Risk Management and Healthcare Policy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development and Validation of an Explainable Machine Learning Model for Prediction of Massive Transfusion in Upper Gastrointestinal Bleeding

Zixi Lin[1,*], Hailiang Zhao[2,*], Yilong Hu[3]

[1]Department of Blood Transfusion, Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu Province, People's Republic of China; [2]Department of Gastroenterology, Affiliated Hospital of Youjiang Medical University for Nationalities; Guangxi Medical and Health Key Cultivation Discipline Construction Project; Guangxi Clinical Medical Research Center for Hepatobiliary Disease, Baise, Guangxi Zhuang Autonomous Region, People's Republic of China; [3]Department of General Surgery of the International Medical Center; The Fourth Affiliated Hospital of Soochow University; Suzhou Dushu Lake Hospital, Suzhou, Jiangsu Province, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yilong Hu, Department of General Surgery of the International Medical Center; The Fourth Affiliated Hospital of Soochow University; Suzhou Dushu Lake Hospital, No. 9 Chongwen Road, Suzhou Industrial District, Suzhou, Jiangsu Province, People's Republic of China, Email yiminhyl@163.com

Background: Upper gastrointestinal bleeding (UGIB) is a medical emergency with high mortality, especially when massive transfusion (MT) is required. Traditional scores like Glasgow-Blatchford provide moderate accuracy but overlook complex variable interactions. We developed and validated an explainable machine learning (ML) model for MT prediction in UGIB, improving precision and interpretability.

Methods: In this retrospective study, 700 UGIB patients from The Affiliated Hospital of Xuzhou Medical University (2021–2025) were divided into training (n=490) and testing (n=210) cohorts. An external validation cohort (n=300) was sourced from The Fourth Affiliated Hospital of Soochow University. From 18 clinical variables, 8 key features were selected using Boruta and LASSO regression. Seven ML algorithms were compared to identify the optimal model, which was then evaluated for discrimination, calibration, and clinical utility. SHapley Additive exPlanations (SHAP) provided interpretability.

Results: The Random Forest (RF) model achieved superior performance with an AUC of 0.862 (95% CI 0.785–0.939) in training, 0.823 (95% CI 0.768–0.879) in testing, and 0.807 (95% CI 0.748–0.866) in external validation. Calibration plots showed strong agreement between predicted and observed probabilities. Decision curve analysis indicated higher net benefit than "treat all" or "treat none" strategies. SHAP analysis ranked impaired mental status, liver cirrhosis, and international normalized ratio (INR) as top predictors, aligning with clinical intuition.

Conclusion: The developed machine learning model demonstrated promising performance in identifying UGIB patients at high risk of massive transfusion. While the model shows potential to assist clinicians in optimizing blood management strategies, further prospective validation is required to confirm its clinical utility in diverse settings.

Keywords: upper gastrointestinal bleeding, massive transfusion, machine learning, SHAP, random forest

Introduction

Upper gastrointestinal bleeding (UGIB) is a major medical emergency, with a notable incidence worldwide, contributing to high hospitalization rates and mortality, particularly in elderly or comorbid patients.[1] Common presentations include hematemesis, melena, and hemodynamic instability, often caused by peptic ulcers, variceal rupture, or erosive gastritis.[2] A severe complication is the need for massive transfusion (MT), defined as ≥10 units of packed red blood cells within 24 hours, which is associated with longer hospital stays, higher costs, and increased mortality.[3,4] Early identification of MT risk is crucial for resource optimization and better outcomes.[5]

https://doi.org/10.2147/RMHP.S573423
Received: 10 October 2025
Accepted: 26 January 2026
Published: 2 February 2026

Predicting MT in UGIB remains difficult due to variable patient factors and presentations. Conventional tools like the Glasgow-Blatchford Score (GBS) and Rockall Score use clinical and lab data to assess bleeding severity, demonstrating moderate predictive accuracy.[6] However, traditional scoring systems were primarily designed to predict mortality rather than massive transfusion and often rely on linear assumptions that oversimplify complex physiological interactions. In contrast, machine learning algorithms excel at capturing non-linear relationships and high-dimensional data patterns, offering a distinct advantage in identifying high-risk patients with greater precision.[7] This highlights the need for advanced, data-driven models with improved precision and interpretability.[8]

Machine learning (ML) algorithms excel in capturing nonlinear relationships and have shown promise in predicting transfusion needs in trauma and surgery, achieving high accuracy via methods like random forest (RF) and gradient boosting.[5] Although ML has shown promise in UGIB, existing studies have notable limitations. For instance, Shung et al developed a robust model that outperformed traditional scores in predicting the need for hospital-based intervention; however, their primary outcome was a composite endpoint (including endoscopy and any transfusion) rather than massive transfusion (MT) specifically, which requires distinct rapid response protocols.[9] Additionally, the opaque "black box" nature of ML models limits clinical trust, as providers need clear insights into predictive factors.[10]

Explainable artificial intelligence techniques, such as SHapley Additive exPlanations (SHAP), address this by quantifying feature contributions globally and locally, linking predictions to key variables like international normalized ratio (INR), hemoglobin, and liver cirrhosis.[11] This enhances transparency and aligns ML with clinical decision-making.

Collectively, prior evidence indicates that ensemble-based machine learning algorithms consistently yield the superior predictive performance for UGIB outcomes compared to linear models. However, the application of these high-performing architectures specifically for massive transfusion remains underexplored. To bridge this gap, in this retrospective study, we developed and validated an explainable ML model to predict MT risk in UGIB patients. Using dual feature selection on 18 clinical variables, we compared seven ML algorithms, selected the optimal one, and interpreted it with SHAP. Internal and external validation were conducted to assess generalizability, calibration, and utility, aiming to offer a practical tool for early risk stratification and transfusion guidance.

Materials and Methods

Study Design and Participants

In this retrospective study, we initially reviewed the medical records of patients with upper gastrointestinal bleeding (UGIB) at The Affiliated Hospital of Xuzhou Medical University between January 1, 2021, and August 31, 2025. After applying exclusion criteria, we identified 700 patients eligible for primary analysis. These patients were randomly allocated in a 7:3 ratio to the Training cohort (n=490) and Testing cohort (n=210) using stratified random sampling based on outcome status to ensure proportional representation of massive transfusion cases in both groups. For external validation of our machine learning model, we included an additional cohort of 300 patients meeting similar selection criteria from The Fourth Affiliated Hospital of Soochow University, spanning January 1, 2021, to August 31, 2025.

This study received approval from the Ethics Committees of The Affiliated Hospital of Xuzhou Medical University and The Fourth Affiliated Hospital of Soochow University, adhering to the Declaration of Helsinki. Due to the retrospective nature of the study and the use of de-identified data, the requirement for informed consent was waived by the institutional review board.

Identification of Research Variables and Participants

We identified 18 clinical factors that may influence the need for massive transfusion in patients with upper gastrointestinal bleeding (UGIB). These factors encompass essential patient characteristics including age (stratified as ≥70 years), sex, body mass index (BMI), and American Society of Anesthesiologists (ASA) classification. Clinical presentation variables evaluated comprise tachycardia (heart rate >100 bpm), impaired mental status, and hypotension (systolic blood pressure <90 mmHg). Comorbidities examined include diabetes mellitus, smoking history, liver cirrhosis, and alcohol consumption. Laboratory parameters assessed consist of white blood cell count (WBC), C-reactive protein (CRP), D-dimer, fibrinogen, hemoglobin (stratified for anemia as <8.0 g/dL), international normalized ratio (INR,

stratified as >1.5), and albumin (stratified for hypoalbuminemia as <2.5 g/dL). All continuous laboratory values were measured using standardized clinical protocols.

Inclusion criteria were as follows: (1) patients diagnosed with UGIB or acute exacerbation of chronic gastrointestinal conditions at our institutions and who received comprehensive clinical management within the same facilities; (2) patients who underwent appropriate interventions such as endoscopy, transfusion, or supportive care; and (3) individuals with complete and accessible clinical data, including age, medical records, and length of hospital stay.

