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.

Multi-Time-Window Landmark Modeling for Dynamic Mortality Prediction in Spontaneous Intracerebral Hemorrhage: A Machine Learning Study Using the MIMIC-IV Database

Chengxin Ma
Shuguang Hospital, Shanghai University of Traditional Chinese Medicine
Jue Zhang
Shuguang Hospital, Shanghai University of Traditional Chinese Medicine
Guowei Shu
Shuguang Hospital, Shanghai University of Traditional Chinese Medicine
Zhen Zhang
Shuguang Hospital, Shanghai University of Traditional Chinese Medicine
Yashengjiang Maimaiti
Huashan Hospital, Fudan University
Zhimin Fei
tommyfei@126.com

Shuguang Hospital, Shanghai University of Traditional Chinese Medicine

Research Article

Keywords: Intracerebral hemorrhage, Machine learning, Landmark analysis, Dynamic prediction, In-hospital mortality, SHAP, XGBoost, MIMIC-IV

Posted Date: June 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9693281/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

Current machine learning models for mortality prediction in spontaneous intracerebral hemorrhage (ICH) employ static, single-time-point designs that cannot capture the progressive accumulation of prognostic information during the intensive care unit (ICU) stay. We developed and validated a multi-time-window landmark modeling framework to dynamically predict in-hospital mortality at clinically meaningful time points and to characterize the temporal evolution of prognostic feature importance.

Methods

Using the MIMIC-IV v3.1 database, we identified 2,298 ICU patients with ICH (in-hospital mortality: 23.8%). Four landmark models — eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Random Forest, and Logistic Regression — were independently trained at 6, 12, 24, and 48 hours after ICU admission using Least absolute shrinkage and selection operator (LASSO)-selected features. Model performance was assessed via area under the receiver operating characteristic curve (AUROC) with DeLong and bootstrap 95% confidence intervals, calibration curves, decision curve analysis, and SHapley Additive exPlanations (SHAP)-based feature importance analysis.

Results

The XGBoost model achieved AUROC values of 0.856 (95% CI: 0.839–0.873), 0.872 (0.855–0.888), 0.864 (0.843–0.881), and 0.857 (0.836–0.879) at the 6h, 12h, 24h, and 48h landmarks, respectively, with the 12h window demonstrating peak discrimination. This inverted-U trajectory was consistent across all four model types. Seven features — Glasgow Coma Scale (GCS) motor minimum, age, maximum glucose, red cell distribution width (RDW) mean, minimum diastolic blood pressure, liver disease, and age < 65 years — were identified as cross-window stable predictors through set intersection of window-specific SHAP rankings. A characteristic temporal shift in feature importance was observed, transitioning from neurological determinants (GCS components) at early windows to systemic and metabolic features (sodium, potassium) at later windows, mapping onto the established pathophysiological trajectory of ICH from primary brain injury to multi-organ complications.

Conclusions

Multi-time-window landmark modeling reveals that mortality prediction in ICH peaks at 12 hours post-admission — representing the optimal balance between information accumulation and patient retention — and that the shifting hierarchy of prognostic features parallels the pathophysiological evolution of ICH. This framework provides a principled approach to dynamic risk assessment in the ICU and offers data-driven insights into the temporal determinants of ICH outcome.

Background

Spontaneous intracerebral hemorrhage (ICH) accounts for 10–20% of all strokes worldwide, yet carries a disproportionately high burden of mortality and long-term disability(1). According to the Global Burden of Disease Study 2021, ICH was responsible for 3.44 million incident cases and 3.31 million deaths globally in 2021, with an age-standardized incidence rate of 40.8 per 100,000 person-years and an age-standardized mortality rate of 39.1 per 100,000 (2). Despite significant declines in age-standardized rates over the past three decades — attributable largely to improved hypertension control — the absolute burden continues to rise in aging populations, particularly in low-

and middle-income countries. In-hospital mortality among ICU-admitted ICH patients ranges from 20% to 40% in contemporary cohorts, and functional independence at six months is achieved by fewer than one-third of survivors (1). The acute phase of ICH management is characterized by rapid, high-stakes clinical decision-making — hemostatic therapy, blood pressure control, neurosurgical consultation, and goals-of-care discussions — all of which depend on accurate and timely prognostic assessment. Early identification of patients at high risk of mortality is therefore critical to guiding treatment intensity and informing family communication.

In recent years, machine learning (ML) models have demonstrated promising performance for mortality prediction in ICH patients using the MIMIC-IV database. Mao et al. developed an XGBoost model with 14 LASSO-selected features in a cohort of 1,596 ICH patients, achieving an area under the receiver operating characteristic curve (AUROC) of 0.907 (3). Chen et al. reported a 7-feature XGBoost model in 1,274 non-traumatic ICH patients (AUROC 0.797), with the additional contribution of external validation (4). Ye et al. incorporated glucose and lipid biomarker panels in an XGBoost model for 2,323 ICH patients, achieving an AUROC of 0.851 (5). Despite these advances, a critical methodological limitation is shared across all existing studies: they employ a static, single-time-point design, extracting features from the entire ICU stay (typically 24 hours or the full admission) and predicting mortality from a fixed reference point.

This static approach is fundamentally misaligned with the clinical reality of ICU care, in which prognostic information accumulates progressively over time. A clinician at the 6-hour mark has access to initial vital signs, neurological examination, and limited laboratory data; by 12 hours, additional laboratory results and imaging are available; by 48 hours, the full trajectory of physiological responses has unfolded. Two interrelated questions arise that static models cannot answer: (1) how does predictive performance evolve as more information becomes available? and (2) which clinical features are most important for prognostication at each stage of the ICU course? Answering these questions requires a dynamic prediction framework that explicitly models time-dependent information accumulation.

The landmark modeling framework, originally developed in the context of dynamic prediction in clinical research (6), provides a natural methodological solution. In a landmark analysis, separate prediction models are built at pre-specified time points (landmarks), each using only information available up to that time and conditioning on the patient having survived to the landmark. This framework has been applied to general ICU mortality prediction using recursive updating (7), but has not been extended to a disease-specific, multi-time-window analysis with comprehensive feature sets in the ICH population.

In this study, we applied a multi-time-window landmark modeling framework to develop and compare four machine learning models — XGBoost, LightGBM, Random Forest, and Logistic Regression — for dynamic in-hospital mortality prediction at four clinically meaningful time points (6, 12, 24, and 48 hours after ICU admission) in a cohort of 2,298 ICH patients from the MIMIC-IV v3.1 database. Using SHAP (SHapley Additive exPlanations) values (8), we further characterized the temporal evolution of feature importance across windows, mapping the shifting prognostic determinants onto the pathophysiological trajectory of ICH from primary brain injury to systemic complications.

Methods

Data Source

This study used data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) version 3.1, a freely accessible electronic health record database comprising over 70,000 ICU admissions at Beth Israel Deaconess Medical Center (BIDMC), Boston, MA, between 2008 and 2019 (9). Access was granted after completion of the Collaborative Institutional Training Initiative (CITI) program and signing of the PhysioNet data use agreement. The

study was conducted in accordance with the PhysioNet Data Use Agreement and did not require separate ethical approval, as MIMIC-IV contains de-identified health data.

Study Population and Outcome

We identified adult patients (≥18 years) with a primary diagnosis of spontaneous intracerebral hemorrhage using ICD-9-CM code 431 (Intracerebral hemorrhage) and ICD-10-CM codes I61.0-I61.9 (Nontraumatic intracerebral hemorrhage). Only the first ICU admission for each patient was included. Patients with secondary diagnoses of traumatic brain injury were not excluded separately, as the ICD codes for nontraumatic ICH specifically denote spontaneous hemorrhage. Participant selection and landmark window allocation are illustrated in Additional file 3: Figure S1.

This yielded a cohort of 2,298 ICU admissions. The primary outcome was in-hospital mortality, defined as death during the index hospital admission, with an overall event rate of 23.8% (547/2,298). For the landmark analysis, patients were further stratified into four sub-cohorts based on survival and continued ICU stay at each landmark time point (6, 12, 24, and 48 hours). The resulting sub-cohort sizes are presented in Table 1, and mortality rates across windows are reported in the Results.

