SC, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Front Oncol. 2023;13:1129380.
7. Mastropietro A, Feldmann C, Bajorath J. Calculation of exact Shapley values for explaining support vector machine models using the radial basis function kernel. Sci Rep. 2023;13(1):19561.
8. Mastropietro A, Bajorath J. Protocol to explain support vector machine predictions via exact Shapley value computation. STAR Protoc. 2024;5(2):103010.
9. Negin Z, Cui K, Gupta A. Trends in Opioid Prescriptions Based on the Medical Information Mart for Intensive Care IV (MIMIC-IV) Database From 2008 to 2022. Cureus. 2026;e116382.
10. Kang C, Li F, Zhao H, et al. Association of statin use with 28-day mortality in the Medical Information Mart for Intensive Care IV database: a retrospective cohort study. J Thorac Dis. 2025;17(1):429–41.
11. Li C, Ren Q, Li X, Han H, Xie K, Wang G. Association between furosemide administration and clinical outcomes in patients with sepsis-associated acute kidney injury receiving renal replacement therapy: a retrospective observational cohort study based on MIMIC-IV database. BMJ Open. 2023;13(7):e074046.
12. Perry LA, Liu Z, Loth J, Penny-Dimri JC, Plummer M, Segal R, Smith J. Perioperative Neutrophil-Lymphocyte Ratio Predicts Mortality After Cardiac Surgery: Systematic Review and Meta-Analysis. J Cardiothorac Vasc Anesth. 2022;36(5):1296–303.
13. Urbanowicz TK, Michalak M, Gąsecka A, Olasińska-Wiśniewska A, Perek B, Rodzki M, Bociański M, Jemielity M. A Risk Score for Predicting Long-Term Mortality Following Off-Pump Coronary Artery Bypass Grafting. J Clin Med. 2021;10(14):3032.
14. Wu XB, Zhong JL, Wang SW, et al. Neutrophil-to-Lymphocyte Ratio Is Associated With Circumferential Wall Enhancement of Unruptured Intracranial Aneurysm. Front Neurol. 2022;13:879882.
15. Park J, Park J, Shin JH, et al. Prognostic Value of the Neutrophil-to-Lymphocyte Ratio before and after Radiotherapy for Anaplastic Thyroid Carcinoma. Cancers (Basel). 2021;13(8):1913.
16. Johnson PJ, Dhanaraj S, Berhane S, Bonnett L, Ma YT. The prognostic and diagnostic significance of the neutrophil-to-lymphocyte ratio in hepatocellular carcinoma: a prospective controlled study. Br J Cancer. 2021;125(5):714–6.
17. Kekeç AF, Çolak TS. Is there a predictive value of the preoperative neutrophil-lymphocyte ratio in terms of intensive care need in geriatric patients who underwent pertrochanteric fracture surgery? Ulus Travma Acil Cerrahi Derg. 2022;28(8):1164–9.
18. Wang YT, Huang YF. Predictive value of critical illness scores for outcomes of patients with community-acquired pneumonia: Analysis of critical care data from the MIMIC-IV database. BMC Pulm Med. 2026;26(1):366.

