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.

Predicting In-Hospital Mortality in ICU Patients with Intracerebral Hemorrhage: A Machine Learning Study Using Dynamic Time-Series Features with Dual-Database Validation

Hanying Gu
Chengde Medical University
Xiuxia Shi
Chengde Central Hospital
Yunpeng Duan
Chengde Central Hospital
Jiangtao Zhang
zhangjiangtao@vip.163.com

Chengde Central Hospital

Article

Keywords: intracerebral hemorrhage, intensive care unit, in-hospital mortality, dynamic prediction, machine learning, time series

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10407148/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

ICH has high ICU mortality, yet conventional scores rely on static data and miss early dynamics. We developed and validated machine learning models using 24-hour time-series features to predict in-hospital mortality.

Methods

This TRIPOD + AI retrospective study used MIMIC-IV for development (n = 1,962) and internal validation (n = 842), and eICU-CRD for external validation (n = 3,470). The first 24 hours were divided into four 6-hour windows; statistical features of eight physiological parameters were extracted per window. MICE and SMOTE addressed missing data and class imbalance. LASSO selected 29 features for four models (LR, RF, XGBoost, LightGBM), benchmarked against a GCS-only baseline. Performance was assessed by AUC, calibration, DCA, and SHAP.

Results

Mortality was 21.3%, 21.5%, and 17.3% across cohorts. Twenty-two of 29 features (76%) were dynamic. LightGBM achieved the best performance: internal AUC 0.823 and external AUC 0.813, significantly exceeding the GCS-only baseline (0.778). Calibration was satisfactory; DCA demonstrated net benefit. SHAP identified GCS and BUN as top predictors, with dynamic vital sign features contributing substantially.

Conclusions

Dynamic physiological trajectories encode prognostic information beyond static snapshots, underscoring the value of continuous ICU monitoring. Our study offers an interpretable, imaging-free tool for early risk stratification, with potential for integration into clinical systems to guide triage and surveillance in high-risk patients.

1. Introduction

Spontaneous intracerebral hemorrhage (ICH) accounts for approximately 10%–15% of all strokes and carries a 30-day mortality rate of 40%–50%, making it one of the leading causes of stroke-related death worldwide [1–4]. ICH patients are critically ill and typically require close monitoring and management in the intensive care unit (ICU). Early and accurate prognostic assessment is essential for formulating individualized treatment strategies, allocating medical resources appropriately, and facilitating effective communication with patients' families—all of which are pivotal to improving patient outcomes [2].

Several ICH prognostic scoring systems are currently used in clinical practice, of which the ICH Score and the FUNC Score are the most widely adopted [5, 6]. Although these tools are straightforward to apply, they rely primarily on clinical and neuroimaging features obtained at admission—such as Glasgow Coma Scale (GCS) score, hematoma volume, and hemorrhage location—and their discriminative performance is limited (AUC typically 0.70–0.80). More importantly, ICH is a dynamically evolving pathological process: secondary brain injury, including hematoma expansion, cerebral edema, and neuroinflammation, progresses continuously over hours to days following onset and is the principal driver of clinical deterioration [7]. Static models based solely on admission "snapshots" are fundamentally incapable of capturing this dynamic trajectory.

With the widespread adoption of electronic health records (EHRs), ICUs now generate vast quantities of high-resolution continuous physiological monitoring data. These time-series signals encode rich information about patients' physiological stability and trends [8, 9]. Blood pressure variability (BPV), for example, reflects autonomic dysfunction and has been established as an independent predictor of in-hospital mortality in ICH patients [10–12]. Yet such dynamic information is almost entirely overlooked by conventional scoring systems. Fully leveraging early ICU time-series data holds the potential to fundamentally overcome the predictive ceiling of static models [9].

Machine learning (ML), particularly gradient boosting ensemble algorithms such as XGBoost and LightGBM [13, 14], offers substantial advantages in handling high-dimensional, nonlinear clinical data. These methods have been applied to ICH prognostication and have demonstrated superior performance over traditional scoring systems [15–18]. However, two major limitations persist in the existing literature. First, most studies continue to rely on static or aggregated admission-time variables, failing to exploit the predictive value of early dynamic physiological data from the ICU. Second, the majority of models have been evaluated only on internal test sets from a single center or database, leaving their generalizability across diverse clinical settings uncertain.

To address these limitations, we hypothesized that integrating dynamic time-series physiological features from the first 24 hours of ICU admission would yield a model with superior predictive performance compared with conventional static approaches. This study makes three key contributions. First, we systematically extract 24-hour time-series features from eight physiological parameters across four 6-hour windows, capturing dynamic changes overlooked by conventional static models. Second, we provide large-scale external validation using an independent multicenter database (eICU-CRD, n = 3,470), demonstrating model generalizability. Third, we apply SHAP analysis to ensure interpretability, revealing that dynamic features contribute substantially to prediction beyond static variables. Accordingly, we developed and rigorously externally validated a series of machine learning models using two large, heterogeneous ICU databases: MIMIC-IV (v2.2) and the eICU Collaborative Research Database (v2.0).

2. Methods

2.1 Study Design and Data Sources

This was a multicenter retrospective cohort study based on two publicly available large-scale critical care databases, reported in accordance with the TRIPOD + AI statement [19]. The MIMIC-IV database (v2.2) contains de-identified health data from patients admitted to the ICUs of Beth Israel Deaconess Medical Center in Boston, USA, between 2008 and 2019, encompassing demographics, vital signs, laboratory results, and diagnostic codes; it was used for model development and internal validation [20]. The eICU Collaborative Research Database (v2.0) is a multicenter database comprising ICU patient data from 208 ICUs across the United States collected between 2014 and 2015, and was used as the external validation cohort [21]. As this study used de-identified publicly available data, it was exempt from institutional review board approval by the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center. All investigators held current CITI Program certification in human subjects research.

2.2 Study Population

Inclusion criteria were: (1) age ≥ 18 years; (2) a confirmed diagnosis of spontaneous ICH—identified in MIMIC-IV using ICD-9 code 431 or ICD-10 codes I61.x, and in eICU using diagnosis description fields containing the terms "intracerebral hemorrhage," "cerebral hemorrhage," or "hemorrhagic stroke"; subarachnoid hemorrhage and subdural/epidural hematoma were excluded from both databases; and (3) ICU length of stay ≥ 24 hours. Exclusion criteria were: (1) traumatic brain hemorrhage; (2) missing critical diagnostic information; (3) death or ICU discharge within 24 hours of admission, precluding complete observation; and (4) missingness exceeding 60% for core physiological variables during the first 24 hours.

2.3 Data Extraction and Feature Engineering

Data were extracted from the first 24 hours following ICU admission and categorized into static and dynamic time-series variables. Static variables included: demographic characteristics (age, sex); comorbidity burden (Elixhauser Comorbidity Index); admission physiological scores (first-recorded SOFA and SAPS II scores); first-available laboratory values (white blood cell count, platelets, hemoglobin, INR, PTT, sodium, potassium, chloride, bicarbonate, creatinine, glucose, and albumin); and treatment variables (use of mechanical ventilation and vasopressors). Dynamic time-series variables were constructed as follows. The first 24 hours were divided into four consecutive 6-hour time windows (W1: 0–6 h, W2: 6–12 h, W3: 12–18 h, W4: 18–24 h). Eight high-frequency physiological parameters were extracted: systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), heart rate (HR), body temperature (Temp), Glasgow Coma Scale (GCS) score, peripheral oxygen saturation (SpO_2), and respiratory rate (RR). Within each time window, seven statistical features were computed for each parameter: mean, standard deviation (SD), coefficient of variation (CV), minimum, maximum, range, and linear slope. Twenty-four-hour aggregate statistics were also calculated for each parameter. Additionally, five blood pressure variability (BPV) indices were derived specifically for SBP:

average real variability (ARV, defined as the mean absolute difference between consecutive measurements), SD, CV, maximum value, and time in target range (TTR, defined as the proportion of SBP readings within 140–180 mmHg). This feature engineering pipeline generated approximately 283 candidate features in total. Physiological measurements were recorded at varying frequencies across the two databases (hourly in MIMIC-IV and every 15–60 minutes in eICU-CRD). For each 6-hour window, we adopted a left-closed, right-open interval definition (e.g., W1: [0h, 6h)). Missing values within a window were linearly interpolated before feature extraction; variables with > 60% missingness were excluded.

2.4 Outcome Definition

The primary outcome was in-hospital mortality, defined as death from any cause during the index hospitalization.

2.5 Missing Data Handling

Variables with a missingness rate exceeding 60% were excluded. For retained variables, missing values were imputed using multiple imputation by chained equations (MICE; m = 1, method=predictive mean matching, maxit = 5, seed = 42) [22]. Residual missing values after MICE were replaced with the training-set median. All imputation parameters were fitted on the training set and applied to the internal validation and external test sets to prevent data leakage. Given the relatively low missing rates for most included variables and the fact that imputation was performed solely for model input, single imputation (m = 1) with predictive mean matching was considered sufficient to ensure stable estimation while preserving computational efficiency.

2.6 Data Partitioning and Class Imbalance Handling

The MIMIC-IV cohort was randomly split in a 7:3 ratio into a training set (n = 1,962) and an internal validation set (n = 842) (set.seed = 2024). The entire eICU cohort (n = 3,470) was used as an independent external validation set. To address class imbalance in the training set, the Synthetic Minority Oversampling Technique (SMOTE; over_ratio = 1, seed = 42) was applied exclusively to the training set [23], yielding a 1:1 ratio of positive to negative cases. No oversampling was performed on the internal validation or external test sets.

2.7 Feature Selection

LASSO logistic regression with L1 regularization was applied to the training set for feature selection [24]. Features were standardized prior to model fitting. The regularization parameter λ was determined by 10-fold cross-validation (type.measure="auc", seed = 42), with the constraint that no more than 30 features were retained; the λ value maximizing cross-validated AUC under this constraint was selected, yielding 29 predictive features.

