Research Square

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

An Interpretable Multimodal Machine Learning Model for Predicting Early Hematoma Progression in Bifrontal Cerebral Contusion

Liang Wu

nydzwl@126.com

General Hospital of Ningxia Medical University

Guoqing Jiang

General Hospital of Ningxia Medical University

TAO Wang

University of South China

Dongpo Su

Xianyang Central Hospital

Yanbin Zhao

Ningxia Medical University

Yixiang Xing

General Hospital of Ningxia Medical University

Hao Xiong

Ningxia Medical University

Qinghua Zhang

General Hospital of Ningxia Medical University

Research Article

Keywords: Bifrontal cerebral contusion, Hematoma progression, Machine learning, Lasso regression, SHAP, Risk prediction, Multimodal features

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10988030/v1

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

Additional Declarations: No competing interests reported.

Abstract

Objective

To develop and validate a risk prediction model for early hematoma progression (HP) in patients with bifrontal cerebral contusions, systematically compare the performance of multiple machine learning algorithms, identify an optimal model with robustness, generalizability, and clinical interpretability, and determine the key factors driving HP.

Methods

This dual-center retrospective study included 1,100 adult patients with radiologically confirmed bifrontal cerebral contusions. The modeling cohort comprised 454 patients from the General Hospital of Ningxia Medical University and was randomly divided into a training set of 318 patients and an internal validation set of 136 patients. An independent cohort of 646 patients from the Second Affiliated Hospital of the University of South China was used for external validation. Demographic characteristics, vital signs, laboratory indicators, and cranial computed tomography features were collected, with HP defined as the outcome. To minimize information leakage, preprocessing, feature selection, model fitting, and hyperparameter optimization were conducted exclusively within the training set. A two-stage feature-selection strategy combining random forest importance ranking and least absolute shrinkage and selection operator regression was used to identify the core predictors, which were subsequently applied unchanged to the internal and external validation cohorts. Thirteen machine learning models were developed, and their discrimination, classification performance, calibration, and clinical utility were evaluated. Shapley Additive Explanations were used to characterize global and patient-level contributions of the predictors to the selected model.

Results

The intersection of features selected by RF and Lasso yielded 10 stable core predictors: time to baseline CT, ventricular compression, sulcal effacement, hematoma density coefficient of variation (Hematoma.DCV), heart rate, white blood cell count, eosinophil count, albumin, monocyte-to-lymphocyte ratio (MLR), and prognostic nutritional index (PNI). Among the 13 models, LightGBM achieved the highest apparent training-set AUC of 0.872 but decreased to 0.652 in the internal validation set, a pattern consistent with limited generalizability and possible overfitting. In contrast, Lasso achieved the numerically highest AUC in the internal validation set, with an AUC of 0.764 (95% CI: 0.677–0.852). In the external cohort, its AUC decreased to 0.694 but remained numerically higher than those of the other models displayed in the external ROC analysis. Because formal pairwise statistical comparisons were not available, these findings should not be interpreted as demonstrating the statistical superiority of Lasso over the other models. External decision curve analysis indicated potential net benefit mainly at low threshold probabilities, whereas calibration was suboptimal (Hosmer–Lemeshow χ^2=33.87, P <

0.001; Brier score = 0.2405), with a tendency to underestimate observed HP risk. SHAP analysis indicated that Hematoma.DCV, time to baseline CT, MLR, sulcal effacement, PNI, heart rate, and ventricular compression were the main drivers of model prediction.

Conclusion

In patients with bifrontal cerebral contusions, the Lasso model showed comparatively stable discrimination and greater structural interpretability than the higher-capacity ensemble models. However, external discrimination was modest, calibration was inadequate, and clinical net benefit was mainly confined to low threshold probabilities. Hematoma density heterogeneity, the timing of baseline CT, systemic inflammatory–immune–nutritional status, and local intracranial compression signs contributed importantly to model predictions. The model should currently be regarded as a supplementary risk-ranking approach rather than a stand-alone screening tool or a source of directly actionable absolute-risk estimates. External recalibration and larger multicenter prospective validation are required before clinical implementation.

1. Introduction

Traumatic brain injury (TBI) is one of the leading causes of death and disability worldwide, with more than 69 million new cases occurring annually[1]. Hematoma progression (HP) secondary to cerebral contusion is a major factor contributing to rapid clinical deterioration and poor neurological outcomes[2, 3]. Bifrontal cerebral contusion represents a distinct subtype of TBI, typically resulting from coup–contrecoup mechanisms. It is characterized by extensive bilateral frontal involvement, complex secondary injury cascades, and a high risk of hemorrhagic and edematous progression[4–7]. Previous studies have shown that patients with bifrontal contusions are more prone to delayed neurological deterioration, progressive mass effect, and poor functional outcomes, and their clinical course differs substantially from that of patients with unilateral frontal injuries[4–7]. Importantly, the main clinical challenge of bifrontal cerebral contusion lies not only in lesion identification but also in the frequent mismatch between initially mild neurological symptoms and subsequent mass effect or delayed deterioration[4–6]. Therefore, subtype-specific risk modeling is of considerable importance for improving diagnostic and therapeutic precision.

Early identification of patients at high risk of HP is essential for optimizing repeat cranial CT strategies, strengthening neurological monitoring, and determining the appropriate timing of surgical intervention[8, 9]. However, conventional clinical prediction models are often based on single linear statistical approaches and may fail to capture nonlinear associations and interactions among multimodal clinical, laboratory, and imaging data. Consequently, their predictive performance and generalizability remain limited [10–14]. In recent years, machine learning has shown promise in risk assessment and clinical decision support for TBI[10–14]. Nevertheless, most existing studies have focused on broad TBI

populations, with endpoints such as mortality, need for neurosurgical intervention, or long-term functional outcome[10–12]. In contrast, systematic comparisons of multiple algorithms, stable feature selection, and clinically interpretable modeling for early HP in patients with bifrontal cerebral contusions —a high-risk subtype with a distinctive tendency toward delayed deterioration—remain limited [4, 9, 13–17].

Our previous multicenter study represented an initial step toward addressing this problem by developing and externally evaluating an interpretable multimodal machine-learning model for delayed hematoma progression in a broad population of patients with frontal lobe contusion [59]. Ten routinely available clinical, laboratory, and CT-derived predictors were identified, including baseline contusion volume, hematoma density-related parameters, hematoma surface area-to-volume ratio, lymphocyte-to-monocyte ratio, Glasgow Coma Scale score, glucose-to-potassium ratio, time from injury to baseline CT, and eosinophil count. The selected support vector machine model achieved an AUROC of 0.801 in the internal test set and 0.724 in the external validation cohort, supporting the feasibility of multimodal risk modeling in frontal lobe contusion [59].

However, that study considered frontal lobe contusion as a broad anatomical category and did not develop a dedicated model for patients with bifrontal involvement. Bifrontal cerebral contusions may constitute a clinically relevant subgroup because bilateral lesions may be associated with more extensive tissue injury, evolving bilateral mass effect, ventricular or sulcal compression, and a mismatch between initially preserved neurological function and subsequent deterioration [4–7]. Accordingly, the performance and calibration of a model developed in a broad frontal-lobe population cannot be assumed to extend directly to the bifrontal subgroup. Dedicated subtype-focused modeling may therefore help determine whether shared predictive domains—such as hematoma heterogeneity, imaging timing, and systemic inflammatory responses—remain informative and whether local compression indicators contribute to risk prediction in patients with bifrontal cerebral contusions.

The present study was designed as a subtype-focused extension of our previous work. We integrated multimodal clinical, laboratory, and CT imaging data from patients with bifrontal cerebral contusions, applied a leakage-controlled two-stage feature-selection strategy combining random-forest importance ranking and Lasso regression, systematically compared 13 machine-learning algorithms, and used SHAP analysis to characterize the contributions of the selected predictors. The primary aim was to develop and externally evaluate a parsimonious and interpretable model for early HP in bifrontal cerebral contusion. A secondary aim was to examine the extent to which the predictive domains identified in our previous broad frontal-lobe study were also represented among the predictors selected in the bifrontal cohort [59].

2. Materials and Methods

2.1 Study Design and Participants

This was a dual-center retrospective prediction-model study involving a total of 1,100 patients with bifrontal cerebral contusions. Consecutive patients admitted to the Department of Neurosurgery, General Hospital of Ningxia Medical University, between August 2020 and October 2025 constituted the modeling cohort (n=454). Before model development, the modeling cohort was randomly divided at a ratio of 7:3 into a training set of 318 patients and an internal validation set of 136 patients. The training set was used for preprocessing-parameter estimation, feature selection, model fitting, and hyperparameter optimization, whereas the held-out internal validation set was used for model evaluation and comparison. Consecutive eligible patients admitted to the Second Affiliated Hospital of the University of South China between January 2020 and December 2025 constituted the independent external validation cohort (n=646). The external cohort was not used for preprocessing-parameter estimation, feature selection, hyperparameter optimization, model fitting, model selection, threshold determination, or coefficient re-estimation. The overall study design, cohort allocation, leakage-controlled model-development pipeline, external validation, and interpretability analysis are summarized in **Figure 1**.

The inclusion criteria were as follows: (1) age ≥18 years; (2) baseline cranial CT completed within 24 h after admission; (3) radiologically confirmed bifrontal cerebral contusion; and (4) availability of the clinical data required for predictor assessment and availability of baseline and follow-up CT images suitable for outcome determination. The exclusion criteria were: (1) penetrating craniocerebral injury, intracranial tumor, cerebral vascular malformation, or other structural intracranial lesions; (2) a history of severe stroke, dementia, or other neurological diseases that could affect outcome assessment; (3) missingness exceeding 30% across the prespecified candidate predictor set or absence of information required to determine HP; (4) emergency surgery immediately after admission, precluding standardized imaging follow-up; and (5) severe heart, liver, kidney, or other major organ failure.

2.2 Ethical Statement

The study protocol was approved by the Ethics Committee of the General Hospital of Ningxia Medical University (approval No. KYLL-2025-2410). Because of the retrospective nature of the study, the requirement for informed consent was waived by the ethics committee. All data were anonymized before analysis.

2.3 Data Collection

Based on previous literature on TBI and hematoma progression and considering clinical availability, four categories of candidate variables were collected: (1) general baseline characteristics, including age, sex, medical history, and admission vital signs; (2) clinical and imaging variables related to neurological injury, including admission Glasgow Coma Scale score, time from injury to first CT examination, baseline hematoma volume, hematoma surface area, edema volume, density-related parameters, and signs of ventricular or sulcal compression; (3) laboratory indicators, including routine blood tests, biochemical parameters, and derived immune-inflammatory indices within 24 h after admission; and (4) qualitative imaging features, including concomitant subarachnoid hemorrhage, epidural hematoma, subdural

hematoma, skull fracture, midline shift, black-hole sign, island sign, and other CT signs. Candidate predictors were defined before model development based on clinical relevance, previous literature, and routine availability[8,15–17]. No variable was selected according to its association with HP in the internal or external validation cohort.

The missing rate of each candidate variable was less than 5%. To minimize information leakage, missing-value imputation and standardization parameters were estimated exclusively from the training set. Missing continuous values were imputed using the corresponding training-set mean. Continuous variables were standardized by Z-score transformation using the means and standard deviations estimated from the training set. The resulting imputation and standardization parameters were subsequently applied unchanged to the internal and external validation cohorts. No preprocessing parameters were re-estimated in either validation dataset.

2.4 Definition of HP and Imaging Measurement

HP was defined as an increase of at least 25% in hematoma volume on follow-up CT compared with baseline CT. Patients were accordingly classified into the HP and non-HP groups.