Exclusion criteria included: (1) patients with a prior diagnosis of chronic gastrointestinal bleeding presenting for elective intervention; (2) individuals previously diagnosed with UGIB who had undergone prior procedures and were now admitted for elective management; (3) patients with concomitant acute disorders, such as severe coagulopathy unrelated to UGIB, acute pancreatitis, or multi-organ failure; (4) those undergoing additional unrelated procedures or with complicating factors like trauma or malignancy requiring separate interventions; and (5) patients with incomplete or missing clinical data.

The diagnosis of massive transfusion was established based on clinical records and defined as the transfusion of ≥10 units of packed red blood cells within the first 24 hours of presentation, characterized by significant hemodynamic instability, persistent bleeding, and the need for aggressive resuscitation.[12] This definition was confirmed by reviewing transfusion logs and patient outcomes by experienced clinicians.

Feature Selection

Variables with a missing rate exceeding 20% were excluded from the analysis. For the remaining variables, missing values were imputed to maximize data utilization. Continuous variables were imputed using the median, and categorical variables were imputed using the mode. To address potential multicollinearity, we conducted a pairwise Pearson correlation analysis among all candidate predictors. Additionally, the LASSO regression model was employed to further mitigate any residual redundancy by shrinking coefficients of correlated features. To ensure fair comparison among algorithms, we applied Z-score standardization to all continuous variables. Crucially, the scaling parameters (mean and standard deviation) were derived solely from the training set and then applied to the validation set to prevent data leakage. To address multicollinearity and identify robust predictors from the initial 18 variables, we implemented a dual-method feature selection approach. First, we employed the Boruta algorithm, a Random Forest-based wrapper method that iteratively compares feature importance against synthetic "shadow" attributes to statistically confirm relevance. This method helps in discerning truly informative features by assessing their significance over multiple iterations. Second, least absolute shrinkage and selection operator (LASSO) regression was applied with 10-fold cross-validation to optimize the regularization parameter (λ), minimizing binomial deviance. This L1-penalized method shrinks non-predictive coefficients to zero. Final variable inclusion required consensus between methods: Only features selected by both Boruta and LASSO were retained. This stringent intersection strategy prioritized generalizability while mitigating overfitting, yielding a refined subset for subsequent modeling.

Model Construction and Comparison

After feature selection, the common variables identified by both the Boruta algorithm and LASSO regression were utilized as input predictors for model development. Seven distinct machine learning models—including Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Extreme Gradient Boosting (XGB), Decision Tree (DT), K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (LGBM)—were constructed to predict the risk of massive transfusion (MT). In this study, we employed a 10-fold cross-validation methodology for model selection. Model performance in the training cohort was comprehensively evaluated using metrics such as the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Comparative analysis revealed that the RF model outperformed the other algorithms across all main metrics, with the highest AUC value (0.862), as well as superior specificity and sensitivity. A pairwise comparison of AUC values across models confirmed the statistical superiority of the RF model. The optimal cut-off point for the RF model was determined through Youden's index. The confusion matrix of the RF model illustrated robust discrimination in classifying patients

within the training cohort. Subsequently, the trained models were evaluated independently on the testing and external validation cohorts. To assess calibration fidelity and clinical utility, we performed calibration curves and decision curve analysis (DCA).

The SHAP to Model Interpretation

Interpreting machine learning models presents significant challenges due to their inherent complexity. The SHAP (Shapley Additive Explanations) method, grounded in game theory, addresses the "black box" nature of these models by quantitatively assessing and ranking the importance of input features in relation to model predictions. By calculating the contribution of each feature to individual and overall prediction outcomes, SHAP offers both local and global interpretability, thereby enhancing the transparency and comprehensibility of machine learning models, such as the Random Forest (RF) model used in this study for predicting massive transfusion risk.

Statistical Analyses

All statistical analyses and data visualizations were performed using R software (version 4.4.2) and JD_DCPM (V6.11, Jingding Medical Technology Co., Ltd). Continuous variables were tested for normality using the Shapiro–Wilk test. Normally distributed data are reported as mean ± standard deviation and compared via Student's *t*-test. Non-normally distributed data are presented as median (interquartile range) and analyzed with the Mann–Whitney *U*-test. Categorical variables are expressed as frequencies (percentages) and compared using Chi-square or Fisher's exact tests (for expected counts <5). For machine learning modeling, the caret and randomForest packages were utilized for hyperparameter tuning and model training. The pROC package was used to compute AUC and confidence intervals. Significance was set at two-sided $p < 0.05$.

Result

Patient Characteristics for Training, Testing, and External Validation Cohorts

A total of 490 patients were included in the training cohort, 210 in the testing cohort, and 300 in the external validation cohort. Baseline characteristics were comparable across cohorts, with no significant differences (all $P > 0.05$; Table 1). Sex distribution was balanced (male: 51.6%, 51.4%, 49.3%), as was the proportion aged ≥70 years (21.4%, 19.0%, 20.7%). Mean BMI ranged from 24.1 ± 3.4 kg/m^2 (testing) to 24.7 ± 3.0 kg/m^2 (validation).

Comorbidities and clinical features showed consistency: smoking history (20.0%, 22.9%, 19.0%), diabetes (14.3%, 17.1%, 14.0%), tachycardia (>100 bpm; 9.8%, 9.5%, 11.0%), and median alcohol consumption (152 [IQR 125–180], 149 [IQR 122–175], 153 [IQR 126–182] units). Laboratory parameters were similar: median WBC (7.1 [IQR 5.2–9.1], 6.9 [IQR 4.8–8.9], 7.3 [IQR 5.4–9.3] × 10^9/L); CRP (5.2 [IQR 2.3–10.0], 4.8 [IQR 2.0–9.4], 5.4 [IQR 2.6–10.9] mg/L); D-dimer (1.04 [IQR 0.84–1.33], 1.00 [IQR 0.79–1.30], 1.09 [IQR 0.87–1.41] μg/mL); fibrinogen (3.04 [IQR 2.51–3.52], 2.96 [IQR 2.46–3.46], 3.11 [IQR 2.57–3.59] μg/mL).

Hematological and hemodynamic indicators were comparable: anemia (hemoglobin <8.0 g/dL; 40.8%, 41.9%, 40.0%); impaired mental status (9.8%, 10.5%, 10.0%); liver cirrhosis (5.3%, 4.8%, 5.0%); INR >1.5 (31.0%, 32.4%, 30.0%); hypotension (systolic BP <90 mmHg; 24.9%, 26.7%, 26.0%); hypoalbuminemia (albumin <2.5 g/dL; 2.2%, 1.9%, 1.7%); ASA classification >2 (25.5%, 24.8%, 26.0%).

The primary outcome of massive transfusion occurred in 19.8% (97/490) of the training cohort, 17.6% (37/210) of the testing cohort, and 20.3% (61/300) of the external validation cohort.

Feature Selection

Feature selection was conducted using least absolute shrinkage and selection operator (LASSO) regression and the Boruta algorithm to identify the most predictive variables for massive transfusion from an initial set of 18 candidate features. The Boruta algorithm, based on random forest importance scoring, confirmed 10 attributes as important through iterative comparisons (Figure 1A).LASSO regression minimized binomial deviance across lambda values, selecting 9 key attributes as coefficients converged (Figure 1B and C). Overlap analysis identified 8 shared features between the two

Table 1 Baseline Characteristics of the Training, Testing, and External Validation Sets