2.8 Model Development

Four machine learning models were constructed based on the 29 selected features:(1) Logistic Regression (LR): Standardized features were used, with zero-variance and highly collinear features (Pearson r > 0.95) removed prior to fitting. Platt scaling was applied on the internal validation set to improve probability calibration. (2) Random Forest (RF) [25]: ntree = 500, mtry=floor(√p), seed = 42. (3) XGBoost: eta = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, min_child_weight = 5, maximum rounds = 500, early stopping rounds = 30. (4) LightGBM: learning_rate = 0.05, num_leaves = 31, feature_fraction = 0.8, bagging_fraction = 0.8, min_data_in_leaf = 20, maximum rounds = 500, early stopping rounds = 30. RF, XGBoost, and LightGBM were trained on unstandardized features. A GCS-only logistic regression model—using the first recorded GCS score as the sole predictor (entered as a negative value to reflect the clinical relationship between lower GCS and higher mortality risk)—was constructed as the reference baseline. Hyperparameters and tuning strategies for all models are detailed in Supplementary Table 2. Hyperparameters for all tree-based models were optimized via grid search combined with 5-fold cross-validation on the training set, with the objective of maximizing the validation AUC. Early stopping with a patience of 30 rounds was applied to prevent overfitting. A multilayer perceptron (MLP; hidden units = 10, weight decay = 0.01, maxit = 500, implemented via the nnet package) was also constructed for reference; its hyperparameter configurations are provided in Supplementary Table 2. As MLP performance was comparable to but not superior to the gradient boosting models, it was not included in the main analysis.

2.9 Model Evaluation

Discrimination: The primary performance metric was the AUC, with 95% CIs estimated by bootstrap resampling (1,000 iterations, seed = 42). The optimal decision threshold was determined by the Youden index, and sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Brier score were calculated accordingly. Calibration:

Calibration curves were constructed using 15 equal-frequency quantile bins, comparing predicted probabilities against observed mortality rates to assess systematic over- or underestimation. Clinical utility: Decision curve analysis (DCA) was used to evaluate the clinical net benefit of the LightGBM model on the external validation set across threshold probabilities of 0–0.5, benchmarked against "treat-all" and "treat-none" reference strategies [26].

2.10 Model Interpretability

For the best-performing LightGBM model, SHAP values were computed on the eICU external test set using the fastshap package (nsim = 50, seed = 42) [27]. The top 15 features ranked by mean absolute SHAP value were visualized in a summary plot displaying global feature importance and the direction of each feature's contribution to predicted outcomes. SHAP dependence plots were generated for key features to reveal nonlinear relationships between feature values and mortality risk.

2.11 Statistical Analysis

All data extraction and preprocessing were performed using SQL (PostgreSQL v12.0). Statistical analyses and model development were conducted in R (v4.5.3), using the glmnet, xgboost, lightgbm, mice, themis, pROC, fastshap, and dcurves packages. Continuous variables are reported as median (interquartile range) or mean ± standard deviation; between-group comparisons were performed using the Mann–Whitney U test or independent-samples t-test. Categorical variables are reported as frequency (percentage) and compared using the chi-squared test or Fisher's exact test. All statistical tests were two-sided, with $p < 0.05$ considered statistically significant.

3. Results

3.1PatientCohortsandBaselineCharacteristics

After applying inclusion and exclusion criteria, 6,274 patients were enrolled: 2,804 from MIMIC-IV (training set: 1,962; internal validation set: 842) and 3,470 from eICU as the external validation cohort (Fig.
1). Baseline demographic, clinical, and laboratory characteristics across the three cohorts are presented in Table 1.

Fig. 1. Patient Selection and Study Design Flowchart.

Patients with intracerebral hemorrhage were identified from MIMIC-IV (derivation) and eICU-CRD (external validation). Exclusion criteria were applied sequentially. The derivation cohort was split 70:30 for training and internal validation.

Abbreviations: ICH, intracerebral hemorrhage; MIMIC-IV, Medical Information Mart for Intensive Care IV; eICU-CRD, eICU Collaborative Research Database.

Table 1. Baseline characteristics of ICH patients across three cohorts.

Values are median (Q1, Q3) or n (%). Between-group comparisons used the Kruskal–Wallis test for continuous variables and chi-squared or Fisher's exact test for categorical variables. "Unknown" denotes missing values and was excluded from percentage calculations. Race/ethnicity classification differs between MIMIC-IV and eICU-CRD.

Abbreviations: BP, blood pressure; MAP, mean arterial pressure; SpO_2, peripheral oxygen saturation; WBC, white blood cell count; BUN, blood urea nitrogen; INR, international normalized ratio; GCS, Glasgow Coma Scale; SD, standard deviation; Q1, first quartile; Q3, third quartile.

Overall, the median patient age was 66.0 years (IQR 54.0–78.0), and 53% were male (3,350/6,274). Patients in the MIMIC-IV cohort were slightly older than those in the eICU cohort (median 68.0 vs. 65.0 years, $p<0.001$). Regarding admission GCS score, the median GCS in the MIMIC-IV training set was 12.0 (IQR 7.0–15.0), significantly lower than the eICU cohort median of 14.0 (IQR 9.0–15.0; $p<0.001$), suggesting a relatively better baseline neurological status in the eICU population. Significant between-cohort differences were also observed in SBP ($p<0.001$) and racial/ethnic composition ($p<0.001$), whereas HR ($p=0.427$) and MAP ($p=0.355$) did not differ significantly across the three groups. Regarding the primary outcome, in-hospital mortality was 21.3% (417/1,962) in the MIMIC-IV training set, 21.5% (181/842) in the internal validation set, and 17.3% (600/3,470) in the eICU external validation set ($p<0.001$).

3.2DynamicPhysiologicalTrajectories

Fig. 2 illustrates the dynamic trajectories of HR, SBP, MAP, SpO_2, RR, and GCS score across four 6-hour time windows in survivors versus non-survivors from both databases (mean ± 95% CI). Non-survivors exhibited greater instability in physiological parameters throughout the 24-hour observation period, particularly in blood pressure and respiratory rate. Between-group differences were already apparent in the earliest time window (W1: 0–6 h) and persisted over time, providing direct visual support for the prognostic value of dynamic time-series features.

Fig. 2. Temporal Trajectories of Vital Signs and Neurological Status by Outcome.

Vital signs were segmented into four consecutive 6-hour windows (W1–W4) within the first 24 hours of ICU admission. Lines represent group means; shaded bands indicate 95% CIs. GCS scores are shown as violin plots with embedded box plots. Significant between-group differences at each window are indicated by asterisks (*P < .05, **P < .01, ***P < .001).

Abbreviations: GCS, Glasgow Coma Scale; HR, heart rate; MAP, mean arterial pressure; RR, respiratory rate; SBP, systolic blood pressure; SpO_2, peripheral oxygen saturation; W1–W4, time windows 1–4 (0–6, 6–12, 12–18, 18–24 h).

3.3FeatureSelection

LASSO regression selected 29 features from approximately 283 candidates (Supplementary Fig. 1). Seven were static features: GCS total score and six laboratory values (BUN, glucose, potassium, INR, WBC, and platelets). The remaining 22 were dynamic time-series features, encompassing variability, extremes, and trend indices of SBP, DBP, HR, RR, body temperature, and MAP across time windows (e.g., SBP SD in W4, SBP Max in W2, SBP Mean/Range in W1, DBP Mean in W1 and W4, DBP CV in W4, HR Mean/Max in W2, HR Max in W3, MAP Mean in W1 and W2, MAP Slope in W3, RR Mean/Min in W3, RR Min in W4) as well as 24-hour aggregate statistics (MAP minimum, MAP slope, RR minimum, and temperature SD). Missing rates for the 29 selected features across the three cohorts are provided in Supplementary Table 1. Laboratory values had missing rates of 5%–15% in MIMIC-IV and 8%–51% in eICU. Two temperature-related dynamic features had extremely high missingness in the eICU cohort (Temp SD 24h: 88.2%; Temp Range W4: 89.6%), reflecting substantial differences in temperature recording practices between the two databases; residual missing values were imputed using training-set medians.

3.4ModelDiscrimination

Complete performance metrics for all four ML models and the GCS-only baseline in both validation cohorts are presented in Table 2, with ROC curves shown in Fig. 3.

Table 2. Predictive performance of machine learning models across validation cohorts.

95% CIs estimated by 1,000-iteration bootstrap resampling. Optimal cutoff determined by the Youden index. GCS-only: logistic regression using GCS total score as the sole predictor; Brier score not reported. DeLong test: LightGBM vs. GCS-only — internal validation p = 0.0002, external validation p < 0.0001.

Abbreviations: AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LR, logistic regression; RF, random forest; GCS, Glasgow Coma Scale; CI, confidence interval.

Fig. 3. Receiver Operating Characteristic Curves for Predicting In-Hospital Mortality.

ROC curves are shown for four machine learning models and GCS-alone in (A) internal validation (MIMIC-IV, n = 842) and (B) external validation (eICU-CRD, n = 3470). AUC values with 95% CIs are displayed in the legend. Dashed line indicates chance-level discrimination.

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; GCS, Glasgow Coma Scale; LightGBM, Light Gradient Boosting Machine; LR, logistic regression; RF, random forest; ROC, receiver operating characteristic; XGB, XGBoost.

In the internal validation set (MIMIC-IV, n=842), all four ML models substantially outperformed the GCS-only baseline (AUC=0.770, 95% CI 0.731–0.805). XGBoost and LightGBM achieved the highest internal validation AUC (both 0.823; XGBoost: 95% CI 0.789–0.854; LightGBM: 95% CI 0.789–0.854), followed by LR (0.813, 95% CI 0.779–0.844) and RF (0.813, 95% CI 0.779–0.844). In the external validation set (eICU, n=3,470), LightGBM achieved the highest AUC (0.813, 95% CI 0.795–0.831), followed by XGBoost (0.811, 95% CI 0.792–0.830). RF (0.810, 95% CI 0.792–0.828) and LR (0.805, 95% CI 0.785–0.823) ranked third and fourth, respectively. All ML models significantly outperformed the GCS-only baseline (AUC=0.778, 95% CI 0.757–0.800). Compared with internal validation, AUC values declined modestly on external validation (LightGBM: Δ=0.010), indicating good cross-database generalizability. At the optimal Youden-index threshold, LightGBM achieved a cutoff of 0.221 on the external validation set, corresponding to sensitivity 79.0% (95% CI 75.9–82.3%), specificity 72.5% (95% CI 70.9–74.1%), accuracy 73.6% (95% CI 72.1–75.0%), PPV 37.5% (95% CI 34.7–40.2%), NPV 94.3% (95% CI 93.3–95.2%), F1 score 0.509 (95% CI 0.479–0.537), and Brier score 0.121 (95% CI 0.115–0.128).

3.5ModelCalibration

Fig. 4 presents calibration curves and decision curve analysis for all four ML models in both validation cohorts. Calibration curves (15 equal-frequency quantile bins, upper panels) for all models remained close to the ideal diagonal (y=x) across both cohorts, with no evidence of systematic over- or underestimation, indicating satisfactory calibration. On external validation, LR achieved the lowest Brier score among the four ML models (0.116); however, its calibration curve showed systematic overestimation at high predicted probability ranges. LightGBM had a slightly higher Brier score (0.121) but demonstrated better calibration across the clinically relevant low-to-moderate probability range. The predicted probability distribution of the LightGBM model on the external validation set is shown in Supplementary Fig. 2.