Hematoma and perihematomal edema volumes were measured using the open-source software 3D Slicer version 5.8.1[18]. Previous studies have shown that software-based semiautomatic measurements are more accurate for irregular hematomas than traditional ABC/2 or Tada formula-based methods[19,20]. Based on baseline and follow-up CT images, semiautomatic segmentation was performed using fixed Hounsfield unit thresholds: 50–100 HU for hematoma regions and 20–40 HU for peri-contusional edema regions. Three-dimensional reconstruction was used to obtain quantitative parameters, including hematoma volume, surface area, mean density, and density standard deviation. Hematoma.DCV was calculated as the standard deviation of hematoma density divided by the mean hematoma density, thereby representing the relative dispersion of CT attenuation values within the segmented hematoma[21,22].

To ensure measurement consistency, all quantitative imaging analyses were independently performed under blinded conditions by two neuroradiologists with more than 5 years of experience. If the difference in continuous measurements exceeded 10% or qualitative assessments were inconsistent, a senior neurosurgeon with more than 10 years of experience made the final adjudication. Interobserver agreement for the quantitative imaging measurements was good, with an overall intraclass correlation coefficient of 0.89.

2.5 Feature Selection Strategy

To minimize information leakage and obtain an unbiased evaluation of model performance, all feature-selection procedures were conducted exclusively within the training set after training-set-based preprocessing. Neither the internal validation set nor the external validation cohort was used to rank, screen, or select predictors.

First, RF-based feature importance was used to rank the prespecified candidate predictors within the training set and identify variables with potentially relevant predictive contributions. Second, Lasso logistic regression was performed within the training set, with the penalty parameter determined by ten-fold cross-validation. Variables with nonzero coefficients at the selected penalty parameter were retained. The intersection of the RF-ranked variables and the Lasso-selected variables constituted the final predictor set.

After feature selection, the identities and definitions of the selected predictors were locked. The same predictor set was then applied unchanged to the internal and external validation cohorts. No feature-selection procedure was repeated in either validation dataset.

2.6 Model Development and Dataset Partitioning

Based on the locked predictor set, 13 machine learning models were evaluated: random forest, kernel support vector machine, logistic regression, nearest-neighbor classification, partial least squares, boosting, neural network, Bayesian classification, discriminant analysis, Lasso logistic regression, adaptive boosting, CATBoost, and LightGBM.

The training set was used exclusively for preprocessing-parameter estimation, feature selection, model fitting, and hyperparameter optimization. The held-out internal validation set was used to compare the prespecified candidate models. Because the final model was selected after reviewing internal validation performance, the internal validation results were regarded as model-selection estimates rather than fully independent estimates of final-model performance. The independent external cohort provided the principal assessment of the transportability of the selected model. No coefficients were re-estimated and no model was refitted in either validation cohort.

2.7 Hyperparameter Optimization

Hyperparameters were optimized using grid search conducted exclusively within the training set. The internal and external validation cohorts were not used for hyperparameter selection or model fitting. The final hyperparameter combinations for all models are provided in the Supplementary Materials.

2.8 Model Evaluation and Interpretability Analysis

Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals where available. Threshold-dependent classification performance was evaluated using accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and the Youden index. Calibration was assessed using calibration plots, the Hosmer–Lemeshow goodness-of-fit test, and the Brier score. Decision curve analysis and clinical impact curves were used to evaluate potential clinical utility across different threshold probabilities.

For external validation, the models developed in the training cohort were applied directly to the independent external cohort without model refitting or coefficient re-estimation. External performance

was interpreted jointly in terms of discrimination, calibration, and estimated net clinical benefit. Because a model may preserve rank discrimination while producing inaccurate absolute probabilities, discrimination and calibration were evaluated as distinct aspects of model performance.

SHAP values were calculated using KernelExplainer to characterize global and patient-level contributions of the predictors to model output**[24]**. Global feature importance, beeswarm plots, dependence plots, and individual-level contribution plots were generated. SHAP values were interpreted as contributions to model predictions rather than evidence of causal relationships.

2.9 Statistical Analysis

All statistical analyses were performed using R version 4.5.1. The principal packages included caret, glmnet, randomForest, e1071, KernelSHAP, pROC, and tidyverse. All tests were two-sided, and $P<0.05$ was considered statistically significant.

The study included 1,100 patients, of whom 242 experienced HP, corresponding to an overall event rate of 22.0%. Although the number of events exceeded commonly cited events-per-predictor heuristics for a regression model containing 10 predictors, these empirical rules do not guarantee an adequate effective sample size for high-capacity machine learning algorithms. In particular, the model-development cohort included 454 patients and the training set included 318 patients. The available training sample was therefore considered more compatible with regularized regression than with complex ensemble or neural-network models.

The overall ratio of HP to non-HP patients was approximately 1:3.55, indicating class imbalance. To preserve the original clinical outcome distribution and avoid potential overfitting introduced by resampling, no additional over-sampling or under-sampling strategy was applied. Model evaluation therefore considered sensitivity, specificity, F1 score, calibration, and external generalizability rather than accuracy alone.

3. Results

3.1 Study Population and Dataset Partitioning

A total of 1,100 patients with bifrontal cerebral contusions were included in the study. The modeling cohort comprised 454 patients treated at the General Hospital of Ningxia Medical University. Of these, 318 patients were assigned to the training set and 136 to the internal validation set. The independent external validation cohort comprised 646 patients treated at the Second Affiliated Hospital of the University of South China. The training set was used for preprocessing, feature selection, model development, and hyperparameter optimization, whereas the internal and external validation cohorts were reserved for model evaluation. No patient was included in more than one cohort, and no model was refitted in the external validation cohort. The study flow and allocation of patients to the training, internal-validation, and external-validation cohorts are shown in **Figure 1**.

3.2 Feature Selection Results

Within the training set, Lasso regression was performed with the penalty parameter determined by cross-validation (**Figure 2A**). Random forest model stability was assessed according to the error rate across the number of trees (**Figure 2B**). The intersection of the predictors retained by random forest and Lasso regression yielded 10 stable core features (**Figure 2C**): time to baseline CT, ventricular compression, sulcal effacement, Hematoma.DCV, heart rate, white blood cell count, eosinophil count, albumin, MLR, and PNI. The selected predictor set was subsequently locked and applied unchanged to the internal and external validation cohorts.

3.3 Comparison of Training Time Across Models

Training time differed markedly among the 13 machine learning models. AdaptiveBoosting required the longest training time, at 78.6 s, followed by BayesMethod 7.0 s, RandomForest 3.5 s, BoostingMethod 2.7 s, and SVM_Kernel 2.4 s. In contrast, DiscriminantModel 0.1 s, LogisticModel 0.2 s, PLSModel 0.2 s, NeighborMethod 0.3 s, and Lasso 0.4 s trained rapidly. Overall, linear and relatively low-complexity models required less computational time. However, computational efficiency alone does not establish readiness for clinical deployment, which additionally requires adequate external calibration, prospective validation, and clinical impact assessment.

3.4 Model Performance in the Training Set

In the training set, the models generally showed high specificity but relatively low sensitivity, although their overall classification profiles varied across algorithms (**Figure 3A**). LightGBM achieved the highest apparent training-set AUC of 0.872 (95% CI: 0.831–0.912), with an accuracy of 0.777, sensitivity of 0.313, specificity of 0.978, F1 score of 0.459, and Youden index of 0.291 (**Figures 3B and 3C**). CATBoost had an AUC of 0.733 (95% CI: 0.671–0.794), Lasso had an AUC of 0.693 (95% CI: 0.629–0.756), DiscriminantModel had an AUC of 0.690 (95% CI: 0.628–0.751), and LogisticModel had an AUC of 0.676 (95% CI: 0.612–0.741). These results represent apparent within-training-set performance and should not be interpreted as evidence of superior generalizability. The complete threshold-dependent classification results for the training set are presented in **Table 1A.**

3.5 Model Performance in the Internal Validation Set

In the internal validation set, classification and discriminative performance varied substantially among models (**Figures 3D–3F**). The AUCs of the 13 models were as follows: RandomForest, 0.595 95% CI: 0.480–0.710; SVM_Kernel, 0.708 95% CI: 0.610–0.806; LogisticModel, 0.733 95% CI: 0.635–0.832; NeighborMethod, 0.664 95% CI: 0.556–0.772; PLSModel, 0.732 95% CI: 0.633–0.830; BoostingMethod, 0.681 95% CI: 0.581–0.780; NeuralNet, 0.703 95% CI: 0.599–0.808; BayesMethod, 0.701 95% CI: 0.602–0.800; DiscriminantModel, 0.741 95% CI: 0.648–0.834; Lasso, 0.764 95% CI: 0.677–0.852; AdaptiveBoosting, 0.680 95% CI: 0.578–0.781; CATBoost, 0.698 95% CI: 0.596–0.799; and LightGBM, 0.652 95% CI: 0.547–0.757.

Among the evaluated models, Lasso achieved the numerically highest internal-validation AUC of 0.764 (95% CI: 0.677–0.852; **Figures 3E and 3F**). However, because formal pairwise AUC comparisons were not performed, this numerical ranking does not establish that Lasso was statistically superior to DiscriminantModel, LogisticModel, PLSModel, or the other candidate algorithms. The threshold-dependent classification results for all candidate models in the internal validation set are summarized in **Table 1B**.

Table 1. Classification performance of the 13 candidate models.

A. Training set

Model	Sensitivity	Specificity	Accuracy	PPV	NPV	F1 score	Youden index
RandomForest	0.232	0.934	0.722	0.605	0.737	0.336	0.167
SVM_Kernel	0.081	0.974	0.703	0.571	0.709	0.142	0.054
LogisticModel	0.323	0.917	0.737	0.627	0.757	0.427	0.240
NeighborMethod	0.253	0.925	0.722	0.595	0.740	0.355	0.178
PLSModel	0.141	0.952	0.706	0.560	0.719	0.226	0.093
BoostingMethod	0.121	0.991	0.728	0.857	0.722	0.212	0.112
NeuralNet	0.242	0.921	0.716	0.571	0.737	0.340	0.163
BayesMethod	0.182	0.947	0.716	0.600	0.727	0.279	0.129
DiscriminantModel	0.273	0.912	0.719	0.574	0.743	0.370	0.185
Lasso	0.182	0.956	0.722	0.643	0.729	0.283	0.138
AdaptiveBoosting	0.162	0.939	0.703	0.533	0.721	0.248	0.100
CATBoost	0.202	0.982	0.746	0.833	0.739	0.325	0.184
LightGBM	0.313	0.978	0.777	0.861	0.766	0.459	0.291

B. Internal validation set

Model	Sensitivity	Specificity	Accuracy	PPV	NPV	F1 score	Youden index
RandomForest	0.250	0.874	0.717	0.400	0.776	0.308	0.124
SVM_Kernel	0.250	0.916	0.748	0.500	0.784	0.333	0.166
LogisticModel	0.281	0.937	0.772	0.600	0.795	0.383	0.218
NeighborMethod	0.219	0.937	0.756	0.538	0.781	0.311	0.156
PLSModel	0.125	0.968	0.756	0.571	0.767	0.205	0.093
BoostingMethod	0.188	0.979	0.780	0.750	0.782	0.300	0.166
NeuralNet	0.031	1.000	0.756	1.000	0.754	0.061	0.031
BayesMethod	0.250	0.937	0.764	0.571	0.788	0.348	0.187
DiscriminantModel	0.281	0.926	0.764	0.563	0.793	0.375	0.208
Lasso	0.250	0.958	0.780	0.667	0.791	0.364	0.208
AdaptiveBoosting	0.219	0.958	0.772	0.636	0.784	0.326	0.177
CATBoost	0.156	0.979	0.772	0.714	0.775	0.256	0.135
LightGBM	0.219	0.947	0.764	0.583	0.783	0.318	0.166

PPV, positive predictive value; NPV, negative predictive value; F1 score, harmonic mean of precision and sensitivity; SVM, support vector machine; PLS, partial least squares; LightGBM, Light Gradient Boosting Machine

3.6 Comparison Between Training and Internal Validation Performance

LightGBM showed a decrease in AUC from 0.872 in the training set to 0.652 in the internal validation set, a pattern consistent with limited generalizability and possible overfitting. The Lasso model yielded an AUC of 0.693 in the training set and 0.764 in the internal validation set. The higher AUC in the internal validation set should not be interpreted as evidence that the model performed intrinsically better in unseen data, because this difference may reflect sampling variability arising from the random split. Based on its numerically highest internal-validation AUC, low computational complexity, regularized structure, and interpretability, Lasso was selected as the final candidate model. Because this selection was informed by internal-validation results, the independent external cohort was regarded as the principal evaluation of the selected model's transportability.