Table 1. Baseline characteristics of ICH patients at the 6-hour landmark

Variable	Overall (n=2279)	Survived (n=1746)	Died (n=533)	p	Valid n	Missing (%)
Demographics						
Age, years	69.0 [57.0–80.0]	68.0 [56.0–80.0]	73.0 [62.0–82.0]	**<0.001**	2279	0.0
Female, n (%)	1061 (46.6)	805 (46.1)	256 (48.0)	0.465	2279	0.0
Comorbidities						
Hypertension, n (%)	1677 (73.6)	1294 (74.1)	383 (71.9)	0.328	2279	0.0
Diabetes, n (%)	531 (23.3)	387 (22.2)	144 (27.0)	0.024	2279	0.0
CHF, n (%)	252 (11.1)	181 (10.4)	71 (13.3)	0.068	2279	0.0
Renal disease, n (%)	226 (9.9)	159 (9.1)	67 (12.6)	0.024	2279	0.0
Liver disease, n (%)	89 (3.9)	44 (2.5)	45 (8.4)	**<0.001**	2279	0.0
Malignancy, n (%)	245 (10.8)	205 (11.7)	40 (7.5)	0.007	2279	0.0
COPD, n (%)	146 (6.4)	106 (6.1)	40 (7.5)	0.279	2279	0.0
Vital Signs (first 6h)						
Heart rate, bpm	85.6 [77.0–95.6]	85.3 [77.0–94.8]	86.6 [77.1–98.7]	0.055	2271	0.4
SBP, mmHg	130.4 [120.2–140.4]	130.7 [120.5–140.4]	129.5 [118.2–141.4]	0.221	2260	0.8
DBP, mmHg	69.3 [61.0–77.8]	70.1 [62.3–78.9]	66.2 [57.7–75.0]	**<0.001**	2259	0.9
SpO_2, %	96.2 [94.1–98.0]	96.3 [94.6–97.9]	95.3 [91.1–98.6]	**<0.001**	2269	0.4
GCS (first 6h)						
GCS total, first	12.0 [7.0–15.0]	14.0 [9.0–15.0]	6.0 [3.0–8.0]	**<0.001**	2235	1.9
GCS eye, first	3.0 [1.0–4.0]	4.0 [2.0–4.0]	1.0 [1.0–2.0]	**<0.001**	2260	0.8
GCS verbal, first	3.0 [1.0–5.0]	4.0 [1.0–5.0]	1.0 [1.0–1.0]	**<0.001**	2255	1.1
GCS motor, first	6.0 [4.0–6.0]	6.0 [5.0–6.0]	4.0 [1.0–5.0]	**<0.001**	2241	1.7
Laboratory (first 6h)						
WBC, $\times 10^9$/L	10.5 [8.2–13.6]	10.0 [8.0–12.7]	12.3 [9.5–15.9]	**<0.001**	1366	40.1
Platelet, $\times 10^9$/L	209.0 [164.0–262.0]	210.0 [170.0–264.0]	197.0 [149.0–254.0]	**<0.001**	1370	39.9
INR	1.1 [1.1–1.3]	1.1 [1.1–1.2]	1.2 [1.1–1.4]	**<0.001**	1282	43.7
Glucose, mg/dL	131.0 [109.0–164.0]	126.0 [107.0–157.0]	151.5 [124.0–189.8]	**<0.001**	1411	38.1
Creatinine, mg/dL	0.9 [0.7–1.1]	0.9 [0.7–1.1]	0.9 [0.7–1.3]	**<0.001**	1413	38.0
BUN, mg/dL	16.0 [12.0–22.0]	16.0 [11.0–21.0]	18.0 [13.0–26.0]	**<0.001**	1415	37.9

Hemoglobin, g/dL	12.5 [11.1–13.7]	12.6 [11.2–13.8]	12.1 [10.6–13.6]	0.003	1367	40.0
Sodium, mmol/L	139.0 [137.0–142.0]	139.0 [137.0–142.0]	140.0 [137.0–142.0]	0.206	1458	36.0
RDW, %	13.7 [13.0–14.6]	13.6 [12.9–14.5]	14.0 [13.3–15.2]	**<0.001**	1365	40.1

Abbreviations: COPD, Chronic obstructive pulmonary disease; SpO2, Peripheral oxygen saturation; GCS, Glasgow Coma Scale; WBC, White blood cell; INR, International normalized ratio.

Data are presented as median [IQR] for continuous variables and n (%) for categorical variables. P values from Mann-Whitney U test (continuous) or chi-square test (categorical), comparing survivors vs. non-survivors. Missing percentages reflect the proportion of patients with no measurement within the 6-hour window (pre-MICE). GCS total = eye + verbal + motor component scores. Significant p (<0.001) shown in bold. The high laboratory missing rates (36–44%) are attributable to the short pre-analytical window — these values decline to <25% by 12h and <15% by 24h (see Additional file 1: Table S1).

Landmark Modeling Framework

The central methodological innovation of this study is the application of a multi-time-window landmark modeling framework. Rather than constructing a single prediction model using data from a fixed time window, we developed four independent prediction models at clinically meaningful time points: 6, 12, 24, and 48 hours after ICU admission.

At each landmark time point *s*, only patients who were still alive and remained in the ICU at time *s* were included in the corresponding sub-cohort. Features were extracted exclusively from the time window spanning ICU admission (intime) to intime + *s* hours, **ensuring that no future information was available to the model at the time of prediction.** Each model independently predicted the fixed endpoint of in-hospital mortality using only information accumulated up to that landmark.

This approach differs from recursive updating strategies (7), in which a single model is repeatedly applied with sliding time windows. In our landmark framework, each model is independently trained and evaluated, allowing direct comparison of predictive performance across time points without dependence on prior predictions.

The landmark sub-cohorts exhibited progressively declining mortality rates — from 23.4% at 6h to 20.4% at 48h — reflecting the conditional-on-survival design: patients who died early were naturally excluded from later windows. This is an inherent feature of the landmark framework rather than a methodological limitation, as each model addresses the distinct clinical question of risk assessment for patients who have survived to that specific time point.

Feature Extraction

Clinical features were extracted from MIMIC-IV tables including $icu.\ chartevents$ (vital signs, GCS), $hosp.\ labevents$ (laboratory values), $mimiciv_derived$ (derived chemistry and vital sign tables), and $hosp.\ admissions/hosp.\ patients$ (demographics). Comorbidity indicators (hypertension, diabetes, COPD, congestive heart failure, renal disease, liver disease, and malignancy) were extracted from ICD-coded diagnosis records using the ich_com or $bties_treatments$ table and treated as static features shared across all time windows.

For each vital sign and laboratory variable, we extracted summary statistics (mean, minimum, maximum, and first recorded value) within each time window. The total GCS score was retained alongside its three component scores — eye, verbal, and motor — with minimum and first values extracted for each, following evidence that individual GCS components carry distinct prognostic information that is partially lost when aggregated into a single total score (10). Specifically, Lee et al. demonstrated that the same GCS total score can correspond to substantially different mortality risks depending on the component breakdown (e.g., GCS=4: E1V1M2 = 48% mortality vs. E1V2M1 = 27%), and that the motor component alone achieves predictive performance approaching that of the full GCS (AUROC 0.817 vs. 0.833 for in-hospital mortality).

SpO2 Artifact Handling

Peripheral oxygen saturation (SpO2) required special handling due to pervasive pulse oximeter probe displacement artifacts. Analysis revealed that 44.7% of patients had their mean SpO2 values depressed by transient readings below 50%, which are physiologically unsustainable without extracorporeal membrane oxygenation (corresponding to PaO2 ≈ 28 mmHg). Among the 152 patients with raw SpO2 mean < 85%, 139 (91.4%) had their values recover to ≥85% after artifact exclusion, confirming these low readings as artifacts.

We therefore recalculated SpO2 mean values by excluding all readings below 50% from the MIMIC-IV vitalsign derived table, using the SQL expression: $AVG(CASEWHENspo2 \geq 50THENspo2ELSENULLEND)$. The SpO2 minimum was excluded entirely as a model feature, as 33% of patients had SpO2_min < 50%, rendering this statistic unreliable.

Data Cleaning

Data cleaning was performed in two stages. First, physiological range filtering was applied at the SQL level based on thresholds from the official MIMIC-IV concept code (MIT-LCP/mimic-code, vitalsign.sql): heart rate 20–300, systolic blood pressure 30–400, diastolic blood pressure 20–300, respiratory rate 4–70, and temperature 30–45°C. Values outside these ranges were set to missing rather than imputed. Laboratory values were not truncated at the SQL level, as extreme values may represent true clinical conditions (e.g., markedly elevated WBC in leukemia); Winsorization at the 1st and 99th percentiles was applied during feature engineering.

Missing Data and Feature Selection

Missing Data Handling

Albumin (74.2% missing) was excluded from the analysis following established practice in MIMIC-IV prediction studies, which commonly employ a 25–30% missingness threshold for variable exclusion (4,11). This high missingness reflects the database structure — only 8.7% of MIMIC-IV chemistry table entries contain albumin values — rather than systematic clinical selection, as albumin is not a routine measurement in acute ICH care.

A critical distinction was made between two classes of missing data in the landmark framework. **Structural missingness** — defined as a variable's absence from the database across all or most patients regardless of time point — was addressed by the 25–30% exclusion threshold. Albumin (74.2% missing across all windows) exemplified this class, reflecting the fact that only 8.7% of MIMIC-IV chemistry records contain albumin measurements. In contrast, **temporal missingness** — defined as missingness driven by the limited observation window at early landmark time points — affected laboratory variables whose collection depends on clinical timing. At the 6h window, several laboratory features exhibited missing rates of 36–44%, substantially exceeding the exclusion threshold.