Fig. 4. Model Calibration and Decision Curve Analysis.

(A, B) Calibration curves compare predicted versus observed mortality probabilities in internal and external validation cohorts; perfect calibration is indicated by the diagonal line. Brier scores are shown for each model. (C, D) Decision curve analysis displays net benefit across a range of threshold probabilities. The "treat all" and "treat none" strategies serve as reference lines.

Abbreviations: DCA, decision curve analysis; LightGBM, Light Gradient Boosting Machine; LR, logistic regression; RF, random forest; XGB, XGBoost.

3.6ClinicalUtility

The lower panels of Fig. 4 present the DCA results for the LightGBM model on the eICU external validation set. Across threshold probabilities of 0–0.5, the net benefit curve of LightGBM exceeded the "treat-all" strategy at the vast majority of thresholds and consistently surpassed the "treat-none" strategy (net benefit=0). This indicates that using the LightGBM model to guide clinical decisions yields greater net benefit than either universal intervention or universal non-intervention across a broad range of clinically relevant risk thresholds.

3.7 Incremental Value of Time-Series Features

To formally quantify the incremental value of dynamic features, we compared a static LightGBM model (trained on the 7 laboratory/clinical features only) against the full model (29 features; 7 static + 22 time-series), with results presented in Table 3. On internal validation, the full model achieved an AUC of 0.823 (95% CI 0.789–0.854) versus 0.817 (95% CI 0.785–0.849) for the static model; the difference was not statistically significant by DeLong test (p=0.523). On external validation, the full model achieved an AUC of 0.813 (95% CI 0.795–0.831) versus 0.808 (95% CI 0.790–0.826) for the static model (DeLong p=0.240). Despite comparable AUC values, the addition of time-series features yielded significant improvements in reclassification and discrimination: internal validation NRI = 0.586 (95% CI 0.417–0.733, p<0.001) and IDI = 0.050 (95% CI 0.025–0.074, p<0.001); external validation NRI = 0.539 (95% CI 0.449–0.623, p<0.001) and IDI = 0.050 (95% CI 0.039–0.061, p<0.001). These results indicate that dynamic time-series features meaningfully improve patient-level risk stratification beyond static variables alone, even when the overall AUC difference is modest.

Table 3. Incremental value of time-series features: static vs. full LightGBM model.

95% CIs estimated by 1,000-iteration bootstrap resampling. Static model: LightGBM trained on 7 laboratory/clinical features only. Full model: LightGBM trained on 29 features (7 static + 22 time-series). NRI and IDI were calculated using the nricens package; DeLong test used for AUC comparison.

Abbreviations: AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; NRI, net reclassification improvement; IDI, integrated discrimination improvement; CI, confidence interval.

3.8 ModelInterpretability

Fig. 5 presents the SHAP feature importance analysis for the LightGBM model on the eICU external test set, displaying the top 15 features ranked by mean absolute SHAP value. Fig. 6 further presents SHAP dependence plots for key features, revealing nonlinear relationships between feature values and mortality risk. SHAP analysis demonstrated that both static and dynamic time-series features contributed substantially to model predictions. Among static features, GCS total score and laboratory values including BUN, platelets, and glucose exhibited high feature importance. Among dynamic features, SBP Range (W1), SBP Max (W2), SBP Mean (W1), DBP Mean (W1), DBP CV (W4), multiple RR statistics across time windows (mean, minimum, and 24-hour aggregate), MAP minimum (24h), and temperature-related aggregate features also ranked prominently in the top 15. Nine of the 15 top-ranked features (60%) were dynamic time-series variables, underscoring the contribution of temporal physiological information to model predictions.

Fig. 5. SHAP Analysis of the Top 15 Features in the LightGBM Model.

(A) Mean absolute SHAP values for the top 15 features, ranked by overall contribution to model output. (B) Beeswarm plot showing the distribution of SHAP values for each feature across all patients in the external validation cohort. Each dot represents one patient; color indicates feature value (yellow = high, purple = low). Positive SHAP values indicate increased predicted mortality risk; negative values indicate decreased risk. GCS Total was the dominant predictor, with lower scores associated with higher mortality risk.

Abbreviations: BUN, blood urea nitrogen; CV, coefficient of variation; DBP, diastolic blood pressure; GCS, Glasgow Coma Scale; INR, international normalized ratio; MAP, mean arterial pressure; RR, respiratory rate; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; WBC, white blood cell count; W1–W4, time windows 1–4 (0–6, 6–12, 12–18, 18–24 h).

Fig. 6. SHAP Dependence Plots for the Top 15 Predictors of In-Hospital Mortality.

Each panel displays the relationship between feature value (x-axis) and its SHAP value (y-axis) for the LightGBM model in the external validation cohort. Positive SHAP values indicate increased mortality risk. Color encodes a secondary interacting feature where applicable. Features are ordered by mean |SHAP| value.

Abbreviations: BUN, blood urea nitrogen; DBP, diastolic blood pressure; GCS, Glasgow Coma Scale; INR, international normalized ratio; MAP, mean arterial pressure; RR, respiratory rate; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; WBC, white blood cell count; W1–W4, time windows 1–4.

4. Discussion

This study developed and externally validated four ML models for predicting in-hospital mortality in ICU patients with ICH using a dual-database design. LightGBM achieved the best external validation performance (AUC 0.813), significantly outperforming the GCS-only baseline (AUC 0.778), with 22 of 29 selected features being dynamic time-series variables. At the optimal threshold of 0.221, the NPV of 94.3% indicates that over 94% of patients classified as low-risk were true survivors, supporting the model's utility in prioritizing limited ICU resources toward genuinely high-risk patients and reducing unnecessary over-treatment. DCA confirmed meaningful net benefit across threshold probabilities of 0.03–0.42, covering the majority of clinical risk-stratification scenarios.

Conventional ICH scoring systems,such as the ICH Score (AUC ~0.84), FUNC Score (AUC ~0.74), and APACHE II (AUC ~0.75–0.80), rely on static admission variables and require neuroimaging [5, 6, 28, 29]. Our model achieves comparable or superior discrimination (external AUC 0.813) using only routinely collected ICU vital signs and laboratory data, without neuroimaging. Critically, all input variables are available from standard monitoring workflows, making the model amenable to automated integration into clinical

information systems (HIS/CIS) without additional clinical burden, an advantage particularly relevant for resource-limited settings.Among recent ML studies, Nie et al. reported an AUC of ~0.85 in MIMIC but lacked external validation [17]. Mao et al. achieved ~0.80 [18]. Huang et al. employed a multicenter design with cross-validation analogous to ours [15]. Miao et al. predicted ICU readmission risk in ICH patients using dual MIMIC databases [30]. Chen et al. developed a CT-based deep learning model for ICH outcome prediction, complementary to our imaging-free approach [31]. Compared with these studies, our principal strengths are the large-scale external validation cohort (eICU-CRD, n=3,470, 208 ICUs across the United States), systematic time-window feature engineering, SHAP interpretability [16, 32], and TRIPOD-compliant reporting.

Trajectory analysis (Fig. 2) revealed that survivors and non-survivors diverged in HR, SBP, MAP, and RR as early as W1 (0–6 h), with differences sustained across all four time windows and consistent between MIMIC-IV and eICU cohorts. This early divergence suggests that vital sign dynamics within the first six hours of ICU admission can serve as actionable warning signals: clinicians observing increasing blood pressure variability, persistently elevated RR, or progressive GCS decline should consider escalating monitoring intensity and initiating multidisciplinary consultation. SHAP analysis identified 9 time-series features among the top 15 predictors. RR Min (W4) ranked seventh (mean |SHAP|=0.021), highlighting the prognostic importance of late-window respiratory status and the need for ongoing ventilatory reassessment. DBP CV (W4) and DBP Mean (W1) reflect blood pressure variability and hypotensive episodes consistent with secondary brain injury pathophysiology [7], while SBP Range (W1) and SBP Max (W2) capture early pressure fluctuation relevant to established BP management targets [10, 11, 33, 34]. These findings suggest that controlling blood pressure variability—not merely its absolute level—represents an important therapeutic target in ICH.

GCS total score (mean |SHAP|=0.173) was the dominant predictor, with mortality risk rising sharply below GCS 8 [5]. For patients with GCS <8, airway protection, intracranial pressure monitoring, and neurosurgical consultation should be prioritized; for borderline scores (8–12), the model's integration of time-series features enables more granular risk stratification to guide monitoring escalation decisions. BUN (mean |SHAP|=0.036) reflects renal impairment and systemic inflammation [35], likely mediated through brain–kidney axis interactions in ICH. Persistently elevated BUN warrants fluid management optimization, early consideration of renal replacement therapy, and vigilance for multi-organ dysfunction. Platelets (mean |SHAP|=0.035), WBC (mean |SHAP|=0.028), and glucose (mean |SHAP|=0.026) highlight the roles of coagulopathy, systemic inflammation, and stress hyperglycemia [7]. These are routinely available parameters, and their prognostic weight supports active correction of thrombocytopenia, glycemic control targeting 7.8–10 mmol/L, and infection surveillance as potentially modifiable interventions.

Although LR achieved a marginally lower Brier score (0.116 vs. 0.121), its calibration curve showed systematic overestimation at high predicted probabilities, whereas LightGBM was better calibrated in the clinically relevant low-to-moderate range. DCA confirmed LightGBM's net benefit advantage across threshold probabilities of 0.03–0.42. For clinical translation, the model can be integrated into the HIS to automatically compute risk scores from the first 24 hours of monitoring data. Patients may be stratified into low-risk (<0.15), intermediate-risk (0.15–0.35), and high-risk (>0.35) tiers, corresponding to standard care, intensified surveillance with multidisciplinary review, and active intervention with family communication, respectively. Concurrent SHAP output identifies the primary drivers of each patient's risk score, enabling a "predict–explain–intervene" workflow that supports targeted clinical decision-making. This model is intended as a decision support tool; final clinical judgment remains with the treating physician, and prospective validation is required before formal implementation.

Our SHAP analysis further quantified the contribution of dynamic features: among the top 15 predictors, 9 (60%) were time-series features, including RR Min (W4), HR Max (W3), DBP CV (W4), and SBP Range (W1). Notably, the 18–24h window features appeared prominently, suggesting that physiological trajectories later in the first day carry independent prognostic information beyond early-window measurements. This finding supports the clinical practice of continued physiological monitoring beyond the initial resuscitation phase and suggests that deterioration patterns, rather than single measurements, may be more informative for risk stratification.