19. Eiriksdottir T, Ardal S, Jonsson BA, et al. Predicting the probability of death using proteomics. Commun Biol. 2021;4(1):758.
20. Sládeček S, Topinková E, Laššánová M. Selection of an Appropriate Tool for Assessing Geriatric Frailty in Clinical Practice: A Narrative Review. Clin Interv Aging. 2026;21:624555.
21. Gallagher SK, Wang J, Lumbard K, Dodd LE, Proschan M. Noninferiority testing with censoring when the event rate is low. Stat Med. 2022;41(25):5102–12.
22. Ji L, Alonzo TA. Using Non-inferiority Test of Proportions in Design of Randomized Non-inferiority Trials with Time-to-event Endpoint with a Focus on Low-event-rate Setting. Clin Trials. 2024;22(2):131–41.
23. Zhou M, Liu CZ, Wei ZQ, et al. An individualized nomogram for predicting progression-free survival in systemic anaplastic large cell lymphoma: a multicenter, retrospective, and internally validated study. Hematology. 2026;31(1):2715239.
24. Wang Y, Zhou X. Independent Risk Factors for Placental Abruption in Preeclampsia and a Preliminary Prediction Model: A Retrospective Study with Internal Validation. Int J Womens Health. 2026;18:626309.
25. Fan D, Zhang L, Miao R, et al. Development and internal validation of a clinical prediction model for 1-year recurrence after first-ever ischemic stroke. Front Neurol. 2026;17:1820699.
26. Chen Y, He M, He M, Yuan Z, Chen X. Malnutrition Severity Drives Mortality in Geriatric Heart Failure: A Multicenter Extreme Gradient Boosting Analysis. Rev Cardiovasc Med. 2026;27(5):48595.
27. Zhang Z, Wu J, Wang J, et al. Extreme gradient boosting-based explainable machine learning model for predicting significant fibrosis in autoimmune hepatitis. QJM. 2026;119(1):34–41.
28. Parthasarathi A, Padukudru S, Arunachal S, et al. The Role of Neutrophil-to-Lymphocyte Ratio in Risk Stratification and Prognostication of COVID-19: A Systematic Review and Meta-Analysis. Vaccines (Basel). 2022;10(8):1233.
29. Civinini A, Leggieri F, Massenzi M, Carulli C, Civinini R, Innocenti M. Validation of the Neutrophil-Lymphocyte Ratio as a Mortality Risk Stratification Marker in Patients with Proximal Femoral Fractures. Biomedicines. 2026;14(3):551.
30. Douillet D, Penaloza A, Mahieu R, et al. Outpatient Management of Patients With COVID-19: Multicenter Prospective Validation of the Hospitalization or Outpatient Management of Patients With SARS-CoV-2 Infection Rule to Discharge Patients Safely. Chest. 2021;160(4):1222–31.
31. Li P, Zhang J, Chen K, Pei Q, Ding M, Wang C. Neutrophil-to-lymphocyte ratio: a potential supportive marker for elderly community-acquired bloodstream infections-a retrospective study. PeerJ. 2025;13:e20449.
32. Zea-Vera R, Ryan CT, Navarro SM, et al. Development of a Machine Learning Model to Predict Outcomes and Cost After Cardiac Surgery. Ann Thorac Surg. 2023;115(6):1533–42.
33. Wändell P, Carlsson AC, Larsson A, et al. TNFR1 is associated with short-term mortality in patients with diabetes and acute dyspnea seeking care at the emergency department. Acta Diabetol. 2020;57(10):1145–50.