3.7 Classification Performance of the Selected Lasso Model

In the training set, the Lasso model had an accuracy of 0.722, sensitivity of 0.182, specificity of 0.956, positive predictive value of 0.643, negative predictive value of 0.729, F1 score of 0.283, and Youden

index of 0.138. Thus, the model correctly classified most non-HP patients but failed to identify a substantial proportion of patients who experienced HP.

The marked imbalance between specificity and sensitivity has important clinical implications. Although the model retained discriminatory information according to the ROC analysis, its low training-set sensitivity indicates that it should not be used as a stand-alone screening or rule-out tool for HP. Instead, the model should be considered a supplementary risk-ranking approach whose output must be interpreted together with serial neurological examinations, repeat CT findings, and clinical judgment.

3.8 Internal ROC, Decision Curve, and Calibration Analyses

ROC analysis showed that LightGBM, CATBoost, Lasso, and DiscriminantModel had relatively high apparent discrimination in the training set. In the internal validation set, Lasso, DiscriminantModel, LogisticModel, and PLSModel yielded comparatively higher AUCs. Decision curve analysis suggested that the selected Lasso model provided a higher estimated net benefit than the treat-all and treat-none strategies over a limited range of threshold probabilities in the internal validation and training sets (**Figures 5A and 5B**, respectively), indicating potential but threshold-dependent clinical utility. The clinical impact curves showed the numbers of patients classified as high risk and the corresponding numbers of observed HP cases across threshold probabilities in the training and internal validation sets (**Figures 4A and 4C**).

Calibration plots suggested approximate agreement between predicted and observed risks in the training and internal validation sets (**Figures 4B and 4D**). However, because calibration intercepts, calibration slopes, and internal Brier scores were not reported, the internal calibration results should be interpreted cautiously and should not be characterized as definitively demonstrating good calibration.

3.9 External Validation

External validation was performed in an independent cohort of 646 patients treated at the Second Affiliated Hospital of the University of South China. Overall, model discrimination was attenuated compared with that observed in the internal validation set of 136 patients. Lasso yielded the numerically highest external AUC of 0.694, followed by LogisticModel at 0.690, DiscriminantModel at 0.689, SVM_Kernel at 0.648, and PLSModel at 0.644. RandomForest, CATBoost, BoostingMethod, LightGBM, NeighborMethod, and NeuralNet yielded AUCs of 0.613, 0.596, 0.586, 0.584, 0.577, and 0.418, respectively. Thus, the Lasso AUC decreased from 0.764 in the internal validation set to 0.694 in the external cohort. Because its AUC was very similar to those of LogisticModel and DiscriminantModel and formal pairwise statistical comparisons were unavailable, the external results support broadly comparable performance among these lower-complexity models rather than the statistical superiority of Lasso. The external ROC curves of the candidate models are shown in **Figure 6C**.

Decision curve analysis showed that several models, including Lasso, provided positive estimated net benefit mainly within a relatively low threshold-probability range of approximately 0.05–0.30 (**Figure 6B**).

At higher thresholds, net benefit approached zero or became negative for most models. The clinical impact curve showed that, at low thresholds, the number of patients classified as high risk was substantially greater than the number who actually experienced HP (**Figure 6A**). The difference narrowed as the threshold increased, but fewer patients were then identified as high risk. These findings illustrate a trade-off between identifying more potential HP cases and increasing false-positive classifications.

External calibration of the Lasso model was inadequate (**Figure 6D**). The Hosmer–Lemeshow test was statistically significant (χ^2=33.87, P<0.001), although this test is sensitive to sample size and the grouping of predicted risks. The Brier score was 0.2405, and the observed event rates were generally higher than the corresponding mean predicted probabilities across several risk strata, indicating systematic underestimation of HP risk. Therefore, although the Lasso model retained modest ability to rank patients according to relative risk, its predicted probabilities should not be interpreted as accurately calibrated absolute risks.

Taken together, the external results indicate that the selected model is not ready for direct threshold-based clinical decision-making. External recalibration and further prospective validation are required before its predicted probabilities can be used to guide follow-up CT, monitoring intensity, or treatment decisions.

3.10 SHAP-Based Interpretability Analysis

The SHAP feature importance bar plot showed that the top 10 variables were Hematoma.DCV 0.0540, time to baseline CT 0.0396, MLR 0.0254, sulcal effacement 0.0216, PNI 0.0193, heart rate 0.0131, ventricular compression 0.0105, albumin 0.0069, white blood cell count 0.0035, and eosinophil count 0.0001(**Figure 7A**). The SHAP beeswarm plot showed that Hematoma.DCV and time to baseline CT had the strongest influence on prediction results, while MLR, PNI, and heart rate also made stable contributions (**Figure 7B**). Dependence plots showed that higher Hematoma.DCV values were generally associated with larger positive contributions to the model output, while the relationship between time to baseline CT and model output was nonlinear. Elevated MLR was also associated with larger positive model contributions, whereas PNI contributed predictive information related to nutritional and immune status (**Figure 7C**). These patterns describe the behavior of the fitted model and should not be interpreted as causal dose–response relationships.

At the individual level, sample 1 had a baseline model output of 0.292 and a final predicted probability of 0.600 after the combined positive and negative contributions of its predictor values (**Figure 7D**). Sample 9 had a final predicted probability of 0.259, which was lower than the baseline output of 0.292 (**Figure 7E**). These examples demonstrate how SHAP can decompose an individual model prediction into predictor-specific contributions. They do not demonstrate that modifying any individual predictor would change the probability of HP.

4. Discussion

This dual-center study included 1,100 patients with bifrontal cerebral contusions, comprising a modeling cohort of 454 patients from the General Hospital of Ningxia Medical University and an independent external validation cohort of 646 patients from the Second Affiliated Hospital of the University of South China. Multimodal clinical, laboratory, and CT imaging information was integrated to predict early HP. Thirteen machine learning models were systematically compared, and feature selection and SHAP-based interpretability analyses were performed using RF combined with Lasso regression. The results showed that although complex ensemble models had stronger within-sample fitting ability in the training set, their generalization performance was unstable. Complex ensemble models achieved higher apparent performance in the training set but showed marked performance attenuation during validation. Lasso achieved the numerically highest AUC in the internal validation set and external cohort; however, external discrimination was modest, and its statistical superiority over similarly performing linear models was not established. Its principal advantages were regularization, structural simplicity, computational efficiency, and transparent predictor contributions. Further analysis suggested that Hematoma.DCV, the baseline CT time window, inflammatory–immune–nutritional status, and local compression signs were the main sources of predictive information. These findings indicate that HP risk in bifrontal cerebral contusion is not determined solely by local lesion burden but reflects the combined effects of local imaging instability, observation time window, and systemic response status[2,3,8,9,13–17].

One of the most noteworthy findings of this study was that LightGBM achieved the highest AUC in the training set but showed a marked decline in the validation set, whereas the Lasso model, despite its less prominent performance in the training set, achieved the best discriminative performance in the validation set. This phenomenon suggests that, under the current data structure, increasing model complexity did not necessarily translate into better generalization. In clinical datasets with limited sample size, relatively insufficient event numbers, and strong correlations among variables, complex models are more likely to learn local noise and random patterns in the training set, resulting in excellent within-sample performance but insufficient external stability[10–14,25–31]. In this dataset, coefficient shrinkage may have reduced model variance and limited the influence of redundant predictors. Nevertheless, these results do not establish that regularized linear models are generally superior to ensemble methods. Rather, they indicate that algorithm capacity should be matched to sample size, predictor dimensionality, correlation structure, and signal-to-noise ratio[25–31].

Lee et al. developed a machine learning model for predicting early hemorrhagic progression in TBI and suggested that complex models such as XGBoost had strong predictive potential[10]. Moyer et al. also found that ensemble algorithms had good discriminative ability when modeling emergency neurosurgical intervention within 24 h in patients with moderate-to-severe TBI [11]. However, those studies largely focused on broader TBI populations, and their modeling endpoints and data structures differed from those of the present study[12–14,25–31]. In contrast, this study found that, in the specific clinical subgroup of bifrontal cerebral contusion, complex models did not translate into better validation performance, whereas the regularized Lasso model showed the most stable performance. This suggests that the benefits of model complexity may be strongly influenced by sample size, event number, and

feature correlation structure. The key implication is that, in real-world clinical prediction tasks, the “most complex model” is not necessarily the “best model”; generalizability, calibration, and interpretability should be valued equally alongside discrimination[12,28–31].

The external validation findings provide a more cautious assessment of model transportability. Although Lasso remained the model with the highest AUC in the external cohort, its AUC decreased from 0.764 in the internal validation set to 0.694 externally, indicating only modest discrimination under a different institutional setting. This attenuation is expected when a prediction model is transferred across centers because differences in patient characteristics, CT acquisition protocols, laboratory measurement procedures, clinical management, and outcome prevalence may alter predictor distributions and predictor–outcome associations. Because the external AUCs of Lasso, logistic regression, and discriminant analysis were very similar, the findings support the competitive transportability of lower-complexity models rather than definitive superiority of Lasso[12,13,28–31].

External validation identified important limitations that were not apparent from discrimination alone. The significant Hosmer–Lemeshow result, Brier score of 0.2405, and calibration plot jointly indicated inadequate agreement between predicted and observed risks, with an overall tendency to underestimate HP risk. The calibration curve suggested that the model generally underestimated the observed probability of HP in the external cohort. This finding highlights the distinction between discrimination and calibration: a model may preserve the ability to rank patients according to relative risk while providing inaccurate absolute risk estimates[25,27–31]. Therefore, the current model may be more appropriate for preliminary relative risk stratification than for directly determining intervention thresholds. Recalibration of the intercept and calibration slope, followed by prospective validation, may improve the accuracy of absolute risk estimates in different clinical settings.

The external DCA and clinical impact curve further showed that clinical usefulness was dependent on the selected threshold probability. Positive net benefit was mainly observed at lower thresholds, where a relatively large number of patients were classified as high risk. Such a strategy may be acceptable when the consequence of missing HP is considered more serious than the burden of additional CT examinations or intensified monitoring. However, the larger difference between the predicted high-risk population and the number of observed HP cases at low thresholds also implies an increased false-positive burden. Consequently, the threshold used in practice should be selected by balancing the consequences of missed HP against radiation exposure, resource use, and unnecessary clinical intervention. However, DCA provides a model-based estimate of clinical utility rather than evidence that model-guided care improves patient outcomes. The observed threshold-dependent net benefit should therefore be regarded as exploratory until confirmed in a prospective impact study**[15,25–27]**.