Variables	Training Cohort N=490	Testing Cohort N=210	Validation Cohort N=300	P value
Male, n (%)	253 (51.6%)	108 (51.4%)	148 (49.3%)	0.841
Age≥70 years, n (%)	105 (21.4%)	40 (19.0%)	62 (20.7%)	0.782
BMI, mean (SD), kg/m^2	24.5 ± 3.2	24.1 ± 3.4	24.7 ± 3.0	0.118
Smoking, n (%)	98 (20.0%)	48 (22.9%)	57 (19.0%)	0.462
Diabetes, n (%)	70 (14.3%)	36 (17.1%)	42 (14.0%)	0.552
Heart rate>100 bpm,n (%)	48 (9.8%)	20 (9.5%)	33 (11.0%)	0.808
Alcohol consumption, n (%)	152 (125–180)	149 (122–175)	153 (126–182)	0.217
WBC,median(IQR),$10^9/L$	7.1 (5.2–9.1)	6.9 (4.8–8.9)	7.3 (5.4–9.3)	0.083
CRP,median(IQR),mg/L	5.2 (2.3–10.0)	4.8 (2.0–9.4)	5.4 (2.6–10.9)	0.101
D-dimer,median(IQR),ug/mL	1.04 (0.84–1.33)	1.00 (0.79–1.30)	1.09 (0.87–1.41)	0.072
Fibrinogen,median(IQR),ug/mL	3.04 (2.51–3.52)	2.96 (2.46–3.46)	3.11 (2.57–3.59)	0.192
Hemoglobin <8.0 g/dL, n (%)	200 (40.8%)	88 (41.9%)	120 (40.0%)	0.902
Impaired mental status, n (%)	48 (9.8%)	22 (10.5%)	30 (10.0%)	0.967
Liver cirrhosis, n (%)	26 (5.3%)	10 (4.8%)	15 (5.0%)	0.976
INR >1.5, n (%)	152 (31.0%)	68 (32.4%)	90 (30.0%)	0.844
SBP<90 mmHg, n (%)	122 (24.9%)	56 (26.7%)	78 (26.0%)	0.254
Albumin<2.5 g/dL, n (%)	11 (2.2%)	4 (1.9%)	5 (1.7%)	0.654
ASA classification>2, n (%)	125 (25.5%)	52 (24.8%)	78 (26.0%)	0.943
MT,n (%)	97 (19.8)	37 (17.6)	61 (20.3)	0.758

Abbreviations: BMI, Body Mass Index; WBC, White Blood Count; CRP, C-reactive protein; INR, International Normalized Ratio; SBP, Systolic Blood Pressure; ASA, American Society of Anesthesiologists; MT, massive transfusion.

methods (Figure 1D), which were retained for downstream modeling to ensure predictive power and parsimony. These features included liver cirrhosis, international normalized ratio (INR), hemoglobin, impaired mental status, systolic blood pressure, heart rate, albumin, and ASA score. As illustrated in Figure S1, the inter-variable correlations were generally weak, with the highest absolute correlation coefficient being 0.35, indicating no severe multicollinearity.

Model Development and Performance

Seven machine learning models were developed using the selected features to predict massive transfusion in patients with upper gastrointestinal bleeding: decision tree (DT), k-nearest neighbors (KNN), light gradient boosting machine (LGBM), naive Bayes (NB), random forest (RF), support vector machine (SVM), and XGBoost. Model hyperparameters were optimized via grid search with 5-fold cross-validation on the training cohort to maximize area under the receiver operating characteristic curve (AUC).

Performance was evaluated on the training cohort, with RF demonstrating the highest discriminative ability (AUC 0.862, 95% CI 0.785–0.939), followed by KNN (AUC 0.807, 95% CI 0.752–0.862) and XGBoost (AUC 0.743, 95% CI 0.678–0.809). Other models showed moderate performance: DT (AUC 0.708, 95% CI 0.641–0.776), SVM (AUC 0.657, 95% CI 0.586–0.728), NB (AUC 0.643, 95% CI 0.571–0.716), and LGBM (AUC 0.602, 95% CI 0.530–0.673) (Figure 2A). Additional metrics for the testing cohort included F1-score (RF: 0.61; KNN: 0.67; DT: 0.55), recall (RF: 0.75; KNN: 0.73; DT: 0.63), precision (RF: 0.51; KNN: 0.63; DT: 0.49), sensitivity (RF: 0.75; KNN: 0.73; DT: 0.63), specificity (RF: 0.85; KNN: 0.73; DT: 0.72), positive predictive value (RF: 0.51; KNN: 0.63; DT: 0.49), and negative predictive value (RF: 0.94; KNN: 0.81; DT: 0.82) (Figure 2B). Figure 2C illustrates the comparison of the area under the ROC curve among the seven machine learning models. The Random Forest (RF) model achieved the highest AUC, indicating the strongest overall discriminative performance. Pairwise comparisons via DeLong's test revealed significant differences, with RF outperforming all others (all $P < 0.05$ except vs KNN, $P = 0.259$), while LGBM and NB exhibited the lowest AUCs ($P < 0.05$ vs RF and KNN) (Figure 2D).Based on superior overall performance, the RF model was selected for testing,external validation and interpretability analysis.

Feature reduction analysis revealed that the top-performing models—k-nearest neighbors (KNN), random forest (RF), XGBoost, and decision tree (DT)—achieved relatively stable discriminatory performance when utilizing eight selected

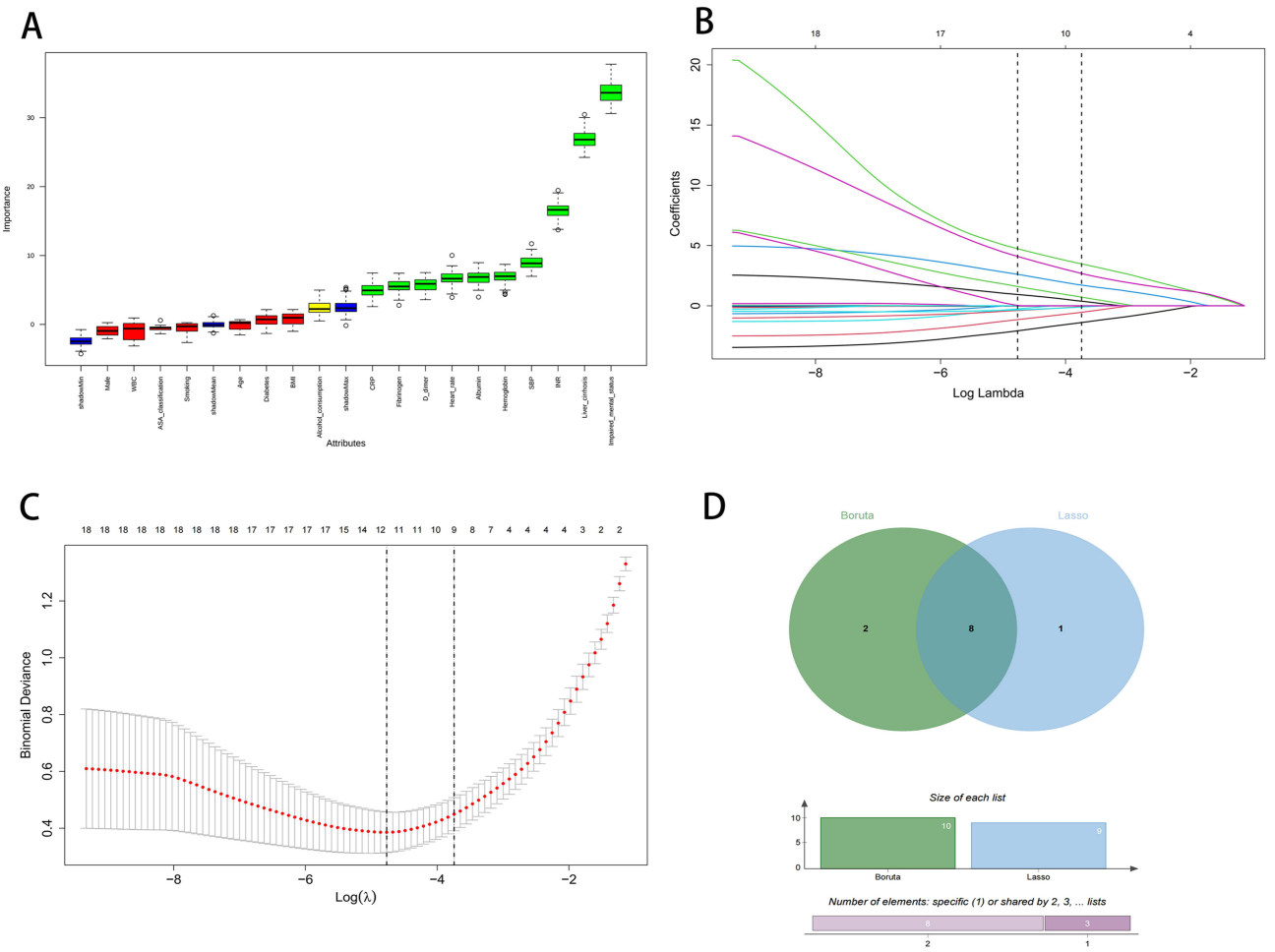

Figure 1 Feature Selection Process Using Boruta and LASSO Regression. (**A**) Boruta Variable Importance Plot. (**B**) LASSO Coefficient Paths. (**C**) LASSO Cross-Validation Curve. (**D**) Venn diagram showing the intersection of 8 consensus features selected by both the Boruta and LASSO algorithms.The overlapping region represents the 8 key features retained for the final model. Labels indicate set sizes: green (Boruta-selected: 10 features), blue (LASSO-selected: 9 features), overlap (shared: 8 features).