Figures

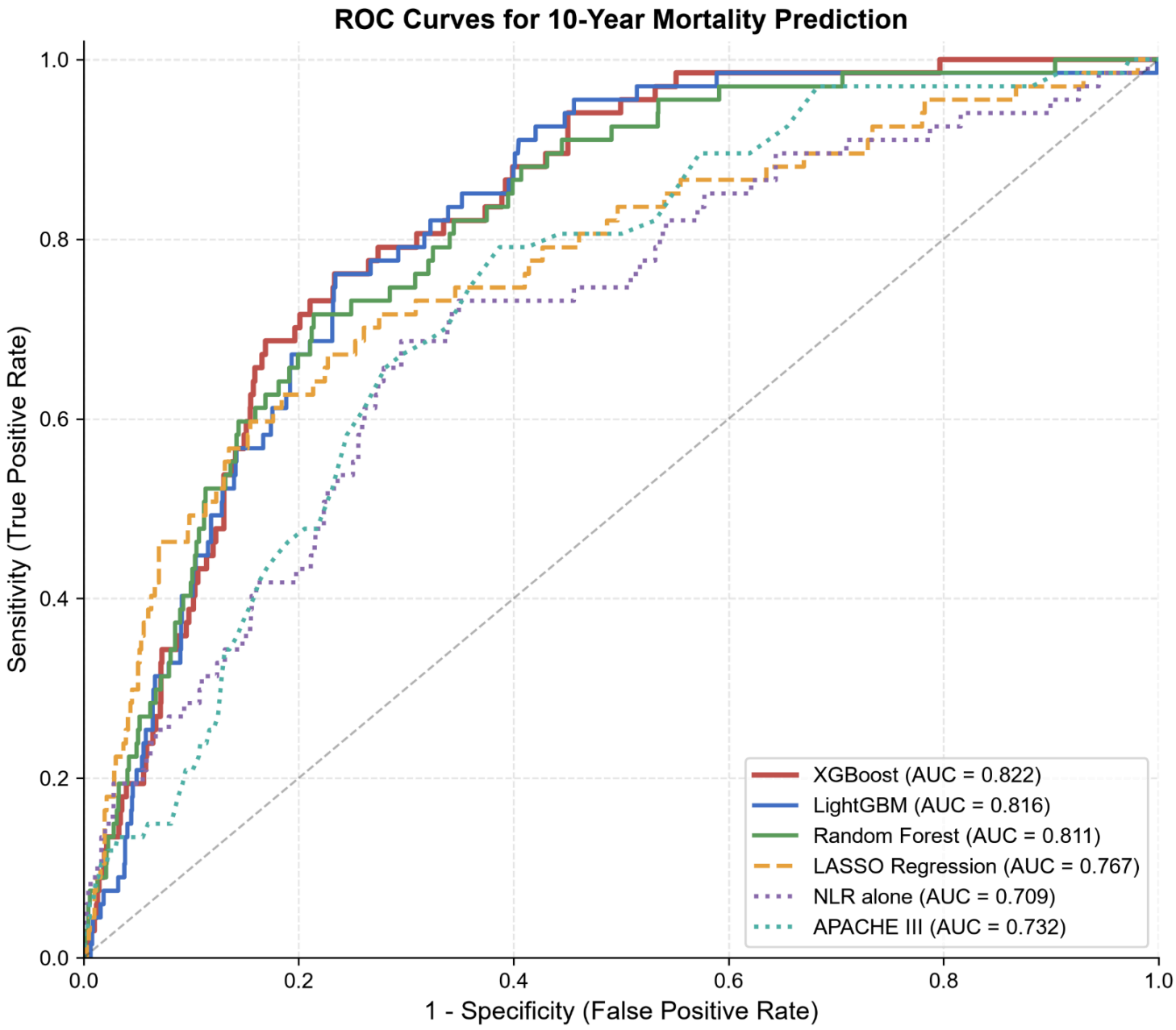


Figure 1

ROC curves of six machine learning models after hyperparameter tuning

Figure 1. ROC curves of six machine learning models after hyperparameter tuning.Receiver operating characteristic (ROC) curves comparing the predictive performance of six machine learning models for 10-year mortality after cardiac surgery.

