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.

Development and Internal Validation of an Interpretable Machine Learning Model for Predicting Postoperative Delirium in Older Adults Undergoing Orthopaedic Surgery: A Prospective Cohort Study

Yongle Pei
Xinjiang Medical University
Weilong Wang
Xinjiang Medical University
Jiawen Sun
Xinjiang Medical University
Wenyan Yuan
Xinjiang Medical University
Mingxia Song
Xinjiang Medical University
Xueyin Wang
Xinjiang Medical University
Yuanyuan Liu
185099332120163.com

The First Affiliated Hospital of Xinjiang Medical University

Research Article

Keywords: postoperative delirium, machine learning, risk prediction model, SHAP analysis, hip and knee replacement and revision surgery

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11256888/v1

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

Additional Declarations: No competing interests reported.

Abstract

Objective

Postoperative delirium (POD) is a common and serious complication following hip or knee replacement and revision surgery in older patients, leading to prolonged recovery, increased healthcare costs, and poor long-term outcomes. Early identification of high-risk patients remains challenging. This study aimed to develop and internally validate an interpretable machine learning (ML)-based risk prediction model for POD using multidimensional perioperative variables and to compare the predictive performance of seven ML algorithms.

Methods

This single-center prospective observational cohort study included 930 older patients (≥ 60 years) who underwent elective hip or knee replacement and revision surgery. POD was assessed using the Confusion Assessment Method (CAM) within 7 days after surgery. The least absolute shrinkage and selection operator (LASSO) regression was used for feature selection. Seven ML algorithms were developed and compared, and SHapley Additive exPlanations (SHAP) was applied for model interpretability.

Results

POD occurred in 30.0% (279/930) of patients. LASSO identified six predictors: preoperative sleep quality (Pittsburgh Sleep Quality Index, PSQI), psychological distress (Kessler Psychological Distress Scale, K10), social support (Groningen Orthopaedic Social Support Scale, GO-SSS), nutritional status (Prognostic Nutritional Index, PNI), comorbidity burden (age-adjusted Charlson Comorbidity Index, aCCI), and intraoperative physiological reserve (Surgical Apgar Score, SAS). The extreme gradient boosting (XGBoost) model performed best, with an area under the curve (AUC) of 0.922 (95% *CI*: 0.886–0.957) and a Brier score of 0.095. SHAP analysis revealed that preoperative sleep quality was the most influential predictor, showing a nonlinear positive association with POD risk and a synergistic interaction with psychological distress.

Conclusion

The XGBoost-based model integrating six readily available predictors demonstrated good predictive performance and potential clinical utility. The SHAP-based interpretability enhances model transparency, supporting early risk identification and personalized perioperative management. External validation is warranted to confirm generalizability.

Introduction

With the accelerating aging of the population and the rising prevalence of sedentary lifestyles and suboptimal weight management, the incidence of hip and knee joint diseases is increasing steadily. Consequently, the number of patients requiring primary joint arthroplasty and revision arthroplasty is projected to rise in the coming decades[1–4]. For elderly patients with multiple comorbidities, both procedures carry considerable risks. Although they improve functional independence and quality of life, the postoperative complication rate remains high, necessitating comprehensive perioperative management[4].

Postoperative delirium (POD) is a common and serious complication of anesthesia and surgery. It is a transient, acute neurocognitive disorder characterized by inattention, fluctuating consciousness, and acute cognitive dysfunction, typically occurring within the first postoperative week. POD is associated with long-term cognitive impairment, prolonged hospital stays, increased readmission rates, and elevated mortality, and it substantially increases healthcare costs[5–8]. Approximately 30%–40% of POD cases are preventable, highlighting the importance of early identification and targeted prevention[9]. However, POD is multifactorial, and early prediction remains challenging[10]. The incidence of POD after primary joint arthroplasty is approximately 18.8%–23.1%, whereas patients undergoing revision surgery may face a substantially higher risk due to longer operative time, greater blood loss, and more severe surgical trauma[11, 12].

Increasing evidence suggests that POD is influenced by multidimensional factors. Preoperative sleep disturbance has been linked to a significantly increased risk of POD, possibly through neurotransmitter imbalance, neuronal degeneration, and systemic inflammation[13]. Psychological distress, such as anxiety and depression, may compromise blood–brain barrier integrity and amplify neuroinflammatory responses[14]. Poor social support has been associated with reduced social interaction and a higher risk of delirium[15]. Malnutrition and inflammation, reflected by the Prognostic Nutritional Index (PNI), may lower the threshold for delirium onset[16]. A higher age-adjusted Charlson Comorbidity Index (aCCI) indicates greater comorbidity burden and reduced physiological reserve[17]. The Surgical Apgar Score (SAS) reflects intraoperative physiological reserve and has been associated with postoperative complications[18]. However, the combined predictive value of these factors for POD in elderly patients undergoing hip or knee replacement and revision surgery remains unclear.

Traditional prediction models based on logistic regression often fail to capture complex nonlinear interactions among predictors and lack interpretability, limiting their clinical utility. Machine learning (ML) offers advantages in handling multivariate, nonlinear clinical data and can extract complex patterns from large datasets[19–23]. However, studies specifically predicting POD in elderly patients undergoing hip or knee replacement and revision surgery remain limited. To achieve personalized risk management, an accurate and interpretable prediction model is needed.

This prospective observational cohort study therefore aimed to develop and internally validate an interpretable ML-based risk prediction model for POD using multidimensional perioperative variables.

Seven ML algorithms were employed, and SHapley Additive exPlanations (SHAP) was applied to enhance model interpretability, ultimately providing a potential tool for early risk stratification and personalized perioperative management.

Materials and methods

Study population

This prospective observational cohort study was approved by the Ethics Committee of the First Affiliated Hospital of Xinjiang Medical University (Approval No. K202508-30). Written informed consent was obtained from all patients. A convenience sampling method was adopted to enroll elderly patients who underwent elective hip or knee replacement and revision surgery at the First Affiliated Hospital of Xinjiang Medical University between September 2025 to September 2026. The inclusion criteria were as follows: (1) age ≥ 60 years; (2) no gender restriction; (3) American Society of Anesthesiologists (ASA) physical status classification I–III; (4) postoperative hospital stay ≥ 7 days; and (5) voluntary participation with written informed consent. The exclusion criteria were: (1) severe hepatic or renal dysfunction; (2) prior diagnosis of psychiatric disorders or long-term use of psychotropic medications; (3) preoperative delirium, or a history of neurological diseases such as stroke, epilepsy, cerebral apoplexy, or Alzheimer's disease; and (4) incomplete clinical data or refusal to cooperate with the investigation.

When constructing predictive models using machine learning methods, at least 10 outcome events per variable (EPV ≥ 10) are required to avoid overfitting[24]. The reported incidence of POD in elderly patients undergoing hip or knee surgery is 18.8%[25]. Considering a 20% loss to follow-up, and with an estimated six variables included in the model, the minimum required sample size was calculated as $10 \times 6 \div 18.8\% \times (1 + 20\%) \approx 383$ cases. Ultimately, 930 patients were enrolled.

Data collection

Demographic data included sex, age, body mass index (BMI), marital status, smoking history, drinking history, educational level, per capita monthly household income, and age-adjusted Charlson comorbidity index (aCCI). Preoperative data included blood lactate concentration, C-reactive protein, white blood cell count, prognostic nutritional index (PNI), where PNI = serum albumin (g/L) + 5 × peripheral blood lymphocyte count ($\times 10^9$/L), and neutrophil-to-lymphocyte ratio (NLR), calculated as neutrophil count / lymphocyte count. Intraoperative data included Surgical Apgar Score (SAS), intraoperative fluid infusion volume, operative time, anesthesia duration, ASA physical status classification, and blood transfusion status (yes/no). Postoperative data included admission to the intensive care unit (ICU) (yes/no), use of patient-controlled analgesia (PCA) pump (yes/no), and length of hospital stay.

Postoperative delirium (POD) was assessed using the Confusion Assessment Method (CAM) within 7 days after surgery. The CAM was originally developed by Inouye et al.[26] in 1990 based on the

diagnostic criteria for delirium from the Diagnostic and Statistical Manual of Mental Disorders, Third Edition (DSM-III). The diagnostic algorithm comprises four features: (1) acute onset and fluctuating course; (2) inattention; (3) disorganized thinking; and (4) altered level of consciousness. A diagnosis of delirium was established when features (1) and (2) were present, along with either feature (3) or feature (4).