This study has several limitations. First, data were drawn exclusively from US ICU databases; generalizability to other populations, particularly Asian cohorts, requires prospective external validation. Second, neuroimaging features (hematoma volume, location) were not incorporated due to the difficulty of systematic EHR extraction. Third, GCS missingness in eICU-CRD (8.8%) and extremely high missingness of temperature-related features (>88%) may affect external validation reliability. Fourth, treatment variables (surgery, anticoagulation reversal) were unavailable; future studies should incorporate these to improve clinical applicability. Fifth, to avoid outcome leakage, ICU length of stay was excluded from model features, as it is not available at the time of prediction in

prospective settings; this exclusion resulted in a minor reduction in discriminative performance. Sixth, this was a retrospective study; prospective validation is needed to confirm real-world effectiveness. Seventh, while our SHAP analysis identified dynamic features as important predictors, we did not perform a formal quantitative decomposition of the variance explained by dynamic versus static features. Future studies may further quantify the relative contribution of temporal features using SHAP-based variance partitioning.

5. Conclusions

This study developed and externally validated machine learning models integrating dynamic time-series physiological features for predicting in-hospital mortality in ICU patients with intracerebral hemorrhage, using a dual-database design comprising MIMIC-IV and eICU-CRD. LightGBM achieved an AUC of 0.813 on a large-scale external validation cohort (n=3,470), significantly outperforming the GCS-only clinical baseline (AUC 0.778) and demonstrating meaningful clinical net benefit across a wide range of threshold probabilities. SHAP analysis identified GCS total score, BUN, and multiple time-series vital sign features as the core predictors, providing a transparent basis for clinical interpretation of model outputs. These findings indicate that machine learning models incorporating dynamic time-series information can offer objective, interpretable decision support for early mortality risk stratification in ICU patients with ICH. Future research should prospectively validate the model and integrate imaging features to further enhance prediction. The current framework, however, already demonstrates that routinely collected ICU time-series data, when properly engineered, can provide robust, interpretable risk stratification without additional imaging or laboratory costs.

Abbreviations

ICH: Intracerebral hemorrhage; ICU: Intensive care unit; ML: Machine learning; AUC: Area under the receiver operating characteristic curve; LR: Logistic regression; RF: Random forest; GCS: Glasgow Coma Scale; SHAP: SHapley Additive exPlanations; LASSO: Least absolute shrinkage and selection operator; MICE: Multiple imputation by chained equations; SMOTE: Synthetic Minority Oversampling Technique; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; MAP: Mean arterial pressure; HR: Heart rate; RR: Respiratory rate; SpO_2: Peripheral oxygen saturation; BPV: Blood pressure variability; BUN: Blood urea nitrogen; DCA: Decision curve analysis; PPV: Positive predictive value; NPV: Negative predictive value; EHR: Electronic health record; IQR: Interquartile range; CI: Confidence interval; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; AI: Artificial intelligence.

Declarations

Acknowledgements

The authors acknowledge the developers of the MIMIC-IV and eICU Collaborative Research Databases for making these datasets publicly available.

Author contributions

HG: writing—original draft, writing—review and editing, conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization. XS: writing—review and editing, conceptualization, supervision, validation. YD: writing—review and editing, conceptualization, data curation, formal analysis, investigation, methodology, supervision, validation. JZ: writing—review and editing, conceptualization, data curation, formal analysis, funding acquisition,investigation, methodology, project administration, supervision, validation.

Funding

The study was funded by the Hebei Provincial Department of Science and Technology (20377764D).

Data availability

The MIMIC-IV dataset used in this study is publicly available from the PhysioNet repository at https://physionet.org/content/mimiciv/. The eICU Collaborative Research Database is publicly available at https://physionet.org/content/eicu-crd/.

Ethics approval and consent to participate

This study used de-identified, publicly available data from the MIMIC-IV and eICU Collaborative Research Databases. Ethical approval and individual patient consent were waived in accordance with the policies of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center Institutional Review Board. All investigators completed the required Collaborative Institutional Training Initiative (CITI) Program certification in human subjects research prior to accessing the data.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke. 2022;17:18-29. http://dx.doi.org/10.1177/17474930211065917.
2. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol. 2021;20:795-820. http://dx.doi.org/10.1016/s1474-4422(21)00252-0.
3. Parry-Jones AR, Krishnamurthi R, Ziai WC, Shoamanesh A, Wu S, Martins SO, et al. World Stroke Organization (WSO): Global intracerebral hemorrhage factsheet 2025. Int J Stroke. 2025;20:145-50. http://dx.doi.org/10.1177/17474930241307876.
4. Song D, Xu D, Li M, Wang F, Feng M, Badr A, et al. Global, regional, and national burdens of intracerebral hemorrhage and its risk factors from 1990 to 2021. Eur J Neurol. 2025;32:e70031. http://dx.doi.org/10.1111/ene.70031.
5. Hemphill JC, 3rd, Bonovich DC, Besmertis L, Manley GT, Johnston SC. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. Stroke. 2001;32:891-7. http://dx.doi.org/10.1161/01.str.32.4.891.
6. Rost NS, Smith EE, Chang Y, Snider RW, Chanderraj R, Schwab K, et al. Prediction of functional outcome in patients with primary intracerebral hemorrhage: the FUNC score. Stroke. 2008;39:2304-9. http://dx.doi.org/10.1161/strokeaha.107.512202.
7. Mayer SA, Rincon F. Treatment of intracerebral haemorrhage. Lancet Neurol. 2005;4:662-72. http://dx.doi.org/10.1016/s1474-4422(05)70195-2.
8. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018;22:1589-604. http://dx.doi.org/10.1109/jbhi.2017.2767063.
9. Lim L, Gim U, Cho K, Yoo D, Ryu HG, Lee HC. Real-time machine learning model to predict short-term mortality in critically ill patients: development and international validation. Crit Care. 2024;28:76. http://dx.doi.org/10.1186/s13054-024-04866-7.
10. Cardona S, Baqai H, Mikdashi F, Aligabi A, Solomon J, Frederick H, et al. Intracranial and Blood Pressure Variability and In-Hospital Outcomes in Intracranial Device-Monitored Patients with Spontaneous Intracerebral Hemorrhage. Neurocrit Care. 2023;39:357-67. http://dx.doi.org/10.1007/s12028-023-01677-6.
11. Mengel A, Siokas V, Buesink R, Roesch S, Laichinger K, Ferizi R, et al. Continuous Blood Pressure Indices During the First 72 Hours and Functional Outcome in Patients with Spontaneous Intracerebral Hemorrhage. Neurocrit Care. 2025;42:839-56. http://dx.doi.org/10.1007/s12028-024-02146-4.
12. Lattanzi S, Cagnetti C, Provinciali L, Silvestrini M. How Should We Lower Blood Pressure after Cerebral Hemorrhage? A Systematic Review and Meta-Analysis. Cerebrovasc Dis. 2017;43:207-13. http://dx.doi.org/10.1159/000462986.
13. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med. 2019;7:152. http://dx.doi.org/10.21037/atm.2019.03.29.
14. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health. 2020;2:e179-e91. http://dx.doi.org/10.1016/s2589-7500(20)30018-2.
15. Huang J, Chen H, Deng J, Liu X, Shu T, Yin C, et al. Interpretable machine learning for predicting 28-day all-cause in-hospital mortality for hypertensive ischemic or hemorrhagic stroke patients in the ICU: a multi-center retrospective cohort study with

internal and external cross-validation. Front Neurol. 2023;14:1185447. http://dx.doi.org/10.3389/fneur.2023.1185447.

16. Wei H, Huang X, Zhang Y, Jiang G, Ding R, Deng M, et al. Explainable machine learning for predicting neurological outcome in hemorrhagic and ischemic stroke patients in critical care. Front Neurol. 2024;15:1385013. http://dx.doi.org/10.3389/fneur.2024.1385013.
17. Nie X, Cai Y, Liu J, Liu X, Zhao J, Yang Z, et al. Mortality Prediction in Cerebral Hemorrhage Patients Using Machine Learning Algorithms in Intensive Care Units. Front Neurol. 2020;11:610531. http://dx.doi.org/10.3389/fneur.2020.610531.
18. Mao B, Ling L, Pan Y, Zhang R, Zheng W, Shen Y, et al. Machine learning for the prediction of in-hospital mortality in patients with spontaneous intracerebral hemorrhage in intensive care unit. Sci Rep. 2024;14:14195. http://dx.doi.org/10.1038/s41598-024-65128-8.
19. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. Bmj. 2024;385:e078378. http://dx.doi.org/10.1136/bmj-2023-078378.
20. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. http://dx.doi.org/10.1038/s41597-022-01899-x.
21. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. http://dx.doi.org/10.1038/sdata.2018.178.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30:377-99. http://dx.doi.org/10.1002/sim.4067.
23. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106. http://dx.doi.org/10.1186/1471-2105-14-106.
24. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16:385-95. http://dx.doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.
25. Shamout F, Zhu T, Clifton DA. Machine Learning for Clinical Outcome Prediction. IEEE Rev Biomed Eng. 2021;14:116-26. http://dx.doi.org/10.1109/rbme.2020.3007816.
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74. http://dx.doi.org/10.1177/0272989x06295361.
27. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020;2:56-67. http://dx.doi.org/10.1038/s42256-019-0138-9.
28. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13:818-29.
29. Parry-Jones AR, Abid KA, Di Napoli M, Smith CJ, Vail A, Patel HC, et al. Accuracy and clinical usefulness of intracerebral hemorrhage grading scores: a direct comparison in a UK population. Stroke. 2013;44:1840-5. http://dx.doi.org/10.1161/strokeaha.113.001009.
30. Miao J, Zuo C, Cao H, Gu Z, Huang Y, Song Y, et al. Predicting ICU readmission risks in intracerebral hemorrhage patients: Insights from machine learning models using MIMIC databases. J Neurol Sci. 2024;456:122849. http://dx.doi.org/10.1016/j.jns.2023.122849.
31. Chen Y, Rivier CA, Mora SA, Torres Lopez V, Payabvash S, Sheth KN, et al. Deep learning survival model predicts outcome after intracerebral hemorrhage from initial CT scan. Eur Stroke J. 2025;10:225-35. http://dx.doi.org/10.1177/23969873241260154.
32. Guan C, Gong A, Zhao Y, Yin C, Geng L, Liu L, et al. Interpretable machine learning model for new-onset atrial fibrillation prediction in critically ill patients: a multi-center study. Crit Care. 2024;28:349. http://dx.doi.org/10.1186/s13054-024-05138-0.
33. Hemphill JC, 3rd, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the Management of Spontaneous Intracerebral Hemorrhage: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2015;46:2032-60. http://dx.doi.org/10.1161/str.0000000000000069.
34. Wang X, Yang J, Moullaali TJ, Sandset EC, Woodhouse LJ, Law ZK, et al. Influence of Time to Achieve Target Systolic Blood Pressure on Outcome After Intracerebral Hemorrhage: The Blood Pressure in Acute Stroke Collaboration. Stroke. 2024;55:849-55. http://dx.doi.org/10.1161/strokeaha.123.044358.