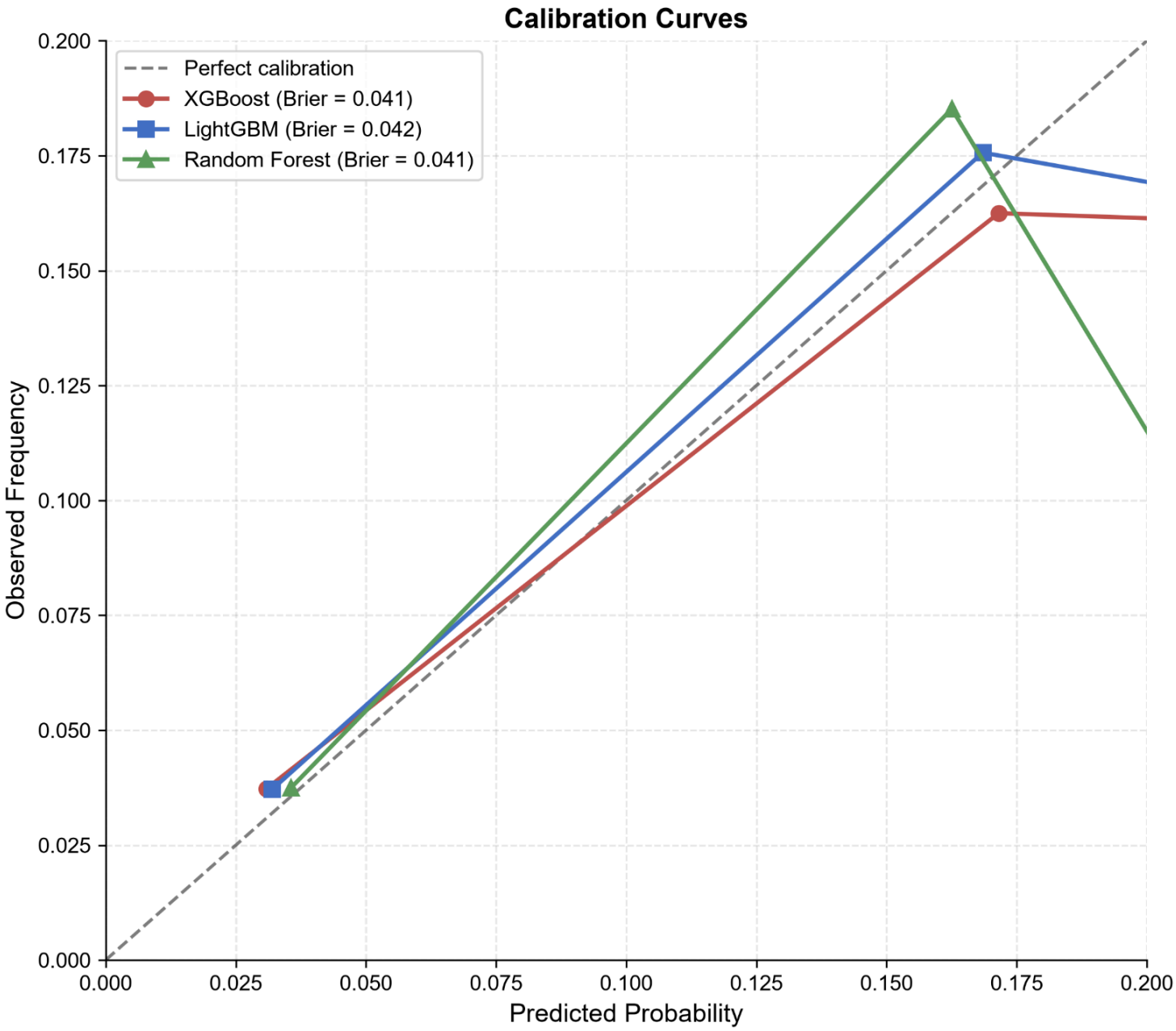


Figure 2

Calibration curves (Brier = 0.041)

Figure 2. Calibration curves of the three best-performing models.Calibration curves for XGBoost, LightGBM, and random forest. The x-axis shows the predicted probability of 10-year mortality, and the y-axis shows the observed mortality frequency. The gray dashed line represents perfect calibration.

