

Early Mortality Prediction in Upper Gastrointestinal Bleeding Using Explainable Machine Learning: Comparison with Established Clinical Risk Scores

Duygu SİDDİKOĞLU

duygu.sddk@gmail.com

Canakkale Onsekiz Mart University Faculty of Medicine

Murat DAŞ

Canakkale Onsekiz Mart University Faculty of Medicine

Ece ÜNAL ÇETİN

Canakkale Onsekiz Mart University Faculty of Medicine

Özge KURTKULAĞI

Canakkale Onsekiz Mart University Faculty of Medicine

Adil Uğur ÇETİN

Mehmet Akif Ersoy State Hospital

Hakan KAYAN

Canakkale Onsekiz Mart University Faculty of Medicine

Senem GÜLER

Canakkale Onsekiz Mart University Faculty of Medicine

Feyza MUTLAY

Canakkale Onsekiz Mart University Faculty of Medicine

Yavuz BEYAZİT

Canakkale Onsekiz Mart University Faculty of Medicine

Research Article

Keywords: upper gastrointestinal bleeding, machine learning, XGBoost, mortality prediction, emergency department, risk stratification, explainable artificial intelligence, SHAP, clinical prediction model, calibration

Posted Date: July 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10145598/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on October 4th, 2026. See the published version at <https://doi.org/10.1186/s12911-026-03883-6>.

Abstract

Background

Early risk stratification in upper gastrointestinal bleeding (UGIB) is essential for emergency department (ED) triage, treatment planning, and mortality reduction. Although conventional clinical scores are widely used, they may not fully capture complex nonlinear interactions among physiological, laboratory, and comorbidity-related predictors. This study evaluated whether explainable machine-learning (ML) models could predict 30-day mortality in UGIB and compared their performance with established clinical risk scores.

Methods

This retrospective single-center cohort study included 719 adult ED presentations with UGIB between January 1, 2015, and January 1, 2026. The primary outcome was all-cause 30-day mortality. ML models were developed using two predictor sets: an early ED predictor set, and a Full model predictor set incorporating transfusion and endoscopic findings. Logistic regression, random forest, gradient boosting, XGBoost, and LightGBM models were evaluated using five-fold stratified group cross-validation. Model performance was assessed using AUROC, AUPRC, F1 score, Brier score, calibration analysis, and decision curve analysis, and compared with the Glasgow-Blatchford Score, Rockall score, AIMS65, and ABC score. Parsimonious models were developed using regularized feature selection, and SHAP analysis was used for model interpretation.

Results

Among 719 presentations, 53 patients (7.4%) died within 30 days. The ED XGBoost model achieved the best overall performance among prespecified models, with an AUROC of 0.789, AUPRC of 0.294, F1 score of 0.355, and Brier score of 0.061. Its performance was similar to the Full XGBoost model. Among traditional scores, ABC showed the highest AUROC (0.742), but the AUROC difference between ED XGBoost and ABC was not statistically significant. A 15-variable Minimal ED XGBoost model achieved an AUROC of 0.814 in exploratory analysis and 0.765 in fully nested sensitivity analysis. SHAP analysis identified lymphocyte count, lactate, potassium, C-reactive protein, respiratory rate, albumin, platelet count, glucose, hemoglobin, and Glasgow Coma Scale as key predictors.

Conclusion

Machine-learning models based on routinely available ED data showed comparable or superior discrimination, better calibration, and greater clinical utility than traditional risk scores for predicting 30-

day mortality in UGIB. A parsimonious explainable ED XGBoost model may provide a practical early risk stratification tool, although external validation is required before clinical implementation.

Introduction

Upper gastrointestinal bleeding (UGIB) remains one of the most frequent causes of emergency department (ED) visits with an incidence of 48–160 per 100,000 persons and over 250,000 admissions annually (1). Despite advances in medical treatment, endoscopic techniques and improvements in supportive care, UGIB is still associated with significant morbidity and mortality (2). The mortality burden is particularly high among elderly patients and those with multiple comorbidities, suggesting that UGIB should not be viewed merely as an acute hemorrhagic event, but rather as a complex clinical condition closely linked to the patient's physiological reserve and systemic response to bleeding (3).

Early risk stratification in UGIB is crucial for guiding resuscitation, determining the timing and urgency of endoscopy, making admission or discharge decisions, predicting the need for intensive care, planning transfusion strategies, and safely identifying low-risk patients suitable for outpatient management (4). Therefore, there is a continuous need for reliable risk prediction models in UGIB to improve triage, guide the timing of urgent endoscopies, and reduce overall mortality rates. For this purpose, various risk stratification systems have been developed, such as the Glasgow-Blatchford Score (GBS), Rockall score, AIMS65, and the recently developed Age, Blood tests and Comorbidities (ABC) score (5–7). While these scores are widely used in clinical practice, they have some significant limitations. Most of these systems rely on combining a limited number of variables using predetermined weights and linear assumptions, which may oversimplify the complex physiological interactions underlying clinical outcomes in UGIB (8–11).

Machine learning (ML) has emerged as a promising approach for clinical risk prediction by capturing complex nonlinear relationships and interactions among demographic characteristics, comorbidities, physiological parameters, and laboratory findings. Several studies have demonstrated the potential of ML models to predict mortality, adverse outcomes, and the need for intervention in patients with gastrointestinal bleeding, with some models outperforming conventional clinical risk scores (12–15). However, superior discrimination alone does not guarantee clinically reliable prediction. Predictive performance may be artificially inflated when models incorporate information unavailable at the intended time of prediction or when preprocessing, feature selection, and validation procedures introduce information leakage (16–18). Moreover, contemporary evaluation of clinical prediction models increasingly emphasizes calibration and clinical utility in addition to discrimination, particularly in settings where risk estimates are intended to guide clinical decision-making (19, 20). These considerations are especially relevant in the ED, where risk stratification must rely on information available at the time of initial assessment and directly support triage and early management decisions (21).

Given these methodological and clinical considerations, we investigated whether ML models could predict 30-day mortality in patients presenting with UGIB and how their performance compared with traditional clinical risk scores, including the GBS, Rockall score, AIMS65, and ABC score. To reflect clinical practice, predictors were defined according to their temporal availability, allowing separate evaluation of an early ED model based on information available at presentation and a Full model that additionally incorporated transfusion and endoscopic findings. We also evaluated whether regularized feature selection could identify a smaller set of early predictors while maintaining predictive performance and used SHapley Additive exPlanations (SHAP) to support model interpretation.

Materials and Methods

Design and population

This retrospective single-center cohort study was conducted in the ED of Çanakkale Onsekiz Mart University (COMU) Hospital, Türkiye. Clinical and laboratory data were obtained from the institutional electronic health record system. The study was approved by the COMU Ethics Committee (approval date: 17.06.2026; decision number: 2026-09-22) and conducted in accordance with the Declaration of Helsinki.

Consecutive adult patients (≥ 18 years) presenting with UGIB between January 1, 2015, and January 1, 2026, were screened for eligibility. UGIB was defined as the presence of hematemesis, coffee-ground vomiting, or melena, with a gastrointestinal bleeding source identified by endoscopy and/or computed tomography (CT). Both variceal and non-variceal bleeding episodes were considered as eligible for inclusion.

Recurrent presentations were retained as separate clinical encounters because they represented distinct bleeding episodes. To prevent patient-level information leakage, all encounters belonging to the same patient were assigned to the same cross-validation fold using an anonymized patient identifier.