Preoperative psychological distress, social support, and sleep quality were assessed using the Kessler Psychological Distress Scale (K10), Groningen Orthopaedic Social Support Scale (GO-SSS), and Pittsburgh Sleep Quality Index (PSQI), respectively. The K10 consists of 10 items, each rated on a 5-point Likert scale (total score 10–50), with higher scores indicating greater psychological distress[27]. The GO-SSS consists of two dimensions (social support and instrumental support), each rated on a 4-point Likert scale (total score 0–36), with higher scores indicating better perceived social support[28]. The PSQI comprises seven dimensions, each scored on a 4-point Likert scale (total score 0–21), with higher scores reflecting poorer sleep quality. A PSQI global score > 7 was defined as indicating the presence of sleep disturbance[29]. All three instruments have been previously published and validated. The Cronbach's α coefficients in this study were 0.844, 0.703, and 0.794, respectively.

A research team was established and divided into two groups with clearly defined roles. All investigators received standardized training covering the study objectives, questionnaire content, communication skills, and methods to avoid subjective bias. Group 1 was responsible for data collection and questionnaire administration. One day before surgery, the investigators explained the study purpose and significance to the patients. After written informed consent was obtained, they demonstrated the questionnaire completion procedures and provided relevant instructions. Patients were assessed using the K10, GO-SSS, and PSQI, while demographic and preoperative data were collected simultaneously. For patients unable to complete the questionnaires independently, the investigators read all items aloud and assisted them in responding. Postoperatively, this group collected intraoperative and postoperative data. Group 2 was specifically tasked with POD assessment using the CAM. They performed POD assessments twice daily, between 9:00–10:00 AM and 7:00–8:00 PM, from postoperative day 1 to day 7. A positive result on either assessment was considered diagnostic of POD. Cases with uncertain diagnoses were reviewed by the principal investigator, who made the final determination. Patients were subsequently classified into the POD group or non-POD group.

Data cleaning and entry were independently performed by two researchers, including data verification, removal of duplicate entries, and verification of outliers by cross-referencing with the original medical records. For missing data, cases with a missing rate exceeding 20% were excluded, and the remaining missing values were handled using complete-case analysis, as the proportion of missing data was low. Upon completion of data entry, all entries were verified and cross-checked by a senior investigator.

Statistical analysis

All statistical analyses and model development were performed using SPSS software (version 26.0), R software (version 4.3.3), and Python software (version 3.11.4). The normality of continuous variables was assessed using the Shapiro–Wilk test. As all continuous variables deviated from a normal distribution, they were presented as medians with interquartile ranges [M (P_{25}, P_{75})], and between-group comparisons were performed using the Mann–Whitney U non-parametric test. Categorical variables were expressed as frequencies and percentages [n (%)] and compared between groups using the χ^2 test, corrected χ^2 test, or Fisher's exact test (when the expected frequency was < 5). A two-sided $P < 0.05$ was considered statistically significant.

Model construction and evaluation

The dataset used for model development was derived from the feature predictors selected by the LASSO regression model. The least absolute shrinkage and selection operator (LASSO) regression model was applied with 10-fold cross-validation to optimize the regularization parameter λ, in order to select the optimal subset of feature predictors while simultaneously addressing multicollinearity among potential predictors. Stratified random cross-validation was employed to split the dataset into a training set and a validation set at a ratio of 7:3, ensuring a balanced distribution of POD cases across both groups. A total of seven predictive models were constructed, including logistic regression (LR), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), support vector machine (SVM), and artificial neural network (ANN).

All models underwent hyperparameter tuning using fivefold cross-validation on the training set. Model performance was evaluated on the validation set using the following metrics: area under the receiver operating characteristic curve (AUC) with its 95% confidence interval, precision, sensitivity, specificity, and F1 score. The optimal model was selected based on the highest AUC, together with good calibration and the highest net clinical benefit.

The SHapley Additive exPlanations (SHAP) method was employed to provide both global and individual interpretations of the optimal model, with positive values indicating an increased risk of POD and negative values indicating a decreased risk. Visualization included global feature importance plots, bee swarm plots, waterfall plots, and force plots. Through these visualizations, global and individual interpretations of the optimal model were achieved, thereby enhancing the model's credibility and clinical acceptability.

Results

Baseline characteristics

A total of 1,025 elderly patients scheduled for elective hip or knee replacement and revision surgery were initially screened, of whom 930 were ultimately included in the final analysis. The patient selection process is illustrated in Fig. 1. During the postoperative hospitalization period, POD occurred in 279

patients (30.0%), with 69.2% (193/279) of cases developing on the first day after surgery. The baseline characteristics of patients in the POD and non-POD groups are summarized in Table 1. Significant differences were observed between the two groups in terms of marital status, smoking history, drinking history, educational level, income, intraoperative blood transfusion, operation time, anesthesia time, ASA classification, postoperative ICU admission, postoperative analgesia use, aCCI score, PNI, CRP, WBC, SAS score, intraoperative fluid volume, age, length of hospital stay, K10 score, GO-SSS score, and PSQI score (all $P < 0.05$).

Table 1
Baseline characteristics of elderly patients undergoing hip or knee replacement and revision surgery, stratified by POD status

Variable	Total (*n* = 930)	Non POD (*n* = 651)	POD (*n* = 279)	*P* value
Sex n (%)				0.629
Male	364 (39.14)	251 (38.56)	113 (40.50)	
Female	566 (60.86)	400 (61.44)	166 (59.50)	
Marital status n (%)				< 0.001
Married	761 (81.83)	567 (87.10)	194 (69.53)	
Divorced	38 (4.09)	24 (3.69)	14 (5.02)	
Widowed/Unmarried	131 (14.09)	60 (9.22)	71 (25.45)	
Smoking history n (%)				< 0.001
No	719 (77.31)	550 (84.49)	169 (60.57)	
Yes	211 (22.69)	101 (15.51)	110 (39.43)	
Drinking history n (%)				< 0.001
No	742 (79.78)	556 (85.41)	186 (66.67)	
Yes	188 (20.22)	95 (14.59)	93 (33.33)	
Educational level n (%)				0.016
Primary school or below	58 (6.24)	43 (6.61)	15 (5.38)	
Junior high school	520 (55.91)	342 (52.53)	178 (63.80)	
High school/Vocational	148 (15.91)	110 (16.90)	38 (13.62)	
College or above	204 (21.94)	156 (23.96)	48 (17.20)	
Income level n (%)				< 0.001
Low	190 (20.43)	106 (16.28)	84 (30.11)	
Middle	398 (42.80)	271 (41.63)	127 (45.52)	
High	301 (32.37)	239 (36.71)	62 (22.22)	
Very high	41 (4.41)	35 (5.38)	6 (2.15)	
Intraoperative blood transfusion n (%)				< 0.001
No	465 (50.00)	361 (55.45)	104 (37.28)	

Variable	Total (*n* = 930)	Non POD (*n* = 651)	POD (*n* = 279)	*P* value
Yes	465 (50.00)	290 (44.55)	175 (62.72)	
Operation time n (%)				< 0.001
< 3 hours	536 (57.63)	440 (67.59)	96 (34.41)	
≥ 3 hours	394 (42.37)	211 (32.41)	183 (65.59)	
Anesthesia time n (%)				< 0.001
< 4 hours	589 (63.33)	471 (72.35)	118 (42.29)	
≥ 4 hours	341 (36.67)	180 (27.65)	161 (57.71)	
ASA classification n (%)				< 0.001
I	161 (17.31)	137 (21.04)	24 (8.60)	
II	667 (71.72)	475 (72.96)	192 (68.82)	
III	102 (10.97)	39 (5.99)	63 (22.58)	
Postoperative ICU admission n (%)				< 0.001
No	844 (90.75)	607 (93.24)	237 (84.95)	
Yes	86 (9.25)	44 (6.76)	42 (15.05)	
Postoperative analgesia use n (%)				< 0.001
No	567 (60.97)	452 (69.43)	115 (41.22)	
Yes	363 (39.03)	199 (30.57)	164 (58.78)	
aCCI score *M (P25, P75)*	3.00 (2.00, 4.00)	3.00 (2.00, 3.00)	4.00 (3.00, 5.00)	< 0.001
Preoperative PNI *M (P25, P75)*	43.04 (37.00, 47.42)	45.10 (39.36, 48.72)	37.35 (34.47, 40.82)	< 0.001
Preoperative CRP *M (P25, P75)*	13.22 (6.51, 29.17)	11.50 (5.35, 21.20)	23.10 (11.32, 45.90)	< 0.001
Preoperative WBC *M (P25, P75)*	6.82 (5.49, 8.56)	6.53 (5.36, 8.35)	8.34 (6.33, 9.42)	< 0.001
SAS score *M (P25, P75)*	5.00 (3.00, 6.00)	6.00 (4.00, 7.00)	3.00 (3.00, 4.00)	< 0.001
Intraoperative fluid volume *M (P25, P75)*	1800.00 (1500.00, 2500.00)	1700.00 (1475.00, 2475.00)	1950.00 (1500.00, 2500.00)	0.014
Age *M (P25, P75)*	67.00 (62.00, 73.00)	66.00 (62.00, 70.00)	72.00 (66.00, 76.00)	< 0.001