It is worth noting that this study did not rely on a single algorithm to identify key variables. Instead, a two-stage feature selection strategy combining RF importance ranking and Lasso shrinkage was used, yielding 10 intersecting core features. These variables highly overlapped with the high-contribution features identified by SHAP analysis. The overlap between the selected predictors and the features

assigned relatively high importance by SHAP indicates that the fitted model relied consistently on these variables. However, this internal methodological concordance does not constitute independent biological validation[13–17,24].

Previous studies on cerebral contusion prediction have emphasized the value of multidimensional modeling. Sheng et al. developed the traumatic brain parenchyma hematoma expansion aid nomogram by combining demographic, imaging, coagulation, inflammatory, and baseline CT timing variables [15]. Zhang et al. reported that admission MLR was associated with hematoma progression after cerebral contusion [16], while combined radiomics–clinical models improved the prediction of contusion progression compared with clinical variables alone [13,14,17]. More recently, our group developed and externally evaluated an interpretable multimodal model for delayed hematoma progression in frontal lobe contusion, identifying baseline contusion burden, hematoma density-related features, CT timing, neurological status, and systemic inflammatory indicators as important predictive domains [59].

This study focused on bifrontal cerebral contusion as a specific pathological subgroup rather than simply including it within general frontal contusion or broad TBI populations. Bifrontal lesions usually indicate stronger contrecoup injury mechanisms, more extensive bilateral brain involvement, and more complex secondary pathological evolution. Therefore, their HP risk may differ from that of general frontal injuries. Previous studies of bifrontal contusions have mainly focused on clinical manifestations, surgical indications, and delayed deterioration. A systematic review by Van de Zande et al. indicated that patients with bifrontal contusions have high clinical fluctuation and complex management needs[4–6]. Earlier studies by Peterson et al. and Statham et al. emphasized the characteristic clinical pattern of bifrontal lesions, in which patients may initially appear relatively lucid but subsequently deteriorate rapidly[4,7,52–55]. Building on these clinical observations, the present study attempted to transform experience-based judgments into a quantitative individualized risk assessment tool through multimodal variable integration. From this perspective, separate modeling for bifrontal contusion is not merely a further subdivision of frontal cerebral contusion but a more refined approach to risk stratification based on lesion distribution, injury mechanism, and clinical management challenges[55].

Our results showed that the key information driving model predictions was concentrated in hematoma internal heterogeneity, early imaging time window, inflammatory–immune status, and local compression signs. This suggests that bifrontal cerebral contusion follows the general rules of hemorrhagic evolution after cerebral contusion while also reflecting more complex local–systemic interactions under bilateral lesion conditions. The clinical difficulty of bifrontal cerebral contusion lies not only in identifying lesions but also in the frequent mismatch between early neurological symptoms and subsequent mass effect or delayed deterioration. Therefore, compared with general cerebral contusion, this subgroup may require early quantitative risk stratification tools more urgently. Future studies with larger samples should directly compare bifrontal and general frontal cerebral contusions to further determine whether subtype-specific modeling improves predictive performance and clinical applicability[2,3,8,9,52–59].

At the feature level, the most important variable in this study was not traditional hematoma volume but Hematoma.DCV. In this study, Hematoma.DCV was defined as the ratio of the standard deviation of hematoma density to the mean hematoma density. It essentially reflects the relative dispersion of CT density distribution within the hematoma, namely the heterogeneity of hematoma components, rather than simple hemorrhage burden. Heterogeneous hematoma density often indicates mixed blood components, coexistence of hemorrhage at different stages, or incomplete coagulation. Pathophysiologically, this may indicate that the lesion is in an unstable state and therefore more likely to expand. In recent years, research on intracerebral hemorrhage has shifted from traditional volume and morphological signs toward quantitative evaluation of density heterogeneity and surface irregularity[21,22,38,42–47]. Ji et al. showed that quantitative shape irregularity and density heterogeneity significantly improved the prediction of hematoma expansion and suggested that DCV may partially replace traditional imaging signs **[22]**. Jeong et al. further found that lower mean HU was associated with increased risk of hematoma expansion**[44]**. Wu et al. demonstrated that density differences in heterogeneous hematomas were associated with poor functional outcomes from the perspective of quantitative HU variation [43,45–47]. In line with these findings, mainly derived from spontaneous intracerebral hemorrhage, our study suggests that quantitative indicators reflecting internal hematoma heterogeneity may also capture lesion instability more sensitively than simple volume burden in patients with bifrontal cerebral contusions. The high SHAP ranking of Hematoma.DCV indicates that density heterogeneity contributed more strongly than hematoma volume to predictions made by the current model. This model-specific finding does not establish the biological or prognostic superiority of DCV and requires confirmation in independent prospective cohorts.

In addition to imaging heterogeneity, time to baseline CT was also a high-weight variable, suggesting that the timing of the first imaging assessment is crucial for risk identification. Hemorrhagic evolution after cerebral contusion is a dynamic process. If baseline CT is performed too early, the true extent of the lesion may not yet be fully visible, and some patients may still be in the phase of hemorrhagic expansion during the following hours. Therefore, the interval from injury to baseline CT should not be regarded merely as procedural information but rather as a key contextual variable determining the evolutionary stage at which the lesion is observed. Previous studies have repeatedly emphasized the importance of the time window. Sheng et al. included baseline CT timing as a core variable in a model for traumatic intraparenchymal hematoma progression and considered it independently valuable for early risk stratification **[15,38–42]**. Morotti et al. also reported in intracerebral hemorrhage expansion research that earlier first imaging was associated with a higher risk of subsequent expansion, with an evident time-dependent relationship[15,38–41]. The present study further supports this concept in the bifrontal cerebral contusion subgroup, suggesting that static imaging findings alone may be insufficient for HP prediction and that incorporating time-window information helps more accurately reflect dynamic lesion evolution. Time to baseline CT may also capture differences in injury recognition, referral pathways, healthcare access, or clinical severity. Its predictive contribution should therefore not be interpreted solely as a biological marker of hematoma evolution.

This study also found that inflammatory–immune–nutritional indicators, including MLR, PNI, albumin, and white blood cell count, played important roles in the model. This suggests that HP is not determined solely by local mechanical injury but may also be closely related to systemic inflammatory response, immune homeostasis imbalance, and nutritional reserve. Elevated MLR may reflect monocyte-related inflammatory activation, indicating more pronounced secondary inflammation and blood–brain barrier injury. PNI and albumin reflect the body's capacity to respond to traumatic stress from the perspective of nutritional and immune reserve. Previous studies provide strong support for this interpretation. Sheng et al. reported that admission MLR was an important independent predictor of hematoma expansion in acute traumatic intraparenchymal hematoma [15]. Zhang et al. subsequently confirmed that higher MLR levels were significantly associated with hematoma progression in patients with cerebral contusions [16]. Consistent with these studies, our study observed a stable contribution of MLR in patients with bifrontal cerebral contusions, suggesting that monocyte-dominated inflammatory responses may play an important role in the early evolution of this subtype. It should be emphasized that SHAP reflects the relative contribution of variables to model discrimination and does not establish causality. Nevertheless, these results suggest that hemorrhagic progression in bifrontal cerebral contusion may result from the combined effects of local lesion instability and systemic response status[32–37]. Meanwhile, ventricular compression and heart rate were also among the important features. The former indicates local mass effect and brain tissue compression, whereas the latter may indirectly reflect pain, stress, hemodynamic fluctuation, and other systemic physiological responses, further supporting this interpretation**[32,36]**.

Most predictors used in the model are potentially obtainable from routine early admission assessments, including vital signs, laboratory tests, and standard cranial CT. Nevertheless, the current model is not ready for routine clinical implementation because its external discrimination was modest, its external calibration was inadequate, and its estimated clinical utility was mainly confined to low threshold probabilities. At present, it should be regarded as a research-stage supplementary risk-ranking approach rather than a replacement for clinical judgment or a stand-alone screening tool. Its relatively transparent structure may facilitate future development of a visual scoring system or electronic decision-support prototype, but only after external recalibration, prospective validation, and formal clinical impact evaluation**[4,52–54]**. To illustrate a potential form of model presentation, a research-stage prototype interface of the Lasso-based hematoma-progression risk calculator was developed (**Figure 8**). This prototype is presented for conceptual demonstration only and should not be used for stand-alone clinical decision-making.

This study has several limitations. First, its retrospective design may have introduced selection bias and information bias. Second, the sample size and number of events were limited relative to the capacity of some ensemble and neural-network algorithms, which may partly explain their marked performance attenuation during validation. Third, patients who underwent emergency surgery immediately after admission or could not complete standardized imaging follow-up were excluded. Consequently, the findings are primarily applicable to hospitalized patients who can undergo sequential CT assessment and may not generalize to the most critically injured patients[27–31].

Fourth, although an independent external cohort was included, external discrimination was only modest and calibration was inadequate. Differences in case mix, event prevalence, CT acquisition, laboratory measurements, referral patterns, and clinical management may have contributed to this attenuation. The significant Hosmer–Lemeshow result, Brier score of 0.2405, and systematic risk underestimation indicate that the current predicted probabilities should not be interpreted as fully calibrated absolute risks[23].

Fifth, the final model was selected after reviewing internal validation performance. The internal validation AUC may therefore be optimistic for the selected model, and the external cohort provides the more relevant estimate of final-model performance. Sixth, formal statistical comparisons of AUCs were not available. Because Lasso, logistic regression, and discriminant analysis achieved similar validation AUCs, the present findings do not establish the statistical superiority of Lasso. Seventh, the Lasso model showed low sensitivity at the reported training-set classification threshold, limiting its suitability as a stand-alone screening or rule-out tool[18–22,58].

Eighth, some quantitative imaging features depended on semiautomatic segmentation. Although blinded dual-reader measurement and consistency assessment were used, reproducibility across different scanners, acquisition protocols, centers, and operators requires further evaluation. Finally, the study did not include a prospective impact analysis. Decision curve analysis alone cannot demonstrate that model-guided care improves CT scheduling, resource use, treatment decisions, or patient outcomes .

In summary, this study indicates that, for HP prediction in patients with bifrontal cerebral contusions, regularized models with more stable generalization ability may have greater clinical application potential than complex ensemble models with better training-set performance. Hematoma internal density heterogeneity, the timing of initial CT examination, and inflammatory–immune–nutritional status are important dimensions influencing prediction. The stable core features selected by RF and Lasso, together with SHAP interpretation, provide new quantitative evidence for early risk identification and precision management in patients with bifrontal cerebral contusions. These findings also suggest that early deterioration risk in this population can be more accurately identified by jointly quantifying local imaging instability and systemic response status.

5. Conclusion

This dual-center retrospective study developed and externally evaluated multimodal prediction models for early HP in patients with bifrontal cerebral contusions. Lasso achieved the numerically highest AUC in the internal validation set and external cohort, but its external discrimination was modest and its predicted probabilities were inadequately calibrated. Because formal pairwise statistical comparisons were unavailable, the findings do not establish the statistical superiority of Lasso over logistic regression, discriminant analysis, or other competing models.

Hematoma.DCV, time to baseline CT, MLR, PNI, heart rate, and ventricular compression contributed importantly to the fitted model's predictions. These results represent predictive associations rather than causal mechanisms. The selected model should currently be regarded as a supplementary risk-ranking approach rather than a stand-alone screening tool or a source of directly actionable absolute-risk estimates. External recalibration, larger multicenter prospective validation, prespecified threshold evaluation, and clinical impact assessment are required before implementation.