Patients aged < 18 years, pregnant patients, patients with trauma-related bleeding, and patients with unavailable 30-day mortality data were excluded. After exclusions, 719 presentations were included in the final analysis.

Data Collection and Predictor Sets

Demographic, clinical, physiological, laboratory, transfusion, and endoscopic data were retrospectively extracted from the electronic health record system. Variable definitions and coding schemes are provided in Additional file 1.

Two predictor sets were defined according to data availability. The ED predictor set included variables available during the initial ED assessment, whereas the Full predictor set additionally incorporated

transfusion status and endoscopic findings. The ED and Full predictor sets comprised 63 and 72 predictors, respectively.

The GBS, ABC score, Rockall score, and AIMS65 were included for comparison as traditional risk scores.

Outcome

The primary outcome was all-cause mortality within 30 days of the index ED presentation. Mortality status, date of death, and cause of death were obtained from the national death notification system. Among the 719 included presentations, 53 (7.4%) deaths occurred within 30 days.

Data Preprocessing

Preprocessing was performed within model pipelines. Missing numeric and categorical variables were imputed using the median and most frequent category, respectively. Numeric variables were standardized and categorical variables were one-hot encoded. All preprocessing steps were fitted within training folds.

Statistical Analysis

Continuous variables were summarized as median (interquartile range) and compared using the Mann–Whitney U test. Categorical variables were summarized as frequency (percentage) and compared using the chi-square test or Fisher’s exact test, as appropriate.

Model discrimination was quantified using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC). Confidence intervals for performance metrics and differences in AUROC between selected models and clinical risk scores were estimated using the percentile bootstrap method with 2,000 patient-level paired bootstrap samples. Two-sided p-values < 0.05 were considered statistically significant.

Model Development

Five supervised learning algorithms were evaluated: logistic regression with L2 regularization, random forest, gradient boosting, XGBoost, and LightGBM. Each algorithm was developed separately using the ED and Full predictor sets. Class-weighted learning was applied where supported to address class imbalance.

Models were internally validated using five-fold stratified group cross-validation. Stratification approximately preserved the distribution of 30-day mortality across folds, while grouping by anonymized patient identifier ensured that recurrent presentations from the same patient were assigned to the same fold, preventing patient-level information leakage.

Preprocessing and model fitting were performed exclusively within the training folds and subsequently applied to the corresponding validation fold. Out-of-fold predictions were aggregated across all encounters and used to estimate internally validated model performance.

Parsimonious Model Development

Elastic Net and LASSO regularized logistic regression models were evaluated using the ED predictor set to identify a reduced set of predictors. Regularization parameters were selected according to cross-validated predictive performance. Predictors retained by the selected regularized model were subsequently used to develop parsimonious versions of logistic regression, random forest, gradient boosting, XGBoost, and LightGBM models.

As a sensitivity analysis, a fully nested feature-selection framework was implemented for the parsimonious modeling strategy. Within each outer training fold, Elastic Net hyperparameters were selected using inner stratified group cross-validation, and predictors with nonzero coefficients were retained. Candidate parsimonious models were then trained using the selected predictors and evaluated exclusively in the corresponding held-out outer validation fold. Predictor-selection stability was summarized as the proportion of outer folds in which each predictor was retained.

Model Performance Assessment

Discrimination was assessed using AUROC, AUPRC, and the maximum F1 score. Calibration was evaluated using the Brier score and calibration plots. Clinical utility was assessed using decision curve analysis. Performance was compared with the GBS, ABC score, Rockall score, and AIMS65.

Model Explainability

The parsimonious ED XGBoost model was examined using SHAP. Predictor importance was summarized using mean absolute SHAP values and visualized using SHAP summary plots.

Software and Reporting

Analyses were performed using Python (version 3.12.13). Data preprocessing, statistical analyses, model development, and explainability analyses were conducted using pandas, SciPy, scikit-learn, XGBoost, LightGBM, and SHAP. Reporting followed the TRIPOD + AI recommendations (22), and the completed TRIPOD + AI checklist is provided in Additional file 2.

Results

The cohort included 719 ED presentations, of which 53 (7.4%) resulted in 30-day mortality. Baseline characteristics according to 30-day mortality status are summarized in Table 1. Compared with survivors, patients with 30-day mortality had lower albumin (3.30 vs 3.59 g/dL; $p < 0.001$), and

lymphocyte count (0.95 vs $1.55 \times 10^9/L$; $p < 0.001$), and higher lactate (2.50 vs 1.70 mmol/L; $p < 0.001$), C-reactive protein (12.00 vs 3.46 mg/dL; $p < 0.001$). All traditional risk scores were higher among patients with 30-day mortality, with the largest difference observed for the ABC score ($p < 0.001$).

Table 1
Baseline characteristics according to 30-day mortality status

Variable	Level	30-day mortality (no = 666)	30-day mortality (yes = 53)	p value
Age, years		73.50 [63–81]	74 [67–84]	0.072*
Sex, n (%)	Male	406 (61.0)	35 (66.0)	0.559**
Systolic blood pressure, mmHg		121 [110-131.75]	116 [105–123]	0.022*
Diastolic blood pressure, mmHg		75 [67–80]	70 [63–80]	0.039*
Heart rate, beats/min		88 [80–97]	90 [80–100]	0.243*
Respiratory rate, /min		17 [15–18]	16 [15–18]	0.880*
Oxygen saturation, %		98 [97–99]	98 [95–98]	0.008*
Glasgow Coma Scale		15 [15–15]	15 [15–15]	< 0.001*
White blood cell count, mm ³ x10 ³		9.22 [6.95–12.01]	10.70 [8.11– 14.81]	0.020*
Hemoglobin, g/dl		8.80 [6.70–11]	9 [6.80–11]	0.887*
Platelet count, mm ³ x10 ³		242 [181-309.75]	247 [178–378]	0.352*
Red cell distribution width, %		15 [13.60–17]	15.30 [14.60–18]	0.011*
Glucose, mg/dL		123 [104-158.75]	148 [114–198]	0.009*
Urea, mg/dL		64 [41-99.50]	70 [47–103]	0.238*
Creatinine, mg/dL		1 [0.79–1.32]	0.99 [0.70–1.40]	0.664*
C-reactive protein, mg/L		3.46 [1–13]	12 [2.90–71]	< 0.001*
Alanine aminotransferase, U/L		11 [8–17]	10 [6–18]	0.502*
Aspartate aminotransferase, U/L		16 [12–22]	17 [11.70–23]	0.404*
Alkaline phosphatase,U/L		71 [53–95]	85 [65–102]	0.014*
Direct bilirubin, mg/dL		0.20 [0.10–0.30]	0.22 [0.20–0.50]	0.004*
Albumin, g/dL		3.59 [3.20-4]	3.30 [2.77–3.60]	< 0.001*
Lactate, mmol/L		1.70 [1.10–2.30]	2.50 [1.70–5.40]	< 0.001*
International normalized ratio		1.04 [1-1.30]	1.20 [1-1.40]	0.020*