Variable	Total (*n* = 930)	Non POD (*n* = 651)	POD (*n* = 279)	*P* value
Length of hospital stay *M (P25, P75)*	19.00 (17.00, 23.00)	18.00 (16.00, 22.00)	21.00 (18.00, 28.00)	< 0.001
Preoperative K10 score *M (P25, P75)*	17.00 (14.00, 26.00)	15.00 (14.00, 19.00)	27.00 (23.00, 31.00)	< 0.001
Preoperative GO-SSS score *M (P25, P75)*	23.00 (18.00, 26.00)	24.00 (20.00, 27.00)	18.00 (16.00, 21.00)	< 0.001
Preoperative PSQI score *M (P25, P75)*	6.00 (4.00, 12.00)	5.00 (3.00, 7.50)	13.00 (10.00, 15.00)	< 0.001

Feature selection

With POD occurrence as the dependent variable, all candidate predictors were incorporated into the LASSO regression model for feature screening. The optimal penalty parameter λ was determined via 10-fold cross-validation (Fig. 2A). Considering both model parsimony and clinical practicality, lambda.1se was selected as the optimal λ value, which retained six features with non-zero coefficients. Figure 2B illustrates the shrinkage trajectories of the variable coefficients as λ varied. The final optimal feature subset comprised preoperative sleep quality (PSQI score), psychological distress (K10 score), social support (GO-SSS score), nutritional status (PNI), comorbidity burden (aCCI score), and intraoperative physiological reserve (SAS score).

Spearman correlation analysis (Fig. 2C) revealed that the correlation coefficients among the selected variables were all at relatively low levels (all $|r| < 0.60$). To further assess multicollinearity, the variance inflation factor (VIF) was calculated for each variable. As shown in Table 2, the VIF values for the six variables were 1.353, 2.208, 1.472, 1.460, 1.488, and 1.695, respectively, all well below the conventional threshold of 5, indicating no significant multicollinearity among the predictors.

Table 2
Collinearity statistics among the six selected predictors

Variable	Tolerance	VIF
SAS score	0.590	1.695
Preoperative PNI	0.685	1.460
aCCI score	0.672	1.488
Preoperative K10 score	0.453	2.208
Preoperative GO-SSS score	0.679	1.472
Preoperative PSQI score	0.739	1.353

Note: VIF, variance inflation factor. All VIF values were well below the conventional threshold of 5, indicating no significant multicollinearity among the predictors.

Model performance

Based on the six selected feature variables, seven different machine learning algorithms were trained and evaluated. ROC curve analysis (Fig. 3A) showed that the XGBoost model achieved the highest AUC among all models, with an AUC of 0.922 (95% *CI*: 0.886–0.957) in the validation set. The AUC values for LR, DT, RF, LightGBM, SVM, and ANN were 0.916, 0.897, 0.921, 0.918, 0.906, and 0.909, respectively (Table 3). The radar plot (Fig. 4A, B) visually presented the comprehensive performance of each model across multiple metrics. XGBoost achieved the highest AUC and specificity, and showed good calibration and the highest net clinical benefit, although its sensitivity and F1 score were not the highest among all models.

Table 3
Performance of seven machine learning models in the validation set

Model	AUC (95% *CI*)	Accuracy	Precision	Sensitivity	Specificity	F1 Score
LR	0.916 (0.877–0.953)	0.874	0.779	0.807	0.903	0.793
DT	0.897 (0.859–0.937)	0.860	0.716	0.880	0.851	0.789
RF	0.921 (0.881–0.957)	0.874	0.786	0.795	0.908	0.790
XGBoost	0.922 (0.886–0.957)	0.863	0.800	0.723	0.923	0.759
LightGBM	0.918 (0.876–0.955)	0.874	0.793	0.783	0.913	0.788
SVM	0.906 (0.864–0.944)	0.853	0.776	0.711	0.913	0.742
ANN	0.909 (0.868–0.949)	0.860	0.762	0.771	0.897	0.766

Calibration analysis, as shown in Fig. 3B, demonstrated that the XGBoost model achieved a Brier score of 0.095 in the validation set, and the calibration curve indicated good agreement between predicted probabilities and observed outcomes. DCA, as presented in Fig. 3C, further suggested that the XGBoost model provided the highest net clinical benefit across the entire range of threshold probabilities.

Model interpretability

To enhance model transparency and clinical acceptability, the SHAP method was applied to interpret the optimal XGBoost model. Global feature importance (Fig. 5A) revealed that preoperative sleep quality (PSQI score) was the most influential predictor, followed by psychological distress (K10 score), social support (GO-SSS score), nutritional status (PNI), comorbidity burden (aCCI score), and intraoperative physiological reserve (SAS score). The beeswarm plot (Fig. 5B) illustrated the direction of each feature's effect: higher PSQI, K10, and aCCI scores, as well as lower GO-SSS, PNI, and SAS scores, were associated with increased POD risk.

At the individual level, waterfall plots (Fig. 5C, D) and force plots (Fig. 5E, F) demonstrated the decision-making process of the model for single-patient risk predictions, providing clinicians with intuitive insights into model reasoning.

To further explore the complex relationships between key predictors and POD risk, SHAP partial dependence plots were generated (Fig. 6). The results showed a nonlinear positive association between preoperative PSQI score and POD risk. When the PSQI score was below 7, its contribution to POD risk remained small and stable; however, once the PSQI score exceeded 10, its contribution increased sharply, suggesting that severely impaired sleep quality represents a critical threshold for substantially elevated POD risk. These thresholds should be interpreted as exploratory rather than prespecified clinical cutoffs.

Moreover, the dependence plots revealed interactive effects between variables. When examining the influence of preoperative PSQI score, the preoperative K10 score (reflecting psychological distress) played a significant moderating role. Among patients with high PSQI scores (> 10), those with concomitant high K10 scores (> 29) exhibited substantially higher POD risk than those with lower K10 scores, indicating a synergistic amplification effect of sleep disturbance and psychological distress on POD risk.

Model stability and clinical utility

The confusion matrices (Fig. 7) demonstrated that the XGBoost model achieved high classification accuracy in both the training and validation sets, with a low proportion of misclassifications, further supporting its stability. In the validation set, the XGBoost model achieved an F1 score of 0.759, indicating favorable overall diagnostic performance.

In terms of clinical application, the six predictors incorporated into the model cover six dimensions—sleep quality, psychological distress, social support, nutritional status, comorbidity burden, and

intraoperative physiological reserve—all of which are routinely available in clinical practice without incurring additional testing costs, thus offering good feasibility for bedside use. Decision curve analysis suggested potential clinical utility across a range of threshold probabilities. These findings indicate that the model may serve as a potential tool for early risk stratification, although further validation using independent external datasets is still needed.

Discussion

POD is a common and serious complication of anesthesia and surgery, with profound implications for patient outcomes. Accurate risk prediction and targeted preventive strategies are therefore essential. In this prospective observational cohort study, we identified six predictors using LASSO regression: preoperative sleep quality (PSQI score), psychological distress (K10 score), social support (GO-SSS score), nutritional status (PNI), comorbidity burden (aCCI score), and intraoperative physiological reserve (SAS score). Among seven machine learning algorithms, XGBoost demonstrated the best discriminative ability and good calibration. The SHAP framework further enhanced model interpretability by delineating the contribution of each predictor. These findings corroborate that delirium onset is multifactorial, involving complex interactions among sleep, psychological, social, nutritional, comorbidity, and intraoperative factors.

The incidence of POD in our cohort was 30.0% (279/930), higher than the 23.05% reported by Yu et al. in patients undergoing primary hip or knee arthroplasty[30]. This difference may be attributable to the inclusion of revision surgery, which often involves adhesiolysis, bone defect management, and prosthesis removal, leading to longer operative time, greater blood loss, and more severe surgical trauma and inflammatory stress. In addition, patients undergoing revision surgery are typically older, have a heavier comorbidity burden, and possess poorer preoperative functional status and cognitive reserve. They may also experience more severe postoperative pain, prolonged immobilization, and sleep disturbance. These factors collectively increase the risk of POD. Clinicians should therefore pay greater attention to early identification and targeted intervention in this high-risk population.

Preoperative sleep quality (PSQI score) was the most important predictor. Previous studies have shown that sleep disruption at home is an independent risk factor for POD, and the risk is further amplified when sleep interruption recurs during hospitalization[31]. Mechanistically, sleep disturbances may interfere with neurotransmitter release, contribute to neuronal degeneration, and exacerbate systemic inflammation and blood–brain barrier disruption[32–35]. These findings suggest that preoperative sleep-promoting interventions—such as sleep hygiene education, noise and light reduction, and pain control—may help reduce POD risk.