features, as evidenced by area under the receiver operating characteristic curve (AUC) values plateauing thereafter (Figure 3A). Specifically, for the RF model, key performance metrics including AUC, sensitivity, specificity, and F1-score demonstrated consistent stability starting from eight features, with minimal fluctuations observed across increasing feature counts up to 18 (Figure 3B). This suggests that an eight-feature subset provides an optimal balance of predictive efficacy and model parsimony, minimizing overfitting while maintaining robust generalizability.

Model Performance on Both the Testing and External Validation Sets

The random forest (RF) model demonstrated robust performance on the testing cohort. The area under the receiver operating characteristic curve (AUC) was 0.823 (95% CI 0.768–0.879), indicating strong discriminative ability (Figure 4A). Calibration was excellent, with a calibration intercept of −0.00 (95% CI −0.31 to 0.31) and a C-statistic of 0.82 (95% CI 0.76–0.87), showing close alignment between predicted probabilities and observed outcomes (Figure 4B).Decision curve analysis (DCA) confirmed the clinical utility of the RF model, revealing superior net benefit across a broad range of threshold probabilities compared to “treat all” and “treat none” strategies (Figure 4C). The model yielded positive net benefit at cost-benefit ratios from 1:100 to 10:1, highlighting its practical value in guiding decisions for massive transfusion prediction.The confusion matrix for the RF model on the testing cohort illustrated its classification accuracy, with 25 true positives, 149 true negatives, 24 false positives, and 12 false negatives (Figure 4D). This vividly underscores the model’s remarkable effectiveness in accurately identifying patients who are at risk, while

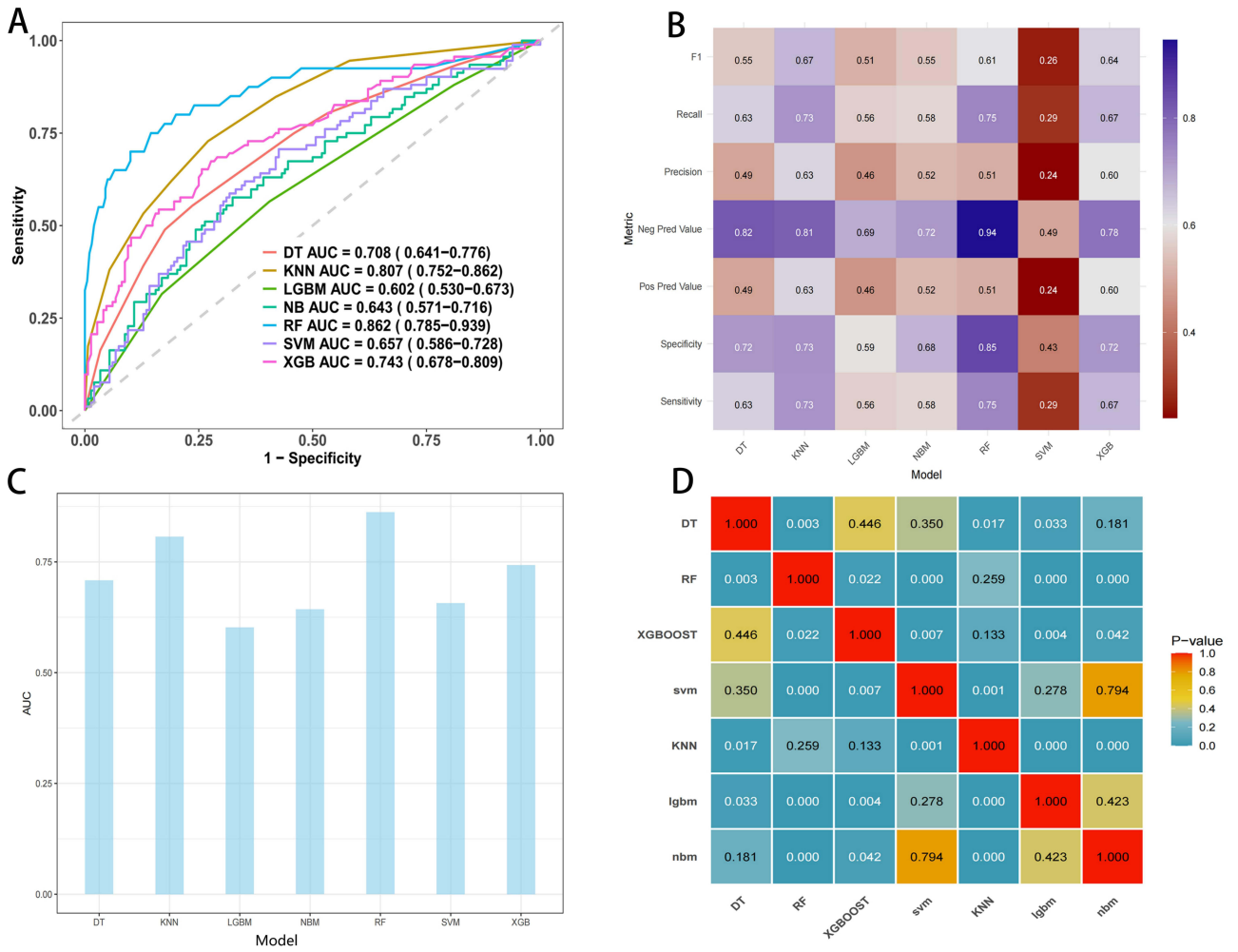

Figure 2 Performance Metrics of Machine Learning Models in the Training Cohort. (**A**) ROC curves and AUC values for the seven model. (**B**) ROC curves and AUC values for the five model. (**C**) AUC Display for Seven Machine Learning Models. (**D**) Pairwise Comparisons of AUCs Among Seven Machine Learning Models.

simultaneously minimizing unnecessary interventions, thereby optimizing treatment strategies and reducing healthcare costs.

On external validation, the AUC was 0.807 (95% CI 0.748–0.866) with C-statistic 0.81 (95% CI 0.74–0.86), calibration intercept ≈ −0.00 (95% CI −0.31 to 0.31) and slope ≈ 1.00 (95% CI 0.72 to 1.28), and higher DCA net benefit (thresholds 0.1–0.8, ratios 1:100 to 100:1; Figure 5A–C). The confusion matrix indicated 47 TP, 208 TN, 31 FP, and 14 FN (Figure 5D). Overall, the model maintained strong discriminative ability, calibration, and clinical utility, excelling in ruling out low-risk cases while identifying high-risk patients for massive transfusion.