Variable	Level	30-day mortality (no = 666)	30-day mortality (yes = 53)	p value
Sodium, mmol/L		139 [136–141]	137 [133–140]	0.022*
Potassium, mmol/L		4.30 [4-4.70]	4.40 [3.96-5]	0.541*
Calcium, mg/dL		8.64 [8.20-9]	8.50 [7.93-9]	0.088*
Melena	Yes	486 (73.0)	33 (62.3)	0.130**
Hematemesis	Yes	211 (31.7)	25 (47.2)	0.031**
Syncope	Yes	50 (7.5)	4 (7.5)	0.999***
Number of chronic diseases	>=2	368 (55.3)	37 (69.8)	
	1	210 (31.5)	14 (26.4)	0.056***
	None	88 (13.2)	2 (3.8)	
Medication use	Yes	579 (86.9)	49 (92.5)	0.343**
Antiplatelet use	Yes	208 (31.3)	11 (20.8)	0.146**
Anticoagulant use	Yes	176 (26.5)	17 (32.1)	0.472**
NSAID use	Yes	91 (13.7)	6 (11.3)	0.780**
Transfusion performed	Yes	455 (68.4)	45 (84.9)	0.018**
Endoscopy performed	Yes	498 (74.9)	39 (73.6)	0.963**
Glasgow-Blatchford score		11 [9–13]	13 [10–14]	0.012*
Rockall score		5 [3–6]	6 [5–7]	< 0.001*
AIMS65 score		1 [1–2]	1 [1–2]	< 0.001*
ABC score		4 [2–6]	7 [5–9]	< 0.001*
Continuous variables are reported as median [IQR], and categorical variables as n (%). *Mann-Whitney U test was used for continuous variables; **Pearson chi-square test or ***Fisher exact test was used for categorical variables, as appropriate.				

Impaired consciousness (GCS < 15) was more frequent among patients with 30-day mortality (18.9% vs 1.2%; $p < 0.001$). Among comorbidities, malignancy (28.3% vs 12.6%; $p = 0.003$), stroke (20.8% vs 8.9%; $p = 0.010$), peripheral arterial disease (7.5% vs 1.2%; $p = 0.008$), and chronic kidney disease (17.0% vs 8.0%; $p = 0.037$) were more common among patients with 30-day mortality. Gastritis and gastric ulcer were the most common endoscopic findings, whereas esophageal varices were more frequent among patients with 30-day mortality (13.2% vs 4.7%; $p = 0.025$).

Model performance is summarized in Table 2. Among the prespecified models, ED XGBoost achieved the best overall performance (AUROC 0.789, AUPRC 0.294, F1 score 0.355, Brier score 0.061). The Full

XGBoost model showed similar performance (AUROC 0.785, AUPRC 0.291, F1 score 0.355, Brier score 0.061).

Table 2
Primary Model Performance

Feature set	Model	AUC	AUPRC	F1	Brier score
ED Model	Logistic Regression	0.693	0.234	0.306	0.139
ED Model	Random Forest	0.776	0.234	0.303	0.106
ED Model	Gradient Boosting	0.765	0.23	0.312	0.069
ED Model	XGBoost	0.789	0.294	0.355	0.061
ED Model	LightGBM	0.745	0.232	0.336	0.093
Full Model	Logistic Regression	0.683	0.222	0.322	0.138
Full Model	Random Forest	0.764	0.249	0.301	0.108
Full Model	Gradient Boosting	0.769	0.232	0.316	0.068
Full Model	XGBoost	0.785	0.291	0.355	0.061
Full Model	LightGBM	0.745	0.233	0.343	0.093
benchmark risk score	GBS	0.603	0.103	0.169	0.345
benchmark risk score	Rockall	0.693	0.205	0.228	0.167
benchmark risk score	AIMS65	0.628	0.129	0.234	0.106
benchmark risk score	ABC	0.742	0.201	0.321	0.142
Comparison	Δ AUROC	95% CI_low	95% CI_high	p value	
ED XGBoost - GBS	0.182	0.116	0.271	< 0.001	
ED XGBoost - Rockall	0.092	0.023	0.177	0.002	
ED XGBoost - AIMS65	0.157	0.083	0.235	< 0.001	
ED XGBoost - ABC	0.042	-0.032	0.122	0.264	

Among the traditional risk scores, ABC showed the highest AUROC (0.742), followed by Rockall (0.693), AIMS65 (0.628), and GBS (0.603). In bootstrap analysis, the AUROC difference between ED XGBoost and ABC was not statistically significant (Δ AUROC 0.042; $p = 0.264$). ROC and precision–recall curves are presented in Fig. 1A and Fig. 1B. The ML models consistently demonstrated superior precision–recall performance compared with the traditional risk scores, with ED XGBoost and Full XGBoost achieving the highest average precision values. ED XGBoost demonstrated lower prediction error than ABC, with a lower Brier score (0.061 vs 0.142) (Fig. 1C; Additional file 3). Decision curve analysis showed positive net

benefit for both ED and Full XGBoost models across the evaluated threshold range, whereas the clinical risk scores provided limited net benefit at higher thresholds (Fig. 1D). Because the ED predictor set contained 63 variables, an exploratory feature-reduction analysis was subsequently performed to evaluate whether predictive performance could be maintained using a smaller set of predictors.

Elastic Net and LASSO were evaluated for predictor selection. The Elastic Net model (C = 0.03, l1_ratio = 0.5) achieved the highest AUROC and was selected for feature reduction (Additional file 4). The selected model retained 15 ED predictors: lymphocyte count, lactate, albumin, Glasgow Coma Scale, hemoglobin, respiratory rate, glucose, C-reactive protein, platelet count, potassium, age, chloride, oxygen saturation, direct bilirubin, and gamma-glutamyl transferase.

Performance of the reduced-predictor models is summarized in Table 3. Minimal ED XGBoost achieved an AUROC of 0.814, AUPRC of 0.316, F1 score of 0.379, and Brier score of 0.060. Minimal ED Logistic Regression achieved an AUROC of 0.809, AUPRC of 0.373, F1 score of 0.398, and Brier score of 0.146.

Table 3
Exploratory Feature-Reduced ED Model Performance

Model or score	AUROC (95% CI)	AUPRC (95% CI)	F1 score	Brier score
Minimal ED Random Forest	0.816 (0.722–0.844)	0.275 (0.162–0.365)	0.323	0.104
Minimal ED XGBoost	0.814 (0.751–0.869)	0.316 (0.208–0.453)	0.379	0.060
Minimal ED LR	0.809 (0.739–0.870)	0.373 (0.253–0.500)	0.398	0.146
Minimal ED Gradient Boosting	0.769 (0.704–0.833)	0.228 (0.145–0.321)	0.333	0.069
Minimal ED LightGBM	0.750 (0.648–0.802)	0.253 (0.147–0.352)	0.333	0.106
ABC score	0.742 (0.666–0.811)	0.201 (0.136–0.307)	0.321	0.142
Comparison	Δ AUROC	95% CI_low	95% CI_high	p value
Minimal ED XGBoost - GBS	0.214	0.098	0.293	0.001
Minimal ED XGBoost - Rockall	0.125	0.051	0.201	0.001
Minimal ED XGBoost - AIMS65	0.187	0.101	0.275	0.001
Minimal ED XGBoost - ABC	0.072	0.004	0.144	0.035

Compared with ABC, Minimal ED XGBoost showed a higher AUROC ($\Delta\text{AUROC} = 0.072$; $p = 0.035$). Calibration and decision curve analyses are shown in Fig. 2. Minimal ED XGBoost demonstrated closer agreement between predicted and observed mortality rates and a lower Brier score than both Minimal ED Logistic Regression and ABC (0.060 vs 0.146 and 0.142, respectively) (Fig. 2A). In decision curve analysis, Minimal ED XGBoost provided higher net benefit than ABC, Rockall, and GBS across most evaluated threshold probabilities (Fig. 2B). Bin-level calibration estimates are provided in Additional file 5. Although Minimal ED Logistic Regression achieved higher AUPRC and F1 values, its markedly higher Brier score suggests less accurate probability estimation despite comparable discrimination.