Psychological distress (K10 score) ranked second. Negative emotional states may compromise blood–brain barrier integrity and amplify neuroinflammatory responses[36–38]. Enhanced psychological assessment, relaxation techniques, and family support should therefore be integrated into perioperative care.

Social support (GO-SSS score) was another important predictor. Low social support may reflect insufficient social interaction, which has been associated with an increased risk of delirium[39, 40]. Encouraging patients to express their needs and engaging family members in psychological support may help mitigate this risk.

Although ranked lower, nutritional status (PNI) and comorbidity burden (aCCI score) also demonstrated predictive value. PNI integrates serum albumin and lymphocyte count, reflecting nutritional–inflammatory status. Malnutrition and inflammation may lower the threshold for delirium[41–44]. The aCCI captures baseline disease burden and physiological reserve, and has been linked to postoperative complications in orthopedic surgery[17, 45–47]. Systematic comorbidity management and nutritional support are therefore recommended.

Intraoperative physiological reserve (SAS score) was the predominant intraoperative factor. The SAS is easily obtained from anesthesia records and reflects stress tolerance. Substantial blood loss can induce cerebral hypoperfusion, hypoxia, and systemic inflammation, thereby promoting delirium[48, 49]. Meticulous intraoperative techniques to minimize blood loss and maintain hemodynamic stability are therefore important.

The strengths of this study include its prospective design, large sample size, comprehensive integration of multidimensional predictors, and the application of SHAP-based interpretability, which addresses the “black-box” limitation of machine learning. However, several limitations should be acknowledged. First, this was a single-center study with internal validation only; external validation in multicenter prospective cohorts is needed to confirm generalizability. Second, although POD was assessed prospectively using the CAM during the first seven postoperative days, we did not extend cognitive follow-up beyond discharge to capture long-term outcomes. Third, the model did not include detailed intraoperative neurophysiological monitoring, anesthetic depth indices, or specific serum biomarkers, which might further improve predictive accuracy. Finally, the full six-predictor model may be less feasible in urgent surgical settings; simplified versions could be considered and prospectively validated in future studies.

Conclusions

In conclusion, this study demonstrated the potential value of machine learning algorithms, particularly XGBoost, in predicting POD in elderly patients undergoing hip or knee replacement and revision surgery. Six key predictors were identified, including preoperative sleep quality (PSQI score), psychological distress (K10 score), social support (GO-SSS score), nutritional status (PNI), comorbidity burden (aCCI score), and intraoperative physiological reserve (SAS score), with preoperative sleep quality consistently ranking as the most important feature. The SHAP-based interpretability enhanced model transparency by quantifying individual risk profiles and elucidating the interplay among sleep quality, psychological distress, social support, nutritional status, comorbidity burden, and physiological reserve in shaping delirium risk. However, this was a single-center study with internal validation only. The model should therefore be regarded as a preliminary risk stratification tool rather than a validated clinical instrument.

External validation in multicenter prospective cohorts is needed before its clinical utility can be fully established.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the principles of the Declaration of Helsinki. The study protocol was approved by the Ethics Committee (Institutional Review Board) of the First Affiliated Hospital of Xinjiang Medical University on August 30, 2025 (Approval No. K202508-30). Written informed consent was obtained from all patients prior to enrollment.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Clinical trial number

Not applicable.

Funding

This work was supported by the “Tianshan Talents” High-Level Talent Program in Medicine and Health (Project No.: TSYC202401B187).

Author Contribution

Y P, W W, and Y L conceived and designed the study. Y P, W W, J S, and W Y collected the data and performed the statistical analysis. Y P and W W drafted the initial manuscript. J S, W Y, M S, X W, and Y L critically revised the manuscript. Y P and Y L provided funds. All authors read the manuscript and approved the final draft.

Acknowledgements

Not applicable.

Data Availability

The dataset used and/or analyzed during this study is available upon reasonable request from the corresponding author.

References

1. Yin X, Wang W, Jiang Q, Weng W. Clinical efficacy analysis of thermal infrared image detection on the intervention of shiquan dabu decoction in patients undergoing hip and knee joint replacement surgery. J Therm Biol. 2025;130:104163. 10.1016/j.jtherbio.2025.104163.
2. Lawford BJ, Kiadaliri A, Englund M, et al. Change in willingness for surgery and risk of joint replacement after an education and exercise program for hip/knee osteoarthritis: a longitudinal cohort study of 55,059 people. PLoS Med. 2025;22(5):e1004577. 10.1371/journal.pmed.1004577.
3. Mao S, Liu Y, Wang F, et al. Design and biomechanical analysis of patientspecific porous tantalum prostheses for knee joint revision surgery. Int J Bioprint. 2023;9(4):735. 10.18063/ijb.735.
4. Di Matteo V, Di Pilla M, La Camera F, Morenghi E, Grappiolo G, Loppini M. Perioperative complications after hip and knee revision arthroplasty in the over 80 years old population: a retrospective observational case-control study. J Clin Med. 2023;12(6):2186. 10.3390/jcm12062186.
5. Lammers-Lietz F, Akyuez L, Boraschi D, et al. Prediction and risk evaluation of delirium after surgery in older patients: development and internal validation of an algorithm from the prospective BioCog cohort study. Br J Anaesth. 2026;136(5):1495–508. 10.1016/j.bja.2026.01.025.
6. Wang B, Wang X, Liang Y, et al. Association of preoperative cerebrospinal fluid sTREM2 concentration with postoperative delirium and 3-year mortality in patients total knee or hip arthroplasty: a prospective cohort study. Int J Surg. 2025;111(12):9372–83. 10.1097/JS9.0000000000003465.
7. Zhang T, Shen T, Li N, et al. Effect of anticholinergic drug burden on postoperative delirium in elderly patients: a nested case-control study. CNS Neurosci Ther. 2026;32(1):e70731. 10.1002/cns.70731.
8. Sadahiro R, Koyama R, Kuchiba A, et al. Medical cost of postoperative delirium after highly invasive cancer resection: a prospective cohort study. Jpn J Clin Oncol. 2026;56(2):187–95. 10.1093/jjco/hyaf172.
9. Hughes CG, Boncyk CS, Culley DJ, et al. American society for enhanced recovery and perioperative quality initiative joint consensus statement on postoperative delirium prevention. Anesth Analg. 2020;130(6):1572–90. 10.1213/ANE.0000000000004641.
10. Wilson JE, Mart MF, Cunningham C, et al. Delirium. Nat Rev Dis Primers. 2020;6(1):90. 10.1038/s41572-020-00223-4.
11. Chen X, Yao W, Liu X, et al. Preoperative geriatric nutritional risk index as a predictor of postoperative delirium in revision arthroplasty: a 10-year retrospective cohort study. Front Med. 2025;12:1626383. 10.3389/fmed.2025.1626383.

12. Peng Z, Wu J, Wang Z, et al. Incidence and related risk factors for postoperative delirium following revision total knee arthroplasty: a retrospective nationwide inpatient sample database study. BMC musculoskelet disord. 2024;25(1):633. 10.1186/s12891-024-07757-8.
13. Han S, He K, Cao L, Li S, Huang L. Activation of LH orexin neurons and VTA dopamine neurons is involved in anesthesia/surgery-induced sleep-wake disorders and delirium-like behaviors in mice. Sci Rep. 2026;16(1):6035. 10.1038/s41598-026-36688-8.
14. Wang Q, Ding G, Li Y. Preoperative anxiety and postoperative delirium: a mini review of the association, mechanisms, and preventive strategies. J Multidiscip Healthc. 2026;19:564668. 10.2147/JMDH.S564668.
15. O'Hanlon S, Baxter M, Hosie A. Postoperative delirium in older patients with cancer: the role of psychological distress and social support. Curr Opin Support Palliat Care. 2022;16(1):38–47. 10.1097/SPC.0000000000000588.
16. Mi X, Jia Y, Song Y, et al. Preoperative prognostic nutritional index value as a predictive factor for postoperative delirium in older adult patients with hip fractures: a secondary analysis. BMC Geriatr. 2024;24(1):21. 10.1186/s12877-023-04629-z.
17. Liu J, Li J, He J, Zhang H, Liu M, Rong J. The age-adjusted charlson comorbidity index predicts post-operative delirium in the elderly following thoracic and abdominal surgery: a prospective observational cohort study. Front Aging Neurosci. 2022;14:979119. 10.3389/fnagi.2022.979119.
18. Li GH, Zhao L, Lu Y, et al. Development and validation of a risk score for predicting postoperative delirium after major abdominal surgery by incorporating preoperative risk factors and surgical apgar score. J Clin Anesth. 2021;75:110408. 10.1016/j.jclinane.2021.110408.
19. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19(1):64. 10.1186/s12874-019-0681-4.
20. Song Y, Gao S, Tan W, Qiu Z, Zhou H, Zhao Y. Multiple machine learnings revealed similar predictive accuracy for prognosis of PNETs from the surveillance, epidemiology, and end result database. J Cancer. 2018;9(21):3971–8. 10.7150/jca.26649.
21. Gambus P, Shafer SL. Artificial intelligence for everyone. Anesthesiology. 2018;128(3):431–3. 10.1097/ALN.0000000000001984.
22. Yosibash Z, Trabelsi N, Buchnik I, et al. Hip fracture risk assessment in elderly and diabetic patients: combining autonomous finite element analysis and machine learning. J Bone Min Res: Off J Am Soc Bone Min Res. 2023;38(6):876–86. 10.1002/jbmr.4805.
23. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43. 10.1136/svn-2017-000101.
24. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016;76:175–82. 10.1016/j.jclinepi.2016.02.031.
25. Hua S, Li C, Wang Y, et al. Establishment of predictive models for postoperative delirium in elderly patients after knee/hip surgery based on total bilirubin concentration: machine learning algorithms.

