acta orthopaedica belgica, 2026, 92, online-only

doi.org/10.52628/92.1.15196

Development and Validation of a Machine Learning Model for Predicting Postoperative Delirium in Elderly Patients with Hip Fracture: A Retrospective Cohort Study

QIAN SONG[1, #], YI CA[2,#], PENG TIAN[3,#], SHUJUN YU[3,#], ZHE HAN[3], JUN WANG[1], YINGUANG ZHANG[3], QIANG DONG[3], AIJUN CHAO[1], JIZHENG ZHANG[4], JIAMING ZHENG[5], RONG-JIAN LU[6]

[1]Department of Osteo-Internal Medicine, Tianjin University Tianjin Hospital, Tianjin, P. R. China. 300122; [2]Department of Radiology, Tianjin University Tianjin Hospital, Tianjin, P. R. China. 300122; [3]Department of Orthopedics, Tianjin University Tianjin Hospital, Tianjin, P. R. China. 300122; [4]Department of Anesthesiology, Tianjin University Tianjin Hospital, Tianjin, P. R. China. 300122; 5. West China School of Medicine, Sichuan University, Sichuan province, P. R. China. 610207; [6]Department of Stomatology, Fifth Medical Center, People's Liberation Army General Hospital, Beijing, P. R. China. 100039.

Correspondence at: Qian Song, M.D., Ph.D., Tel: 86-13920433136 – E-mail: songqian66882020@163.com; Peng Tian, M.D., Ph.D., Tel: 86-15822710991 – E-mail: tianpeng007@foxmail.com; Shujun Yu, M.D., Ph.D., Tel: 86-13752293602 – E-mail: Yu2415198089@163.com; Jizheng Zhang, M.D., Ph.D., Tel: 86-13702016782 – E-mail: kevin19801210zjz@163.com; Rong-jian Lu M.D., Ph.D. – E-mail: lujian9806@sohu.com

ABSTRACT

Postoperative delirium (POD) is a common and serious complication in elderly hip fracture patients, including those undergoing surgery for femoral neck fractures and intertrochanteric fractures, and is associated with poor clinical outcomes. Early identification of at-risk individuals remains challenging with conventional methods. To develop and validate machine learning models for predicting POD using preoperative variables, and to identify key risk factors, we conducted a retrospective study of 400 patients aged ≥65 years undergoing hip fracture surgery. Five machine learning algorithms were developed and validated using 70-30 split with 10-fold cross-validation. Results demonstrated that POD incidence was 20.0% (80/400). Significant predictors included age (OR=1.06, 95%CI:1.02-1.10), cognitive impairment (OR=2.85, 95%CI:1.70-4.78), hypoalbuminemia (OR=2.32, 95%CI:1.45-3.71), and preoperative waiting time (OR=1.18, 95%CI:1.06-1.32). The Random Forest model demonstrated superior performance (AUC=0.89, accuracy=0.83), outperforming other algorithms (XGBoost AUC=0.87, SVM AUC=0.84, Logistic Regression AUC=0.82, Decision Tree AUC=0.79). Variable importance analysis consistently identified cognitive impairment, hypoalbuminemia, and age as the most prominent predictors across all models. In conclusion, machine learning models, particularly Random Forest, effectively predict POD risk using routine preoperative data. Within our study cohort, machine learning models, particularly Random Forest, showed potential for predicting POD risk using routine preoperative data upon internal validation. The consistent identification of key predictors enables targeted prevention strategies for high-risk elderly hip fracture patients. The broader applicability of the model requires confirmation through external validation in future studies. The consistent identification of key predictors enables targeted prevention strategies for high-risk elderly hip fracture patients.

Keywords: Postoperative delirium; Hip fracture; Machine learning; Elderly patients; Prediction model; Random Forest.

INTRODUCTION

Hip fractures represent a significant and growing global health burden among elderly populations, with annual incidence rates continuing to rise in tandem with aging demographics worldwide[1,2]. These injuries commonly signify a critical juncture in the functional independence of older adults, heralding a marked decline in mobility and overall quality of life[3]. Within this vulnerable patient population, postoperative delirium (POD) emerges as one of the most common and devastating complications, affecting between 15-50% of hip fracture surgery patients[4]. The clinical and economic implications of POD are profound, being independently associated with prolonged hospitalization, increased institutionalization rates, persistent cognitive deterioration, accelerated functional decline, elevated mortality, and substantially higher healthcare expenditures[5].