However, because these same variables had acceptable missing rates at later windows (e.g., 17–18% at 24h), the elevated early-window missingness was attributed to the short pre-analytical interval — blood draws simply had not yet occurred — rather than to any systematic unavailability in the database. These temporally missing variables were therefore retained across all windows, with their missing values handled uniformly through Multiple Imputation by Chained Equations (MICE), which borrows information from correlated variables to produce plausible imputations. This two-class approach ensured that exclusion decisions were driven by data availability rather than by the study design itself. Per-window missing rates for all candidate variables are provided in Additional file 1: Table S1. This approach is supported by evidence that the proportion of missing data should not be used as the sole criterion for decisions on multiple imputation (12), and by established guidance emphasizing that the imputation model should include all variables predictive of missingness — in our case, the richly measured vital signs (99.6% complete) and GCS components (98% complete) served as strong auxiliary anchors for laboratory imputation (13).

Remaining missing values were imputed using MICE with Bayesian Ridge regression as the base estimator, implemented via scikit-learn's IterativeImputer (max_iter=10, random_state=42) (14). Categorical variables with missing values were imputed using mode imputation.

Feature Selection

Feature selection was performed independently for each landmark time window using a two-stage approach. In the first stage, evidence-based pre-screening reduced the initial 87 variables to 42 candidates by excluding: identifiers and outcome variables; variables with excessive missingness (albumin); redundant aggregate scores (ICH Score components, SOFA score); highly collinear laboratory max/first values (Pearson $r > 0.87$ with mean); redundant vital sign aggregations; and treatment variables excluded to avoid collider bias (see Model Development section).

In the second stage, LASSO regression with 5-fold stratified cross-validation was applied, using the 1-standard-error (1-SE) rule to select the most parsimonious model within one standard error of the minimum cross-validated mean squared error (15). This yielded window-specific feature sets of varying sizes (6h: 13 features; 12h: 22 features; 24h: 20 features; 48h: 22 features), reflecting the increasing information available at later time points.

Three recently proposed composite indices — estimated plasma volume status (ePVS), glucose-to-potassium ratio (GPR), and platelet-to-WBC ratio (PWR) — were evaluated as candidate features in the first screening stage. All three were eliminated by LASSO regularization (coefficients shrunk to zero), consistent with their constituent variables already being represented in the candidate pool.

Model Development

Four classification algorithms were trained independently at each landmark time window:

1. **XGBoost** (primary model): Hyperparameters were optimized via two-stage grid search — a coarse search (162 combinations) followed by a fine search (19,683 combinations per window) — optimizing for AUROC under 5-fold stratified cross-validation. The optimal configurations consistently favored shallow trees (max_depth 2–6), low learning rates (0.015–0.06), and strong regularization (min_child_weight 5–12, subsample 0.5–0.8). Scale_pos_weight was dynamically set to the inverse class ratio (n_negative/n_positive ≈ 3.2–3.9) to address class imbalance. XGBoost served as the primary model for SHAP-based interpretability analysis (8).

2. **LightGBM** (sensitivity analysis): Hyperparameters were optimized via a two-stage grid search strategy identical to XGBoost — a coarse search followed by a fine-grained search around the best coarse parameters — optimizing for

AUROC under 5-fold stratified cross-validation. Scale_pos_weight was dynamically set to the inverse class ratio. LightGBM was considered consistent with XGBoost if the AUROC difference was < 0.02.

3. **Random Forest** (tree-based baseline): Configured with class_weight='balanced', n_estimators=200, and max_depth=6, serving as a bagging-based tree model reference.

4. **Logistic Regression** (traditional baseline): Fitted with L2 regularization (C=1.0, solver=lbfgs, max_iter=2000) on standardized features, with class_weight='balanced'. This provided a clinically familiar reference framework and enabled direct comparison between linear and non-linear modeling approaches.

Treatment variable exclusion: Treatment variables (mannitol use, vasopressor use, anticoagulant therapy, mechanical ventilation, renal replacement therapy, and surgical intervention) were deliberately excluded from the primary models. These variables constitute colliders in the causal pathway — treatment decisions are influenced by disease severity and simultaneously affect outcomes — and their inclusion would introduce collider bias that inflates apparent predictive performance. Future work by our group will specifically incorporate treatment variables within a causal inference framework to disentangle their prognostic effects.

Deep learning exclusion: Deep learning approaches (RNN/LSTM/Transformer) were not employed for three reasons: (1) the cohort size (n=2,298) is insufficient for reliable deep learning training, as comparable time-series DL studies have used samples exceeding 40,000 (16,17); (2) ICU time-series data are highly sparse, with most laboratory tests measured only 1–3 times within 24 hours, meaning the majority of "time steps" in RNN-based models would be LOCF-imputed values rather than genuine observations; and (3) our primary objective was clinical interpretability via SHAP, which is more mature and reliable in the traditional ML framework.

Model Evaluation

Model performance was evaluated using 5-fold stratified cross-validation with out-of-fold predictions, assessed across four dimensions:

Discrimination: The primary metric was the AUROC, reported with both DeLong 95% confidence intervals and bootstrap 95% confidence intervals (1,000 resamples) to provide dual validation of the uncertainty estimates. Additional discrimination metrics included sensitivity, specificity, precision, F1 score, accuracy, and Brier score.

Calibration: Calibration curves were generated by plotting observed event rates against predicted probabilities across deciles of predicted risk. Quantitative calibration was assessed using the calibration intercept (ideal = 0, indicating no systematic over- or under-prediction) and calibration slope (ideal = 1, indicating correct spread of predicted probabilities).

Clinical utility: Decision curve analysis (DCA) was performed to evaluate whether model-guided clinical decisions would yield net benefit — i.e., whether the reduction in false positives (overtreatment) at a given risk threshold outweighs the cost of missed true positives (undertreatment) — relative to two default strategies: intervening on all patients or intervening on none (18). The model is considered clinically useful if its net benefit curve lies above both default strategies across a clinically relevant range of decision thresholds. The threshold probability range was displayed from 0 to 0.40, reflecting the clinically plausible range for acute ICH management decisions. Thresholds beyond 0.40 correspond to scenarios where no clinician would defer intervention — it is implausible that a patient or clinician would demand a mortality risk exceeding 40% before initiating neurocritical care interventions — and net benefit estimates become unreliable due to the low number of patients exceeding such extreme thresholds (19).

Interpretability: SHAP (SHapley Additive exPlanations) values (8) were computed for the XGBoost model at each landmark window using TreeExplainer, generating both summary bar plots (global feature importance) and beeswarm plots (feature-level value distributions). Cross-window comparison of SHAP feature rankings was performed to identify stable predictors and characterize the temporal evolution of feature importance. Features appearing among the top-ranked contributors (i.e., within the full set of LASSO-selected features for that window) across all four landmark windows were designated as "cross-window stable features," with their exact count determined by set intersection of the four window-specific feature sets rather than an arbitrary threshold.

Statistical Analysis and Reporting

All analyses were performed using Python 3.10.20 with scikit-learn 1.7.2, XGBoost 1.7.6, LightGBM 4.6.0, and SHAP 0.49.1 libraries. The source code is publicly available as described in the Data Availability statement. Statistical significance was set at $\alpha = 0.05$. Continuous variables were compared using the Mann-Whitney U test; categorical variables were compared using the chi-square test. This study was reported following the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence) guidelines (20).

Results

Study Population and Baseline Characteristics

A total of 2,298 ICU patients with ICH were included in the study. In-hospital mortality was 23.8% overall, declining progressively across landmark windows — from 23.4% (533/2,279) at 6h, to 22.4% (498/2,222) at 12h, to 21.3% (423/1,982) at 24h, and to 20.4% (313/1,537) at 48h — reflecting the conditioning-on-survival design. The decreasing denominator across windows reflects the progressive exclusion of patients who died or were discharged from the ICU before the landmark. Baseline characteristics of the 6h landmark cohort are presented in Table 1.

Model Discrimination

The four machine learning models achieved consistently strong discrimination across all landmark windows. The XGBoost model achieved AUROC values of 0.856 (95% CI: 0.839–0.873), 0.872 (0.855–0.888), 0.864 (0.843–0.881), and 0.857 (0.836–0.879) at the 6h, 12h, 24h, and 48h landmarks, respectively. Full performance metrics for XGBoost are presented in Table 2; performance of LightGBM, Random Forest, and Logistic Regression is provided in Additional file 2: Table S2. Performance followed an inverted-U trajectory — rising from 6h to a peak at 12h, then gradually declining through 24h and 48h — a pattern consistently observed across all four model types (Figure 1). The LightGBM sensitivity analysis demonstrated performance comparable to XGBoost (ΔAUROC $\leq$ 0.011), confirming the robustness of the findings to the choice of gradient boosting implementation.

Table 2. XGBoost model performance across four landmark time windows

Time Window	n	AUROC (95% CI)	Sensitivity	Specificity	PPV	F1 Score	Brier Score	Calibration Slope	Calibration Intercept
6h	2279	0.856 (0.839–0.873)	0.792	0.770	0.513	0.622	0.151	0.981	-1.074
12h	2222	**0.872 (0.855–0.888)**	0.787	0.805	0.540	0.640	0.137	0.948	-0.975
24h	1982	0.864 (0.843–0.881)	0.780	0.781	0.493	0.604	0.145	0.925	-1.111
48h	1537	0.857 (0.836–0.879)	0.696	0.840	0.528	0.600	0.129	0.897	-0.809

XGBoost model performance across four landmark time windows. AUROC = area under the receiver operating characteristic curve; 95% CI by DeLong method. Sensitivity, specificity, PPV, and F1 score calculated at the optimal threshold determined by Youden's index. Brier score ranges from 0 (perfect) to 1 (worst). Calibration slope (ideal = 1) and intercept (ideal = 0) derived from decile-based calibration curves. Bold = best-performing window (12h). Performance of LightGBM, Random Forest, and Logistic Regression in Additional file 2: Table S2.