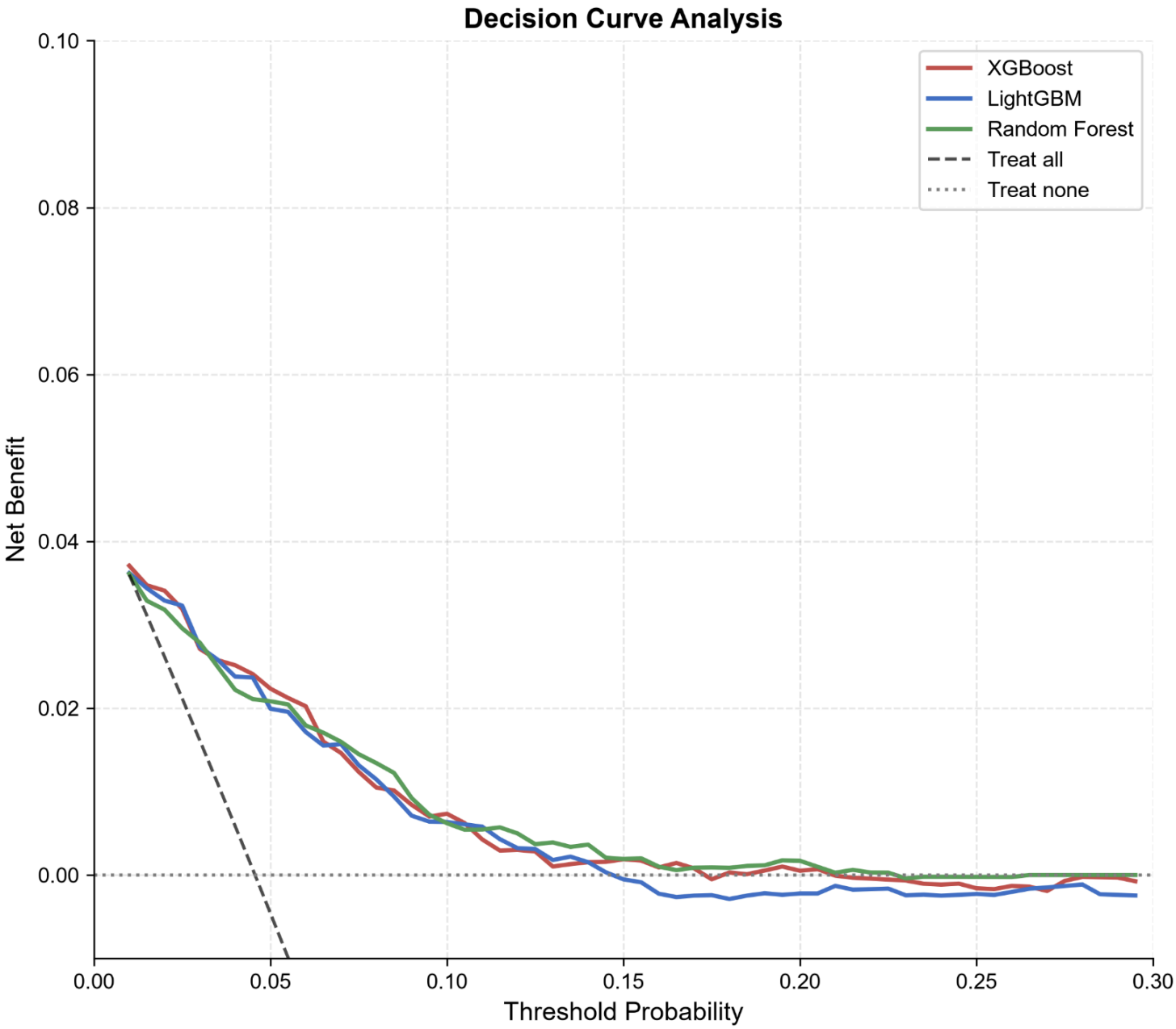


Figure 3

Decision curve analysis

Figure 3. Decision curve analysis for the three best-performing models.Decision curve analysis evaluating the clinical net benefit of XGBoost, LightGBM, and random forest across threshold probabilities from 0% to 30%. The "treat all" and "treat none" strategies are shown as reference lines. All three models demonstrated positive net benefit from approximately 5% to 30% threshold probability.

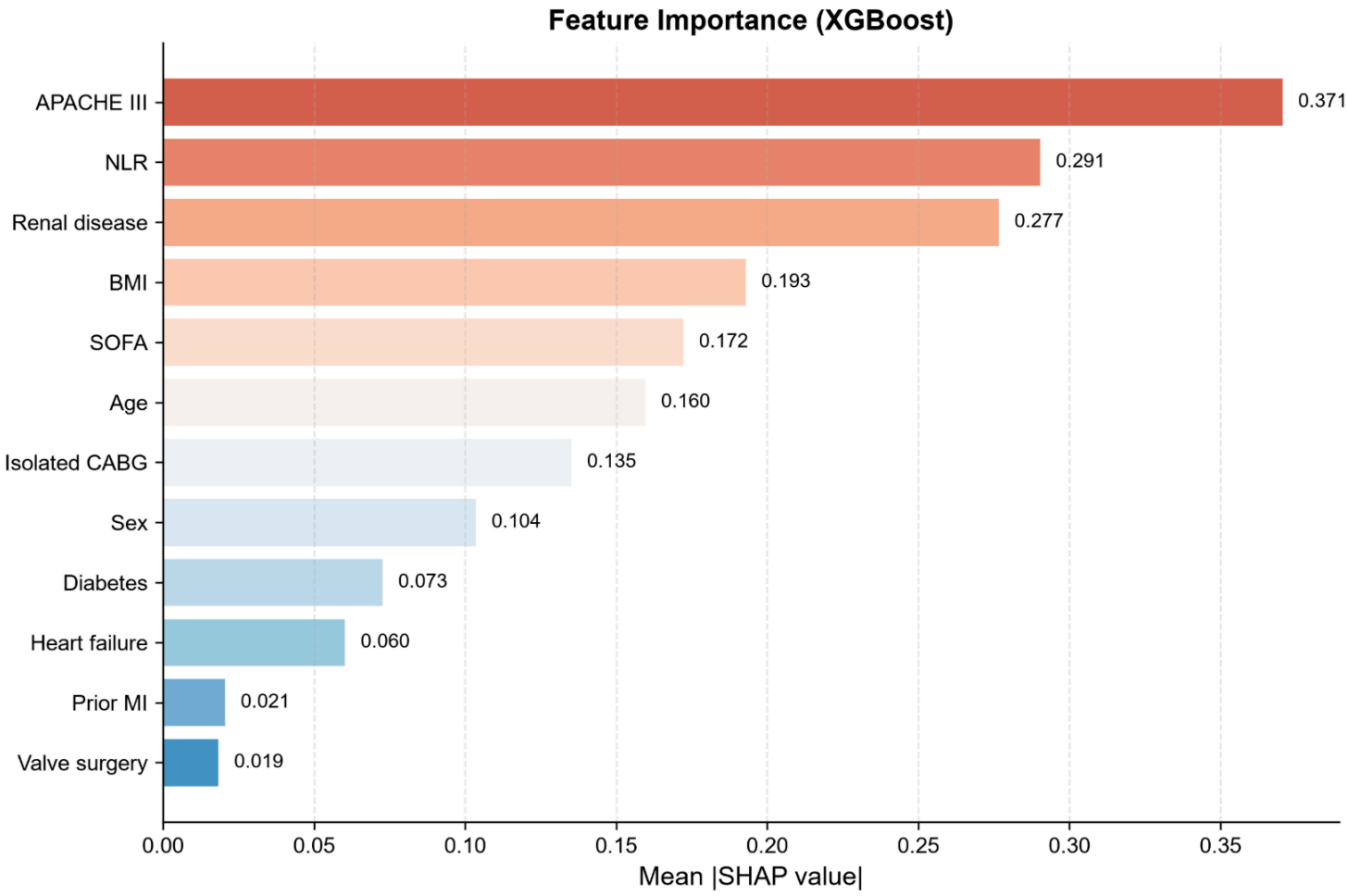


Figure 4

SHAP feature importance ranking (XGBoost, test set)