BMC anesthesiol. 2025;25(1):375. 10.1186/s12871-025-03259-9.

26. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. Ann Intern Med. 1990;113(12):941–8. 10.7326/0003-4819-113-12-941.
27. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med. 2002;32(6):959–76. 10.1017/s0033291702006074.
28. van den Akker-Scheek I, Stevens M, Spriensma A, van Horn JR. Groningen orthopaedic social support scale: validity and reliability. J Adv Nurs. 2004;47(1):57–63. 10.1111/j.1365-2648.2004.03065.x.
29. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res. 1989;28(2):193–213. 10.1016/0165-1781(89)90047-4.
30. Yu R, Yang M, Chen J, Zhang F. The relationship between preoperative serum indirect bilirubin and postoperative delirium in geriatric patients undergoing joint replacement. PLoS ONE. 2025;20(3):e0320719. 10.1371/journal.pone.0320719.
31. Todd OM, Gelrich L, MacLullich AM, Driessen M, Thomas C, Kreisel SH. Sleep disruption at home as an independent risk factor for postoperative delirium. J Am Geriatr Soc. 2017;65(5):949–57. 10.1111/jgs.14685.
32. Hume NE, Zerfas I, Wong A, et al. Clinical impact of the implementation strategies used to apply the 2013 pain, agitation/sedation, delirium or 2018 pain, agitation/sedation, delirium, immobility, sleep disruption guideline recommendations: a systematic review and meta-analysis. Crit Care Med. 2024;52(4):626–36. 10.1097/CCM.0000000000006178.
33. Qiu Y, Huang X, Huang L, et al. 5-HT(1A) receptor antagonist improves behavior performance of delirium rats through inhibiting PI3K/akt/mTOR activation-induced NLRP3 activity. IUBMB Life. 2016;68(4):311–9. 10.1002/iub.1491.
34. Guo H, Li LH, Lv XH, et al. Association between preoperative sleep disturbance and postoperative delirium in elderly: a retrospective cohort study. Nat Sci Sleep. 2024;16:389–400. 10.2147/NSS.S452517.
35. Devinney MJ, VanDusen KW, Kfouri JM, et al. The potential link between obstructive sleep apnea and postoperative neurocognitive disorders: current knowledge and possible mechanisms. Can J Anaesth = J Can anesth. 2022;69(10):1272–87. 10.1007/s12630-022-02302-4.
36. Welcome MO. Cellular mechanisms and molecular signaling pathways in stress-induced anxiety, depression, and blood-brain barrier inflammation and leakage. Inflammopharmacology. 2020;28(3):643–65. 10.1007/s10787-020-00712-8.
37. Brinn A, Stone J. Neutrophil-lymphocyte ratio across psychiatric diagnoses: a cross-sectional study using electronic health records. BMJ open. 2020;10(7):e036859. 10.1136/bmjopen-2020-036859.

38. Deng Y, Dong J, Pan C, et al. Intraoperative methylene blue infusion reduces postoperative delirium in patients undergoing pancreatic surgery: a randomized controlled clinical trial. J Clin Anesth. 2026;108:112060. 10.1016/j.jclinane.2025.112060.
39. Tang X, Yv H, Wang F, et al. The relationship between suboptimal social networks and postoperative delirium: the PNDABLE study. Front Aging Neurosci. 2022;14:851368. 10.3389/fnagi.2022.851368.
40. Raghuraman S, Richards E, Mahmoud A, et al. Understanding rehabilitation and support needs after an episode of delirium: a qualitative thematic analysis of interviews with older people with delirium, family carers and healthcare professionals. BMC Geriatr. 2025;25(1):547. 10.1186/s12877-025-06196-x.
41. Wang W, Yao W, Tang W, Li Y, Lv Q, Ding W. Association between preoperative albumin levels and postoperative delirium in geriatric hip fracture patients. Front Med. 2024;11:1344904. 10.3389/fmed.2024.1344904.
42. Ritter C, Tomasi CD, Dal-Pizzol F, et al. Inflammation biomarkers and delirium in critically ill patients. Crit Care. 2014;18(3):R106. 10.1186/cc13887.
43. Chen J, Zheng C, Zhong J, et al. Preoperative prognostic nutritional index is useful factor for predicting postoperative delirium after primary total joint arthroplasty. BMC musculoskelet disord. 2021;22(1):778. 10.1186/s12891-021-04626-6.
44. Yu C, Chen L, Lei X, Xu Z, Zhao H, Chen C. Comparative analysis of four nutritional scores in predicting delirium in ICU patients. Front Nutr. 2025;12:1482150. 10.3389/fnut.2025.1482150.
45. Amit P, Marya SKS. Age-adjusted charlson comorbidity index as a novel guideline for patient selection between unilateral versus bilateral simultaneous total knee arthroplasty. Arch Orthop Trauma Surg. 2022;142(4):657–63. 10.1007/s00402-021-03841-z.
46. Marya SKS, Amit P, Singh C. Impact of charlson indices and comorbid conditions on complication risk in bilateral simultaneous total knee arthroplasty. Knee. 2016;23(6):955–9. 10.1016/j.knee.2016.05.013.
47. Feinkohl I, Janke J, Slooter AJC, et al. Metabolic syndrome and the risk of postoperative delirium and postoperative cognitive dysfunction: a multi-centre cohort study. Br J Anaesth. 2023;131(2):338–47. 10.1016/j.bja.2023.04.031.
48. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ. An apgar score for surgery. J Am Coll Surg. 2007;204(2):201–8. 10.1016/j.jamcollsurg.2006.11.011.
49. Deng Y, Zhang T, Xie H, Zeng J. Perioperative blood loss is a risk factor for postoperative delirium in geriatric hip fracture patients: a retrospective study. Front Med. 2025;12:1617891. 10.3389/fmed.2025.1617891.