Additional calibration statistics are presented in Additional file 6. Among all evaluated models, the Minimal ED XGBoost model demonstrated the calibration intercept closest to the ideal value of zero, whereas the XGBoost-based models generally exhibited calibration slopes closer to the ideal value of one than the Minimal ED Logistic Regression and ABC models. These findings were broadly consistent with the calibration plots, bin-level calibration estimates, and Brier scores, further supporting the favorable calibration performance of the Minimal ED XGBoost model.

SHAP analysis of the Minimal ED XGBoost model identified lymphocyte count, lactate, potassium, C-reactive protein, respiratory rate, albumin, platelet count, glucose, hemoglobin, and Glasgow Coma Scale as the most important predictors (Fig. 3). Lower lymphocyte count and albumin, higher lactate and C-reactive protein, and lower Glasgow Coma Scale values were associated with higher predicted mortality risk.

In a fully nested sensitivity analysis designed to account for feature-selection optimism, the Minimal ED XGBoost model achieved an AUROC of 0.765, AUPRC of 0.303, F1 score of 0.344, and Brier score of 0.061. The number of selected predictors ranged from 4 to 16 across outer folds. Glasgow Coma Scale, lactate, and lymphocyte count were retained in all five outer folds, while C-reactive protein and albumin were retained in four folds and glucose in three folds (Additional file 7). These findings indicate that the apparent performance gain observed in the exploratory parsimonious analysis was partly attributable to feature-selection optimism, while the reduced-predictor strategy retained reasonable discrimination and favorable calibration under a more conservative validation framework.

Discussion

In this study, we investigated the ability of ML models to predict 30-day mortality in 719 patients with UGIB presenting to the ED of a university hospital. The ED predictor set consisted of 63 variables available during the initial ED assessment, including demographic characteristics ($n = 2$), clinical presentation variables ($n = 6$), comorbidities ($n = 14$), laboratory parameters ($n = 30$), medication-related variables ($n = 5$), and physiological measurements ($n = 6$). Our findings can be summarized as follows: (i) among the previously defined models, the ED XGBoost model showed the best overall performance (AUROC 0.789, AUPRC 0.294, Brier score 0.061); (ii) the performance of the ED model, consisting of variables obtained at the initial admission, was found to be almost identical to that of the Full model,

which included endoscopic findings and transfusion data (AUROC 0.789 vs. 0.785); (iii) the 15-variable Minimal ED XGBoost model demonstrated favorable performance in the exploratory feature-reduced analysis (AUROC 0.814), while fully nested validation provided a more conservative estimate (AUROC 0.765) and helped assess feature-selection optimism; (iv) SHAP analysis identified lymphocyte count, lactate, potassium, C-reactive protein, respiratory rate, albumin, platelet count, glucose, and hemoglobin among the most influential predictors; and (v) this model not only showed better discrimination than traditional scoring systems but also provided superior calibration and clinical benefit.

Early risk stratification represents a cornerstone in the clinical management of patients presenting with UGIB (23). Accurate risk identification, particularly within ED settings, directly guides critical triage decisions, including endoscopic timing, the required level of care, transfusion strategies, and discharge planning (24, 25). Especially in recent years, with the increasing burden on healthcare systems, the reliable distinction between low- and high-risk patients has become even more crucial (26). While various risk scores, such as the Rockall score, GBS, AIMS65, and ABC score, are routinely employed in clinical practice to predict outcomes in UGIB, most of these systems rely on combining a limited number of variables using predefined weights (5, 10, 27). In contrast, mortality following UGIB results from the complex interaction of numerous factors, including bleeding severity, age, comorbidities, inflammation, organ reserve, nutritional status, and treatment response (28). Therefore, ML approaches capable of modeling nonlinear relationships have attracted increasing attention in recent years for the early identification of high-risk patients with UGIB (29).

One of the most striking findings of the present study is that the ED-XGBoost model significantly outperformed conventional risk scoring systems, including the Rockall score (AUROC = 0.693), AIMS65 score (AUROC = 0.628), and GBS (AUROC = 0.603), in predicting 30-day mortality among patients with UGIB. It is noteworthy that although GBS and the Rockall score are among the most commonly used and guideline-endorsed risk stratification tools for patients with UGIB (30), they showed relatively limited performance in predicting mortality. Although the ABC score showed the best performance among the traditional scoring systems (AUROC = 0.742), the difference between the ABC score and the ED-XGBoost model was not statistically significant (Δ AUROC = 0.042, $p = 0.264$). This finding is consistent with previous studies suggesting that ML models may provide comparable or improved prognostic performance compared with conventional UGIB risk scores for mortality or adverse-outcome prediction (12–14). More broadly, recent UGIB studies have also applied ML approaches to support risk stratification and intervention prediction (15). The superiority of the ML models was also apparent in the precision–recall analysis, which is particularly informative in imbalanced datasets such as the present cohort, where 30-day mortality occurred in only 7.4% of presentations. Taken together, the consistently higher AUROC and average precision values observed for the ML models suggest that these approaches may capture clinically relevant patterns that are not fully represented by traditional risk scoring systems.

This study successfully demonstrated that the ED model, consisting only of variables obtained at the time of ED admission, performed nearly identically to the Full model, which included endoscopic findings and transfusion data. While the AUROC was 0.785 for the Full model, it was 0.789 for the ED model.

Similarly, the Brier score was 0.061 in both models. Although this result may be considered unexpected at first glance, it suggests that mortality following UGIB is not determined exclusively by bleeding source characteristics. This finding is consistent with previous studies demonstrating that mortality in UGIB is frequently influenced by advanced age, comorbidity burden, hemodynamic compromise, and physiological deterioration rather than by bleeding-related factors alone (31, 32). The limited contribution of endoscopic characteristics to our model suggests that physiological reserve, inflammatory status, and overall clinical condition at presentation largely determine the subsequent clinical course. Accordingly, much of the prognostic information associated with mortality may already be acquired during the initial ED evaluation.

Given that mortality following UGIB appears to be largely determined by factors reflecting physiological reserve, inflammatory status, and the systemic response to acute hemorrhagic stress, it is reasonable to expect that models based on a smaller number of highly informative variables could achieve comparable or even superior prognostic performance. In this context, the 15-variable Minimal ED XGBoost model demonstrated the highest discrimination performance in the exploratory feature-reduced analysis (AUROC 0.814). However, in a fully nested sensitivity analysis designed to account for feature-selection optimism, discrimination decreased to an AUROC of 0.765 while calibration performance remained unchanged. This finding suggests that part of the apparent performance gain in the exploratory analysis was attributable to optimism introduced during feature selection.

Moreover, this model demonstrated significantly better discrimination than the ABC score (Δ AUROC = 0.072; $p = 0.035$), indicating that an increased number of predictors does not necessarily improve predictive performance. Variables that are redundant or contain little prognostic information may introduce noise into the model, increase the risk of overfitting, and reduce generalizability (33). In contrast, more parsimonious models consisting of biologically meaningful and prognostically relevant variables may yield more stable results and be easier to implement in clinical practice.