Figure 4. SHAP feature importance ranking for the XGBoost model.Bar plot showing the mean absolute SHAP value for each of the 12 features included in the XGBoost model.

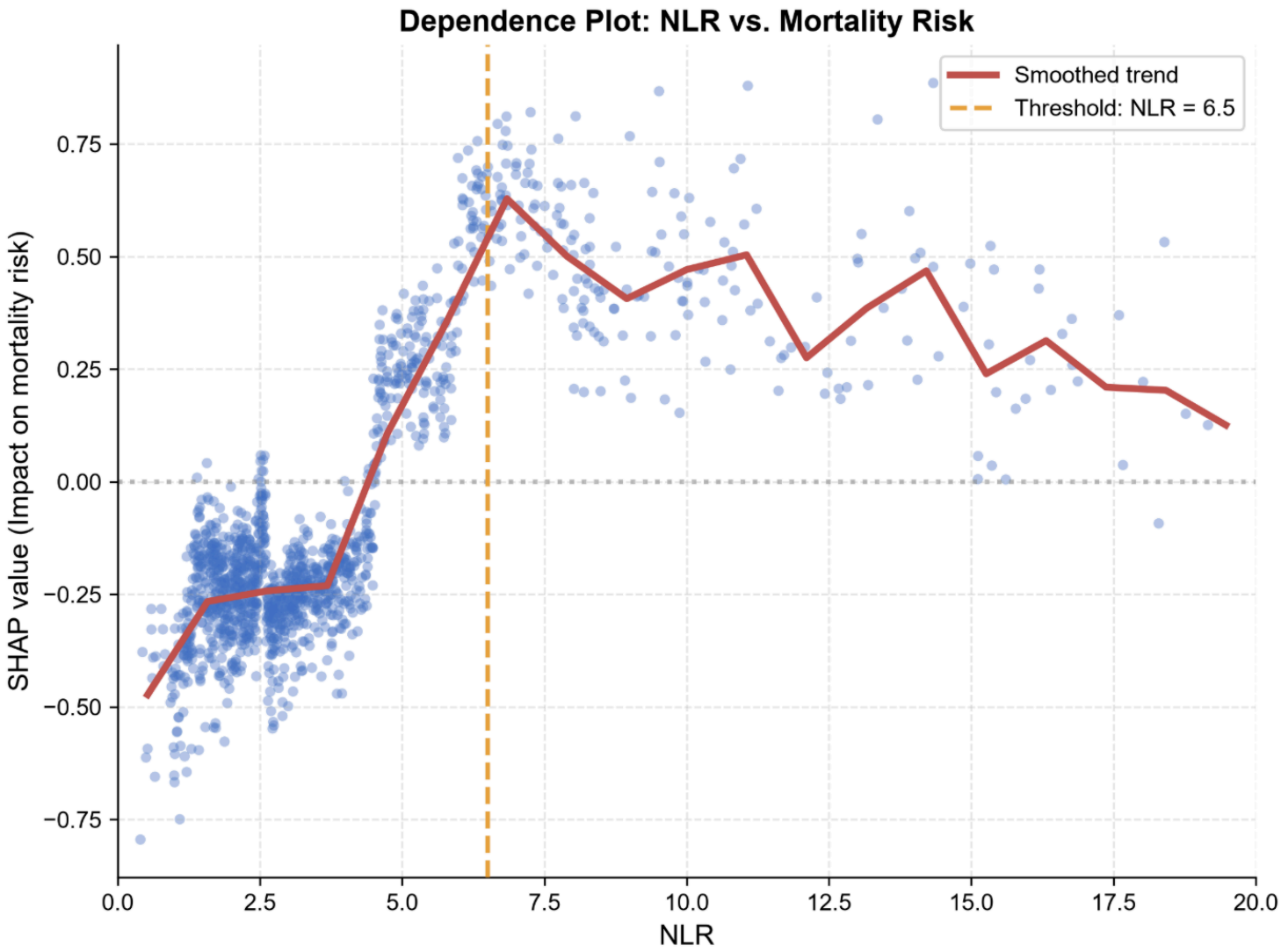


Figure 5

SHAP dependence plot on test set (threshold at 6.5)

Figure 5. SHAP dependence plot for NLR.Scatter plot showing the relationship between NLR values (x-axis) and SHAP values (y-axis) for individual patients in the test set. Each blue dot represents one patient. The red line is the smoothed trend, and the yellow dashed line marks the threshold at NLR = 6.5.