35. Hoste EAJ, Kellum JA, Selby NM, Zarbock A, Palevsky PM, Bagshaw SM, et al. Global epidemiology and outcomes of acute kidney injury. Nat Rev Nephrol. 2018;14:607-25. http://dx.doi.org/10.1038/s41581-018-0052-0.

Tables

Table 1
Baseline characteristics of ICH patients across three cohorts.

Characteristic	Overall N = 6274[1]	MIMIC Training N = 1,962[1]	MIMIC Validation N = 842[1]	eICU External N = 3,470[1]	p-value[2]
Age (years)					< 0.001
Median (Q1, Q3)	66.0 (54.0, 78.0)	68.0 (57.0, 79.0)	68.0 (56.0, 79.0)	65.0 (53.0, 77.0)	
Mean ± SD	64.9 ± 16.1	66.4 ± 15.6	66.5 ± 15.1	63.6 ± 16.5	
Sex					0.851
Male	3,350 (53%)	1,058 (54%)	449 (53%)	1,843 (53%)	
Female	2,923 (47%)	904 (46%)	393 (47%)	1,626 (47%)	
Unknown	1	0	0	1	
Race/Ethnicity					< 0.001
White	4,163 (66%)	1,153 (59%)	513 (61%)	2,497 (72%)	
Black/African American	733 (12%)	208 (11%)	82 (9.7%)	443 (13%)	
Hispanic/Latino	281 (4.5%)	62 (3.2%)	27 (3.2%)	192 (5.5%)	
Asian/Pacific Islander	230 (3.7%)	91 (4.6%)	29 (3.4%)	110 (3.2%)	
Other/Unknown	867 (14%)	448 (23%)	191 (23%)	228 (6.6%)	
GCS Total Score					< 0.001
Median (Q1, Q3)	13.0 (8.0, 15.0)	12.0 (7.0, 15.0)	13.0 (7.0, 15.0)	14.0 (9.0, 15.0)	
Mean ± SD	11.3 ± 4.2	10.6 ± 4.3	10.8 ± 4.2	11.9 ± 4.0	
Unknown	309	4	1	304	
Heart Rate (bpm)					0.427
Median (Q1, Q3)	79.6 (70.4, 89.6)	79.3 (70.6, 88.9)	80.4 (71.3, 89.8)	79.6 (70.0, 89.9)	
Mean ± SD	80.6 ± 14.4	80.4 ± 13.8	81.0 ± 13.9	80.7 ± 14.8	
Unknown	56	2	1	53	
Systolic BP (mmHg)					< 0.001
Median (Q1, Q3)	131.0 (120.9, 140.7)	129.4 (119.0, 138.9)	129.4 (119.0, 138.1)	132.2 (122.4, 142.4)	
Mean ± SD	130.7 ± 15.5	129.1 ± 15.4	128.5 ± 16.6	132.0 ± 15.2	
Unknown	211	117	40	54	
Diastolic BP (mmHg)					< 0.001
Median (Q1, Q3)	68.0 (61.3, 76.3)	70.0 (62.7, 78.0)	69.1 (62.0, 78.5)	66.8 (60.2, 75.0)	
Mean ± SD	68.9 ± 11.3	70.4 ± 11.4	70.2 ± 12.2	67.7 ± 10.9	
Unknown	211	117	40	54	

Values are median (Q1, Q3) or n (%). Between-group comparisons used the Kruskal–Wallis test for continuous variables and chi-squared or Fisher's exact test for categorical variables. "Unknown" denotes missing values and was excluded from percentage calculations. Race/ethnicity classification differs between MIMIC-IV and eICU-CRD.

Abbreviations: BP, blood pressure; MAP, mean arterial pressure; SpO_2, peripheral oxygen saturation; WBC, white blood cell count; BUN, blood urea nitrogen; INR, international normalized ratio; GCS, Glasgow Coma Scale; SD, standard deviation; Q1, first quartile; Q3, third quartile.

Characteristic	Overall N = 6274[1]	MIMIC Training N = 1,962[1]	MIMIC Validation N = 842[1]	eICU External N = 3,470[1]	p-value[2]
MAP (mmHg)					0.355
Median (Q1, Q3)	85.2 (78.3, 92.5)	85.3 (78.7, 92.2)	84.8 (78.1, 91.9)	85.3 (78.2, 92.9)	
Mean ± SD	85.8 ± 10.8	85.7 ± 10.1	85.4 ± 10.9	85.9 ± 11.1	
Unknown	67	10	3	54	
SpO2 (%)					< 0.001
Median (Q1, Q3)	97.3 (95.9, 98.8)	97.1 (95.9, 98.6)	97.2 (95.9, 98.7)	97.5 (96.0, 98.9)	
Mean ± SD	97.2 ± 2.0	97.1 ± 1.9	97.1 ± 1.9	97.3 ± 2.0	
Unknown	75	2	2	71	
Respiratory Rate (breaths/min)					0.044
Median (Q1, Q3)	18.2 (16.2, 20.7)	18.3 (16.5, 20.6)	18.4 (16.6, 20.4)	18.1 (16.0, 20.8)	
Mean ± SD	18.8 ± 3.8	18.7 ± 3.2	18.8 ± 3.1	18.8 ± 4.3	
Unknown	199	9	2	188	
Glucose (mg/dL)					0.514
Median (Q1, Q3)	128.0 (108.0, 158.0)	129.0 (107.0, 160.0)	129.0 (106.0, 157.0)	127.0 (108.0, 156.0)	
Mean ± SD	142.9 ± 60.2	145.0 ± 65.3	142.1 ± 60.4	141.8 ± 57.0	
Unknown	448	100	43	305	
Creatinine (mg/dL)					0.004
Median (Q1, Q3)	0.9 (0.7, 1.1)	0.9 (0.7, 1.1)	0.9 (0.7, 1.1)	0.8 (0.7, 1.1)	
Mean ± SD	1.1 ± 1.2	1.1 ± 1.2	1.1 ± 1.0	1.1 ± 1.2	
Unknown	441	98	43	300	
Sodium (mEq/L)					0.007
Median (Q1, Q3)	139.0 (137.0, 142.0)	139.0 (137.0, 142.0)	139.0 (137.0, 142.0)	139.0 (137.0, 141.0)	
Mean ± SD	139.2 ± 4.7	139.3 ± 4.6	139.2 ± 4.8	139.1 ± 4.7	
Unknown	410	92	41	277	
Potassium (mEq/L)					< 0.001
Median (Q1, Q3)	3.9 (3.6, 4.2)	3.9 (3.6, 4.3)	3.9 (3.6, 4.3)	3.8 (3.5, 4.2)	
Mean ± SD	3.9 ± 0.6	4.0 ± 0.6	4.0 ± 0.7	3.9 ± 0.5	
Unknown	420	97	42	281	
WBC (×10³/μL)					0.023

Values are median (Q1, Q3) or n (%). Between-group comparisons used the Kruskal–Wallis test for continuous variables and chi-squared or Fisher's exact test for categorical variables. "Unknown" denotes missing values and was excluded from percentage calculations. Race/ethnicity classification differs between MIMIC-IV and eICU-CRD.

Abbreviations: BP, blood pressure; MAP, mean arterial pressure; SpO_2, peripheral oxygen saturation; WBC, white blood cell count; BUN, blood urea nitrogen; INR, international normalized ratio; GCS, Glasgow Coma Scale; SD, standard deviation; Q1, first quartile; Q3, third quartile.

Characteristic	Overall N = 6274[1]	MIMIC Training N = 1,962[1]	MIMIC Validation N = 842[1]	eICU External N = 3,470[1]	p-value[2]
Median (Q1, Q3)	10.4 (7.9, 13.5)	10.3 (7.9, 13.2)	10.0 (7.9, 12.9)	10.6 (8.0, 13.7)	
Mean ± SD	11.5 ± 9.4	11.3 ± 8.3	11.3 ± 12.2	11.6 ± 9.2	
Unknown	540	105	47	388	
Hemoglobin (g/dL)					0.038
Median (Q1, Q3)	12.3 (10.8, 13.6)	12.2 (10.8, 13.6)	12.1 (10.9, 13.4)	12.4 (10.8, 13.7)	
Mean ± SD	12.2 ± 2.1	12.2 ± 2.1	12.0 ± 2.0	12.2 ± 2.2	
Unknown	565	104	48	413	
BUN (mg/dL)					0.013
Median (Q1, Q3)	16.0 (11.0, 22.0)	16.0 (12.0, 22.0)	16.0 (12.0, 22.0)	15.0 (11.0, 22.0)	
Mean ± SD	19.5 ± 15.0	19.6 ± 14.8	19.5 ± 13.8	19.5 ± 15.4	
Unknown	438	99	43	296	
INR					< 0.001
Median (Q1, Q3)	1.1 (1.1, 1.3)	1.1 (1.1, 1.3)	1.2 (1.1, 1.3)	1.1 (1.0, 1.3)	
Mean ± SD	1.2 ± 0.4	1.2 ± 0.4	1.3 ± 0.6	1.2 ± 0.4	
Unknown	2,163	287	113	1,763	
Platelets (×10³/μL)					0.991
Median (Q1, Q3)	207.0 (163.0, 257.0)	208.0 (161.5, 258.0)	205.0 (162.0, 263.0)	207.0 (165.0, 255.0)	
Mean ± SD	215.1 ± 84.7	215.6 ± 85.0	217.0 ± 89.5	214.2 ± 83.4	
Unknown	562	106	48	408	
In-hospital Mortality	1,198 (19%)	417 (21%)	181 (21%)	600 (17%)	< 0.001

Values are median (Q1, Q3) or n (%). Between-group comparisons used the Kruskal–Wallis test for continuous variables and chi-squared or Fisher's exact test for categorical variables. "Unknown" denotes missing values and was excluded from percentage calculations. Race/ethnicity classification differs between MIMIC-IV and eICU-CRD.

Abbreviations: BP, blood pressure; MAP, mean arterial pressure; SpO_2, peripheral oxygen saturation; WBC, white blood cell count; BUN, blood urea nitrogen; INR, international normalized ratio; GCS, Glasgow Coma Scale; SD, standard deviation; Q1, first quartile; Q3, third quartile.