Figures

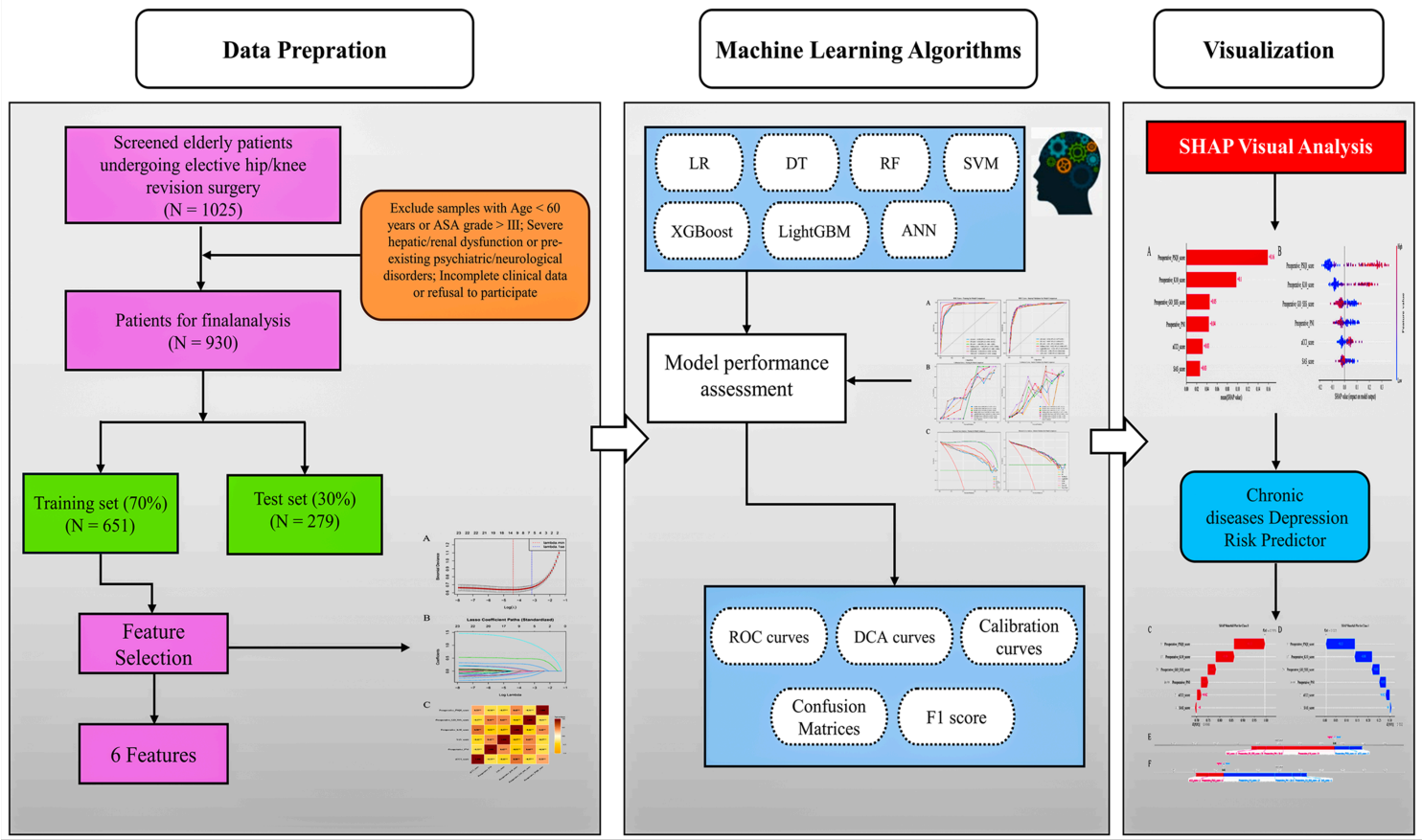


Figure 1

Flowchart of patient enrollment and model development. LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine; ANN, Artificial Neural Network; POD, Postoperative Delirium; CAM, Confusion Assessment Method.

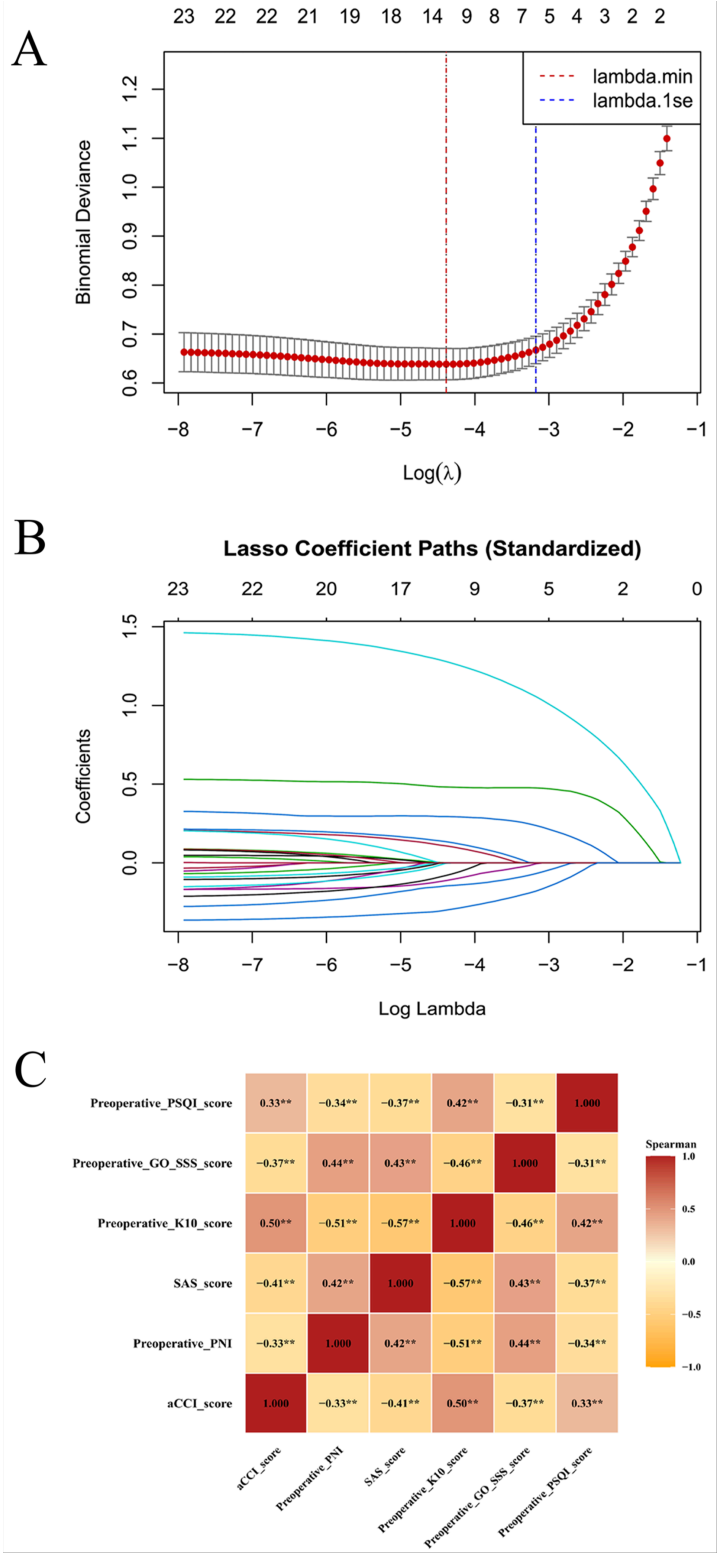


Figure 2

Feature selection using LASSO regression and correlation analysis. **(A)** Coefficient trajectories of candidate variables as the regularization parameter λ varies. **(B)**Tenfold cross-validation curves for tuning parameter selection in the LASSO model; the left and right vertical dashed lines indicate the λ value at minimum binomial deviance (lambda.min) and the λ value within one standard error of the

minimum (lambda.1se), respectively. **(C)** Spearman correlation heatmap among the six selected predictors.