The multifactorial pathogenesis of POD involves complex interactions between patient vulnerability factors and precipitating insults, including surgical stress, anesthesia exposure, and perioperative physiological perturbations[6]. This complexity presents substantial challenges for accurate risk prediction using conventional statistical methods. Traditional risk assessment tools, while clinically accessible, frequently demonstrate limited predictive performance in heterogeneous elderly populations due to their inability to adequately capture nonlinear relationships and higher-order interactions among multiple risk factors[7,8]. Furthermore, existing delirium prediction models incorporate intraoperative and postoperative variables, limiting their utility for preoperative risk stratification and early intervention planning[9].

The emerging application of machine learning (ML) in clinical prediction represents a paradigm shift in prognostic modeling, offering distinct advantages in handling complex, high-dimensional data and detecting subtle patterns that may elude conventional approaches[10,11]. ML algorithms can automatically model intricate interactions among numerous clinical variables without requiring pre-specified hypotheses, potentially uncovering novel risk associations and improving predictive accuracy[11]. Within perioperative medicine, various studies have begun exploring ML for various outcome predictions, yet its application specifically for POD prediction in hip fracture patients remains notably underexplored[12-14]. Current literature reveals a significant gap in comprehensive comparisons of multiple ML architectures specifically optimized for this clinical context, particularly using exclusively preoperative variables to facilitate early risk assessment[15].

Additionally, the interpretability of ML models---understanding which factors drive predictions---represents a critical aspect for clinical implementation[16]). While studies have developed POD prediction models, few have systematically compared feature importance across multiple algorithms or validated their clinical utility in real-world settings[17]. The integration of explainable artificial intelligence (XAI) techniques with robust ML models could potentially bridge the gap between predictive accuracy and clinical interpretability, fostering greater trust and adoption among healthcare providers[18,19].

This study aims to address these research gaps by developing and validating multiple machine learning models for predicting POD in elderly hip fracture patients using readily available preoperative clinical data. We specifically seek to compare the performance of diverse ML algorithms---including ensemble methods, support vector machines, and traditional logistic regression---against conventional approaches, while simultaneously identifying and ranking key predictive features through advanced interpretability methods. By leveraging exclusively preoperative variables, our models aim to facilitate early identification of high-risk patients, enabling timely implementation of targeted preventive strategies and optimized resource allocation in the perioperative care pathway.

METHODS

Study Design and Population

We conducted a retrospective cohort study of 400 patients aged 65 years or older who underwent surgical treatment for hip fracture at a tertiary academic medical center between January 2021 and July 2025. Patients were excluded if they presented with preoperative delirium, were managed non-surgically, or had incomplete medical records that precluded accurate data extraction. The study was approved by the institutional ethics committee with a waiver for informed consent due to its retrospective nature.

Data Collection and Predictor Variables

Data were systematically collected from electronic health records using a standardized protocol. The primary outcome was postoperative delirium (POD), defined according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria. POD ascertainment followed a two-step process to maximize accuracy: (1) initial screening using ICD-9/10 codes for delirium (ICD-9: 293.0, 293.1; ICD-10: F05, F05.9, R41.0), followed by (2) structured chart review for all code-flagged patients and a 20% random sample of non-flagged patients. Two clinicians, blinded to predictor variables and trained in DSM-5 and Confusion Assessment Method (CAM) criteria, independently reviewed nursing notes (including Nursing Delirium Screening Scale (Nu-DESC) documentation when available), psychiatric consultations, and physician progress notes. Disagreements were resolved by a third senior geriatrician. It is important to note that no standardized delirium screening instrument (such as CAM or Nu-DESC) was routinely administered prospectively in this retrospective study. POD diagnosis was based solely on the two-step process described above. Nursing notes, including any available Nu-DESC documentation, were reviewed as part of the chart review but did not constitute a prospective screening

protocol. Disagreements were resolved by a third senior geriatrician. Data on postoperative administration of antipsychotics (e.g., olanzapine) or other medications for delirium symptoms were collected separately for descriptive purposes only and were not used to define the POD outcome. This strict separation ensures that the definition of the POD outcome is based entirely on clinical diagnosis (DSM-5 criteria via chart review) and is independent of any treatment decisions or documentation of pharmacological management. Predictor variables encompassed demographic characteristics, preexisting comorbidities, laboratory parameters, and surgical factors. Specifically, we collected data on age, gender, hypertension, diabetes, coronary artery disease, chronic kidney disease, cognitive impairment, hypoalbuminemia, electrolyte disorders, anemia, preoperative waiting time, and surgery type. A composite measure of total comorbidity burden was also calculated.