Abbreviations

AUC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; DCA, decision curve analysis; DCV, density coefficient of variation; GCS, Glasgow Coma Scale; HP, hematoma progression; HU, Hounsfield unit; Lasso, least absolute shrinkage and selection operator; LightGBM, Light Gradient Boosting Machine; MLR, monocyte-to-lymphocyte ratio; PLS, partial least squares; PNI, prognostic nutritional index; RF, random forest; ROC, receiver operating characteristic; SHAP, Shapley Additive Explanations; SVM, support vector machine; TBI, traumatic brain injury.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and applicable institutional and national ethical standards. The study protocol was reviewed and approved by the Ethics Committee of the General Hospital of Ningxia Medical University, Yinchuan, China (approval No. KYLL-2025-2410).

The Second Affiliated Hospital, Hengyang Medical School, University of South China participated as the external validation center and provided institutional authorization for the use of deidentified retrospective data in accordance with its local procedures for multicenter retrospective research. Because this study involved only retrospective analysis of deidentified clinical and imaging data, the requirement for individual informed consent was waived by the Ethics Committee of the General Hospital of Ningxia Medical University and was accepted by the participating external center in accordance with its local institutional requirements. All data were deidentified before analysis.

Consent for publication

All authors read and agree to submit this manuscript to your journal.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare no conflicts of interest relevant to this study.

Funding

This study was supported by the Natural Science Foundation of Ningxia, China (No. 2025AAC030868, awarded to Liang Wu). The funding bodies had no role in the design of the study; the collection, analysis, or interpretation of the data; the preparation of the manuscript; or the decision to submit the manuscript for publication.

Authors' contributions

Liang Wu: Conceptualization, methodology, formal analysis, software, investigation, data curation, validation, visualization, writing—original draft, and funding acquisition.

Guoqing Jiang: Conceptualization, methodology, formal analysis, supervision, project administration, validation, and writing—review and editing. Tao Wang: nvestigation, case identification, retrospective data collection, and coordination of the external validation cohort. Dongpo Su: Investigation, data verification, data curation, and writing—review and editing. Yanbin Zhao: Investigation, patient recruitment, data collection, and data curation. Yixiang Xing: Investigation, imaging assessment, and data curation. Hao Xiong: Investigation, imaging assessment, and data curation. Qinghua Zhang: Investigation, imaging assessment, and data curation. All authors interpreted the findings, critically reviewed the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Additional declarations:

Originality: The authors confirm that this manuscript has not been published elsewhere and is not under consideration by another journal.

Artificial Intelligence (AI) Statement: During the preparation of this work, the authors used no AI-assisted technologies or large language models (LLMs) for generating content, data analysis, or drafting the manuscript. The authors take full responsibility for the content and integrity of the manuscript.

Reporting Checklist: The authors confirm that they have followed the TRIPOD+AI statement for reporting clinical prediction models. A completed TRIPOD+AI checklist is provided as a supplementary file with this submission.

References

1. Dewan MC, Rattani A, Gupta S, Baticulon RE, Hung YC, Punchak M, et al. Estimating the global incidence of traumatic brain injury. J Neurosurg. 2018;130(4):1080–1097. doi:10.3171/2017.10.JNS17352.

2. Kurland D, Hong C, Aarabi B, Gerzanich V, Simard JM. Hemorrhagic progression of a contusion after traumatic brain injury: a review. J Neurotrauma. 2012;29(1):19–31. doi:10.1089/neu.2011.2122.
3. Fletcher-Sandersjöö A, Tatter C, Tjerkaski J, Bartek J, Maegele M, Nelson DW, et al. Time course and clinical significance of hematoma expansion in moderate-to-severe traumatic brain injury: an observational cohort study. Neurocrit Care. 2023;38(1):60–70. doi:10.1007/s12028-022-01609-w.
4. Van De Zande N, Manivannan S, Sharouf F, Shastin D, Abdulla M, Chumas PD, et al. Demographics, presentation, and clinical outcomes after traumatic bifrontal contusions: a systematic review. Neurosurg Rev. 2020;43(3):977–986. doi:10.1007/s10143-019-01098-0.
5. Statham PF, Johnston RA, Macpherson P. Delayed deterioration in patients with traumatic frontal contusions. J Neurol Neurosurg Psychiatry. 1989;52(3):351–354. doi:10.1136/jnnp.52.3.351.
6. Peterson EC, Chesnut RM. Talk and die revisited: bifrontal contusions and late deterioration. J Trauma. 2011;71(6):1588–1592. doi:10.1097/TA.0b013e31822b791d.
7. Gao L, Wu X, Hu J, Jin Y, Han X, Wu X, et al. Intensive management and prognosis of 127 cases with traumatic bilateral frontal contusions. World Neurosurg. 2013;80(6):879–888. doi:10.1016/j.wneu.2013.01.020.
8. Adatia K, Newcombe VFJ, Menon DK. Contusion progression following traumatic brain injury: a review of clinical and radiological predictors, and influence on outcome. Neurocrit Care. 2021;34(1):312–324.
9. Jirlow U, Hossain I, Korhonen O, Depreitere B, Rostami E. Cerebral contusions—pathomechanism, predictive factors for progression and historical and current management. Brain Spine. 2024;4:103329.
10. Lee HS, Kim JH, Son J, Park H, Choi J. Machine learning models for predicting early hemorrhage progression in traumatic brain injury. Sci Rep. 2024;14(1):11690.
11. Moyer JD, Lee P, Bernard C, Henry L, Lang E, Cook F, et al. Machine learning-based prediction of emergency neurosurgery within 24 h after moderate to severe traumatic brain injury. World J Emerg Surg. 2022;17(1):42.
12. Bark D, Boman M, Depreitere B, Wright DW, Lewén A, Enblad P, et al. Refining outcome prediction after traumatic brain injury with machine learning algorithms. Sci Rep. 2024;14(1):8036.
13. He H, Liu J, Li C, Guo Y, Liang K, Du J, et al. Predicting hematoma expansion and prognosis in cerebral contusions: a radiomics-clinical approach. J Neurotrauma. 2024;41(11–12):1337–1352. doi:10.1089/neu.2023.0410.
14. Shih YJ, Liu YL, Chen JH, Ho CH, Yang CC, Chen TY, et al. Prediction of intraparenchymal hemorrhage progression and neurologic outcome in traumatic brain injury patients using radiomics score and clinical parameters. Diagnostics. 2022;12(7):1677.
15. Sheng J, Chen W, Zhuang D, Li T, Yang J, Cai S, et al. A clinical predictive nomogram for traumatic brain parenchyma hematoma progression. Neurol Ther. 2022;11(1):185–203.
16. Zhang H, Duan X, Zhang Y, Zhuang G, Cao D, Meng W, et al. Association between monocyte-to-lymphocyte ratio and hematoma progression after cerebral contusion. Neurocrit Care.

2024;40(3):953–963.
17. Zhang L, Zhuang Q, Wu G, Yu J, Shi Z, Yuan Q, et al. Combined radiomics model for prediction of hematoma progression and clinical outcome of cerebral contusions in traumatic brain injury. Neurocrit Care. 2022;36(2):441–451.
18. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30(9):1323–1341. doi:10.1016/j.mri.2012.05.001.
19. Xu X, Chen X, Zhang J, Zheng Y, Sun G, Yu X, et al. Comparison of the Tada formula with software Slicer: precise and low-cost method for volume assessment of intracerebral hematoma. Stroke. 2014;45(11):3433–3435. doi:10.1161/STROKEAHA.114.007095.
20. Fletcher-Sandersjöö A, Lewén A, Hånell A, Nelson DW, Maegele M, Svensson M, et al. Volumetric assessment of traumatic intracranial hematomas: is ABC/2 reliable? J Neurotrauma. 2024;41(23–24):2545–2553.
21. Morotti A, Boulouis G, Dowlatshahi D, Li Q, Barras CD, Delcourt C, et al. Standards for detecting, interpreting, and reporting noncontrast computed tomographic markers of intracerebral hemorrhage expansion. Ann Neurol. 2019;86(4):480–492. doi:10.1002/ana.25563.
22. Ji Z, Hao Y, Gao B, Zhang X, Zhang Y, Jia J, et al. Quantitative shape irregularity and density heterogeneity predict hematoma expansion in patients with intracerebral hemorrhage. Ann Clin Transl Neurol. 2025;12(10):2068–2076. doi:10.1002/acn3.70141.
23. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. doi:10.1136/bmj.m441.
24. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774.
25. Ye H, Jiang Y, Wu Z, Ruan Y, Shen C, Xu J, et al. A comparative study of a nomogram and machine learning models in predicting early hematoma expansion in hypertensive intracerebral hemorrhage. Acad Radiol. 2024;31(12):5130–5140. doi:10.1016/j.acra.2024.05.035.
26. Duan C, Liu F, Gao S, Zhao J, Niu L, Li N, et al. Comparison of radiomic models based on different machine learning methods for predicting intracerebral hemorrhage expansion. Clin Neuroradiol. 2022;32(1):215–223.
27. Liu Y, Zhao F, Niu E, Chen L. Machine learning for predicting hematoma expansion in spontaneous intracerebral hemorrhage: a systematic review and meta-analysis. Neuroradiology. 2024;66(9):1603–1616.
28. Pease M, Arefan D, Barber J, Yuh E, Puccio A, Hochberger K, et al. Outcome prediction in patients with severe traumatic brain injury using deep learning from head CT scans. Radiology. 2022;304(2):385–394. doi:10.1148/radiol.212181.
29. Tran AT, Zeevi T, Haider SP, Abou Karam G, Berson ER, Tharmaseelan H, et al. Uncertainty-aware deep-learning model for prediction of supratentorial hematoma expansion from admission non-

contrast head computed tomography scan. NPJ Digit Med. 2024;7(1):26. doi:10.1038/s41746-024-01007-w.

30. Zhou Z, Chen W, Yu R, Chen Y, Li X, Zhou H, et al. HE-Mind: a model for automatically predicting hematoma expansion after spontaneous intracerebral hemorrhage. Eur J Radiol. 2024;176:111533. doi:10.1016/j.ejrad.2024.111533.
31. Yalcin C, Abramova V, Terceño M, Oliver A, Silva Y, Lladó X. Hematoma expansion prediction in intracerebral hemorrhage patients by using synthesized CT images in an end-to-end deep learning framework. Comput Med Imaging Graph. 2024;117:102430. doi:10.1016/j.compmedimag.2024.102430.
32. Wang H, Liu Y, Yuan J, Wang Y, Yuan Y, Liu Y, et al. Development and validation of a nomogram for predicting mortality in patients with acute severe traumatic brain injury: a retrospective analysis. Neurol Sci. 2024;45(10):4931–4956.
33. Chu H, Huang C, Zhou Z, Tang Y, Dong Q, Guo Q. Inflammatory score predicts early hematoma expansion and poor outcomes in patients with intracerebral hemorrhage. Int J Surg. 2023;109(3):266–276. doi:10.1097/JS9.0000000000000191.
34. Zhang D, Zhuang D, Li T, Liu X, Zhang Z, Zhu L, et al. An analysis of neutrophil-to-lymphocyte ratios and monocyte-to-lymphocyte ratios with six-month prognosis after cerebral contusions. Front Immunol. 2024;15:1336862.
35. Wang R, Xu J, He M. Blood leukocyte-based clusters in patients with traumatic brain injury. Front Immunol. 2025;15:1504668.
36. He J, Zhang Y, Li T, Deng H, Wang P, Chong W, et al. Glucose-albumin ratio as a new biomarker for predicting mortality after intracerebral hemorrhage. Neurosurg Rev. 2023;46(1):94.
37. Ji Z, Ye W, Wen X, Zhao X, Li N. Predicting intracerebral hemorrhage expansion with inflammation indices, non-contrast computed tomography signs and computed tomography angiography spot sign. Neuropsychiatr Dis Treat. 2024;20:1879–1887. doi:10.2147/NDT.S475550.
38. Morotti A, Boulouis G, Dowlatshahi D, Li Q, Shamy M, Salman RAS, et al. Intracerebral haemorrhage expansion: definitions, predictors, and prevention. Lancet Neurol. 2023;22(2):159–171.
39. Al-Shahi Salman R, Frantzias J, Lee RJ, Lyden PD, Battey TWK, Ayres AM, et al. Absolute risk and predictors of the growth of acute spontaneous intracerebral haemorrhage: a systematic review and meta-analysis of individual patient data. Lancet Neurol. 2018;17(10):885–894. doi:10.1016/S1474-4422(18)30253-9.
40. Dowlatshahi D, Brouwers HB, Demchuk AM, Hill MD, Aviv RI, Ufholz LA, et al. Predicting intracerebral hemorrhage growth with the spot sign: the effect of onset-to-scan time. Stroke. 2016;47(3):695–700. doi:10.1161/STROKEAHA.115.012012.
41. Morotti A, Dowlatshahi D, Boulouis G, Al-Ajlan F, Demchuk AM, Aviv RI, et al. Predicting intracerebral hemorrhage expansion with noncontrast computed tomography: the BAT score. Stroke. 2018;49(5):1163–1169. doi:10.1161/STROKEAHA.117.020138.