The variables highlighted in the SHAP analysis provided important insights into the pathophysiological processes associated with mortality in UGIB. In this context, lactate was among the most influential mortality predictors in our ML model, with significantly higher levels observed in non-survivors than in survivors (2.50 vs. 1.70 mmol/L, $p < 0.001$). Lactate is generally known as a biochemical indicator of tissue hypoperfusion and impaired oxygen delivery (34), and elevated lactate levels have been consistently associated with adverse outcomes in critically ill patients, including those with gastrointestinal bleeding (35, 36). Consistent with these observations, Seo et al. (37) showed that ML models, particularly random forest, could predict adverse events and mortality in initially stable non-variceal UGIB, with lactate identified among the key predictive variables alongside standard laboratory markers. Similarly, Liu et al. (38) reported that ML models incorporating lactate among other clinical variables achieved high prognostic accuracy for risk stratification in patients with UGIB. Supporting these findings, both the XGBoost model and SHAP analysis in our study identified lactate as a pivotal predictor of mortality.

One important insight gained from our research is the significant impact of inflammation on mortality. Deceased patients had significantly lower lymphocyte counts (0.95 vs. $1.55 \times 10^9/L$), higher CRP levels (12.0 vs. 3.46 mg/dL) and lower albumin levels (3.30 vs. 3.59 g/dL) with all comparisons reaching statistical significance. The importance assigned to these variables by the SHAP analysis is consistent with their known biological and clinical relevance (39–40). Low albumin levels not only reflect acute systemic stress but also underlying frailty, chronic inflammation, and impaired physiological reserve, which may limit the capacity to compensate for acute blood loss and contribute to an increased risk of mortality (11, 24, 41). In addition, lymphocytopenia and elevated CRP levels reflect the systemic inflammatory response and dysregulated immune function that frequently accompany critical illness (42, 43). Together, these findings suggest that mortality following UGIB may be influenced not only by the severity of bleeding itself but also by the patient's underlying inflammatory and immune-nutritional status, which may impair recovery from acute hemorrhagic stress.

This study has several methodological strengths. Predictors were defined according to their temporal availability, enabling separate evaluation of an early ED model and an expanded Full model. To minimize information leakage, variables unavailable at the intended time of prediction were excluded from the corresponding models. Traditional risk scoring systems were evaluated separately rather than used as predictors, preprocessing procedures were fitted exclusively within training folds, and recurrent presentations from the same patient were assigned to the same cross-validation fold. These design choices are consistent with the concerns raised by Kapoor and Narayanan regarding leakage-related overestimation of ML performance and with TRIPOD + AI recommendations emphasizing transparent reporting of predictor timing, preprocessing, and validation procedures (16, 22). Predictor selection and hyperparameter optimization for the parsimonious models were evaluated within a fully nested stratified group cross-validation framework. These procedures reduced the risk of feature-selection optimism and provided a more conservative assessment of parsimonious model performance. Model evaluation was not limited to discrimination metrics alone; AUPRC, calibration plots, Brier scores, and decision curve analysis were used to assess performance in the context of class imbalance, probability calibration, and potential clinical utility. Finally, regularized predictor selection together with SHAP-based interpretation supported the development of a parsimonious and explainable early prediction model.

This study has several limitations. First, this is a single-center retrospective study, which may limit its generalizability. Second, the number of mortality events can be considered as relatively small ($n = 53$), which may affect the robustness of the comparisons and contributed to variability in the observed results. Third, although rigorous internal validation procedures were applied, external validation in independent cohorts was not performed. Finally, SHAP analysis provides insight into model behavior but does not establish causal relationships.

In conclusion, ML models based on demographic, clinical, and laboratory data available at the time of emergency department admission demonstrated significantly better calibration and clinical utility while achieving superior or comparable discrimination compared with traditional risk scores in patients with UGIB. A parsimonious ED XGBoost model based on routinely available admission variables provided an

interpretable and clinically practical alternative with favorable calibration and competitive predictive performance. Although discrimination decreased under fully nested validation compared with the initial exploratory analysis, the model retained good overall performance and identified clinically plausible predictors of mortality. Our findings suggest that mortality following UGIB is largely determined by the patient's physiological reserve, inflammatory status, and systemic response to hemorrhagic stress. Following external validation, such models may support earlier identification of high-risk patients and improve clinical decision-making in UGIB.

Abbreviations

ABC: Age, Blood tests and Comorbidities score

AIMS65: Albumin, INR, Mental status, Systolic blood pressure, Age >65 years

AUPRC: Area under the precision–recall curve

AUROC: Area under the receiver operating characteristic curve

CRP: C-reactive protein

DCA: Decision curve analysis

ED: Emergency department

GBS: Glasgow-Blatchford score

GCS: Glasgow Coma Scale

ML: Machine learning

ROC: Receiver operating characteristic

SHAP: Shapley Additive Explanations

UGIB: Upper gastrointestinal bleeding

XGBoost: Extreme Gradient Boosting

Declarations

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy and institutional data protection regulations but are available from the corresponding author on reasonable request and with permission from the relevant institutional authorities. The Python analysis

code used for model development and evaluation is likewise available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

The study was approved by the Çanakkale Onsekiz Mart University Clinical Research Ethics Committee (approval date: 17.06.2026; decision number: 2026-09-22). Due to the retrospective nature of this cohort study utilizing anonymized electronic health record data, the requirement for informed consent was waived by the Çanakkale Onsekiz Mart University Clinical Research Ethics Committee. The study was conducted in strict accordance with the ethical principles of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing financial or personal interests that could inappropriately influence or bias the conduct or reporting of this research.

3. Funding

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

4. Authors' contributions

D.S. Conceptualization, methodology, software, formal statistical analysis, data curation, visualization, and writing – original draft preparation. M.D. Data collection, conceptualization, clinical validation, project administration, and review & editing. E.U.C., Ö.K., A.U.Ç., H.K., S.G. and F.M. Data collection, clinical validation. Y.B, Data collection, clinical validation, project administration, and writing– review & editing. All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

5. Acknowledgements

Not applicable.

References

1. Long B, Gottlieb M. Emergency medicine updates: Upper gastrointestinal bleeding. Am J Emerg Med. 2024;81:116–23. 10.1016/j.ajem.2024.04.052.
2. Aljarad Z, Mobayed BB. The mortality rate among patients with acute upper GI bleeding (with/without EGD) at Aleppo University Hospital: A retrospective study. Ann Med Surg (Lond).