Model Interpretability

To elucidate the decision-making process of the random forest model, SHapley Additive exPlanations analysis was employed, providing both global and local interpretability. The mean absolute SHAP values highlighted the relative importance of the eight selected features, with impaired mental status emerging as the most influential, followed by liver cirrhosis and international normalized ratio. Other contributors, in descending order of impact, included systolic blood pressure, hemoglobin, albumin, heart rate, and American Society of Anesthesiologists score (Figure 6A). The SHAP summary plot further illustrated the directional impact of each feature on model predictions (Figure 6B). Specifically, higher values of impaired mental status, liver cirrhosis, and international normalized ratio were associated with positive SHAP contributions, thereby increasing the predicted risk of massive transfusion. In contrast, higher values of systolic

https://doi.org/10.2147/RMHP.S573423

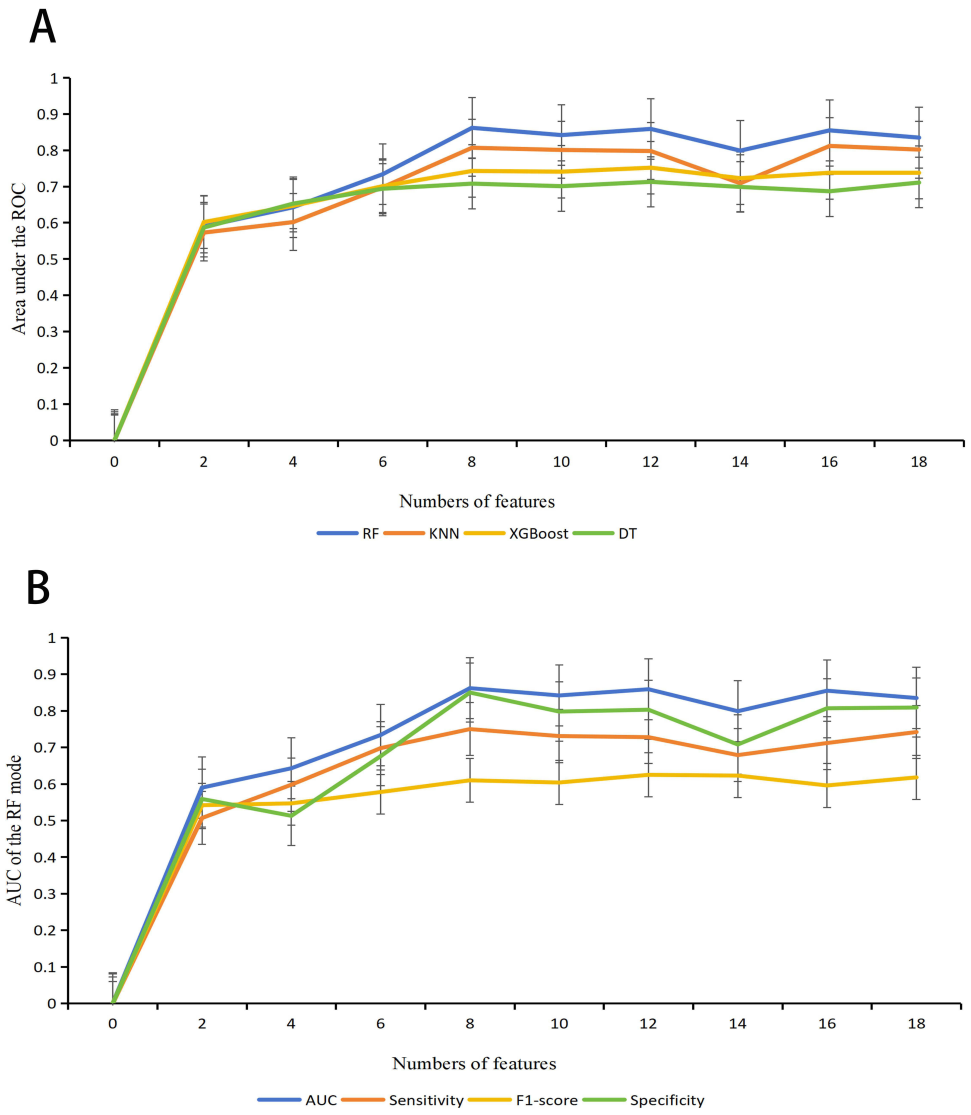

Figure 3 Impact of Feature Reduction on Model Performance. (**A**) AUCs of the Top 4 Best-Performing ML Models with Varied Numbers of Features. (**B**) AUC, Sensitivity, Specificity, and F1-Score for the RF Model with Varying Numbers of Features. Across the top-performing models, AUC values stabilize at 8 features, with the RF model achieving optimal and consistent performance in AUC, sensitivity, specificity, and F1-score at this point.

blood pressure, hemoglobin and albumin corresponded to negative SHAP contributions, decreasing the predicted risk, which implies that lower levels of these features elevate risk. Features such as heart rate, and American Society of Anesthesiologists score also exhibited positive SHAP contributions, with higher values increasing risk, albeit with relatively smaller overall influence compared to the top features. This interpretable framework underscores the model's transparency, facilitating clinical adoption by linking predictions to physiologically relevant risk factors.

Discussion

In this multicenter retrospective study, we developed and validated an explainable machine learning (ML) model for predicting massive transfusion (MT) risk in patients with upper gastrointestinal bleeding (UGIB). Utilizing a dual feature selection strategy combining the Boruta algorithm and least absolute shrinkage and selection operator (LASSO) regression, we identified 8 consensus variables from an initial set of 18 clinical features. Seven ML algorithms were compared, with the random forest (RF) model demonstrating superior performance, achieving an area under the receiver operating characteristic curve (AUC) of 0.823 (95% CI 0.768–0.879) in the internal testing cohort (n=210) and 0.807 (95% CI 0.748–0.866) in the external validation cohort (n=300). The model showed strong performance on the training