Table 2
Predictive performance of machine learning models across validation cohorts.

Model	Dataset	AUC (95% CI)	Cutoff	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 Score (95% CI)	Brier Score (95% CI)
LR	Internal validation	0.813 (0.779–0.844)	0.235	76.2% (70.0–82.3%)	77.5% (74.3–80.4%)	77.2% (74.3–79.9%)	48.1% (42.5–54.1%)	92.3% (90.0–94.4%)	0.590 (0.536–0.642)	0.132 (0.118–0.146)
RF	Internal validation	0.813 (0.779–0.844)	0.341	80.7% (74.5–86.2%)	70.2% (66.7–73.5%)	72.4% (69.4–75.4%)	42.6% (37.2–47.9%)	93.0% (90.7–95.1%)	0.557 (0.504–0.611)	0.142 (0.131–0.154)
XGBoost	Internal validation	0.823 (0.789–0.854)	0.248	84.0% (77.9–89.2%)	69.0% (65.4–72.4%)	72.2% (69.1–74.9%)	42.6% (37.2–47.8%)	94.0% (91.7–96.1%)	0.565 (0.509–0.614)	0.137 (0.123–0.151)
LightGBM	Internal validation	0.823 (0.789–0.854)	0.237	79.0% (72.7–84.9%)	72.2% (68.7–75.4%)	73.6% (70.5–76.4%)	43.7% (38.1–49.2%)	92.6% (90.2–94.9%)	0.563 (0.507–0.614)	0.136 (0.122–0.151)
LR	External validation	0.805 (0.785–0.823)	0.167	74.0% (70.4–77.5%)	73.9% (72.2–75.5%)	73.9% (72.3–75.4%)	37.2% (34.3–40.0%)	93.1% (92.0–94.2%)	0.495 (0.465–0.523)	0.116 (0.109–0.123)
RF	External validation	0.810 (0.792–0.828)	0.339	77.8% (74.5–80.9%)	72.1% (70.4–73.7%)	73.1% (71.5–74.6%)	36.8% (34.1–39.4%)	94.0% (92.9–94.9%)	0.500 (0.472–0.526)	0.130 (0.124–0.135)
XGBoost	External validation	0.811 (0.792–0.830)	0.247	79.3% (76.0–82.7%)	70.8% (69.1–72.5%)	72.2% (70.7–73.8%)	36.2% (33.6–38.8%)	94.2% (93.2–95.2%)	0.497 (0.470–0.524)	0.127 (0.120–0.134)
LightGBM	External validation	0.813 (0.795–0.831)	0.221	79.0% (75.9–82.3%)	72.5% (70.9–74.1%)	73.6% (72.1–75.0%)	37.5% (34.7–40.2%)	94.3% (93.3–95.2%)	0.509 (0.479–0.537)	0.121 (0.115–0.128)
GCS-only	Internal validation	0.770 (0.731–0.805)	—	80.1% (74.2–86.1%)	62.8% (59.1–66.4%)	66.5% (63.2–69.5%)	37.1% (32.3–41.9%)	92.0% (89.5–94.5%)	0.507 (0.457–0.556)	—
GCS-only	External validation	0.778 (0.757–0.800)	—	79.3% (76.1–82.6%)	63.9% (62.2–65.8%)	66.6% (65.1–68.2%)	31.5% (29.3–33.8%)	93.7% (92.6–94.8%)	0.451 (0.426–0.477)	—

95% CIs estimated by 1,000-iteration bootstrap resampling. Optimal cutoff determined by the Youden index. GCS-only: logistic regression using GCS total score as the sole predictor; Brier score not reported. DeLong test: LightGBM vs. GCS-only — internal validation $p = 0.0002$, external validation $p < 0.0001$.

Abbreviations: AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; LR, logistic regression; RF, random forest; GCS, Glasgow Coma Scale; CI, confidence interval.

Table 3
Incremental value of time-series features: static vs. full LightGBM model.

Model	Dataset	AUC	Sensitivity	Specificity	NPV	Brier Score
Static LightGBM	Internal validation (MIMIC-IV)	0.817 (0.785–0.849)	85.6% (80.3–90.8%)	65.8% (62.3–69.5%)	94.4% (92.2–96.5%)	0.137 (0.124–0.149)
Full LightGBM		0.823 (0.789–0.854)	79.0% (72.7–84.9%)	72.2% (68.7–75.4%)	92.6% (90.2–94.9%)	0.136 (0.122–0.151)
Static LightGBM	External validation (eICU-CRD)	0.808 (0.790–0.826)	70.8% (67.1–74.3%)	78.2% (76.7–79.8%)	92.8% (91.7–93.8%)	0.123 (0.116–0.129)
Full LightGBM		0.813 (0.795–0.831)	79.0% (75.9–82.3%)	72.5% (70.9–74.1%)	94.3% (93.3–95.2%)	0.121 (0.115–0.128)

95% CIs estimated by 1,000-iteration bootstrap resampling. Static model: LightGBM trained on 7 laboratory/clinical features only. Full model: LightGBM trained on 29 features (7 static + 22 time-series). NRI and IDI were calculated using the nricens package; DeLong test used for AUC comparison.

Abbreviations: AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; NRI, net reclassification improvement; IDI, integrated discrimination improvement; CI, confidence interval.

Figures

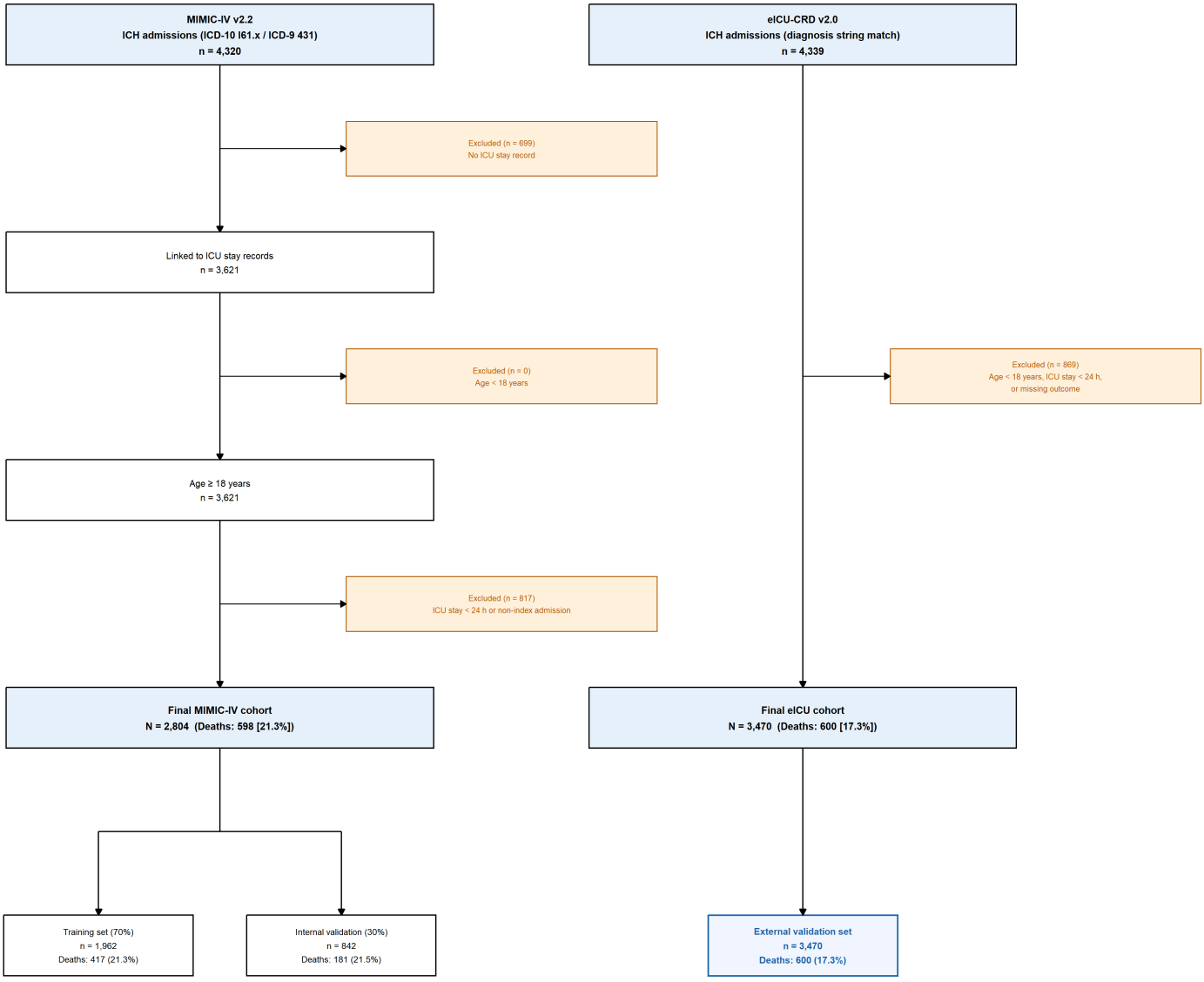


Figure 1

Patient Selection and Study Design Flowchart.

Patients with intracerebral hemorrhage were identified from MIMIC-IV (derivation) and eICU-CRD (external validation). Exclusion criteria were applied sequentially. The derivation cohort was split 70:30 for training and internal validation.

Abbreviations: ICH, intracerebral hemorrhage; MIMIC-IV, Medical Information Mart for Intensive Care IV; eICU-CRD, eICU Collaborative Research Database.