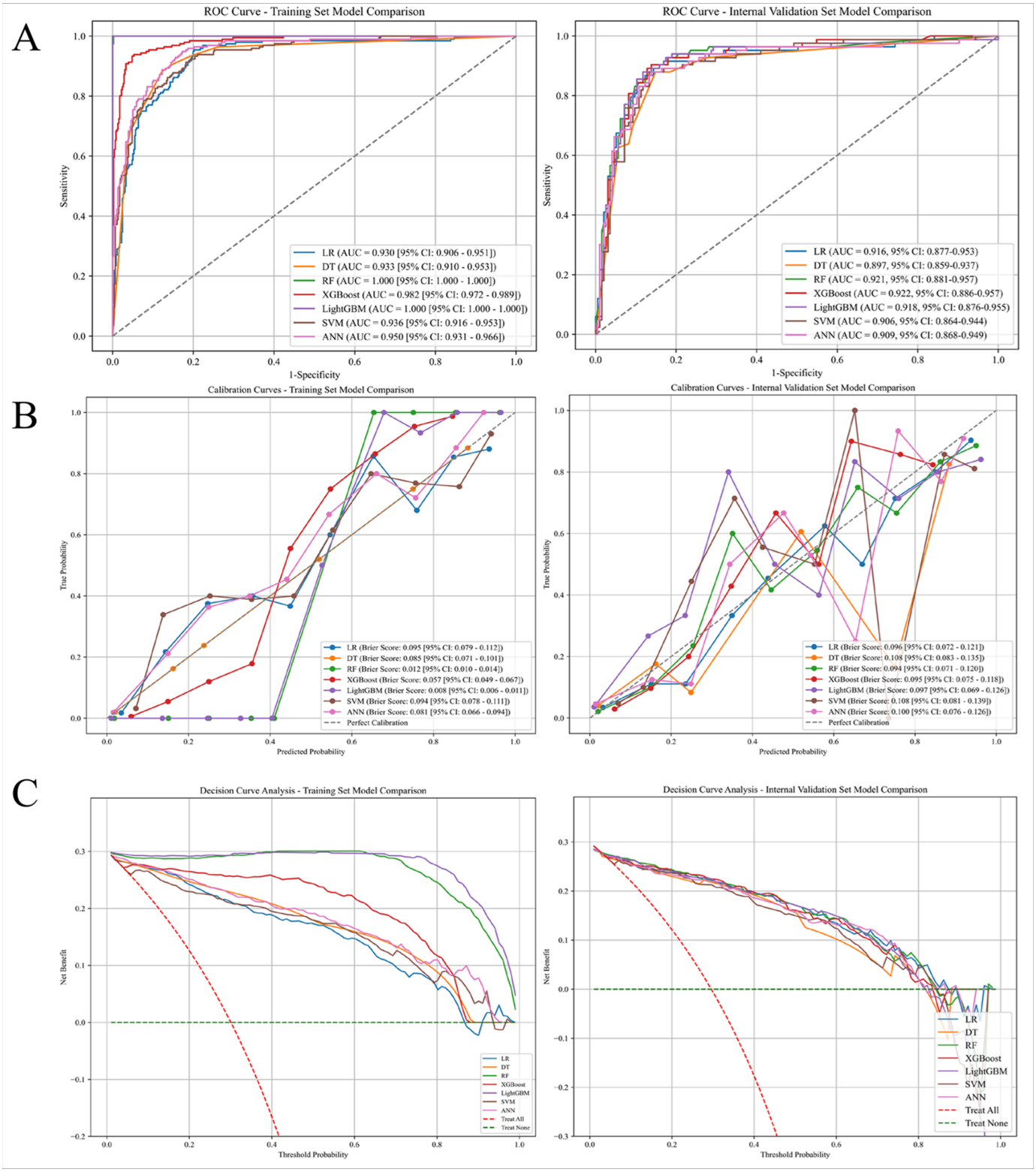


Figure 3

Performance evaluation of seven machine learning models in the validation set. (A) Receiver operating characteristic (ROC) curves with corresponding area under the curve (AUC) values and 95% confidence

intervals. (B) Calibration curves comparing predicted probabilities with observed outcomes; the diagonal dashed line represents perfect calibration. (C) Decision curve analysis (DCA) evaluating the net clinical benefit across a range of threshold probabilities; the "Treat All" and "Treat None" lines represent reference strategies.

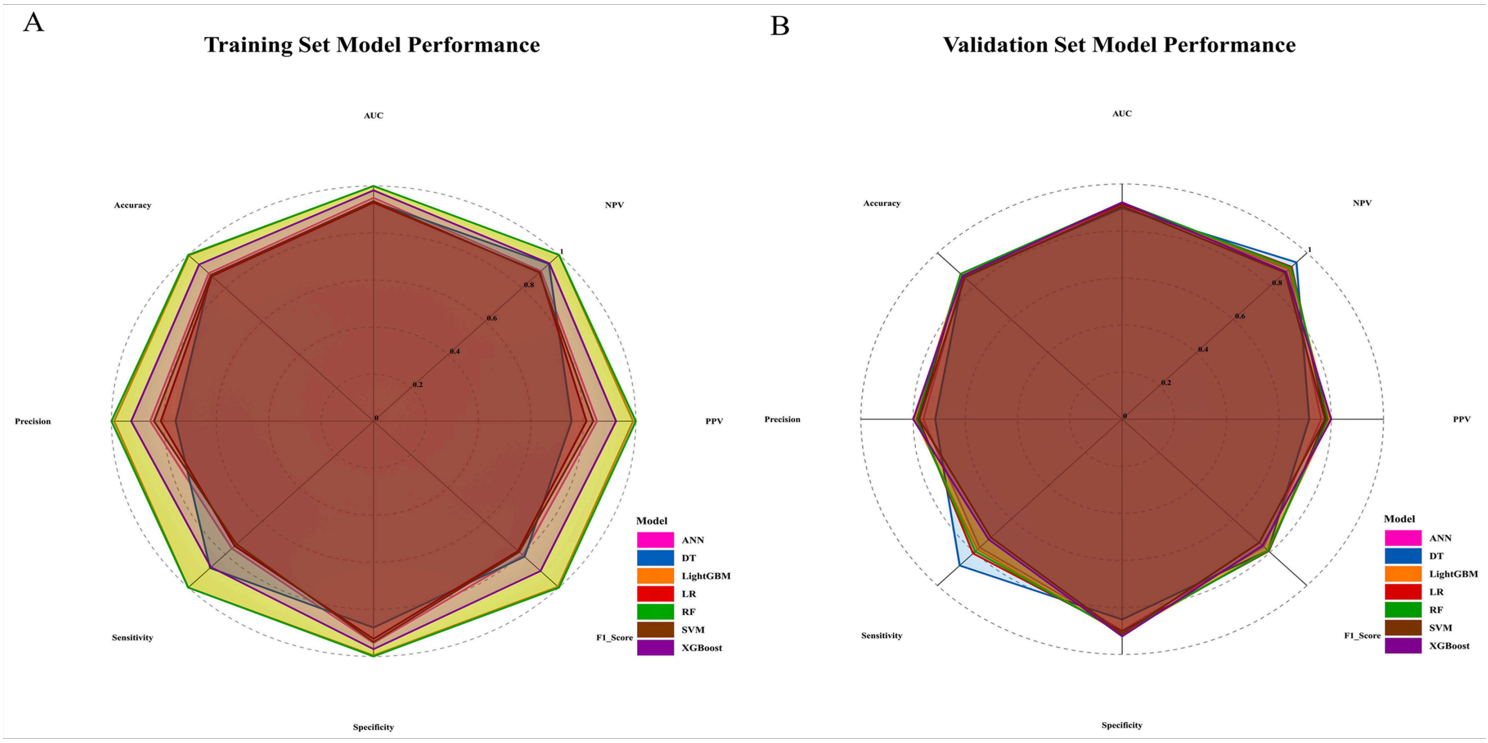


Figure 4

Radar plot comparing the comprehensive performance of seven machine learning models across multiple metrics, including accuracy, precision, sensitivity, specificity, and F1 score. Each axis represents a performance metric, and a larger area indicates better overall performance.

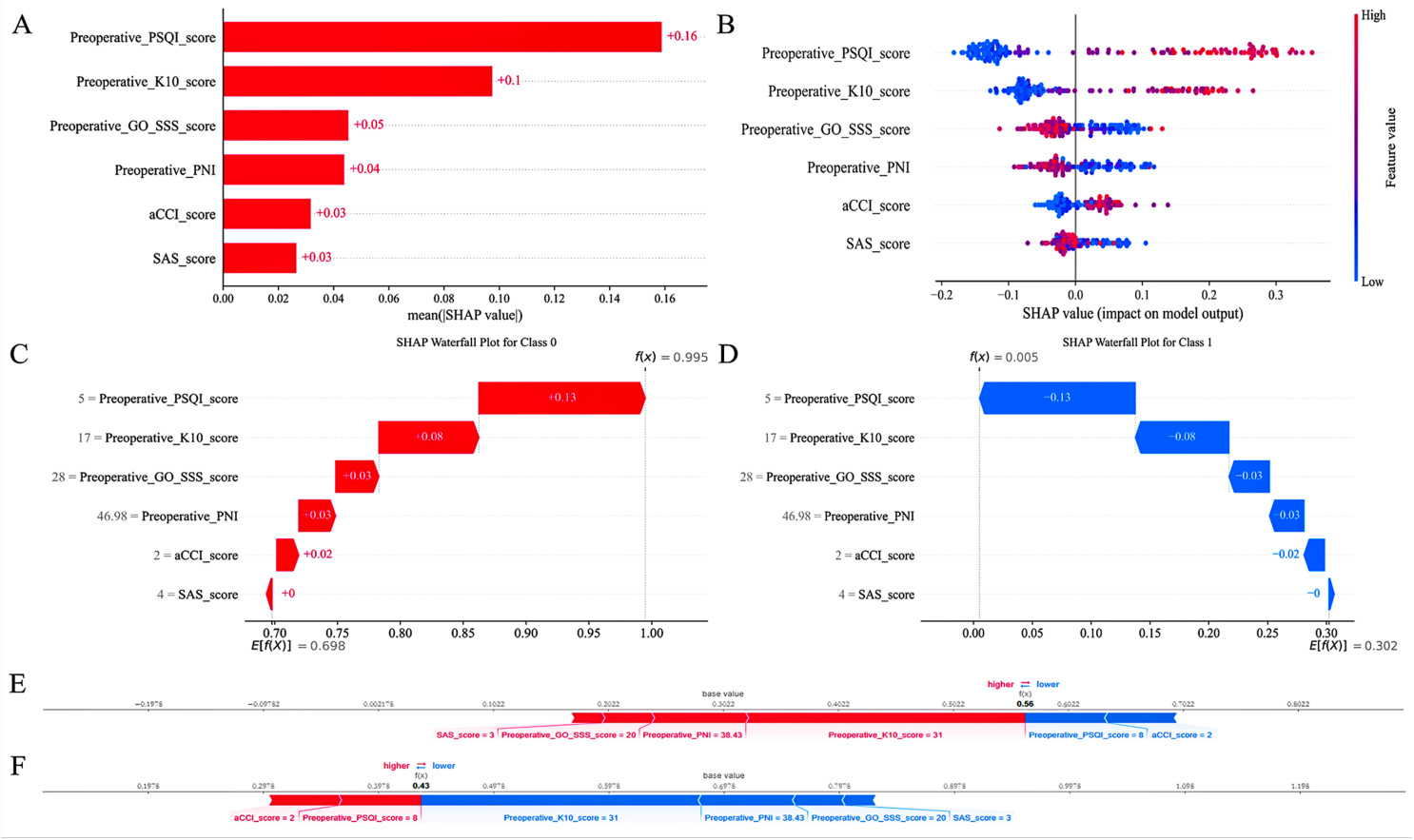


Figure 5

SHAP-based interpretability analysis of the optimal XGBoost model. (A) Global feature importance ranked by mean absolute SHAP values. (B) Beeswarm plot illustrating the distribution and direction of each feature's effect on model predictions; each dot represents a patient, with color indicating feature value (red: high; blue: low), and the position on the x-axis reflects the SHAP value (positive values indicate increased POD risk). (C & D) Waterfall plots demonstrating the cumulative contribution of each variable from the baseline prediction (base value) to the final output (f(x)) for two representative individual samples. (E & F) Force plots visually illustrating the positive and negative contributions of each feature for the same two individual samples, providing an intuitive understanding of how specific patient characteristics drive the model's risk prediction.