Figure 1. Model discrimination across four landmark time windows

AUROC (with DeLong 95% CI error bars) for XGBoost, LightGBM, Random Forest, and Logistic Regression at the 6h, 12h, 24h, and 48h landmarks. The 12h window achieves peak discrimination (XGBoost: 0.872 [0.855–0.888]) across all four model types.

Model Calibration

Calibration curves are shown in Figure 2, with quantitative calibration metrics in Table 3. XGBoost demonstrated the best overall calibration, with calibration slopes of 0.98, 0.95, 0.93, and 0.90 at the 6h, 12h, 24h, and 48h windows, respectively, approaching the ideal value of 1. Calibration intercepts were modestly negative (−0.81 to −1.07), reflecting a slight tendency toward over-estimation of absolute mortality risk. Logistic Regression showed slightly less favorable calibration (slopes 0.87–0.93), while LightGBM and Random Forest exhibited greater departures from ideal calibration (see Table 3 for full metrics).

Figure 2. Calibration curves for four machine learning models at each landmark time window

Panels: a) 6h, b) 12h, c) 24h, d) 48h. Observed event rates vs. predicted probabilities (deciles). Dashed diagonal = perfect calibration. Note: XGBoost calibration slopes (0.90–0.98) approach the ideal value of 1; however, the points lie systematically below the diagonal because the calibration intercept is negative (−0.81 to −1.07), reflecting modest

over-estimation of absolute risk. This is expected in imbalanced prediction tasks and does not impair the relative risk ranking (AUC).

Table 3. Calibration metrics for four machine learning models across four landmark time windows

Model	Window	Intercept (ideal = 0)	Slope (ideal = 1)	Brier Score (lower is better)
XGBoost	6h	−1.074	**0.981**	0.151
XGBoost	12h	−0.975	**0.948**	0.137
XGBoost	24h	−1.111	0.925	0.145
XGBoost	48h	−0.809	0.897	0.129
LightGBM	6h	−0.993	0.916	0.148
LightGBM	12h	−0.749	0.800	0.132
LightGBM	24h	−0.859	0.822	0.137
LightGBM	48h	−0.774	0.806	0.131
Random Forest	6h	−0.885	1.020	0.143
Random Forest	12h	−0.845	1.147	0.137
Random Forest	24h	−0.876	1.109	0.139
Random Forest	48h	−0.784	1.316	0.137
Logistic Reg.	6h	−1.163	0.931	0.156
Logistic Reg.	12h	−1.215	0.922	0.149
Logistic Reg.	24h	−1.268	0.890	0.156
Logistic Reg.	48h	−1.308	0.870	0.157

Calibration intercept and slope derived from decile-based calibration curves. Negative intercept indicates systematic overestimation of absolute risk; slope <1 indicates under-spread predictions. Brier score ranges from 0 (perfect) to 1 (worst). Bold = best slope per window (closest to ideal value of 1).

SHAP Feature Importance and Cross-Window Analysis

SHAP analysis of the XGBoost model identified 32 unique features selected by LASSO across the four windows. Through set intersection of the four window-specific feature sets, **seven cross-window stable features** were identified: GCS motor minimum, age, maximum glucose, RDW mean, minimum diastolic blood pressure, liver disease, and age < 65 years — features that retained prognostic importance irrespective of the assessment time point.

The more instructive finding was the temporal shift in feature importance rankings (Table 4, Figure 3). At the 6h and 12h windows, neurological features — GCS components (motor, eye, verbal, total) — dominated the importance landscape. By the 48h window, systemic and metabolic features ascended in the SHAP rankings: maximum sodium (rank 5) and maximum potassium (rank 9) became prominent contributors. Platelet dynamics exhibited a distinctive

window-dependent pattern, with thrombocytopenia prominent at 12h (platelet mean, rank 5) and reactive thrombocytosis emerging at 48h (platelet maximum, rank 3).

Table 4. SHAP feature importance (mean |SHAP|) of the XGBoost model across four landmark time windows

Feature	6h	12h	24h	48h
gcs_total_min	0.551	0.454	0.547	-
gcs_motor_min	**0.380**	**0.464**	**0.272**	**0.570**
gcs_total_first	0.250	0.316	0.154	-
age	**0.202**	**0.292**	**0.367**	**0.210**
gcs_motor_first	0.196	-	0.160	0.227
wbc_min	0.192	0.116	-	0.162
rdw_mean	**0.190**	**0.142**	**0.111**	**0.130**
glu_max	**0.125**	**0.124**	**0.280**	**0.314**
dbp_min	**0.120**	**0.168**	**0.164**	**0.115**
age_lt65	**0.093**	**0.045**	**0.068**	**0.073**
age_ge80	-	0.027	-	-
cr_max	0.083	0.117	0.086	-
glu_min	0.073	0.133	-	-
liver_disease	**0.063**	**0.012**	**0.009**	**0.012**
plt_mean	-	0.238	-	-
inr_min	-	0.199	0.118	0.133
lnr_mean	-	-	0.086	-
wbc_mean	-	0.140	-	-
k_max	-	0.128	0.230	0.182
spo2_mean	-	0.126	0.143	0.110
gcs_eye_min	-	0.126	0.161	0.441
na_max	-	0.122	0.225	0.286
na_min	-	-	-	0.040
hr_mean	-	0.120	-	0.110
hr_max	-	0.055	0.027	0.069
plt_min	-	-	0.253	-
wbc_max	-	-	0.135	-
plt_max	-	-	-	0.335
rr_mean	-	-	-	0.184
cr_mean	-	-	-	0.139
dbp_max	-	-	-	0.113

sbp_min	-	-	-	0.088

Values represent mean |SHAP| (mean absolute Shapley value) of the XGBoost model. Empty cells indicate the feature was not selected by LASSO for that specific time window. Bold italic = cross-window stable feature present in all four windows.

Figure 3. Cross-window SHAP feature importance of the XGBoost model

Heatmap display: features (rows) × time windows (columns). Color intensity = mean |SHAP| value. The seven cross-window stable features (red labels, left) remain prognostic across all windows, while the shift from neurological (GCS components) to systemic-metabolic (sodium, potassium) features mirrors the ICH pathophysiological trajectory. Per-window bar plots and beeswarm plots provided in Additional file 4, 5: Figures S2–S3. The temporal trajectory of key feature importance is tracked in Figure 4.

GCS motor minimum ranked 2 at 6h, 1 at 12h, 4 at 24h (behind GCS total minimum, age, and maximum glucose), and 1 at 48h. The temporary decline at 24h coincides with the peak importance of age and metabolic markers. Neurological features maintained absolute dominance of SHAP importance at all windows (GCS motor minimum = 0.570 at 48h, the single highest value across all windows), while metabolic and systemic features joined as secondary contributors at later windows — reflecting the pathophysiological extension from primary brain injury to multi-organ involvement rather than a transfer of prognostic dominance. The temporal trajectory of feature importance is summarized in Figure 4.

Figure 4. Temporal trajectory of SHAP feature importance across landmark time windows

Line plot of mean |SHAP| for six representative features across the 6h, 12h, 24h, and 48h landmarks. Solid lines = cross-window stable features; dashed lines = features selected in a subset of windows. Neurological features (GCS motor, GCS eye) in warm colors maintain absolute importance at all windows; metabolic features (glucose, sodium, potassium) in cool colors ascend as secondary contributors at later windows, reflecting the addition of systemic complications alongside enduring neurological primacy.

Decision Curve Analysis

Decision curve analysis (Additional file 6: Figure S4) demonstrated that XGBoost yielded positive net benefit across clinically relevant threshold probabilities (~10–40%) at all four landmark windows, indicating that model-guided clinical decisions would result in more favorable net outcomes than either default strategy over this range. The sensitivity, specificity, and PPV values reported in Table 2 reflect the Youden-optimal threshold, which provides a balanced trade-off for standardized reporting; in clinical practice, the decision threshold may be adjusted — lower thresholds for screening, higher thresholds for definitive treatment decisions — depending on the clinical context and the consequences of false positives versus false negatives.

Discussion

Principal Findings

This study applied a multi-time-window landmark modeling framework to develop and evaluate four machine learning models for dynamic in-hospital mortality prediction in a cohort of 2,298 ICU patients with spontaneous intracerebral hemorrhage. The XGBoost model achieved robust discrimination across all windows, with AUROC values ranging from 0.856 (6h) to 0.872 (12h). Two principal findings emerged from the landmark framework.

First, predictive performance followed an inverted-U trajectory, peaking at the 12-hour landmark (AUROC 0.872) — a pattern consistently observed across all four model types. This finding identifies the 12-hour mark as the optimal balance point between information accumulation and patient retention for mortality prediction in ICH.