 https://doi.org/10.2147/RMHP.S573423

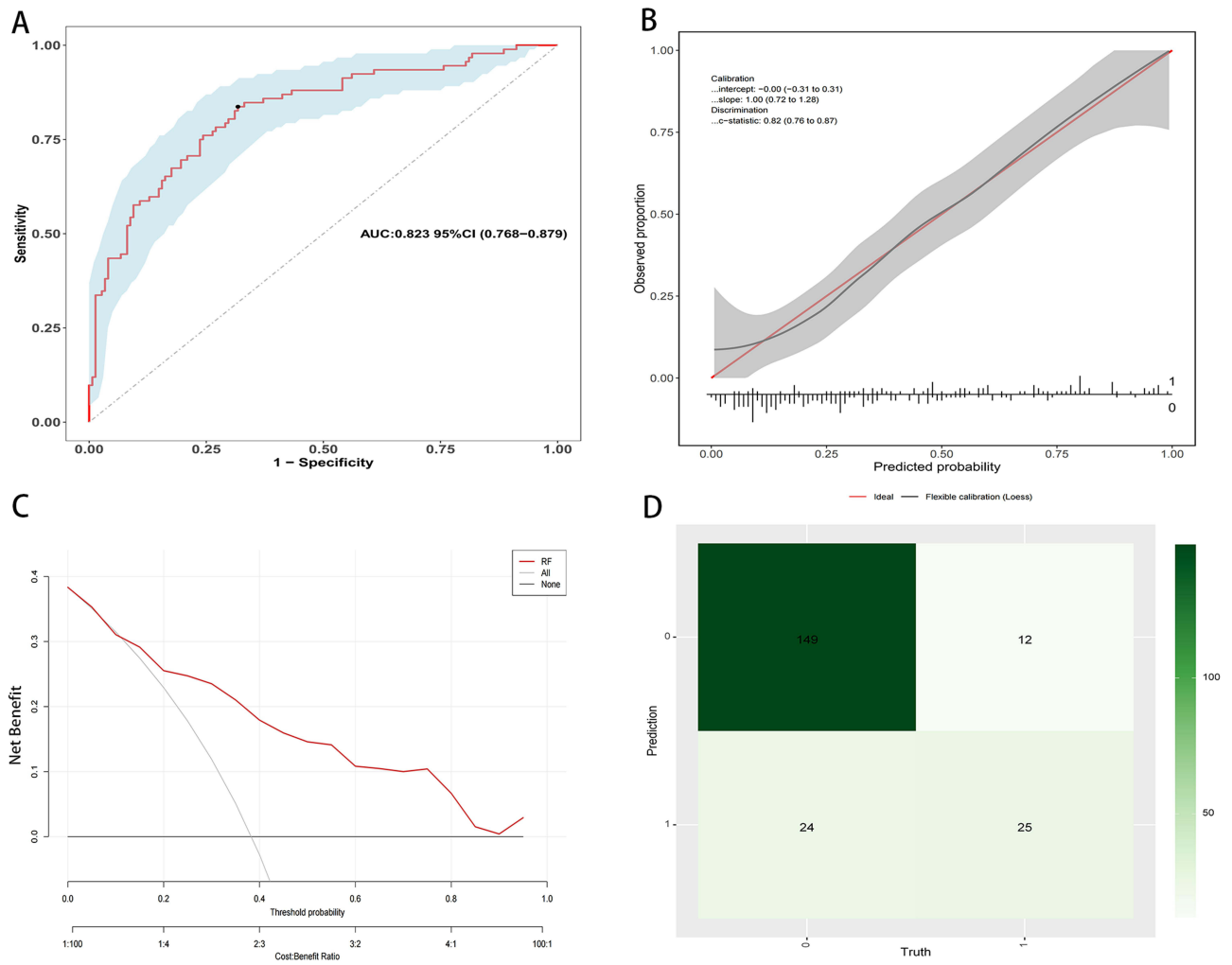

Figure 4 RF mode evaluation in testing cohort. (**A**) Testing set ROC curves and AUC values. (**B**) Calibration curve of the testing set. (**C**) Analysis of decision curves for the test cohort. (**D**) Confusion matrix of the RF model in testing set.

cohort (AUC 0.862), with minimal overfitting evident in validation. Calibration was excellent, with intercepts near 0 and slopes near 1 in both validation cohorts, while decision curve analysis (DCA) indicated substantial net clinical benefit across threshold probabilities of 0.1–0.8. SHapley Additive exPlanations (SHAP) analysis enhanced interpretability, identifying impaired mental status, liver cirrhosis, and international normalized ratio (INR) as the most influential predictors. These results highlight the efficacy of explainable ML in providing robust, generalizable risk stratification for MT in UGIB, potentially surpassing traditional methods and supporting timely clinical decisions.

Our findings build on prior applications of ML in hemorrhage prediction, where models like RF have shown advantages in handling complex, nonlinear data interactions compared to logistic regression. For example, studies in trauma and surgical cohorts have reported AUC values of 0.80–0.90 for transfusion risk models using RF or gradient boosting, aligning with our RF model's validation performance.[13] However, UGIB-specific ML models remain underrepresented; a recent single-center study using RF for rebleeding prediction achieved an AUC of approximately 0.82 but was limited by lack of external validation and interpretability.[14] In contrast, our multicenter design, encompassing 1000 patients across two institutions, along with rigorous external validation, addresses these shortcomings. The RF model's superiority over alternatives in our comparative analysis underscores its robustness for binary classification tasks in heterogeneous UGIB populations. Furthermore, conventional tools such as the Glasgow-Blatchford Score (GBS) and Rockall Score typically yield AUCs of 0.70–0.80 for bleeding severity but lack specificity for MT and integration of

https://doi.org/10.2147/RMHP.S573423

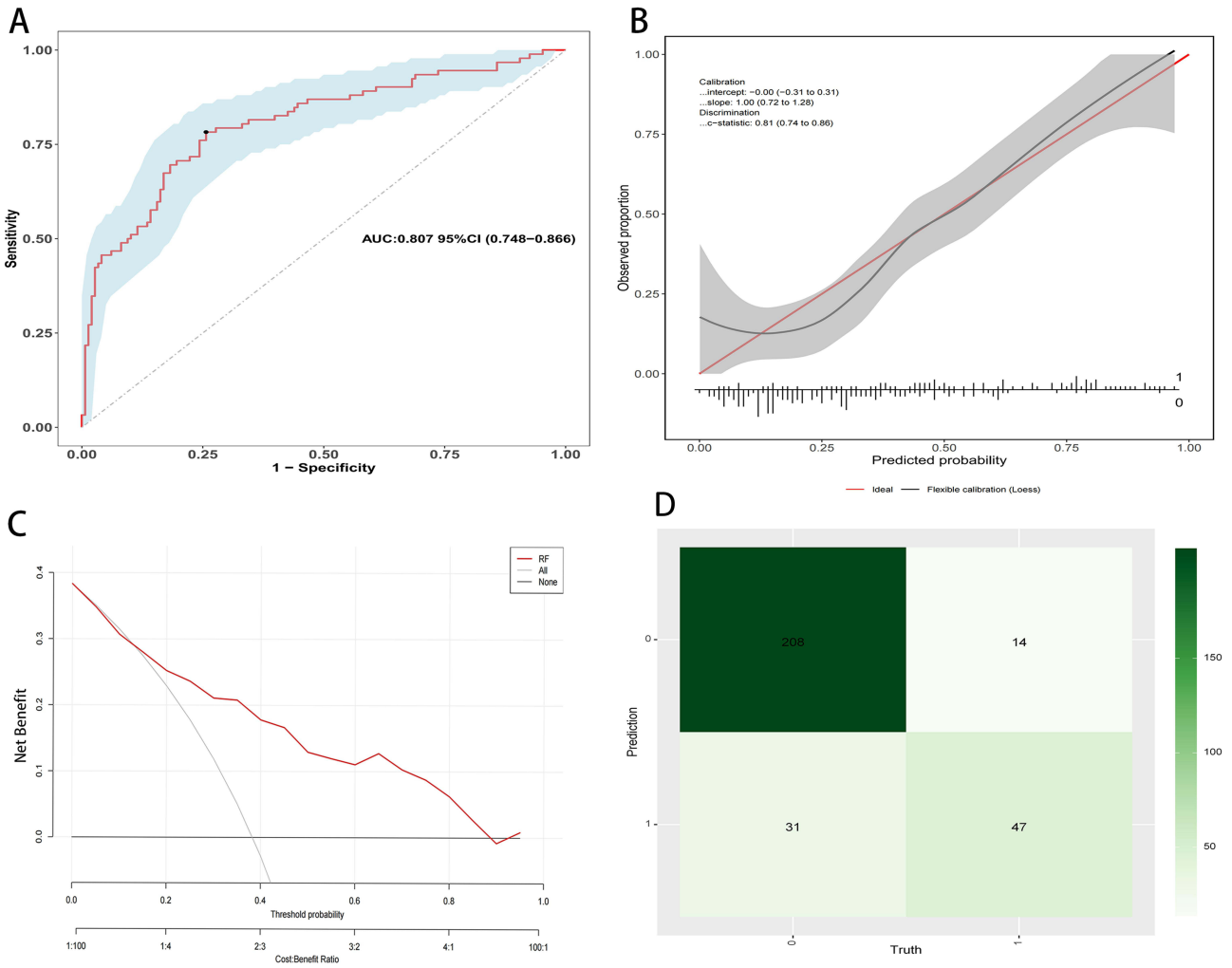

Figure 5 RF mode evaluation in external validation cohort. (**A**) External validation set ROC curves and AUC values. (**B**) Calibration curve of the external validation set. (**C**) Decision curve analysis of the external validation set (**D**) Confusion matrix of the RF model in the external validation set.

dynamic factors like mental status or INR.[9,15] The enhanced discriminative power of our model suggests that ML can better capture multifaceted UGIB pathophysiology, including hemodynamic instability and coagulopathy.