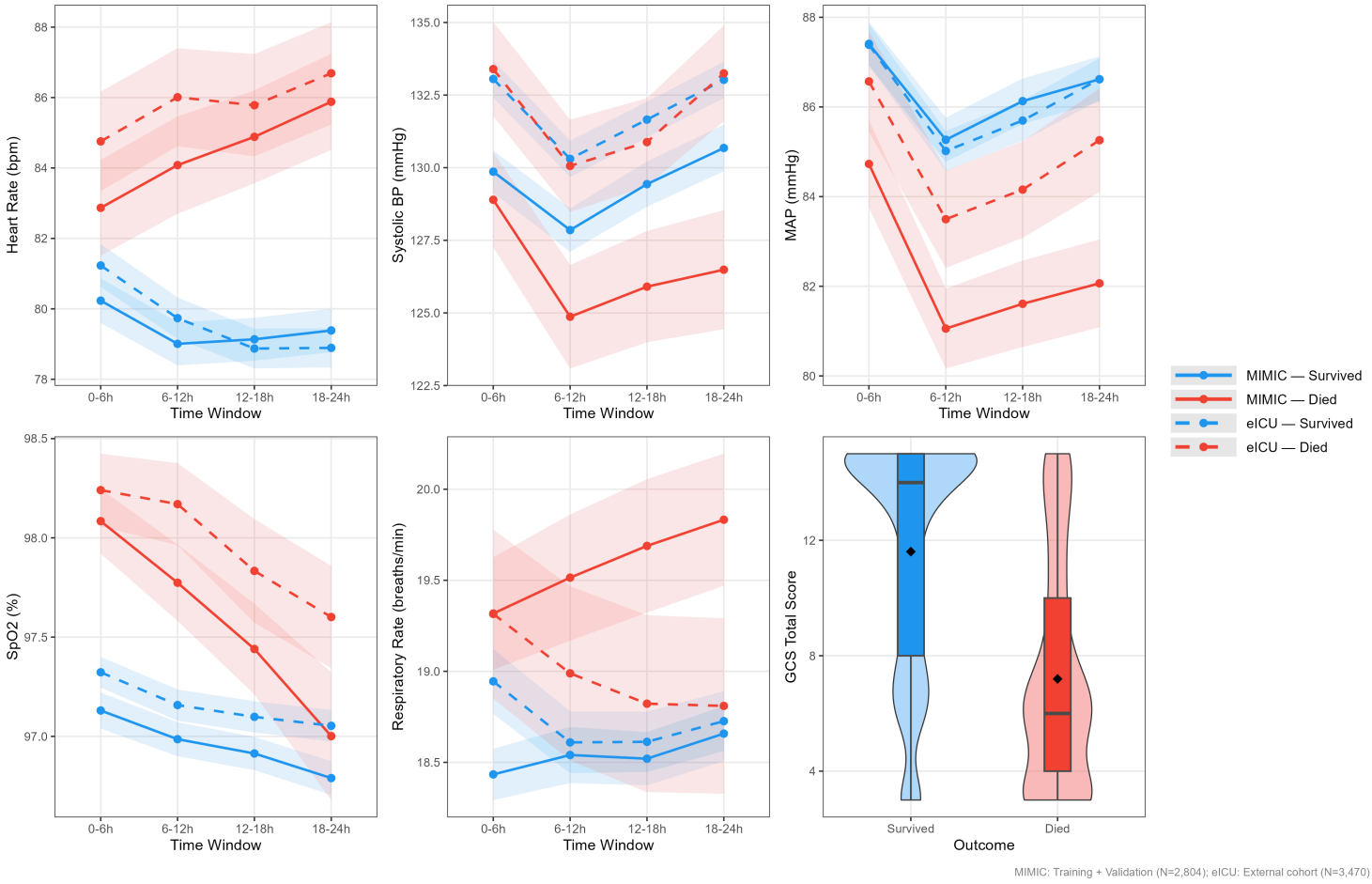


Figure 2

Temporal Trajectories of Vital Signs and Neurological Status by Outcome.

Vital signs were segmented into four consecutive 6-hour windows (W1–W4) within the first 24 hours of ICU admission. Lines represent group means; shaded bands indicate 95% CIs. GCS scores are shown as violin plots with embedded box plots. Significant between-group differences at each window are indicated by asterisks (*P < .05, **P < .01, ***P < .001).

Abbreviations: GCS, Glasgow Coma Scale; HR, heart rate; MAP, mean arterial pressure; RR, respiratory rate; SBP, systolic blood pressure; SpO_2, peripheral oxygen saturation; W1–W4, time windows 1–4 (0–6, 6–12, 12–18, 18–24 h).

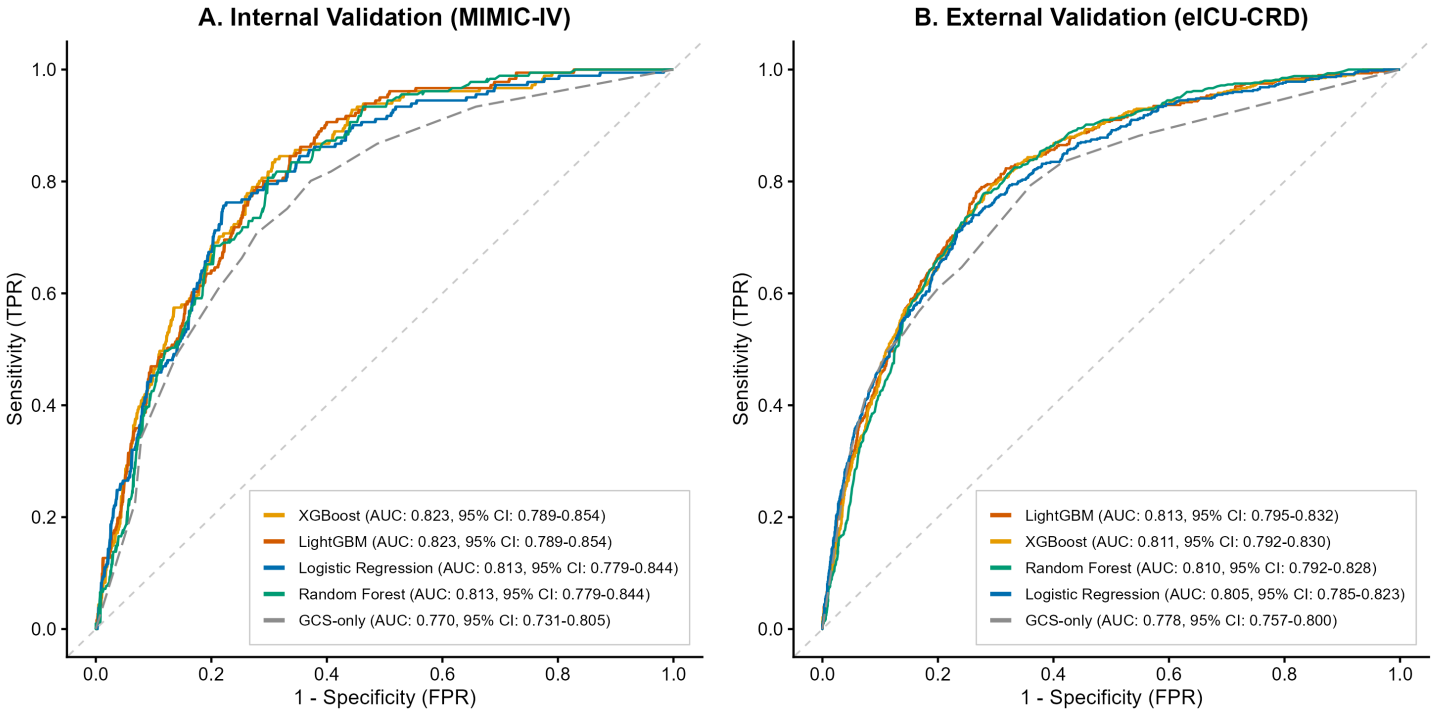


Figure 3

Receiver Operating Characteristic Curves for Predicting In-Hospital Mortality.

ROC curves are shown for four machine learning models and GCS-alone in (A) internal validation (MIMIC-IV, n = 842) and (B) external validation (eICU-CRD, n = 3470). AUC values with 95% CIs are displayed in the legend. Dashed line indicates chance-level discrimination.

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; GCS, Glasgow Coma Scale; LightGBM, Light Gradient Boosting Machine; LR, logistic regression; RF, random forest; ROC, receiver operating characteristic; XGB, XGBoost.

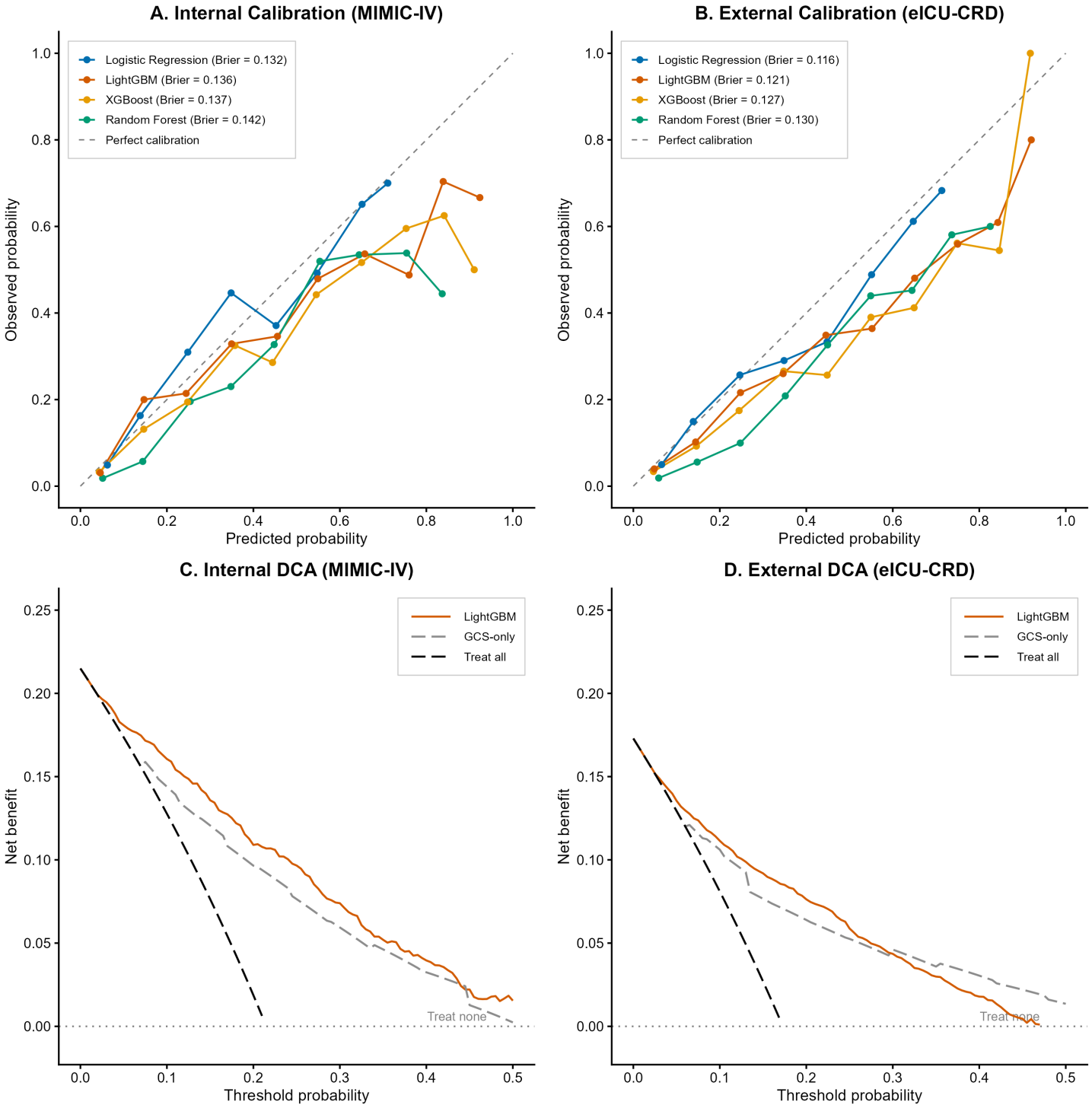


Figure 4

Model Calibration and Decision Curve Analysis.