42. Boulouis G, Morotti A, Brouwers HB, Charidimou A, Jessel MJ, Auriel E, et al. Association between hypodensities detected by computed tomography and hematoma expansion in patients with intracerebral hemorrhage. JAMA Neurol. 2016;73(8):961–968. doi:10.1001/jamaneurol.2016.1218.
43. Morotti A, Arba F, Boulouis G, Charidimou A. Noncontrast CT markers of intracerebral hemorrhage expansion and poor outcome: a meta-analysis. Neurology. 2020;95(14):632–643. doi:10.1212/WNL.0000000000010660.
44. Jeong HG, Bang JS, Kim BJ, Bae HJ, Han MK. Hematoma Hounsfield units and expansion of intracerebral hemorrhage: a potential marker of hemostatic clot contraction. Int J Stroke. 2021;16(2):163–171. doi:10.1177/1747493019895703.
45. Wu Q, Huang L, Chen N, Ren S, Ye F, Zhao X, et al. Quantification of Hounsfield unit difference in heterogeneous hematoma predicts poor functional outcomes in acute intracerebral hemorrhage. Sci Rep. 2025;15(1):33007. doi:10.1038/s41598-025-18409-9.
46. Yu Z, Ma L, Zheng J, Guo R, Li M, Wang X, et al. Comparison of hematoma density heterogeneity and ultraearly hematoma growth in predicting hematoma expansion in patients with spontaneous intracerebral hemorrhage. J Neurol Sci. 2017;379:44–48. doi:10.1016/j.jns.2017.05.049.
47. Yu L, Zhao M, Lin Y, Zeng J, He Q, Zheng Y, et al. Noncontrast computed tomography markers associated with hematoma expansion: analysis of a multicenter retrospective study. Brain Sci. 2023;13(4):608. doi:10.3390/brainsci13040608.
48. Oge DD, Topcuoglu MA, Gocmen R, Arsava EM. The dynamics of hematoma surface regularity and hematoma expansion in acute intracerebral hemorrhage. J Clin Neurosci. 2020;74:160–163.
49. Chen J, Xu T, Chen Y, Wang X, Shen H, Zhang Q, et al. The patterns of morphological change during intracerebral hemorrhage expansion: a multicenter retrospective cohort study. Front Med. 2022;8:774632.
50. Fletcher-Sandersjöö A, Svedung Wettervik T, Tatter C, Tjerkaski J, Nelson DW, Maegele M, et al. Absolute contusion expansion is superior to relative expansion in predicting traumatic brain injury outcomes: a multicenter observational cohort study. J Neurotrauma. 2024;41(5–6):705–713. doi:10.1089/neu.2023.0274.
51. Smith PD, Shukla I, Azam F, Trautmann D, Gee E, Korb M, et al. Predictive factors for traumatic cerebral contusion volume, expansion, and outcomes. J Neurosurg. 2025;142(6):1616–1624. doi:10.3171/2024.8.JNS241051.
52. Kumari MM, Senthilkumar TPJ, Ravi Y. A study to evaluate prognostic factors and define a critical volume for early surgery in patients with bifrontal brain contusions. Indian J Neurosurg. 2024;13(1):44–52. doi:10.1055/s-0043-1774813.
53. Ratan R, Mohindra S, Tripathi M, Singla R, Chauhan R. Factors determining the requirement of surgical intervention and prognosis in cases of traumatic bifrontal contusions: a prospective observational study. Surg Neurol Int. 2023;14:438. doi:10.25259/SNI_754_2023.
54. Li Z, Li B, Qiu P, Jiang J. Surgical treatment of traumatic bifrontal contusions: when and how? World Neurosurg. 2016;93:261–269. doi:10.1016/j.wneu.2016.06.021.

55. Wong GKC, Ngai K, Poon WS, Zheng VZY, Yu C. Cognitive outcomes of patients with traumatic bifrontal contusions. Acta Neurochir Suppl. 2018;126:63–65. doi:10.1007/978-3-319-65798-1_14.
56. Lv Y, Luo Z, Jin X, Lu Z, Han P. A novel CT-based edema grading system combined with machine learning for precise prognostic prediction in traumatic brain injury. Eur J Med Res. 2025;30(1):1094. doi:10.1186/s40001-025-03441-2.
57. Chen D, Zhan L, Shi X, Zhu X, Xie J, Qiu J, et al. Severity prediction in patients with oedema in cerebral contusion using deep learning from computed tomography scans. Clin Radiol. 2026;92:107113. doi:10.1016/j.crad.2025.107113.
58. Bodanapally UK, Shanmuganathan K, Issa G, Dreizin D, Li G, Sudini K, et al. Dual-energy CT in hemorrhagic progression of cerebral contusion: overestimation of hematoma volumes on standard 120-kV images and rectification with virtual high-energy monochromatic images after contrast-enhanced whole-body imaging. AJNR Am J Neuroradiol. 2018;39(4):658–662. doi:10.3174/ajnr.A5558.
59. Jiang G, Liu X, Zhang Q, Wang T, Zhang C, Niu Z, Sun S, Sun H, Zhao Y, Wu L. An interpretable multimodal model for early prediction of delayed hematoma progression in frontal lobe contusion: a machine learning approach. Front Neurosci. 2026;20:1897735. doi:10.3389/fnins.2026.1897735.

Figures

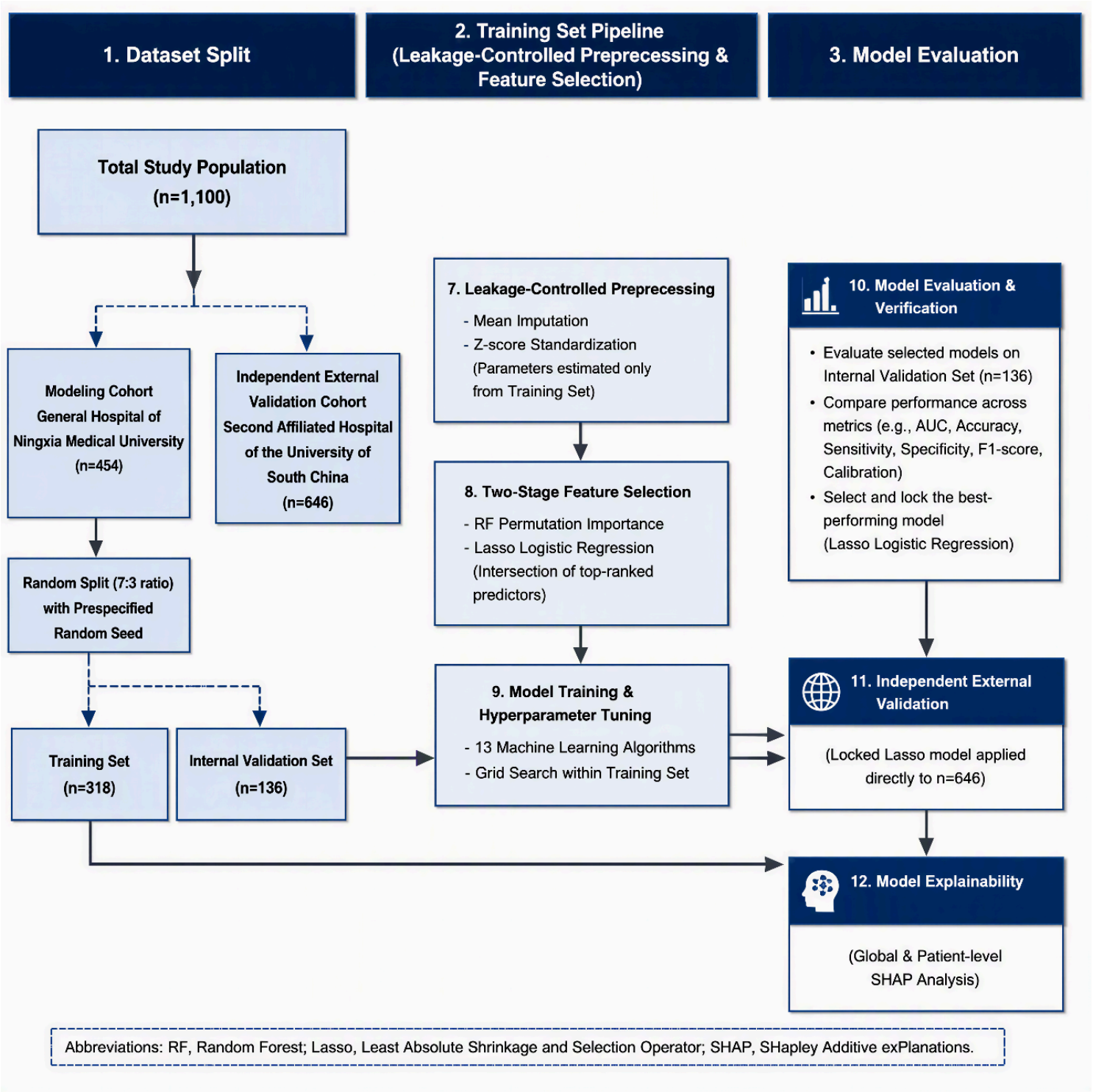


Figure 1

Study design, model-development pipeline, external validation, and model interpretation.