- 2021;71:102958. 10.1016/j.amsu.2021.102958.
3. Vara-Luiz F, Mendes I, Palma C, Mascarenhas P, Simas D, Gomes P, Ruge Gonçalves A, Simão I, Teixeira M, Ramos Lopes S, Côrte-Real F, Antónia Duarte M, Bravo C, Patita M, Nunes G, Pinto-Marques P, Fonseca J. Upper gastrointestinal bleeding differences between older and younger adults: should bleeding in non-cirrhotic patients be considered a geriatric syndrome? *Ther Adv Gastroenterol*. 2025;18:17562848251343416.
 4. Meltzer AC, Klein JC. Upper gastrointestinal bleeding: patient presentation, risk stratification, and early management. *Gastroenterol Clin North Am*. 2014;43(4):665–75.
 5. Işık Nİ, Beştemir A, Çelik GK, Demirci AE, Dönmez S, Yatmaz Y, Yavuz Y, Dogan H, Yigit Y. Comparative Performance Analysis of Aims65, Pre-Endoscopic Rockall, Glasgow-Blatchford, and Modified Age Blood Test Comorbidity Scores in Emergency Department Non-Variceal Upper Gastrointestinal Bleeding. *J Emerg Med*. 2025;76:50–9. 10.1016/j.jemermed.2025.03.016.
 6. Morarasu BC, Sorodoc V, Haisan A, Morarasu S, Bologa C, Haliga RE, Lionte C, Marciuc EA, Elsiddig M, Cimpoesu D, Dimofte GM, Sorodoc L. Age, blood tests and comorbidities and AIMS65 risk scores outperform Glasgow-Blatchford and pre-endoscopic Rockall score in patients with upper gastrointestinal bleeding. *World J Clin Cases*. 2023;11(19):4513–30.
 7. Rout G, Sharma S, Gunjan D, Kedia S, Nayak B, Shalimar. Comparison of various prognostic scores in variceal and non-variceal upper gastrointestinal bleeding: A prospective cohort study. *Indian J Gastroenterol*. 2019;38(2):158–66.
 8. Di Gioia G, Sangineto M, Paglia A, Cornacchia MG, Parente F, Serviddio G, Romano AD, Villani R. Limits of pre-endoscopic scoring systems in geriatric patients with upper gastrointestinal bleeding. *Sci Rep*. 2024;14(1):20225.
 9. Alali AA, Boustany A, Martel M, Barkun AN. Strengths and limitations of risk stratification tools for patients with upper gastrointestinal bleeding: a narrative review. *Expert Rev Gastroenterol Hepatol*. 2023 Jul-Dec;17(8):795–803.
 10. Bardakçı O, Sıddıkoğlu D, Akdur G, Şimşek G, Atalay Ü, Das M, Akdur O, Beyazit Y. Prediction of adverse outcomes using non-endoscopic scoring systems in patients over 80 years of age who present with the upper gastrointestinal bleeding in the emergency department. *Ulus Travma Acil Cerrahi Derg*. 2022;28(1):39–47.
 11. Cumaoglu MO, Vural A, Alpaslan M, Dolanbay T, Kulu AR, Sezgin E, Yanmaz SR. Inflammatory indices as independent predictors of in-hospital mortality in acute upper gastrointestinal bleeding. *BMC Gastroenterol*. 2026;26(1):127. 10.1186/s12876-026-04625-8.
 12. Shung DL, Simonov M, Gentry M, Au B, Laine L. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: a systematic review. *Dig Dis Sci*. 2019;64:2078–87. 10.1007/s10620-019-05645-z.
 13. Deshmukh F, Merchant SS. Explainable machine learning model for predicting GI bleed mortality in the intensive care unit. *Am J Gastroenterol*. 2020;115:1657–68. 10.14309/ajg.0000000000000632.

14. Shung DL, Au B, Taylor RA, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology*. 2020;158:160–7. 10.1053/j.gastro.2019.09.009.
15. Raghareutai K, Tanchotsrinon W, Sattayalertyanyong O, Kaosombatwattana U. Development and validation of a machine learning model to predict hemostatic intervention in patients with acute upper gastrointestinal bleeding. *BMC Med Inf Decis Mak*. 2025;25:145. 10.1186/s12911-025-02969-x.
16. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4:100804. 10.1016/j.patter.2023.100804.
17. Davis SE, Matheny ME, Balu S, Sendak MP. A framework for understanding label leakage in machine learning for health care. *J Am Med Inf Assoc*. 2024;31:274–80. 10.1093/jamia/ocad178.
18. Shetty AP, Shetty S, Hegde P, et al. A leakage-controlled and SHAP-driven machine learning framework for paediatric respiratory disease classification using Indian hospital EHR data. *BMC Med Inf Decis Mak*. 2026;26:198. 10.1186/s12911-026-03493-2.
19. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. 10.1186/s12916-019-1466-7.
20. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26:565–74. 10.1177/0272989X06295361.
21. Emakhu J, Brooks RC, Adrianto I, et al. Early emergency department decision support for heart failure hospitalization using triage-level unstructured and structured data: a retrospective cohort study. *BMC Med Inf Decis Mak*. 2026;26:224. 10.1186/s12911-026-03481-6.
22. 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. <https://doi.org/10.1136/bmj-2023-078378>.
23. Srygley FD, Gerardo CJ, Tran T, Fisher DA. Does this patient have a severe upper gastrointestinal bleed? *JAMA*. 2012;307(10):1072–9. 10.1001/jama.2012.253.
24. Das M, Kurtkulagi O, Unal Cetin E, Kayan H, Guler S, Demirer Aydemir F, Mutlay F, Çetin AU, Beyazit Y. 30-day mortality prediction in acute upper gastrointestinal bleeding: Incremental value of the prognostic nutritional index with ABC and Rockall scores. *Biomol Biomed*. 2026;26(9):1614–23.
25. Mohammadyari F, Aflak A, ZareDini M, Irandoost R, Ghaffari S, Gharib A, Roshanaei M, Tahavori M, Olangian-Tehrani S. Risk stratification and scoring systems in upper and lower gastrointestinal bleeding: review of performance and limitations in the emergency department. *Front Med (Lausanne)*. 2025;12:1564015.
26. Jin X, Wang X, Mao P. Early Identification of High-Risk Factors for Upper Gastrointestinal Bleeding. *Evid Based Complement Alternat Med*. 2022; 2022:5641394. 10.1155/2022/5641394
27. Kim S, Lee S, Kim I, Hou JU, Kim BW. Assessment of the Prognostic Value of the ABC Score in Non-Variceal UGIB Showing Superiority to Traditional Scores, With Room for Improvement: A Retrospective Observational Study. *Korean J Helicobacter Up Gastrointest Res*. 2025;25(4):355–62.

28. Liu C, Li J. Risk factors and prognosis analysis of gastrointestinal bleeding combined with hemorrhagic shock in patients in intensive care unit: A retrospective study. *PLoS ONE*. 2026;21(5):e0348276. 10.1371/journal.pone.0348276.
29. Ustaolioğlu İ, Ak R. Predicting Mortality in Non-Variceal Upper Gastrointestinal Bleeding: Machine Learning Models Versus Conventional Clinical Risk Scores. *J Clin Med*. 2025;14(20):7425.
30. Gralnek IM, Stanley AJ, Morris AJ, Camus M, Lau J, Lanas A, Laursen SB, Radaelli F, Papanikolaou IS, Cúrdia Gonçalves T, Dinis-Ribeiro M, Awadie H, Braun G, de Groot N, Udd M, Sanchez-Yague A, Neeman Z, van Hooft JE. Endoscopic diagnosis and management of nonvariceal upper gastrointestinal hemorrhage (NVUGIH): European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2021. *Endoscopy*. 2021;53(3):300–32. 10.1055/a-1369-5274.
31. Cengiz A, Köksal Y, Aydin SB. Emergency Department Predictors of Mortality and Adverse Outcomes in Upper Gastrointestinal Bleeding: A 5-Year Retrospective Cohort Study. *J Emerg Med*. 2026;84:17–27. 10.1016/j.jemermed.2026.01.023.
32. Renukaprasad AK, Narayanaswamy S. A Comparative Analysis of Risk Scoring Systems in Predicting Clinical Outcomes in Upper Gastrointestinal Bleed. *Cureus*. 2022;14(7):e26669.
33. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health*. 2020;8(1):e000262.
34. Zeng F, Du L, Ling L. Lactate level as a predictor of outcomes in patients with acute upper gastrointestinal bleeding: A systematic review and meta-analysis. *Exp Ther Med*. 2024;27(3):113.
35. El-Kersh K, Chaddha U, Sinha RS, Saad M, Guardiola J, Cavallazzi R. Predictive Role of Admission Lactate Level in Critically Ill Patients with Acute Upper Gastrointestinal Bleeding. *J Emerg Med*. 2015;49(3):318–25.
36. Shah A, Chisolm-Straker M, Alexander A, Rattu M, Dikdan S, Manini AF. Prognostic use of lactate to predict inpatient mortality in acute gastrointestinal hemorrhage. *Am J Emerg Med*. 2014;32(7):752–5.
37. Seo DW, Yi H, Park B, Kim YJ, Jung DH, Woo I, Sohn CH, Ko BS, Kim N, Kim WY. Prediction of Adverse Events in Stable Non-Variceal Gastrointestinal Bleeding Using Machine Learning. *J Clin Med*. 2020;9(8):2603.
38. Liu Z, Jiang G, Zhang L, Shrestha P, Hu Y, Zhu Y, Li G, Xiong Y, Zhan L. The future of critical care: AI-powered mortality prediction for acute variceal gastrointestinal bleeding and acute non-variceal gastrointestinal bleeding patients. *Front Med (Lausanne)*. 2025;12:1580094.
39. Keskin G, Çakmak A, Çalışkan MU. Machine Learning-Based Prediction of Long-Term Mortality in STEMI Patients Using Clinical, Laboratory, and Inflammatory-Metabolic Indices. *J Clin Med*. 2026;15(5):1800.
40. Li J, Yang J. Development and validation of a novel nutrition-inflammation prognostic score for predicting 30-day mortality in critically ill stroke patients. *Front Nutr*. 2025;12:1658896.
41. Capurso C, Lo Buglio A, Bellanti F, Serviddio G. C-Reactive Protein/Albumin Ratio vs. Prognostic Nutritional Index as the Best Predictor of Early Mortality in Hospitalized Older Patients, Regardless