Second, SHAP analysis revealed a characteristic temporal evolution of feature importance, with seven features — GCS motor minimum, age, maximum glucose, RDW mean, minimum diastolic blood pressure, liver disease, and age < 65 years — identified as cross-window stable predictors through set intersection of the four window-specific feature rankings. Beyond these stable features, a transition from neurological dominance at early windows to systemic-metabolic determinants at later windows was observed, mapping onto the established pathophysiological trajectory of ICH from primary brain injury to multi-organ complications.

Comparison with Existing Prediction Models

All existing ML-based prediction models for ICH using the MIMIC-IV database have employed static, single-time-point designs (3–5). This observation extends beyond the ICH-specific literature. A recent systematic review and meta-analysis of 68 machine learning studies spanning all stroke types (approximately 1.76 million patients) found that virtually all models relied on static, single-time-point prediction, with the authors explicitly noting the near-absence of dynamic, time-varying prediction approaches despite their clinical importance(21). Mao et al. developed an XGBoost model using 24-hour aggregate features in a cohort of 1,596 ICH patients, achieving an AUROC of 0.907 [3]. Chen et al. reported an XGBoost model with 7 features in 1,274 non-traumatic ICH patients (AUROC 0.797), with the notable contribution of external validation [4]. Recently, Ye et al. published a study of 2,323 ICH patients from MIMIC-IV, with a focus on glucose and lipid biomarkers as predictive features (AUROC 0.851) (5).

The present study differs from these prior works in a fundamental aspect: rather than asking "can we predict mortality?" at a single time point, the landmark framework asks "how does predictive performance evolve over time, and which features matter most at each stage?" The 12-hour peak AUROC (0.872) is numerically lower than the highest reported static-model AUROC (Mao et al., 0.907), but this gap should be interpreted in the context of a key methodological difference: Mao et al. included treatment-related variables (mannitol use, anticoagulant therapy, vasopressor use, mechanical ventilation, surgical intervention) as model features. These variables constitute colliders in the causal pathway — treatment decisions are influenced by disease severity and simultaneously affect outcomes — and their inclusion can inflate apparent predictive performance through collider bias. Our models, which deliberately excluded treatment variables, achieved an AUROC range of 0.856 — 0.872 under the more informative dynamic landmark framework, remaining within the clinically useful range. This choice of tree-based ensemble methods is further supported by a recent metaregression of 68 stroke prediction studies, which found that random forest models maintained sustained predictive performance over longer time horizons, whereas logistic regression showed gradual decline — a pattern the authors attributed to the inability of static linear models to capture evolving clinical trajectories(21).

In the broader dynamic prediction landscape, Britsch et al. recently applied a recursive updating strategy to predict 48-hour mortality in a general ICU population (7). While their approach shares the motivation of dynamic risk assessment, the present study differs in three critical respects. First, our landmark framework uses independently

trained models at each window, allowing cross-window comparisons without conditioning on prior predictions. Second, we focus specifically on the ICH population, where the temporal dynamics of neurological injury and systemic complications follow a disease-specific trajectory. Third, our fixed endpoint (in-hospital mortality) maintains clinical interpretability, whereas the sliding 48-hour window in Britsch et al. predicts a different endpoint at each assessment.

Beyond the specific modeling approach, a broader trend toward temporal analysis in ICH prognostication has recently emerged. Wang et al. applied latent class growth modeling to GCS trajectories over 120 hours (n = 5,443, pooled from MIMIC-IV, eICU-CRD, and a Chinese center), identifying five GCS trajectory classes associated with distinct mortality risks (22). Zuo et al. similarly employed latent class analysis to characterize 36-hour blood glucose trajectories in ICH patients from MIMIC-IV (23). While these trajectory studies share with our work the recognition that clinical variables evolve over time and that this evolution carries prognostic information, they differ methodologically in two important respects: they analyze a single variable's trajectory with traditional regression, whereas our landmark framework simultaneously models the full panel of LASSO-selected features using machine learning, enabling the identification of feature importance shifts across time windows that single-variable trajectory methods cannot reveal. To our knowledge, the present study represents the first application of a multi-time-window landmark ML framework with a comprehensive feature set for dynamic mortality prediction in ICH.

The 12-Hour Peak: Clinical Interpretation

The AUROC differences across landmark windows are modest (ΔAUROC ≤0.016), consistent with the reporting standard in the ICH prediction literature where model comparison is based on point estimates and confidence intervals rather than formal statistical testing (3,4). The selection of the 12-hour window as the recommended decision point is supported by both the descriptive discrimination peak and its clinical actionability — by 12 hours, sufficient information has accumulated for meaningful risk stratification, while the window remains early enough to guide intervention decisions. The inverted-U performance trajectory — rising from 6h to 12h, then declining through 24h and 48h — reflects two countervailing forces. Between 6h and 12h, the accumulation of clinical information is substantial: the feature set expands from 13 to 22 LASSO-selected variables, incorporating important laboratory values (platelets, INR, potassium) that become available during this window. This information gain accompanies the improvement in discrimination from AUROC 0.856 to 0.872.

Between 12h and 48h, the incremental information gain diminishes (24h: 20 features; 48h: 22 features), while the patient composition shifts progressively toward a more homogeneous moderate-risk profile. Notably, patients who die within the first 12 hours — typically those with catastrophic hemorrhages and extreme physiological derangements, who are relatively easier to identify — are excluded from later windows. The remaining long-stay survivors present a more nuanced risk spectrum, where discrimination depends on detecting subtle systemic deterioration rather than gross neurological devastation. This compositional shift increases the difficulty of the prediction task, as reflected in the declining AUROC from 0.872 (12h) to 0.864 (24h) to 0.857 (48h).

From a clinical standpoint, the 12-hour landmark represents a practical decision point. By 12 hours, acute interventions (hemostatic therapy, blood pressure control, neurosurgical consultation) have been initiated, initial imaging is available, and sufficient laboratory data have accumulated to inform a refined risk assessment. A model prediction at this juncture can support clinical communication with families and guide decisions about escalation versus de-escalation of care intensity. Earlier prediction (6h), while possible, carries greater uncertainty due to information sparsity. Later prediction (24–48h) offers diminishing returns and is less actionable, as critical care trajectories are often already established by this point.

Taken together, the landmark framework supports a three-phase clinical interpretation of the dynamic predictors of ICH outcome. At the **6-hour window**, information is insufficient for definitive prognostication, and predictions should be treated as preliminary assessments only. The **12-hour window** represents the optimal decision point, where predictions achieve peak discrimination and can inform surgical decisions, family communication, and treatment intensity adjustments. At **24–48 hours**, the model shifts focus toward the prevention of systemic complications — hypernatremia, acute kidney injury, and electrolyte disturbances — rather than re-evaluation of the primary neurological insult(24). This phased interpretation provides clinicians with a structured approach to dynamic risk assessment throughout the acute ICU course.

SHAP Feature Evolution: From Neurological Injury to Systemic Homeostatic Failure

The cross-window SHAP analysis provides, to our knowledge, the first quantitative demonstration of the temporal evolution of prognostic determinants in ICH patients during the acute ICU phase. The seven cross-window stable features — led by GCS motor minimum — represent a core set of prognostic indicators that retain importance irrespective of the assessment time point, consistent with their well-established roles in ICH outcome.

The more instructive finding, however, is the temporal shift observed beyond these stable features. At the 6h and 12h windows, the feature importance landscape is dominated by neurological features — GCS components (motor, eye, verbal, total) — reflecting the primacy of primary brain injury as the driver of early mortality. By the 48h window, systemic and metabolic features ascend in the SHAP rankings: maximum sodium (rank 5), maximum potassium (rank 9) become prominent contributors. Platelet dynamics also exhibit a distinctive window-dependent pattern, with thrombocytopenia prominent at 12h (platelet mean, rank 5) and reactive thrombocytosis emerging at 48h (platelet maximum, rank 3).

This transition from neurological to systemic-metabolic determinants maps directly onto the established pathophysiological trajectory of severe ICH, in which primary brain injury from the initial hemorrhage gives way to secondary injury processes — perihematomal edema, inflammation, and blood-brain barrier disruption — followed by systemic complications in patients who survive the acute phase (1,25). Importantly, neurological features (GCS components) maintained absolute SHAP dominance at all windows — GCS motor minimum reached its single highest value at 48h (0.570) — indicating that the primary neurological insult remains the overriding prognostic determinant throughout the ICU course. Rather than a transfer of predictive power, the rising importance of sodium, and potassium at later windows reflects the addition of systemic complications as secondary prognostic signals alongside the enduring primacy of neurological status. This pattern is visualized in Figure 4, which tracks the temporal trajectory of key feature importance across the four landmark windows.

This finding carries clinical implications beyond prediction. It suggests that after the first 24 hours, the focus of prognostication — and potentially intervention — should shift from the primary neurological insult to the prevention and management of systemic complications. The SHAP-derived feature trajectory provides a data-driven map of this pathophysiological transition.