SHAP interpretability revealed clinically intuitive feature contributions, with impaired mental status as the top predictor, reflecting its association with severe hypoperfusion and shock in UGIB. Liver cirrhosis and elevated INR ranked next, consistent with evidence linking portal hypertension, coagulopathy, and anticoagulation to exacerbated bleeding and transfusion needs.[16,17] Other key features, such as low systolic blood pressure, hemoglobin, and albumin, along with high heart rate and American Society of Anesthesiologists (ASA) score, align with markers of acute blood loss, tissue hypoperfusion, and overall frailty.[18,19] These findings help overcome the "black box" problem in machine learning, where models can seem mysterious and hard to explain.[20] This allows doctors to connect predictions to factors they can change, such as quickly fixing blood clotting issues or stabilizing a patient's blood flow, and could reduce complications from massive transfusions like overload in the circulatory system.[21] The model's high negative predictive value and DCA net benefits support its utility for ruling out low-risk patients, optimizing resource allocation in emergency settings, and minimizing overtreatment.[22,23] In resource-limited environments, this could translate to improved survival by facilitating proactive blood product preparation without unnecessary interventions.[24–26] From a practical clinical workflow perspective, this explainable ML model holds significant potential to serve as an automated triage tool within the Emergency Department. Upon a patient's arrival, the eight accessible variables could be auto-populated from the Electronic Health Record or quickly input by triage nurses. For patients flagged as "high-risk" for

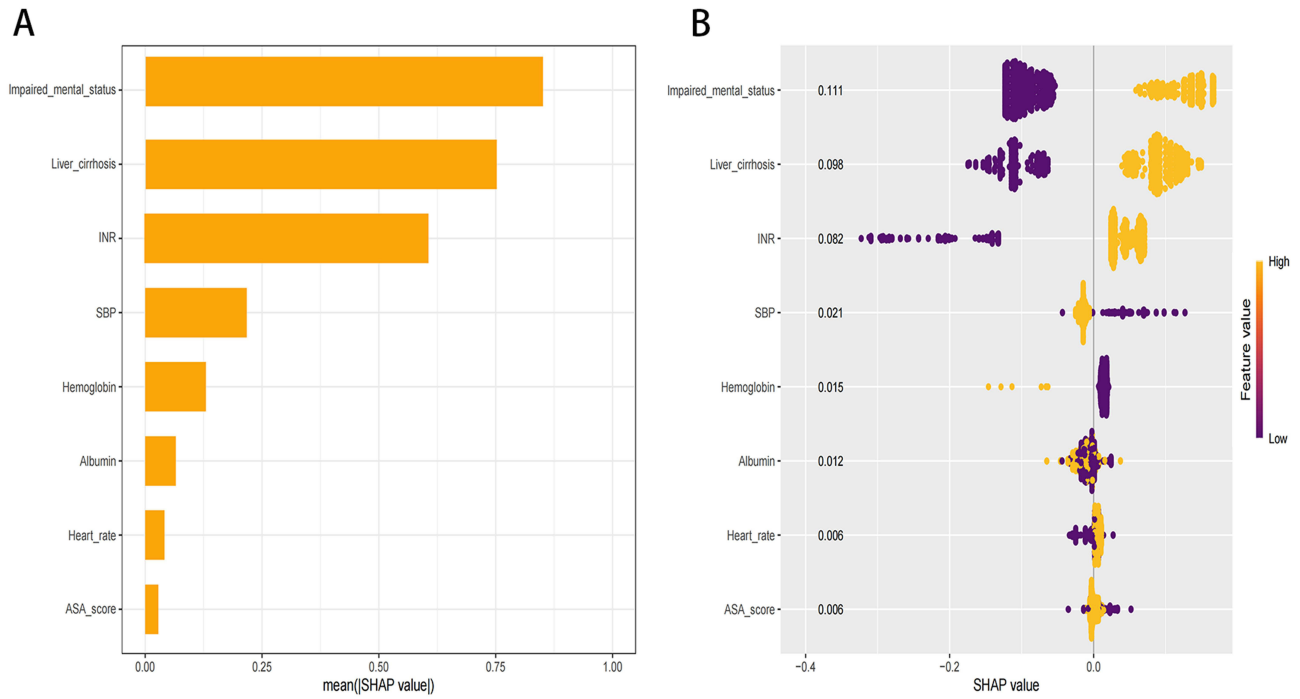

Figure 6 SHAP analysis of the RF model. (**A**) Mean absolute SHAP values corresponding to each clinical attribute. (**B**) SHAP values depicting the impact of various clinical characteristics on the model's output.

massive transfusion, the model could trigger an automated alert to the blood bank to prepare cross-matched blood products immediately, bypassing the typical delay associated with manual ordering. Furthermore, the interpretability provided by SHAP values aids in tailoring early intervention strategies. For instance, if the model identifies elevated INR and liver cirrhosis as dominant risk drivers for a specific patient, clinicians can prioritize the early administration of plasma or prothrombin complex concentrates alongside red blood cells, rather than solely focusing on volume resuscitation. This shift from reactive to proactive management is critical, as reducing the time to the first transfusion unit is strongly correlated with improved survival in hemorrhagic shock.

A key strength of this study is the external validation performed in an independent hospital; however, we acknowledge potential differences in demographics and clinical practices that could influence model generalizability.Firstly, referral biases may lead to variations in patient case-mix. For instance, tertiary referral centers often manage patients with higher acuity, such as those with advanced liver cirrhosis or complex malignancies, compared to community hospitals. Secondly, and perhaps more importantly, clinical variations in transfusion practices exist between institutions. Although guidelines recommend restrictive transfusion strategies for UGIB, the specific threshold for triggering a Massive Transfusion Protocol may depend on local resources, blood bank availability, and physician preferences. Notably, despite these potential disparities, our model demonstrated robust performance in the external cohort (AUC=0.807). This suggests that the selected predictors capture the core physiological derangements associated with severe hemorrhage, rendering the model resilient to institution-specific variations and supporting its potential for broader clinical application.

Despite these advancements, limitations must be acknowledged. The retrospective design, reliant on electronic medical records, may introduce selection bias or missing data, though mitigated by strict inclusion criteria and data preprocessing. The analysis may be influenced by potential unmeasured confounding factors, including detailed pre-hospital medication history or variations in endoscopic timing. Although we employed multivariate feature selection methods (LASSO and Boruta) to mitigate these effects, the possibility of residual confounding remains. Our cohorts, totaling 1,000 patients from Chinese hospitals, may not fully represent global populations with varying ethnicities, comorbidities, or healthcare practices; international validation is warranted. The moderate sample size restricted subgroup analyses, and while SHAP provides associational insights, it does not infer causality. Finally, real-time implementation requires prospective testing to evaluate

https://doi.org/10.2147/RMHP.S573423

workflow integration and long-term outcomes. While retrospective analysis confirms theoretical accuracy, prospective validation is indispensable to test the model's robustness under the dynamic constraints of real-time emergency care where data availability may be fragmented. Such studies are also crucial for evaluating human factors, particularly whether clinicians trust and adhere to algorithmic alerts or if practical barriers like alert fatigue diminish the tool's utility. Ultimately, only prospective assessment can definitively prove whether this AI integration translates into tangible clinical benefits, including reduced time-to-transfusion and improved survival rates compared to standard practice. We acknowledge the potential concern regarding the class imbalance in our dataset, where massive transfusion events occurred in approximately 20% of cases. While severe imbalance can lead to models that favor the majority class, our analysis suggests the model remained robust.First, by rigorously reducing the feature space to 8 consensus predictors, we maintained a favorable Events Per Variable (EPV) ratio of >12 in the training set (97 events/8 features), surpassing the widely accepted threshold of 10 required to minimize overfitting and estimation bias.Second, performance metrics in the external validation cohort confirmed that the model did not ignore the minority class; it achieved a sensitivity of 77% (47/61) and a high negative predictive value, ensuring that high-risk patients were effectively identified.Finally, the 20% prevalence in our cohort aligns with real-world clinical epidemiology for UGIB. Preserving this natural distribution allows for better calibration and more realistic probability estimates in clinical practice compared to models trained on artificially balanced datasets.

In summary, this study developed and externally validated a machine learning-based nomogram for predicting massive transfusion in patients with UGIB. By integrating eight routine clinical and laboratory variables, the model demonstrated robust discrimination and calibration in both the training and external validation cohorts. Rather than replacing clinical judgment, this tool is intended to serve as an adjunctive decision aid, potentially helping physicians to identify high-risk patients earlier and allocate blood resources more rationally. Although the results are encouraging, large-scale, multicenter prospective studies are warranted to further verify the model's generalizability and to assess its actual impact on patient outcomes before widespread clinical adoption.