of Admitting Diagnosis. *Nutrients*. 2025;17(17):2907. 10.3390/nu17172907.

42. Jiang J, Du H, Su Y, Li X, Zhang J, Chen M, Ren G, He F, Niu B. Nonviral infection-related lymphocytopenia for the prediction of adult sepsis and its persistence indicates a higher mortality. *Med (Baltim)*. 2019;98:e16535.
43. Sproston NR, Ashworth JJ. Role of C-Reactive Protein at Sites of Inflammation and Infection. *Front Immunol*. 2018;9:754.

Figures

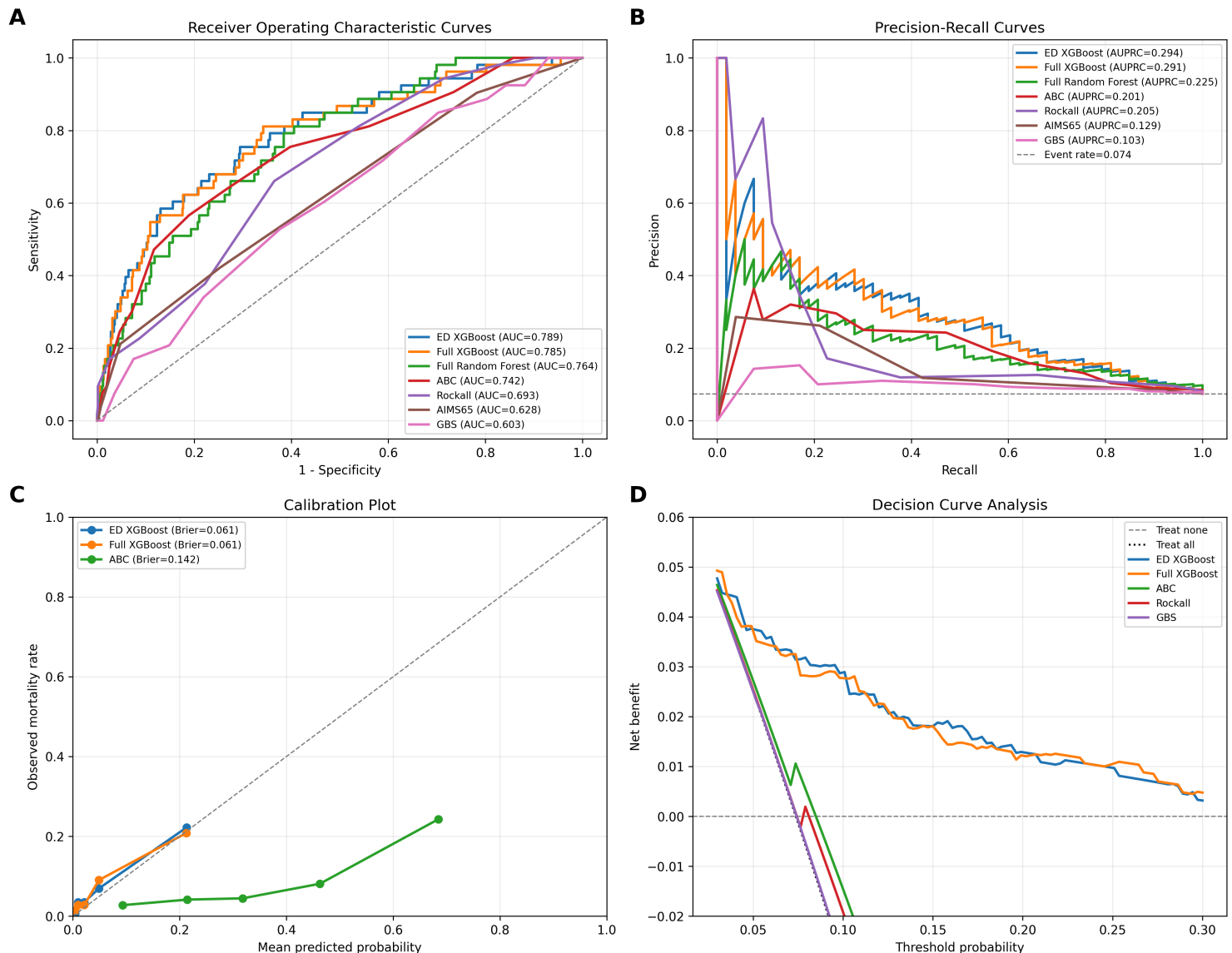


Figure 1

Performance comparison of machine-learning models and traditional risk scores for predicting 30-day mortality in patients with upper gastrointestinal bleeding.

(A) Receiver operating characteristic curves.

(B) Precision–recall curves.

(C) Calibration plots.

(D) Decision curve analysis.

Abbreviations: AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve.