GCS Motor Component in Context

The prominence of the GCS motor component warrants careful interpretation. This finding aligns with the broader neurotrauma literature: the IMPACT study identified GCS motor as the most prognostically informative GCS

component in traumatic brain injury (26), and a recent large-scale analysis demonstrated that the motor component alone achieves predictive performance approaching that of the full GCS (AUROC 0.817 vs. 0.833 for in-hospital mortality) (10).

It is critical to emphasize that the prominence of the motor component in our models does not imply that motor score is superior to total GCS. Rather, in a high-dimensional feature set where GCS total and its three components (motor, eye, verbal) are simultaneously present, LASSO regularization, which selects among collinear variables, preferentially retained the component carrying the largest independent prognostic signal — the motor score. This is a feature of the regularization algorithm, not a clinical judgment about GCS scoring. The clinical advantage of the motor component is practical: motor response remains assessable in intubated or tracheostomized patients, in whom the verbal component is unobtainable, making it a particularly robust bedside prognostic indicator in the ICU setting.

The Landmark Conditioning-on-Survival: Methodological Feature, Not Flaw

An important methodological consideration concerns the progressive decline in mortality rates across landmark windows — from 23.4% at 6h to 20.4% at 48h — which reflects the conditioning-on-survival inherent to the landmark design. We argue that this is not a methodological limitation but rather the correct definition of the clinical question: each model addresses "what is the mortality risk for a patient who has survived to this time point?", which is precisely the information needed for real-time clinical decision-making.

Two observations support this interpretation. First, patients who die at very early time points are typically those with catastrophic hemorrhages and extreme feature values; their exclusion from later windows makes the prediction task harder, not easier, as the remaining patients present a more nuanced risk spectrum. The fact that the 12h model achieves peak discrimination despite this increased difficulty suggests that the information accumulation gain substantially outweighs the compositional change penalty. Second, analyzing a common patient subset across windows — while methodologically "clean" — would address a different question (marginal value of additional data in a fixed cohort) rather than the clinically relevant question (dynamic risk assessment for the patient currently in the ICU). The clinician at the 12-hour bedside needs a model calibrated for patients who have survived to 12 hours, not a model restricted to those who will ultimately survive beyond 48 hours.

Limitations

Several limitations of this study should be acknowledged. First, our models lack external validation. While we employed rigorous internal validation (5-fold stratified cross-validation with both DeLong and bootstrap confidence intervals), generalizability to other institutions, populations, and clinical practices remains to be established. A planned external validation study using an independent database will address this gap. Notably, Chen et al. demonstrated the feasibility of external validation for MIMIC-IV-derived ICH prediction models (4).

Second, the study is retrospective and based on a single-center database (BIDMC), which may limit generalizability to centers with different demographic compositions, treatment protocols, or ICU admission thresholds. Third, the MIMIC-IV database does not contain reliable imaging-derived features — particularly hematoma volume, intraventricular hemorrhage extent, and midline shift — which are established prognostic factors in ICH (27). The absence of these features may limit the absolute performance ceiling of our models and represents an inherent constraint of electronic health record-based prediction in ICH.

Fourth, treatment variables (mannitol, vasopressors, mechanical ventilation, surgical intervention) were deliberately excluded from the primary models to avoid collider bias. While this conservative approach may result in a lower AUROC ceiling compared with studies that include treatment features, it yields estimates of intrinsic prognostic value that are not artificially inflated by post-treatment information. Future work by our group will specifically incorporate treatment variables within a causal inference framework to disentangle their prognostic effects.

Importantly, the declining mortality rates across landmark windows (23.4% to 20.4%) reflect the conditioning-on-survival design and have been addressed in detail above. Two additional methodological choices should be noted for transparency: the exclusion of deep learning approaches, and the use of grid search without nested cross-validation. These were deliberate design decisions driven by sample size considerations and alignment with the existing ICH prediction literature, as detailed in the Methods; alternative approaches remain open for future investigation.

Conclusions

This multi-time-window landmark modeling framework reveals two key dimensions of mortality prediction in ICH: first, the 12-hour post-admission mark represents the optimal balance point between accumulating clinical information and patient retention for dynamic risk assessment; second, the temporal evolution of SHAP-derived feature importance — from neurological injury markers at early windows to systemic metabolic determinants at later windows — provides a data-driven map of the pathophysiological progression from primary brain injury to multi-organ complications. These findings demonstrate that dynamic, time-specific prediction is both feasible and clinically informative in the ICH ICU setting. External validation and prospective evaluation are warranted to establish the clinical utility of landmark-based dynamic prediction models in routine practice.

Abbreviations

1-SE
One standard error
AUROC
Area under the receiver operating characteristic curve
BIDMC
Beth Israel Deaconess Medical Center
CI
Confidence interval
CITI
Collaborative Institutional Training Initiative
COPD
Chronic obstructive pulmonary disease
DCA
Decision curve analysis
GCS
Glasgow Coma Scale
ICD-9-CM/10-CM
International Classification of Diseases, 9th/10th Revision, Clinical Modification
ICH
Intracerebral hemorrhage
ICU

Intensive care unit
IMPACT
International Mission for Prognosis and Analysis of Clinical Trials (in TBI)
INR
International Normalized Ratio
IQR
Interquartile range
LASSO
Least Absolute Shrinkage and Selection Operator
LightGBM
Light Gradient Boosting Machine
LSTM
Long Short-Term Memory
MICE
Multiple Imputation by Chained Equations
MIMIC-IV
Medical Information Mart for Intensive Care IV
ML
Machine learning
LOCF
Last Observation Carried Forward
PPV
Positive predictive value
RDW
Red cell distribution width
RNN
Recurrent neural network
SHAP
SHapley Additive exPlanations
SOFA
Sequential Organ Failure Assessment
SpO2
Peripheral oxygen saturation
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence
WBC
White blood cell
XGBoost
eXtreme Gradient Boosting

Declarations

Ethics approval and consent to participate

The MIMIC-IV database is a publicly available, de-identified database. Access was approved by the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC) Institutional Review Boards. The study was conducted in accordance with the Declaration of Helsinki. The requirement for informed consent was waived due to the retrospective, de-identified nature of the data.

Consent for publication

Not applicable.

Availability of data and materials

The MIMIC-IV v3.1 database is publicly available via PhysioNet (https://physionet.org/content/mimiciv/). Access requires completion of a data use agreement and CITI "Data or Specimens Only Research" training. The analysis code will be made publicly available upon acceptance. Requests for code access during the review process can be directed to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

[This work was supported by the Wu Jieping Medical Foundation Clinical Research Special Fund (Grant No. 320.6750.2024-25-4). The present study on mortality prediction in ICH patients falls within the broader scope of the funded project on neurological critical care.]

Authors' contributions

[C.M. conceived and designed the study, performed data extraction and cleaning, developed the machine learning models, conducted the statistical analysis, generated all figures and tables, and drafted the manuscript. J.Z. and G.S. contributed to clinical interpretation of results. Z.Z. contributed to data quality checking. Y.M. reviewed the methodological framework. Z.F. supervised the study and critically revised the manuscript. All authors read and approved the final manuscript.]

Acknowledgements

We acknowledge the MIT Laboratory for Computational Physiology and the BIDMC for creating and maintaining the MIMIC-IV database.

References

1. Sheth KN. Spontaneous intracerebral hemorrhage. N Engl J Med. 2022 Oct 27;387(17):1589–96. doi:10.1056/NEJMra2201449
2. Xu L, Wang Z, Wu W, Li M, Li Q. Global, regional, and national burden of intracerebral hemorrhage and its attributable risk factors from 1990 to 2021: results from the 2021 Global Burden of Disease Study. BMC Public Health. 2024 Sep 6;24(1). doi:10.1186/s12889-024-19923-7
3. 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 Jun 20;14(1):14195. doi:10.1038/s41598-024-65128-8