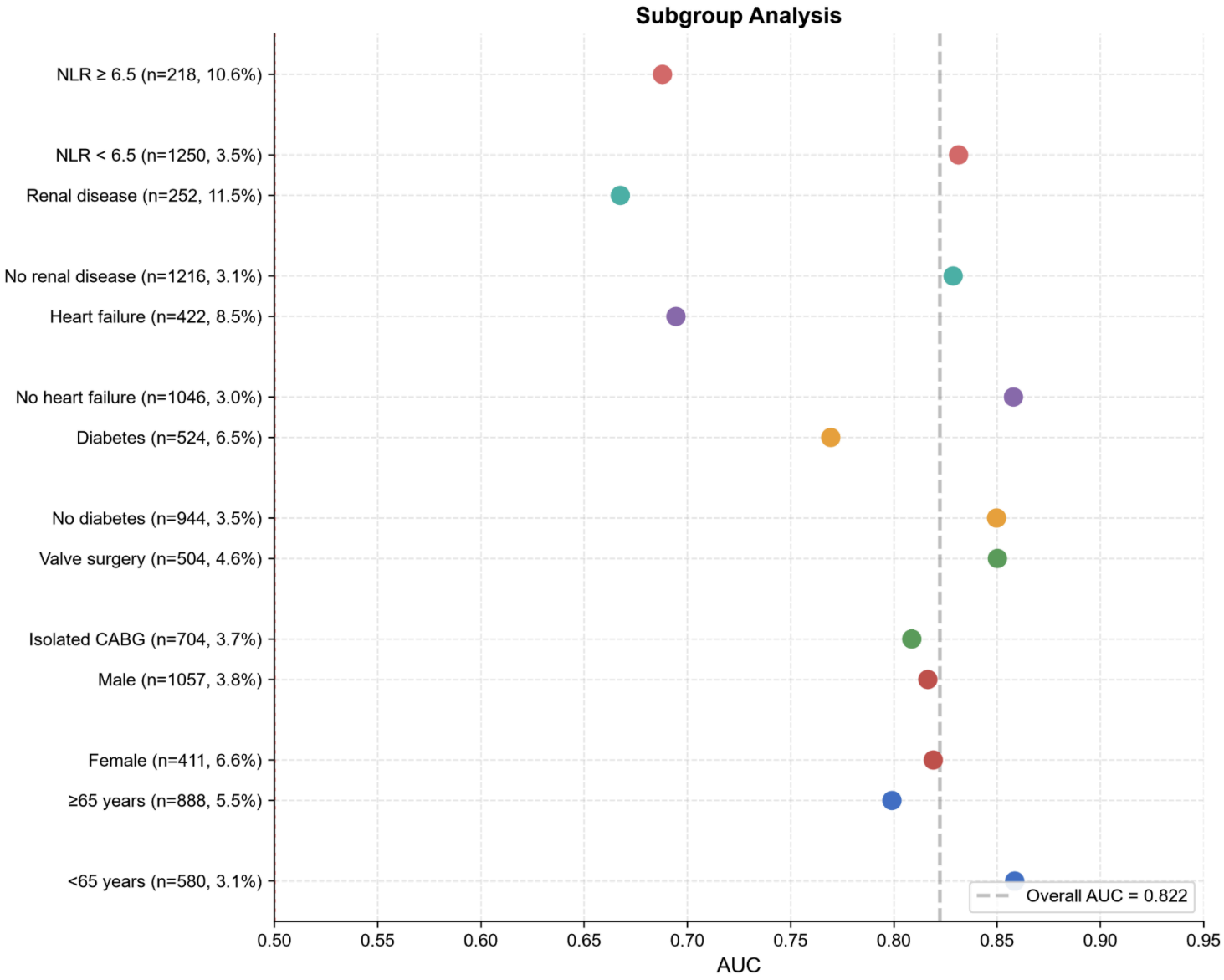


Figure 6

Subgroup analysis forest plot

Figure 6. Subgroup analysis of model performance.Forest plot showing the AUC of the XGBoost model across 14 clinical subgroups. The gray dashed line represents the overall AUC (0.822). Subgroups include NLR level, renal disease, heart failure, diabetes, surgery type, sex, and age.