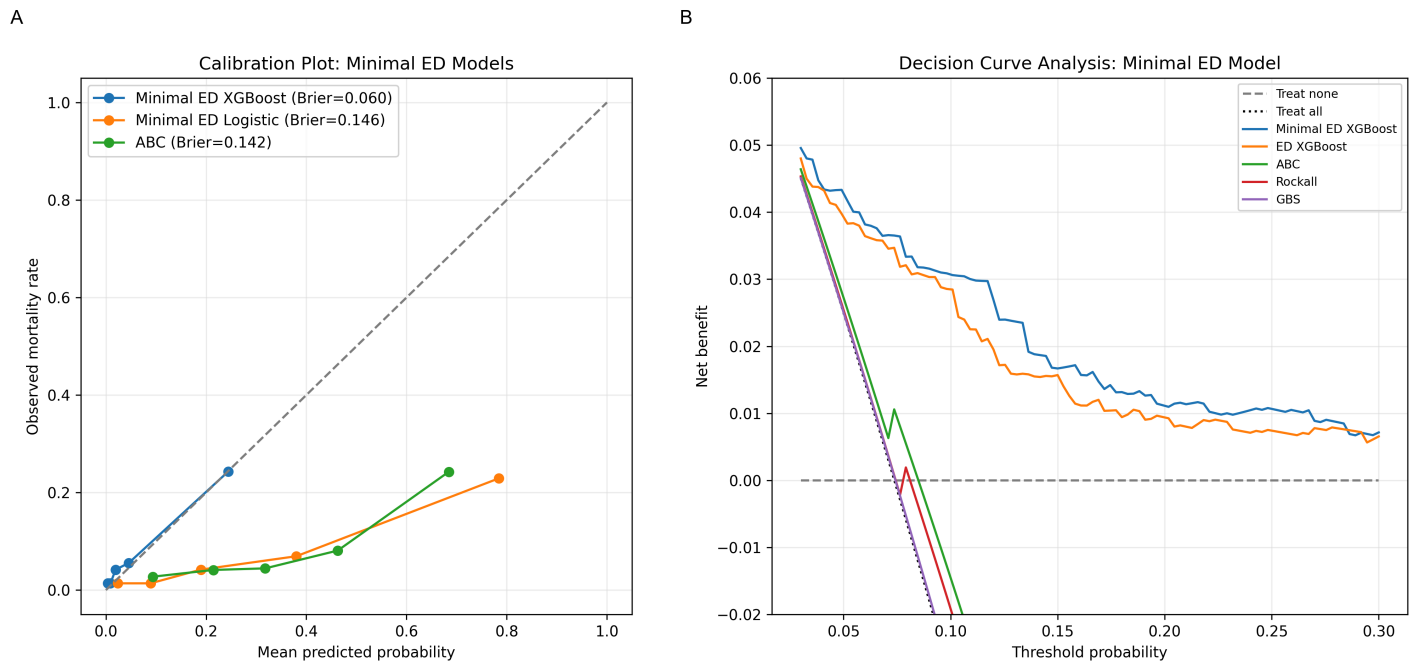


Figure 2

Calibration and clinical utility of the parsimonious models for predicting 30-day mortality in patients with upper gastrointestinal bleeding.

(A) Calibration plots comparing the agreement between predicted and observed mortality probabilities for the Minimal ED XGBoost model, Minimal ED Logistic Regression model, and ABC score.

(B) Decision curve analysis showing the net clinical benefit of the Minimal ED XGBoost model, Minimal ED Logistic Regression model, and traditional risk scores across a range of threshold probabilities.

Abbreviations: ED, emergency department; ABC, Age, Blood tests, and Comorbidities score.

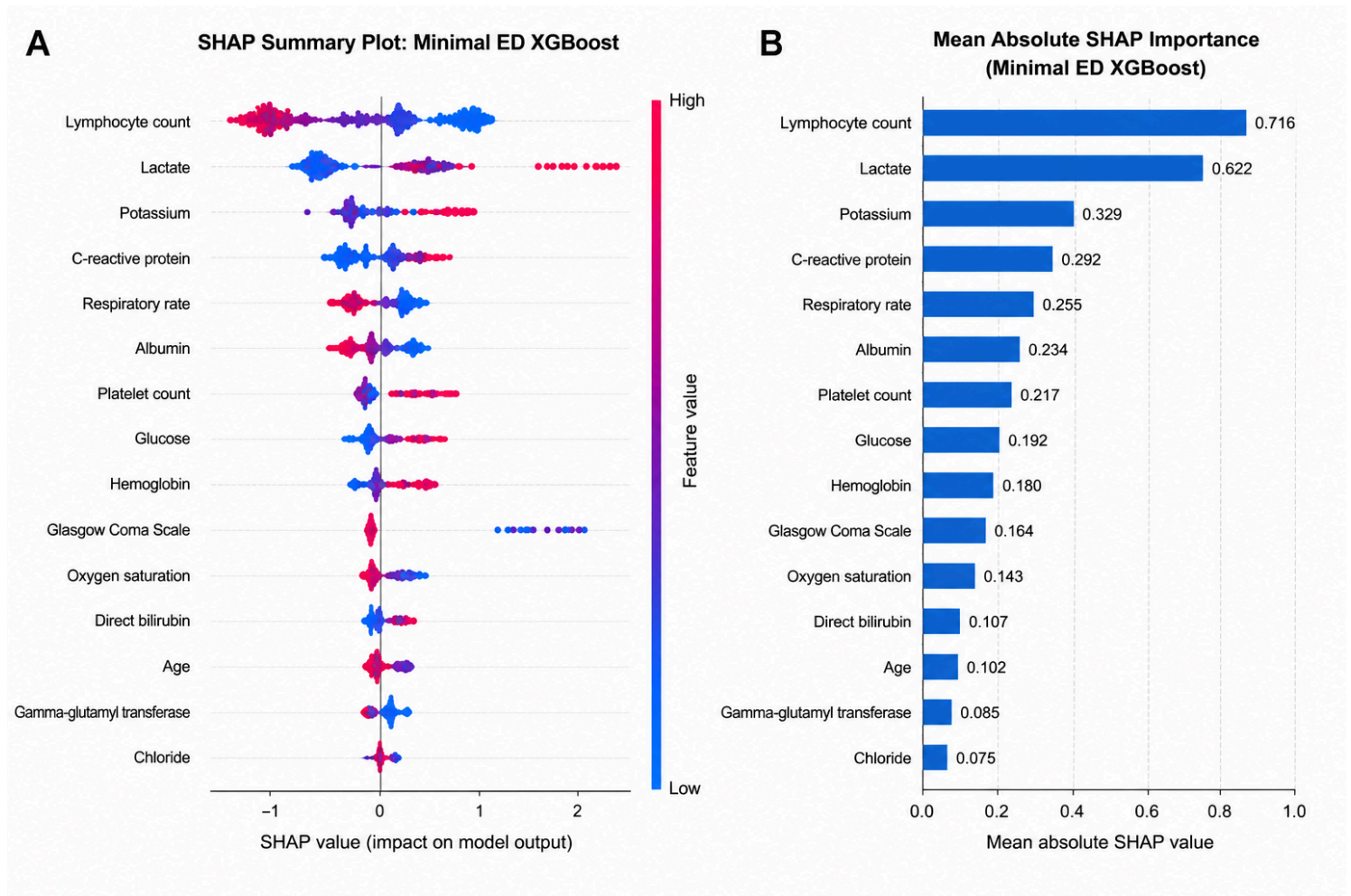


Figure 3

SHAP analysis of the Minimal ED XGBoost model.

(A) SHAP summary plot demonstrating the distribution of SHAP values for the 15 selected predictors across all patient presentations. Predictor importance is ranked according to mean absolute SHAP values. Higher SHAP values indicate a greater contribution to increased predicted mortality risk.

(B) Mean absolute SHAP values representing the global importance of predictors in the Minimal ED XGBoost model.

Abbreviations: SHAP, Shapley Additive Explanations; ED, emergency department; GCS, Glasgow Coma Scale; CRP, C-reactive protein.

Supplementary Files

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

- [Additionalfile1.DataDictionary.xlsx](#)
- [Additionalfile2TRIPODAIChecklist.docx](#)

- [Additionalfile3.BinLevelCalibrationEstimatesforEDXGBoostandABC.xlsx](#)
- [Additionalfile4.Hyperparametertuningresultsforregularizedlogisticregressionmodels.xlsx](#)
- [Additionalfile5.BinLevelCalibrationEstimatesforMinimalmodelsandABC.xlsx](#)
- [Additionalfile6Calibrationstatistics.xlsx](#)
- [Additionalfile7FullyNestedSensitivityAnalysis.xlsx](#)