Abbreviations

AUC, Area under receiver operating characteristic curve; UGIB, Upper gastrointestinal bleeding; MT, massive transfusion; ML, Machine learning; LASSO, Least absolute shrinkage and selection operator; SHAP, SHapley Additive exPlanations; ASA, American Society of Anesthesiologists; DCA, Decision curve analysis; INR, International normalized ratio; RF, Random Forest; SVM, Support Vector Machine; NB, Naive Bayes; XGB, Extreme Gradient Boosting; DT, Decision Tree; KNN, K-Nearest Neighbors; LGBM, Light Gradient Boosting Machine; BMI, Body mass index; WBC, White blood cell count; CRP, C-reactive protein; SBP, Systolic blood pressure.

Data Sharing Statement

All original data are available from the corresponding author (Yilong Hu) upon reasonable request.

Ethical Approval and Consent to Participate

This study adhered to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board of the Ethics Committee of The Affiliated Hospital of Xuzhou Medical University and The Fourth Affiliated Hospital of Soochow University. The informed consent is waived by the ethics committee because this is a retrospective design study. Patient confidentiality and data privacy were strictly safeguarded throughout the study.

Acknowledgments

This study was generously supported by Jingding Medical Tech, to whom we extend our sincere gratitude. We especially thank them for providing authorization and technical support for the JD_DCPM software. The team at Jingding Medical Tech offered invaluable assistance in data processing.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically

reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Sarajlic P, Simonsson M, Jernberg T, Bäck M, Hofmann R. Incidence, associated outcomes, and predictors of upper gastrointestinal bleeding following acute myocardial infarction: a SWEDEHEART-based nationwide cohort study. *Eur Heart J Cardiovasc Pharmacother*. 2022;8(5):483–491. doi:10.1093/ehjcvp/pvab059
2. Kanjee Z, Asombang AW, Berzin TM, Burns RB. How would you manage this patient with nonvariceal upper gastrointestinal bleeding? Grand rounds discussion from Beth Israel Deaconess Medical Center. *Ann Intern Med*. 2021;174(6):836–843. doi:10.7326/M21-1206
3. Lee SM, Lee G, Kim TK, et al. Development and validation of a prediction model for need for massive transfusion during surgery using intraoperative hemodynamic monitoring data. *JAMA Netw Open*. 2022;5(12):e2246637. doi:10.1001/jamanetworkopen.2022.46637
4. Li Q, Chen G, Li Q. Multi-task machine learning for transfusion decision support in acute upper gastrointestinal bleeding: a novel ensemble approach with clinical validation. *J Transl Med*. 2025;23(1):979. doi:10.1186/s12967-025-06995-1
5. El-Menyar A, Naduvilekandy M, Asim M, Rizoli S, Al-Thani H. Machine learning models predict triage levels, massive transfusion protocol activation, and mortality in trauma utilizing patients hemodynamics on admission. *Comput Biol Med*. 2024;179:108880. doi:10.1016/j.compbiomed.2024.108880
6. Shung DL, Lin JK, Laine L. Achieving value by risk stratification with machine learning model or clinical risk score in acute upper gastrointestinal bleeding: a cost minimization analysis. *Am J Gastroenterol*. 2024;119(2):371–373. doi:10.14309/ajg.0000000000002520
7. Lou SS, Liu H, Lu C, Wildes TS, Hall BL, Kannampallil T. Personalized surgical transfusion risk prediction using machine learning to guide preoperative type and screen orders. *Anesthesiology*. 2022;137(1):55–66. doi:10.1097/ALN.0000000000004139
8. Herrin J, Abraham NS, Yao X, et al. Comparative effectiveness of machine learning approaches for predicting gastrointestinal bleeds in patients receiving antithrombotic treatment. *JAMA Netw Open*. 2021;4(5):e2110703. doi:10.1001/jamanetworkopen.2021.10703
9. Shung DL, Chan CE, You K, et al. Validation of an electronic health record–based machine learning model compared with clinical risk scores for gastrointestinal bleeding. *Gastroenterology*. 2024;167(6):1198–1212. doi:10.1053/j.gastro.2024.06.030
10. Zheng NS, Keloth VK, You K, et al. Detection of gastrointestinal bleeding with large language models to aid quality improvement and appropriate reimbursement. *Gastroenterology*. 2025;168(1):111–120.e4. doi:10.1053/j.gastro.2024.09.014
11. Bai Z, Lin S, Sun M, et al. Machine learning based CAGIB score predicts in-hospital mortality of cirrhotic patients with acute gastrointestinal bleeding. *Npj Digit Med*. 2025;8(1):489. doi:10.1038/s41746-025-01883-w
12. Gauss T, Richards JE, Tortù C, et al. Association of early norepinephrine administration with 24-hour mortality among patients with blunt trauma and hemorrhagic shock. *JAMA Netw Open*. 2022;5(10):e2234258. doi:10.1001/jamanetworkopen.2022.34258
13. Yu YD, Lee KS, Man Kim J, et al. Artificial intelligence for predicting survival following deceased donor liver transplantation: retrospective multi-center study. *International Journal of Surgery*. 2022;105:106838. doi:10.1016/j.ijsu.2022.106838
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(1):160–167. doi:10.1053/j.gastro.2019.09.009
15. Grdinic AG, Radovanovic S, Gleditsch J, et al. Developing a machine learning model for bleeding prediction in patients with cancer-associated thrombosis receiving anticoagulation therapy. *Journal of Thrombosis and Haemostasis*. 2024;22(4):1094–1104. doi:10.1016/j.jtha.2023.12.034
16. Ma Y, Luo M, Guan G, Liu X, Cui X, Luo F. An explainable predictive machine learning model of gangrenous cholecystitis based on clinical data: a retrospective single center study. *World J Emerg Surg*. 2025;20(1):1. doi:10.1186/s13017-024-00571-6
17. Ning C, Ouyang H, Xiao J, et al. Development and validation of an explainable machine learning model for mortality prediction among patients with infected pancreatic necrosis. *eClinicalMedicine*. 2025;80:103074. doi:10.1016/j.eclinm.2025.103074
18. He J, Wang X, Zhu P, et al. Identification and validation of an explainable early-stage chronic kidney disease prediction model: a multicenter retrospective study. *eClinicalMedicine*. 2025;84:103286. doi:10.1016/j.eclinm.2025.103286
19. Hong Y, Chen X, Wang L, Zhang F, Zeng Z, Xie W. Machine learning prediction of metabolic dysfunction-associated fatty liver disease risk in American adults using body composition: explainable analysis based on SHapley additive exPlanations. *Front Nutr*. 2025;12:1616229. doi:10.3389/fnut.2025.1616229
20. Mueller SC, Patil P, Levy JI, et al. Quantifying aviation-related contributions to ambient ultrafine particle number concentrations using interpretable machine learning. *Environ Sci Technol*. 2025:acs.est.5c07989. doi:10.1021/acs.est.5c07989
21. Liang C, Liu L, Zhao T, et al. Predicting visual acuity after retinal vein occlusion anti-VEGF treatment: development and validation of an interpretable machine learning model. *J Med Syst*. 2025;49(1):57. doi:10.1007/s10916-025-02190-3
22. Han C, Yang G, Wen H, et al. Development and validation of a quick screening tool for predicting neck pain patients benefiting from spinal manipulation: a machine learning study. *Chin Med*. 2025;20(1):74. doi:10.1186/s13020-025-01131-z
23. Hasan M, Wu W, Zhao X. SHAP-driven feature analysis approach for epileptic seizure prediction. *J Med Syst*. 2025;49(1):77. doi:10.1007/s10916-025-02211-1
24. Luo T, Yan M, Zhou M, et al. Improved prognostication of overall survival after radiotherapy in lung cancer patients by an interpretable machine learning model integrating lung and tumor radiomics and clinical parameters. *Radiol med*. 2024;130(1):96–109. doi:10.1007/s11547-024-01919-3

25. Chen Y, Long T, Wang M, et al. Prospective cohort study integrating plasma proteomics and machine learning for early risk prediction of prostate cancer. *Int J Surg*. 2025;111(9):6123–6134. doi:10.1097/JS9.0000000000002805
26. Lu K, Huang Z, Liang S, et al. A physiology-based trigger score to guide perioperative transfusion of allogeneic red blood cells: a multicentre randomised controlled trial. *Transfus Med*. 2022;32(5):375–382. doi:10.1111/tme.12883