4. Chen Y, Liu X, Yan M, Wan Y. Death risk prediction model for patients with non-traumatic intracerebral hemorrhage. BMC Med Inform Decis Mak. 2025 Jan 22;25(1):35. doi:10.1186/s12911-025-02865-4
5. Ye Q, Xue T, Zhang Y, Xu Y, He Y, Song J, et al. Predicting mortality in non-traumatic intracerebral hemorrhage with glucose and lipid data. Front Endocrinol (Lausanne). 2026 Apr 29;17. doi:10.3389/fendo.2026.1781456
6. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. Scand J Stat. 2007;34(1):70–85. doi:10.1111/j.1467-9469.2006.00529.x
7. Britsch S, Britsch M, Lindner S, Hahn L, Schneider-Lindner V, Helbing T, et al. An interpretable machine learning algorithm enables dynamic 48-hour mortality prediction during an ICU stay. Commun Med. 2025 Oct 15;5:426. doi:10.1038/s43856-025-01192-z PubMed PMID: 41094092; PubMed Central PMCID: PMC12528449.
8. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017. p. 4765–74.
9. 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 Jan 3;10(1):1. doi:10.1038/s41597-022-01899-x
10. Lee SJ, Chen YL, Wu TH, Liu CY, Wang CH, Tsai CH, et al. Comparison of Glasgow coma scale, motor component, eye component, and simplified motor scale for predicting trauma outcomes: a 13-year multicenter retrospective cohort study. BMC Emerg Med. 2025 May 30;25:86. doi:10.1186/s12873-025-01246-4 PubMed PMID: 40442594; PubMed Central PMCID: PMC12123894.
11. Sun Y, He Z, Ren J, Wu Y. Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest: a retrospective analysis of MIMIC-IV database based on machine learning. BMC Anesthesiol. 2023 May 25;23(1):178. doi:10.1186/s12871-023-02138-5
12. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019 Jun;110:63–73. doi:10.1016/j.jclinepi.2019.02.016
13. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–99. doi:10.1002/sim.4067
14. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393. doi:10.1136/bmj.b2393 PubMed PMID: 19564179.
15. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction [Internet]. 2nd ed. New York: Springer; 2009. Available from: https://doi.org/10.1007/978-0-387-84858-7 doi:10.1007/978-0-387-84858-7
16. Deng Y, Liu S, Wang Z, Wang Y, Jiang Y, Liu B. Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients. Front Med (Lausanne). 2022 Sep 28;9. doi:10.3389/fmed.2022.933037
17. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018 Apr 17;8(1):6085. doi:10.1038/s41598-018-24271-9
18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov;26(6):565–74. doi:10.1177/0272989X06295361
19. Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. doi:10.1186/s41512-019-0064-7
20. 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 Apr 16;385:e078378. doi:10.1136/bmj-2023-078378

21. Chen Y, Ou Z, Deng Y, Luo A, Li X, Yang Y, et al. Predictive value of machine learning for poststroke mortality risk: systematic review and Meta-Analysis. J Med Internet Res. 2026 Apr 2;28:e83821. doi:10.2196/83821
22. Wang J. Association of latent class growth model-derived Glasgow Coma Scale trajectories with in-hospital mortality in intracerebral hemorrhage. Zhonghua Yi Xue Za Zhi. 2026;106(5):427–33. PubMed PMID: 41611275.
23. Zuo H, Zuo X, Zhang Y, Zheng W. Prognostic value of blood glucose trajectories in critically ill patients with intracerebral hemorrhage: a retrospective cohort study. PLoS One. 2026 Feb 24;21(2):e0342745. doi:10.1371/journal.pone.0342745
24. Husain-Syed F, Takeuchi T, Neyra JA, Ramírez-Guerrero G, Rosner MH, Ronco C, et al. Acute kidney injury in neurocritical care. Crit Care. 2023 Sep 3;27(1):341. doi:10.1186/s13054-023-04632-1
25. Magid-Bernstein J, Girard R, Polster S, Srinath A, Romanos S, Awad IA, et al. Cerebral hemorrhage: pathophysiology, treatment, and future directions. Circ Res. 2022 Apr 15;130(8):1204–29. doi:10.1161/CIRCRESAHA.121.319949
26. Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, et al. IMPACT database of traumatic brain injury: design and description. J Neurotrauma. 2007 Feb;24(2):239–50. doi:10.1089/neu.2006.0036
27. Hemphill JC, Bonovich DC, Besmertis L, Manley GT, Johnston SC. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. Stroke. 2001 Apr;32(4):891–7. doi:10.1161/01.STR.32.4.891

Figures

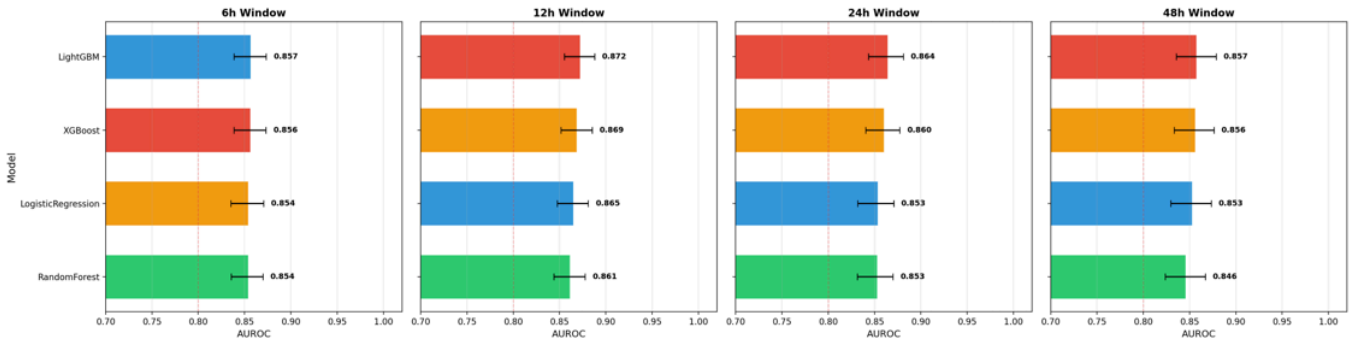


Figure 1

Model discrimination across four landmark time windows

AUROC (with DeLong 95% CI error bars) for XGBoost, LightGBM, Random Forest, and Logistic Regression at the 6h, 12h, 24h, and 48h landmarks. The 12h window achieves peak discrimination (XGBoost: 0.872 [0.855–0.888]) across all four model types.

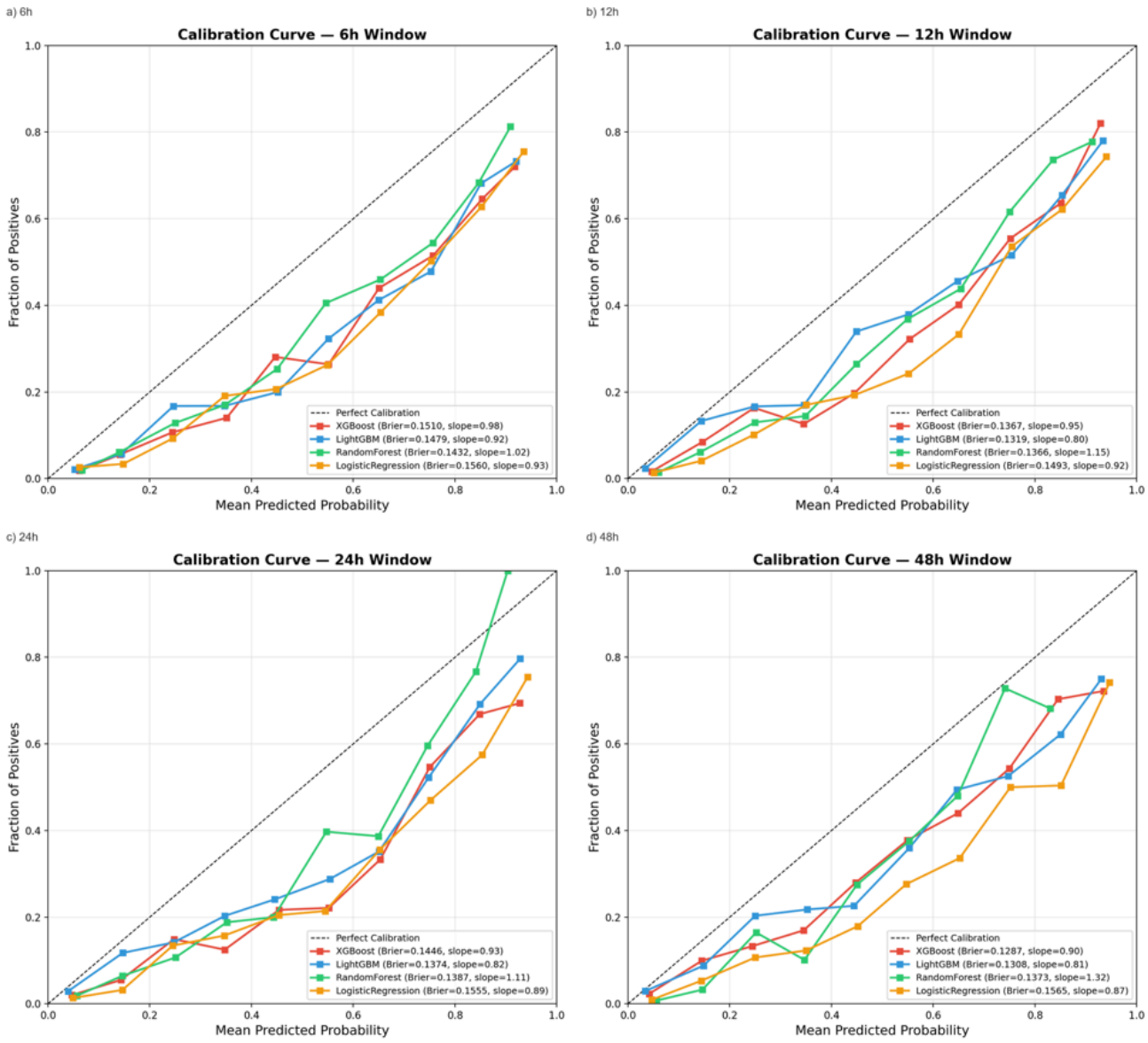


Figure 2

Calibration curves for four machine learning models at each landmark time window

Panels: a) 6h, b) 12h, c) 24h, d) 48h. Observed event rates vs. predicted probabilities (deciles). Dashed diagonal = perfect calibration. Note: XGBoost calibration slopes (0.90–0.98) approach the ideal value of 1; however, the points lie systematically below the diagonal because the calibration intercept is negative (−0.81 to −1.07), reflecting modest over-estimation of absolute risk. This is expected in imbalanced prediction tasks and does not impair the relative risk ranking (AUC).