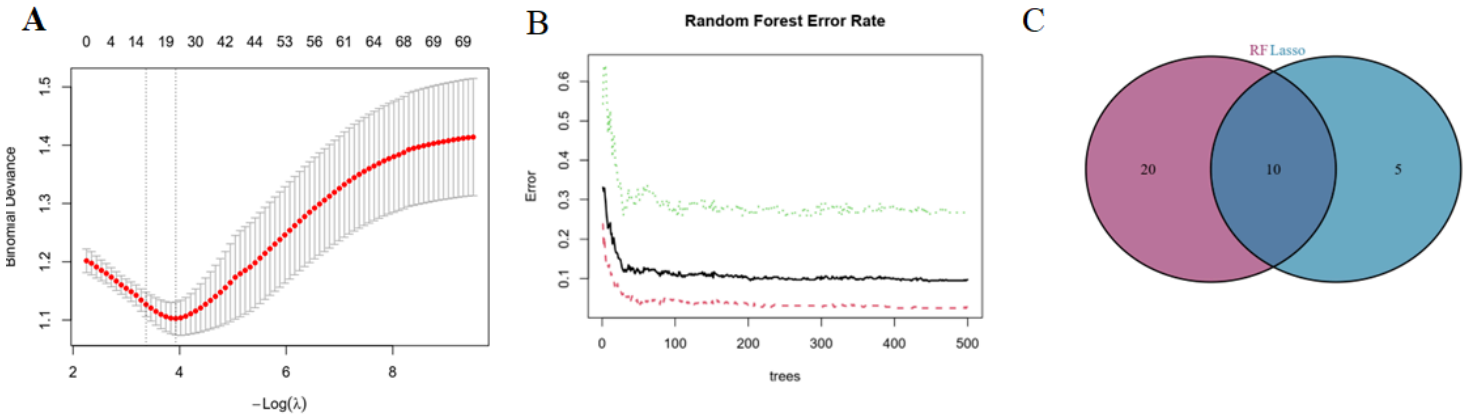


Figure 2

Two-stage feature selection. (A) Cross-validation curve for selection of the Lasso penalty parameter; (B) random forest error rate as a function of the number of trees; (C) intersection of predictors selected by random forest and Lasso regression.

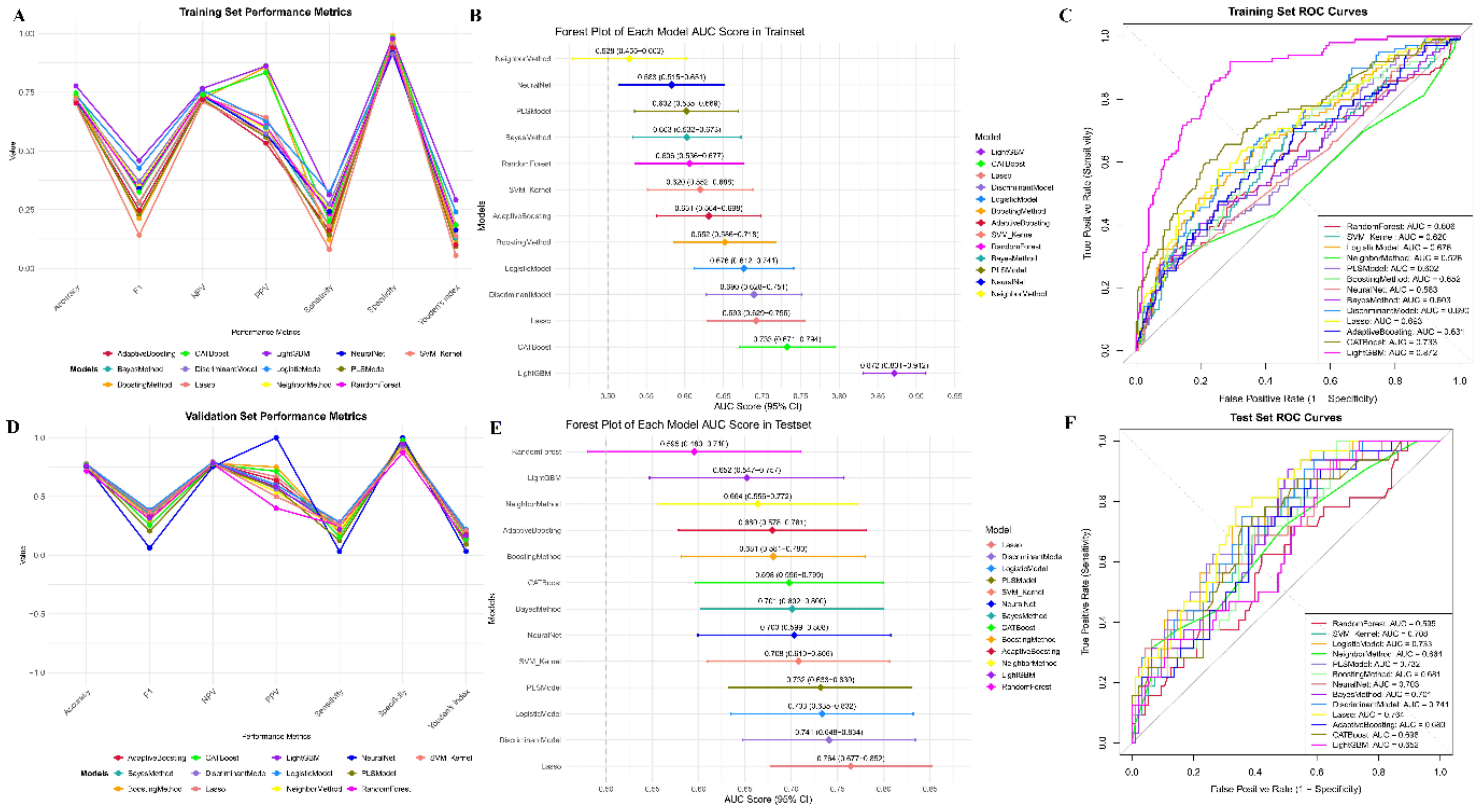


Figure 3

Comparative performance of the 13 candidate models in the training and internal validation sets. (A) Training-set classification metrics; (B) training-set AUC estimates with 95% confidence intervals; (C) training-set ROC curves; (D) internal-validation classification metrics; (E) internal-validation AUC estimates with 95% confidence intervals; (F) internal-validation ROC curves.

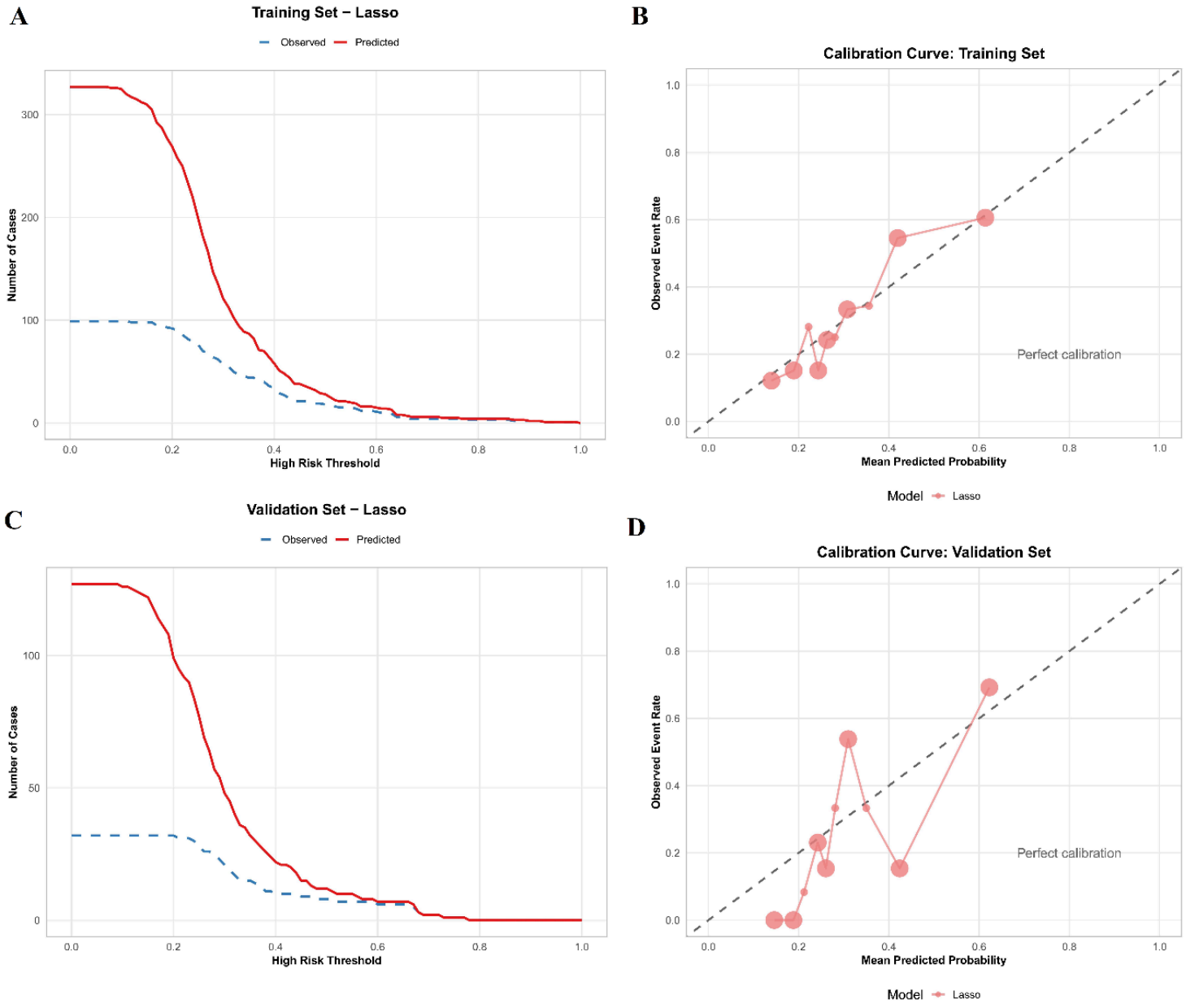


Figure 4

Clinical impact and calibration of the selected Lasso model. (A) Clinical impact curve in the training set; (B) calibration curve in the training set; (C) clinical impact curve in the internal validation set; (D) calibration curve in the internal validation set.

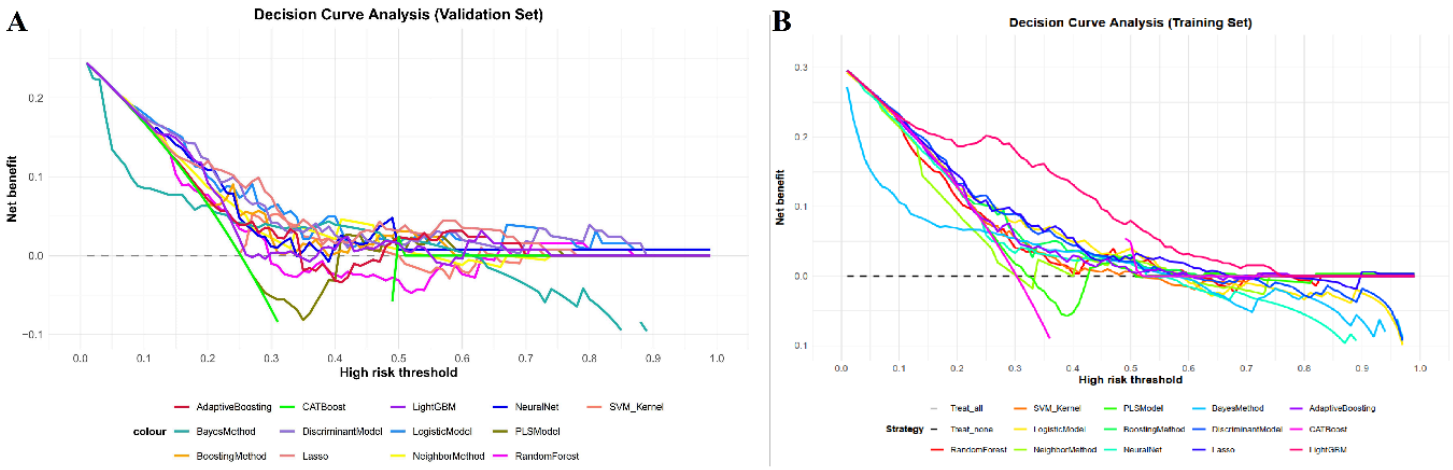


Figure 5

Decision curve analysis and confusion matrices in the training and internal validation sets. (A) decision curve analysis in the internal validation set ;(B) Decision curve analysis in the training set.