Data Preprocessing and Model Development

This study followed a standardized machine learning development workflow to ensure a rigorous and leakage-free evaluation. The dataset was first split into a model development set and a hold-out internal validation set. All subsequent data-dependent steps—including handling of missing values, normalization, and hyperparameter tuning—were conducted exclusively within the development set to prevent any information from the validation set influencing model development. The final model, configured with optimal parameters identified through cross-validation on the development set, was then trained on the entire development set and evaluated only once on the isolated validation set to obtain an unbiased performance estimate.

To ensure a rigorous and leak-proof evaluation, the dataset was first randomly split into a training set (70%, n=280) and a hold-out test set (30%, n=120) using stratified sampling based on the POD outcome. All subsequent steps were performed exclusively on the training set to define preprocessing parameters and develop models. For variables with <20% missing values, multiple imputation by chained equations (MICE) with 10 iterations was applied to the training data. The mean imputed values from the training set were then used to impute missing values in the test set. Continuous variables were standardized using Z-score normalization, with the mean and standard deviation calculated from the training set. No feature selection was performed prior to modeling; all variables listed in the Predictor Variables section were used.

Clarification on Predictor Inclusion: The univariate logistic regression analysis (results in Table II) was performed solely to provide a conventional clinical description of individual associations with POD. It was not used for feature selection, and no filtering based on statistical significance was applied. To allow the machine learning models to evaluate all potential relationships, including complex interactions, all variables listed in the Predictor Variables section were included as input features for model training. This approach prevents information leakage from outcome-driven feature selection.

Five machine learning algorithms were implemented: Logistic Regression (LR) with L2 regularization, Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) with a radial basis function kernel. For each algorithm, hyperparameter tuning was conducted strictly within the training set via a grid search coupled with 10-fold cross-validation, optimizing for the area under the ROC curve (AUC). The searched hyperparameter grids included: for RF - number of trees: [100, 300, 500], max depth: [5, 10, 15]; for XGBoost - learning rate: [0.01, 0.1], max depth: [3, 6]; for SVM - C: [0.1, 1, 10], gamma: [‘scale’, ‘auto’]. The model configuration achieving the highest mean cross-validated AUC was selected.

Model Evaluation and Validation

The final model for each algorithm was produced by retraining the optimal hyperparameter configuration on the entire preprocessed training set. This locked model was then applied only once to the preprocessed hold-out internal validation set for final, unbiased evaluation. Model performance was assessed on the internal validation set using the area under the receiver operating characteristic curve (AUC) with 95% confidence interval (DeLong’s method), accuracy, sensitivity, specificity, and F1-score. Model calibration was evaluated in detail on the test set. Beyond the Brier score (0.11 for the primary RF model), we report the calibration slope (1.02) and intercept (-0.03), indicating excellent agreement between predicted probabilities and observed outcomes. The Hosmer-Lemeshow goodness-of-fit test further supported good calibration ($p = 0.42$). Clinical utility was assessed via decision curve analysis across probability thresholds from 0 to 0.5.

Reporting Guideline

This study was conducted and reported in accordance with the key elements of the Transparent Reporting

of a multivariable prediction model for Individual Prognosis Or Diagnosis – Artificial Intelligence (TRIPOD-AI) statement.

Statistical Analysis

Data distribution was tested using the Shapiro-Wilk test. As appropriate, patient characteristics were described using mean ± standard deviation and median (interquartile range [IQR]) for continuous variables, and frequency and percentage for categorical variables. Independent t-tests or Mann-Whitney U tests were applied for continuous variables with normal or non-normal distribution, respectively. Categorical variables were analyzed using the Pearson Chi-squared test. Univariate logistic regression analysis was performed to assess the unadjusted associations between individual preoperative variables and postoperative delirium, providing a conventional clinical summary of potential risk factors. All variables listed in the Predictor Variables section, irrespective of their univariate statistical significance, were included as input features in the machine learning models. To enhance model performance, we conducted zero-mean normalization of the data. All statistical analyses were calculated using SPSS software (version 26.0; SPSS Inc., Chicago, IL, USA) and Python 3.7.6 (Python Software Foundation, http://python.org).

RESULTS

Baseline Characteristics and Delirium Prevalence

The final analytical cohort comprised 400 consecutive elderly patients who underwent surgical management for hip fractures. Postoperative delirium (POD) was identified in 80 patients, corresponding to an incidence of 20.0%. As detailed in Table I, baseline characteristics differed significantly between patients who developed POD and those who did not.

Patients in the delirium group were older and had a lower body mass index. They also had a higher prevalence of specific comorbidities, including cognitive impairment (present in over one-third of delirium patients), chronic kidney disease, heart failure, and preoperative pneumonia.