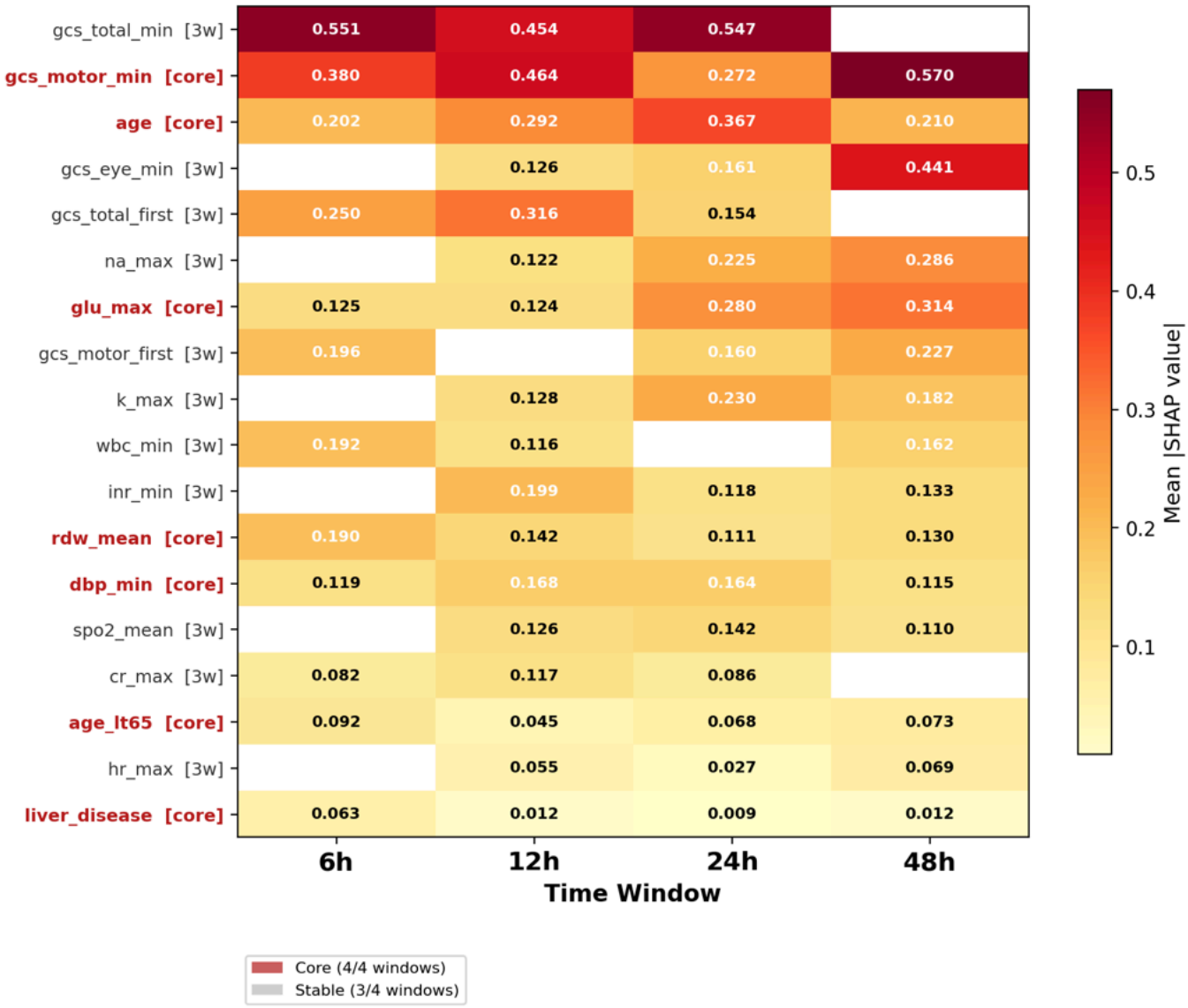


Figure 3

Cross-window SHAP feature importance of the XGBoost model

Heatmap display: features (rows) × time windows (columns). Color intensity = mean |SHAP| value. The seven cross-window stable features (red labels, left) remain prognostic across all windows, while the shift from neurological (GCS components) to systemic-metabolic (sodium, potassium) features mirrors the ICH pathophysiological trajectory. Per-window bar plots and beeswarm plots provided in Additional file 4, 5: Figures S2–S3. The temporal trajectory of key feature importance is tracked in Figure 4.

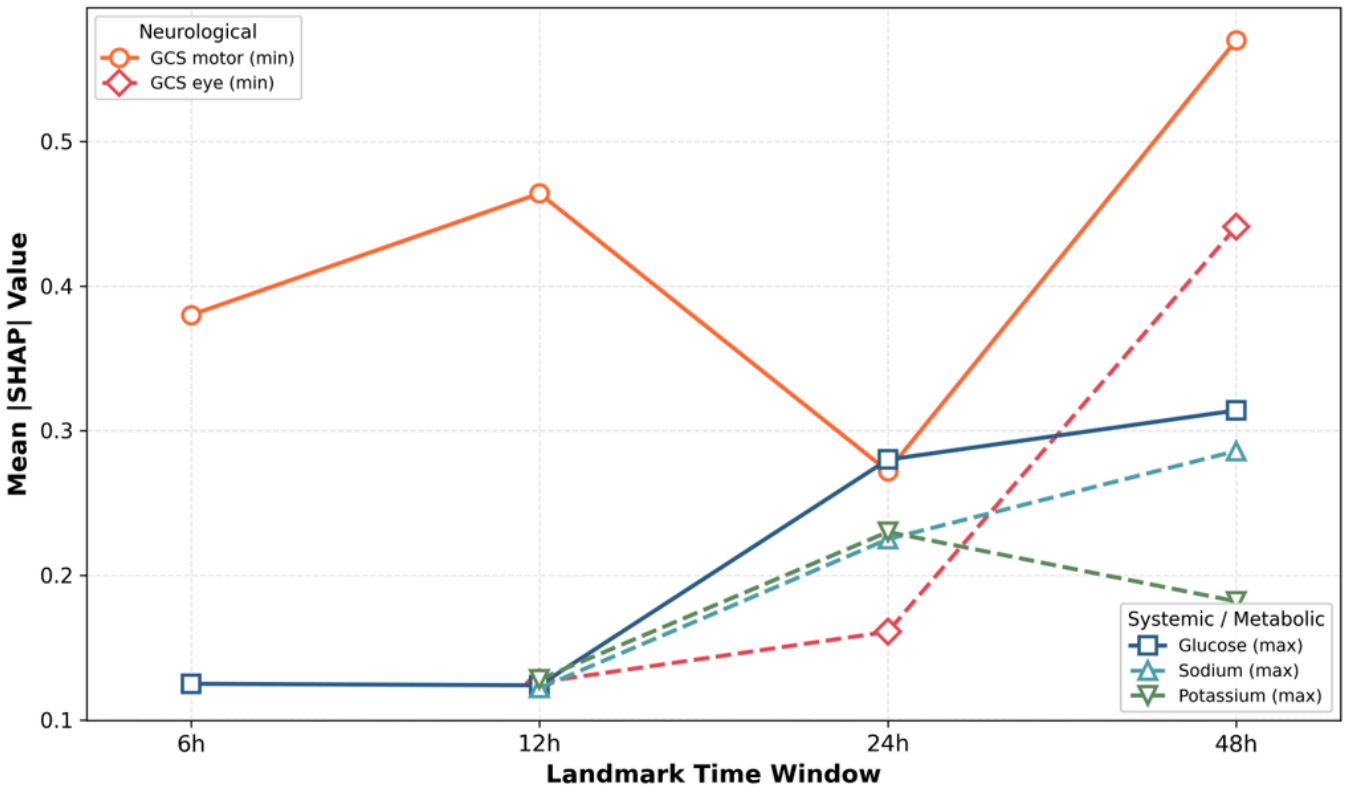


Figure 4

Temporal trajectory of SHAP feature importance across landmark time windows

Line plot of mean |SHAP| for six representative features across the 6h, 12h, 24h, and 48h landmarks. Solid lines = cross-window stable features; dashed lines = features selected in a subset of windows. Neurological features (GCS motor, GCS eye) in warm colors maintain absolute importance at all windows; metabolic features (glucose, sodium, potassium) in cool colors ascend as secondary contributors at later windows, reflecting the addition of systemic complications alongside enduring neurological primacy.

Supplementary Files

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

- Additionalfile4FigureS2.docx
- Additionalfile3FigureS1.docx
- Additionalfile6FigureS4.docx
- Additionalfile5FigureS3.docx
- Additionalfile1TableS1.docx
- Additionalfile7TRIPODAIChecklist.docx
- Additionalfile2TableS2.docx