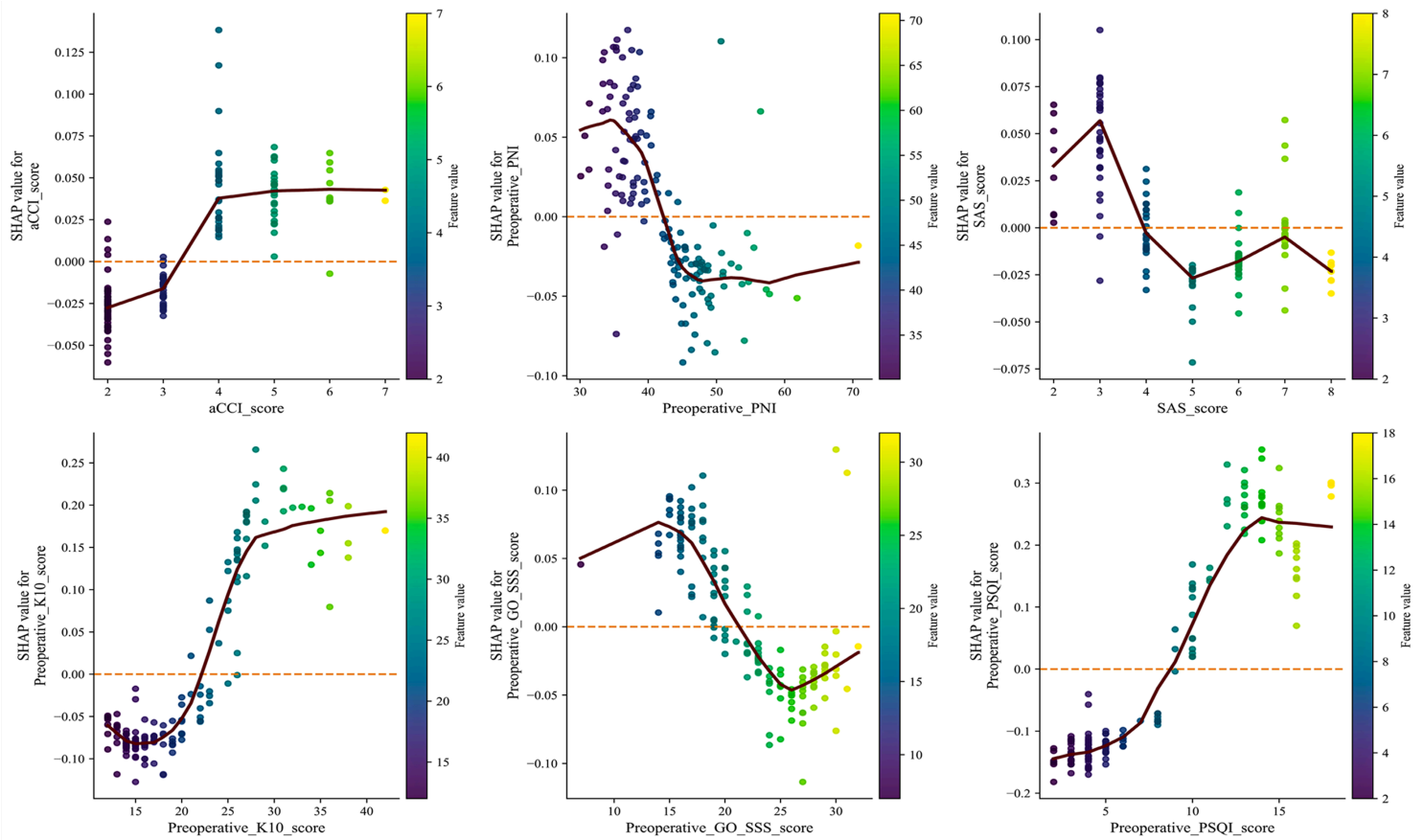


Figure 6

SHAP partial dependence plots revealing the nonlinear relationships and interaction effects between key predictors and POD risk. The plots illustrate how changes in the preoperative PSQI score affect model predictions, with the moderating role of the preoperative K10 score shown through color gradation.

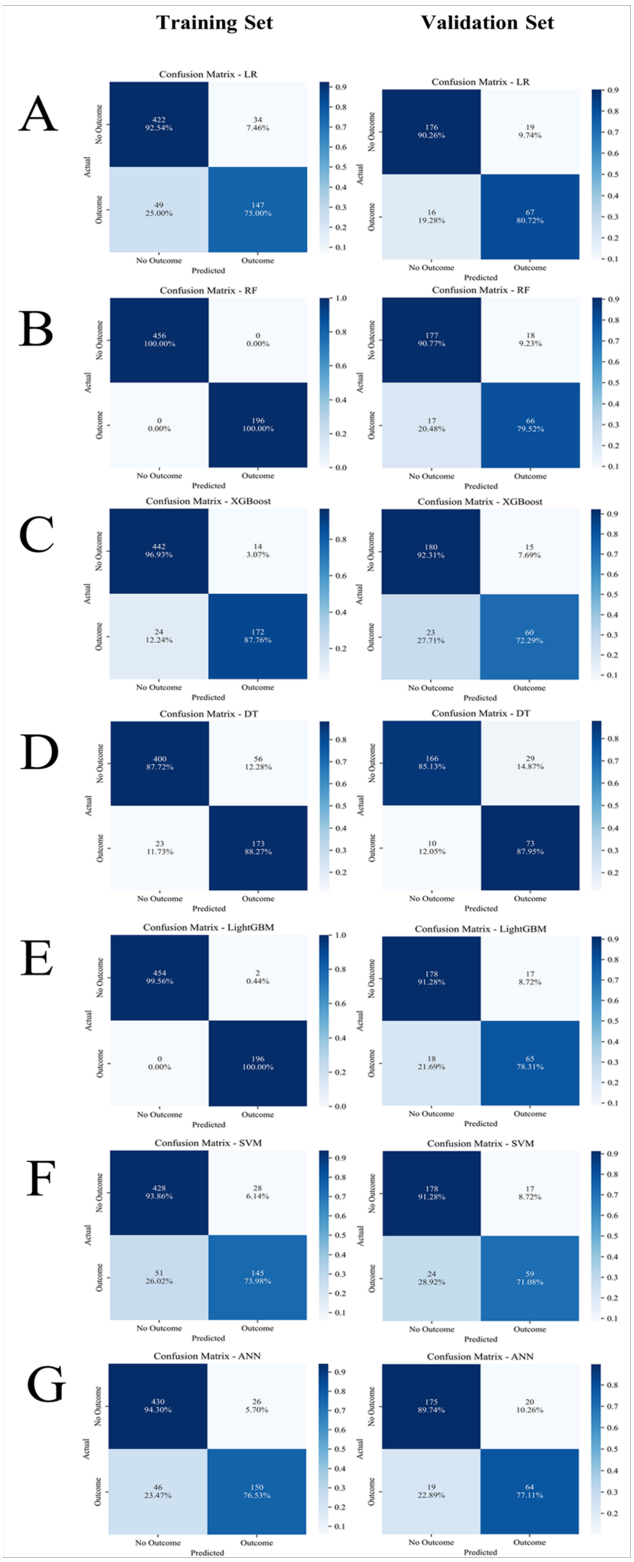


Figure 7

Confusion matrices of seven machine learning models across the training and validation sets. Panels A–G correspond to LR, RF, XGBoost, DT, LightGBM, SVM, and ANN, respectively. In each panel, the confusion matrix for the training set and that for the validation set are presented, with true positives, false positives, true negatives, and false negatives shown for each dataset, enabling direct comparison of classification accuracy and error distribution between the two datasets across models.