Laboratory findings at admission also differed between groups. The delirium group had significantly lower levels of nutritional markers (prealbumin, albumin, total protein) and higher levels of inflammatory and coagulation markers (fibrinogen, D-dimer). They also exhibited perturbations in white blood cell counts, signs of anemia, worse renal function parameters, higher fasting blood glucose, and more frequent electrolyte imbalances. Functional status, as measured by the Barthel Index, was significantly lower in the delirium group preoperatively. Finally, the mean preoperative waiting time was significantly longer for patients who developed POD.

In summary, patients who developed POD presented preoperatively with a distinct profile characterized by older age, greater comorbidity burden, poorer nutritional and laboratory markers, reduced functional independence, and longer surgical wait times.

Clinical Decision Pathway Based on the Predictive Model

Figure 1 illustrates the proposed clinical decision pathway for implementing the Random Forest prediction model in the care of new geriatric hip fracture patients. Upon admission, key preoperative data—including age, cognitive status, and albumin levels—are collected. These data are then input into the validated Random Forest model, which calculates a probability for postoperative delirium (POD). Patients are stratified into two management pathways based on a predefined risk threshold. Those identified as high-risk (probability > threshold) immediately trigger a multimodal prevention protocol. This protocol encompasses targeted interventions such as nutritional optimization, cognitive prehabilitation, efforts to minimize preoperative waiting time, and multidisciplinary consultation. Following these preparatory interventions, the patient proceeds to surgery with a tailored, risk-adapted management plan.

Conversely, patients identified as low-risk (probability < threshold) proceed directly to surgery with standard perioperative care, thereby avoiding unnecessary interventions and optimizing resource allocation. This pathway enables early, personalized risk stratification and facilitates timely preventive measures for the most vulnerable patients.

Univariate Analysis of Risk Factors for Postoperative Delirium

Univariate logistic regression analysis was performed to identify preoperative factors associated with the development of postoperative delirium, with the comprehensive results detailed in Table II. The analysis revealed a wide spectrum of significant predictors spanning demographic, comorbidity, laboratory, functional, and surgical domains. Among the strongest identified risk factors were advanced age, with each additional year conferring a 5.8% increase in delirium odds (OR=1.058, 95% CI: 1.017-1.099, p=0.005),

Table I. — Baseline Characteristics of Study Population (n=400).