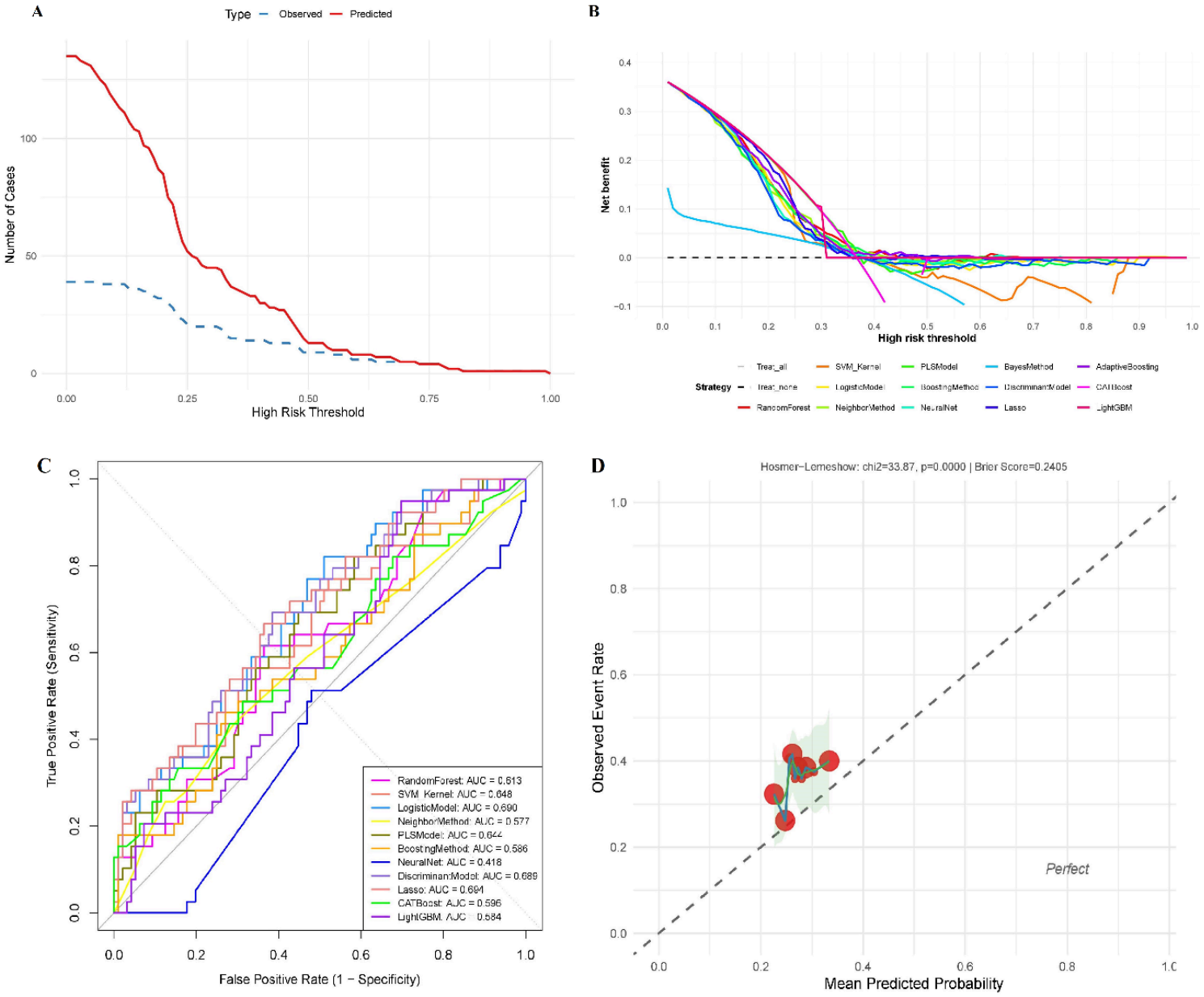


Figure 6

External validation of the candidate models and the selected Lasso model. (A) Clinical impact curve; (B) decision curve analysis; (C) ROC curves of the candidate models; (D) calibration curve of the selected Lasso model.

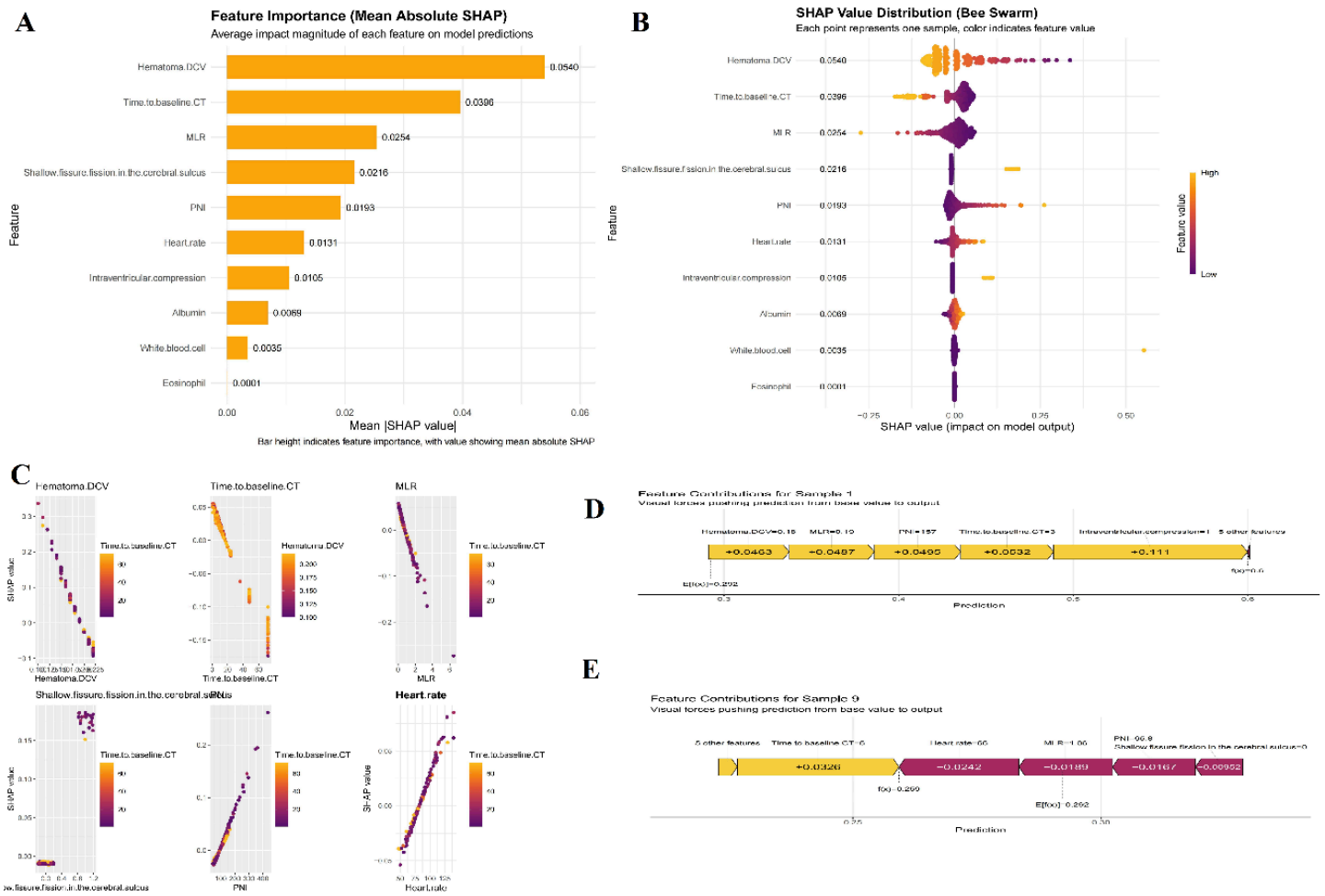


Figure 7

Global and patient-level SHAP interpretation of the selected Lasso model. (A) Feature importance based on mean absolute SHAP values; (B) SHAP beeswarm plot; (C) SHAP dependence plots for the leading predictors; (D) patient-level feature contributions for sample 1; (E) patient-level feature contributions for sample 9.

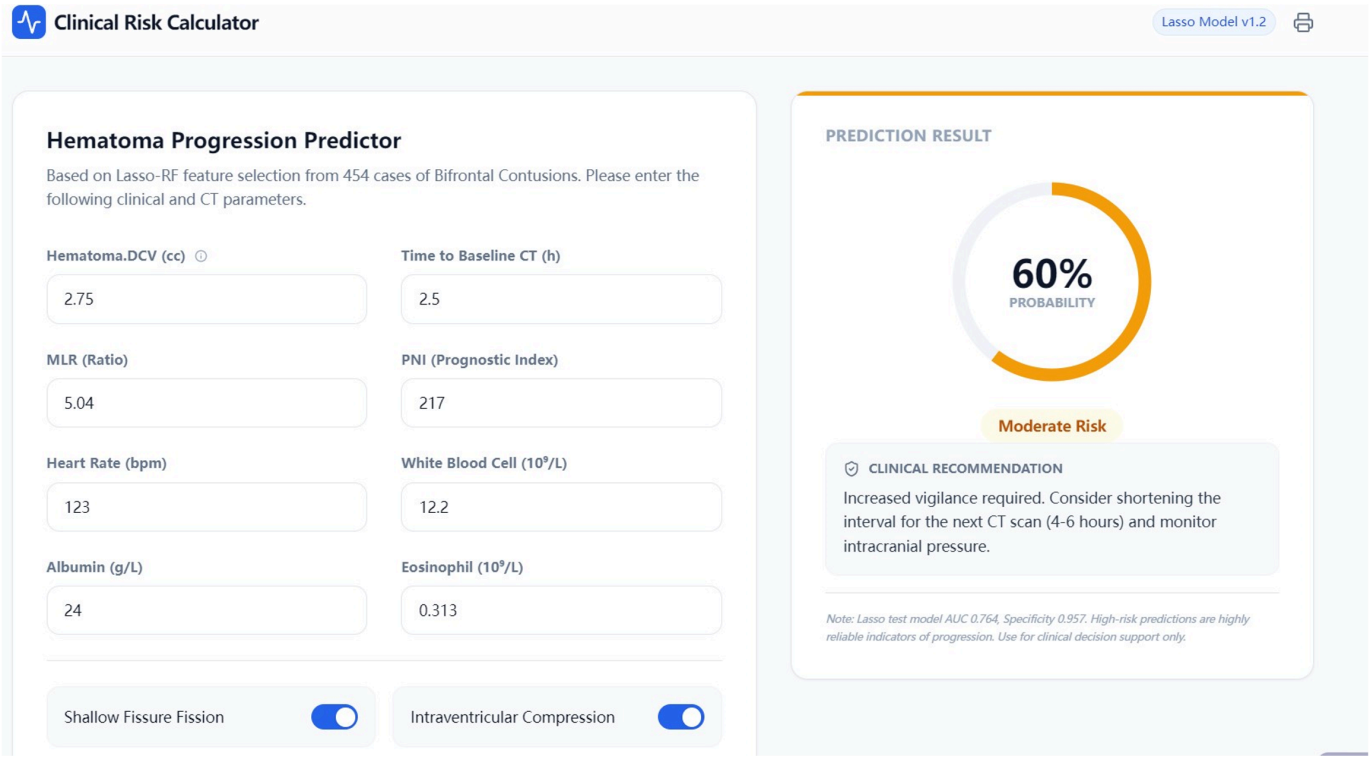


Figure 8

Prototype interface of the Lasso-based hematoma-progression risk calculator.