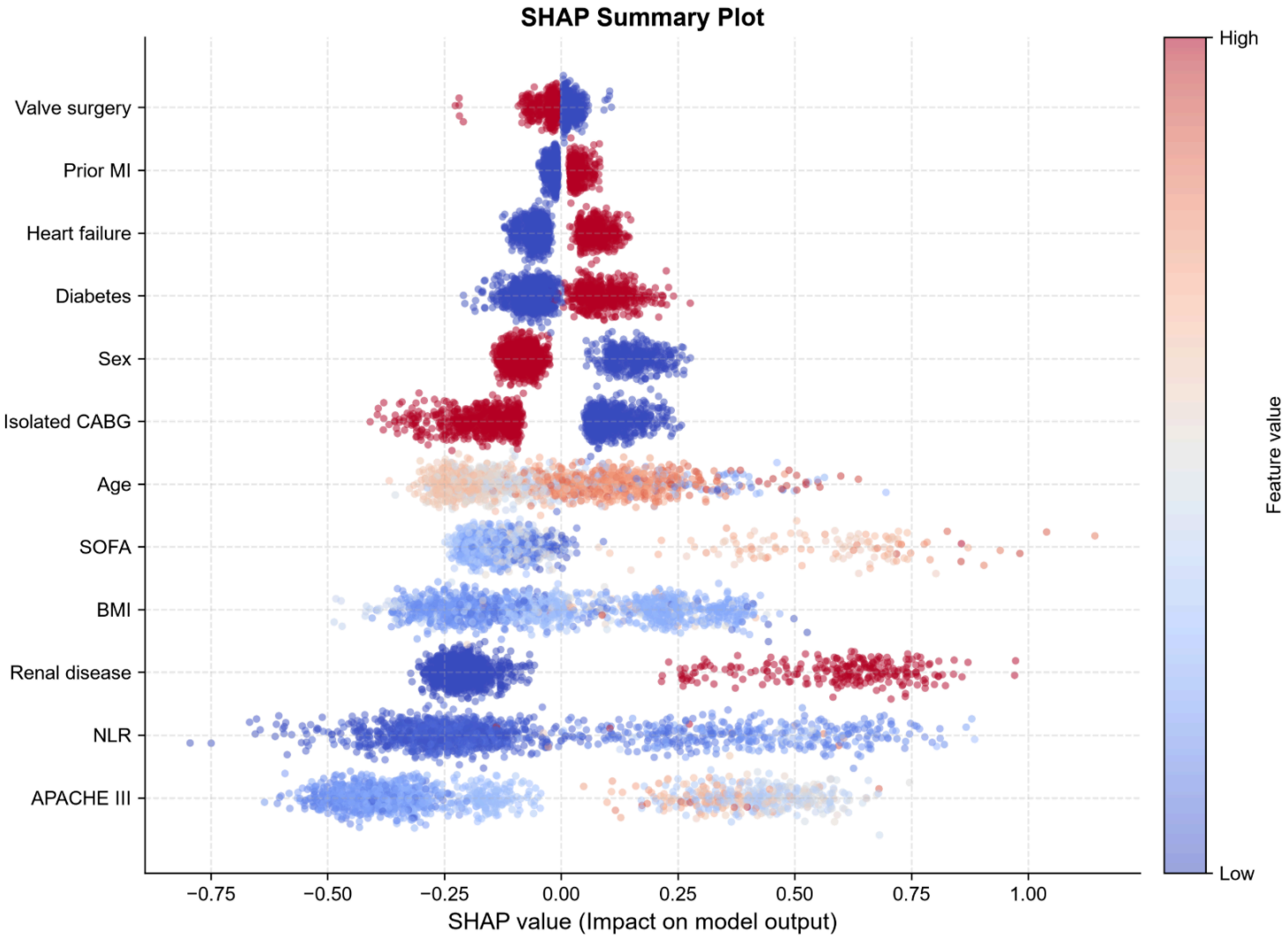


Figure 7

SHAP beeswarm plot (test set)

Figure 7. SHAP beeswarm plot for the XGBoost model.Beeswarm plot showing the distribution of SHAP values for each feature across all patients in the test set. Each dot represents one patient, colored by feature value (red = high, blue = low). Positive SHAP values increase predicted mortality risk, while negative values decrease it. Features are ordered by importance from top to bottom.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1SupplementaryTables.docx