(A, B) Calibration curves compare predicted versus observed mortality probabilities in internal and external validation cohorts; perfect calibration is indicated by the diagonal line. Brier scores are shown for each model. (C, D) Decision curve analysis displays net benefit across a range of threshold probabilities. The "treat all" and "treat none" strategies serve as reference lines.

Abbreviations: DCA, decision curve analysis; LightGBM, Light Gradient Boosting Machine; LR, logistic regression; RF, random forest; XGB, XGBoost.

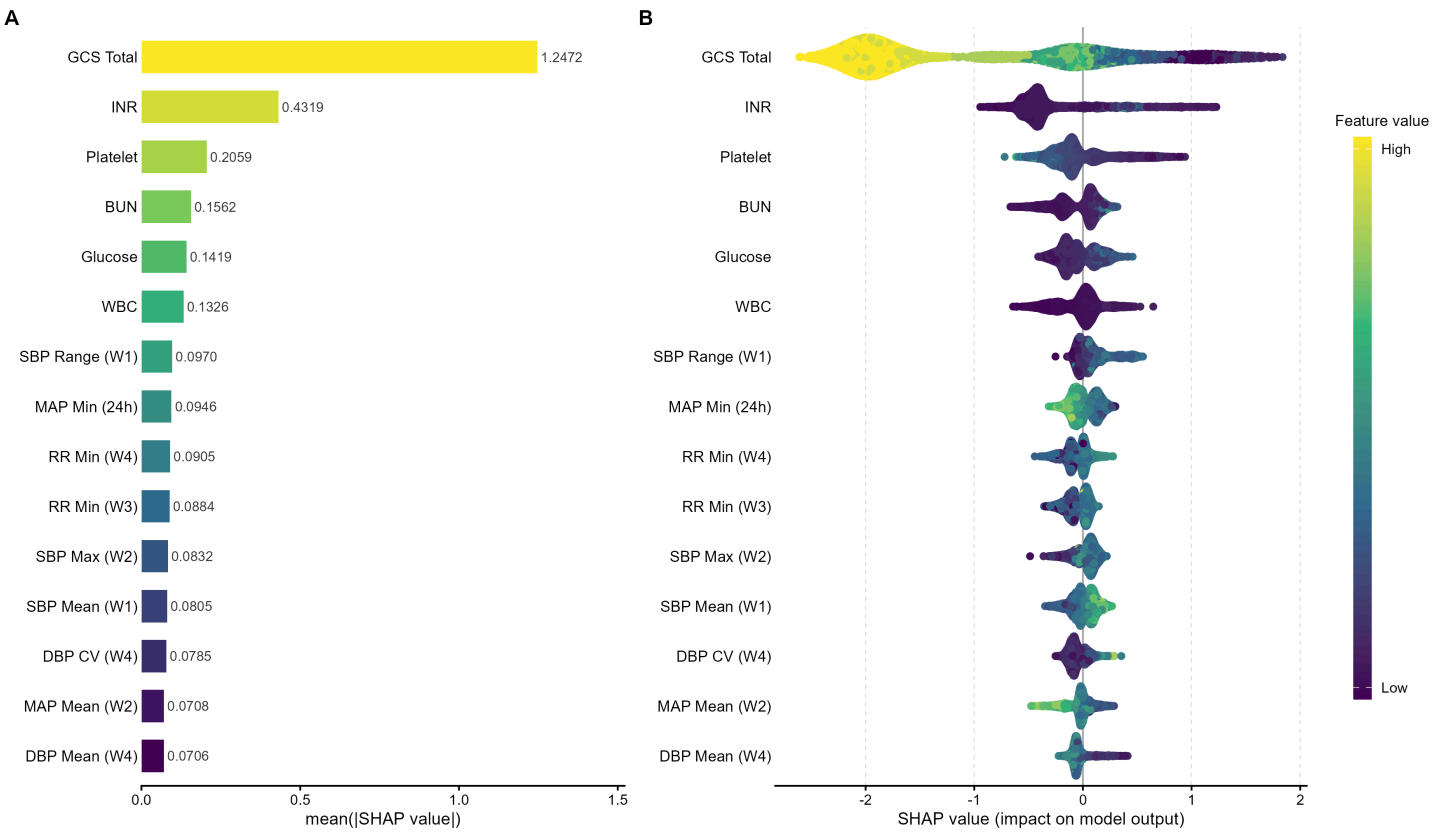


Figure 5

SHAP Analysis of the Top 15 Features in the LightGBM Model.

(A) Mean absolute SHAP values for the top 15 features, ranked by overall contribution to model output. (B) Beeswarm plot showing the distribution of SHAP values for each feature across all patients in the external validation cohort. Each dot represents one patient; color indicates feature value (yellow = high, purple = low). Positive SHAP values indicate increased predicted mortality risk; negative values indicate decreased risk. GCS Total was the dominant predictor, with lower scores associated with higher mortality risk.

Abbreviations: BUN, blood urea nitrogen; CV, coefficient of variation; DBP, diastolic blood pressure; GCS, Glasgow Coma Scale; INR, international normalized ratio; MAP, mean arterial pressure; RR, respiratory rate; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; WBC, white blood cell count; W1–W4, time windows 1–4 (0–6, 6–12, 12–18, 18–24 h).

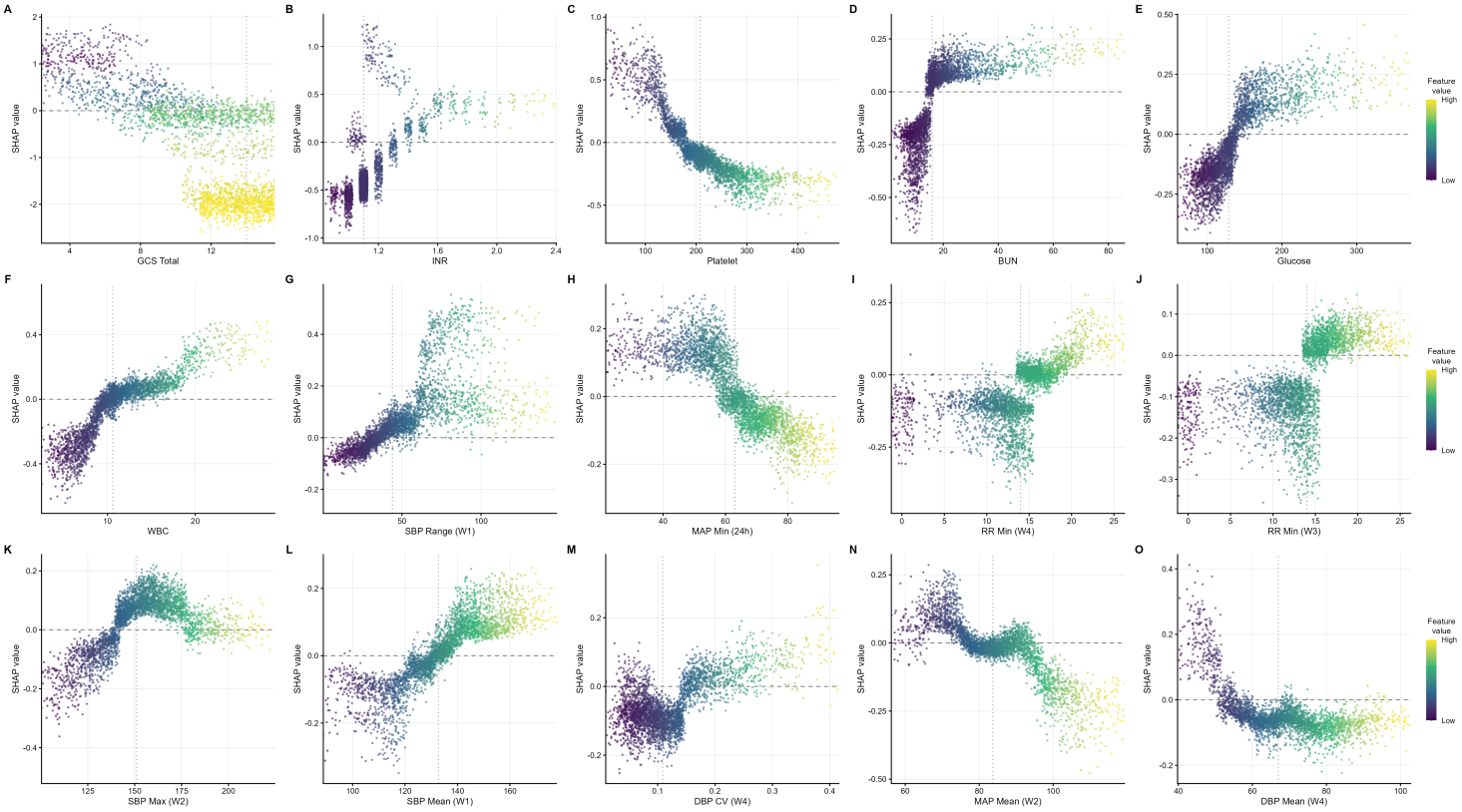


Figure 6

SHAP Dependence Plots for the Top 15 Predictors of In-Hospital Mortality.

Each panel displays the relationship between feature value (x-axis) and its SHAP value (y-axis) for the LightGBM model in the external validation cohort. Positive SHAP values indicate increased mortality risk. Color encodes a secondary interacting feature where applicable. Features are ordered by mean |SHAP| value.

Abbreviations: BUN, blood urea nitrogen; DBP, diastolic blood pressure; GCS, Glasgow Coma Scale; INR, international normalized ratio; MAP, mean arterial pressure; RR, respiratory rate; SBP, systolic blood pressure; SHAP, SHapley Additive exPlanations; WBC, white blood cell count; W1–W4, time windows 1–4.

Supplementary Files

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

- SFig1LASSO.png
- SFig2ProbabilityDistribution.png
- STable1Missing.docx
- STable2Hyperparams.docx