Characteristic	Overall (n=400)	Delirium Group (n=80)	Non-Delirium Group (n=320)	P-value
Demographics				
Age, years	78.5 ± 8.2	82.3 ± 7.1	76.8 ± 8.5	**<0.001**
Male gender	180 (45.0%)	32 (40.0%)	148 (46.3%)	0.312
BMI, kg/m²	23.8 ± 3.5	22.9 ± 3.2	24.0 ± 3.6	**0.018**
Comorbidities				
Hypertension	272 (68.0%)	60 (75.0%)	212 (66.3%)	0.122
Diabetes	130 (32.5%)	33 (41.3%)	97 (30.3%)	0.061
Coronary artery disease	151 (37.8%)	36 (45.0%)	115 (35.9%)	0.130
Chronic kidney disease	60 (15.0%)	18 (22.5%)	42 (13.1%)	**0.034**
Cognitive impairment	75 (18.8%)	28 (35.0%)	47 (14.7%)	**<0.001**
COPD	45 (11.3%)	12 (15.0%)	33 (10.3%)	0.233
Heart failure	52 (13.0%)	16 (20.0%)	36 (11.3%)	**0.037**
Pulmonary nodule	65 (16.3%)	18 (22.5%)	47 (14.7%)	0.089
Fatty liver	42 (10.5%)	10 (12.5%)	32 (10.0%)	0.512
Kidney stone	20 (5.0%)	6 (7.5%)	14 (4.4%)	0.256
Gallbladder stone	28 (7.0%)	8 (10.0%)	20 (6.3%)	0.229
Pneumonia	76 (19.0%)	24 (30.0%)	52 (16.3%)	**0.006**
Thrombus of lower limb	22 (5.5%)	6 (7.5%)	16 (5.0%)	0.398
Laboratory Findings				
Prealbumin, mg/L	185.2 ± 45.6	156.8 ± 38.9	192.4 ± 44.2	**<0.001**
Total protein, g/L	66.8 ± 6.2	63.5 ± 5.8	67.5 ± 6.1	**<0.001**
Albumin, g/L	36.8 ± 4.9	33.2 ± 4.5	37.6 ± 4.7	**<0.001**
Globulin, g/L	27.5 ± 4.1	25.8 ± 3.9	27.9 ± 4.0	**<0.001**
A/G ratio	1.35 ± 0.22	1.29 ± 0.20	1.36 ± 0.22	**0.009**
AST, U/L	24.6 ± 8.3	26.8 ± 9.1	24.0 ± 8.0	**0.008**
ALT, U/L	19.2 ± 7.5	18.8 ± 7.2	19.3 ± 7.6	0.592
AST/ALT	1.32 ± 0.41	1.45 ± 0.44	1.29 ± 0.39	**0.002**
ALP, U/L	85.3 ± 28.7	92.6 ± 31.2	83.5 ± 27.9	**0.010**
GGT, U/L	32.5 ± 18.9	35.2 ± 20.1	31.8 ± 18.5	0.152
Total bilirubin, umol/L	15.8 ± 6.2	16.2 ± 6.5	15.7 ± 6.1	0.526
Direct bilirubin, umol/L	5.2 ± 2.3	5.4 ± 2.5	5.1 ± 2.2	0.328
Indirect bilirubin, umol/L	10.6 ± 4.5	10.8 ± 4.7	10.5 ± 4.4	0.598
FBG, mmol/L	6.82 ± 1.95	7.25 ± 2.10	6.72 ± 1.89	**0.028**
BUN, mmol/L	7.35 ± 3.12	8.26 ± 3.45	7.12 ± 2.98	**0.004**
Scr, umol/L	78.6 ± 25.3	85.2 ± 28.7	76.9 ± 24.1	**0.011**
BUN/Scr	0.098 ± 0.032	0.102 ± 0.035	0.097 ± 0.031	0.225
CPK, U/L	156.8 ± 89.4	168.5 ± 95.2	153.9 ± 87.6	0.187
LDH, U/L	258.6 ± 68.3	272.4 ± 72.1	255.2 ± 66.9	**0.041**
Cholinesterase, U/L	7524 ± 1856	6985 ± 1723	7658 ± 1872	**0.004**
Blood uric acid, umol/L	335.6 ± 88.7	328.9 ± 92.4	337.4 ± 87.8	0.442
Serum sodium, mmol/L	138.5 ± 3.8	137.2 ± 4.1	138.8 ± 3.6	**0.006**
Serum kalium, mmol/L	3.85 ± 0.42	3.78 ± 0.45	3.87 ± 0.41	0.078
Serum calcium, mmol/L	2.18 ± 0.15	2.15 ± 0.16	2.19 ± 0.15	**0.038**
Serum chlorine, mmol/L	103.8 ± 3.2	102.9 ± 3.5	104.0 ± 3.1	**0.008**
WBC, $\times 10^9$/L	8.8 ± 3.1	9.5 ± 3.4	8.6 ± 3.0	**0.009**

Table I. — Baseline Characteristics of Study Population (n=400) - continued.

Characteristic	Overall (n=400)	Delirium Group (n=80)	Non-Delirium Group (n=320)	P-value
Neutrophils, $\times10^9$/L	6.95 ± 2.68	7.62 ± 2.85	6.78 ± 2.60	**0.010**
Lymphocyte, $\times10^9$/L	1.08 ± 0.45	0.92 ± 0.41	1.12 ± 0.46	**<0.001**
Monocyte, $\times10^9$/L	0.43 ± 0.18	0.46 ± 0.20	0.42 ± 0.17	0.089
Eosinophilic granulocyte, $\times10^9$/L	0.08 ± 0.06	0.07 ± 0.05	0.08 ± 0.06	0.215
Basophilic granulocyte, $\times10^9$/L	0.02 ± 0.02	0.02 ± 0.02	0.02 ± 0.02	0.874
RBC, $\times10^{12}$/L	4.05 ± 0.62	3.88 ± 0.65	4.09 ± 0.60	**0.005**
Hemoglobin, g/L	124.6 ± 18.5	116.8 ± 17.9	126.4 ± 18.2	**<0.001**
Hematocrit	0.37 ± 0.05	0.35 ± 0.05	0.38 ± 0.05	**<0.001**
MCV, fl	91.2 ± 5.8	90.8 ± 6.1	91.3 ± 5.7	0.512
MCH, pg	30.7 ± 2.3	30.5 ± 2.5	30.8 ± 2.2	0.328
MCHC, g/L	336.8 ± 10.5	335.2 ± 11.3	337.2 ± 10.2	0.152
Platelet count, $\times10^9$/L	198.6 ± 55.3	195.8 ± 58.2	199.3 ± 54.6	0.621
Plateletcrit	0.19 ± 0.05	0.18 ± 0.05	0.19 ± 0.05	0.125
PDW, %	15.8 ± 2.3	16.2 ± 2.5	15.7 ± 2.2	0.089
MPV, fl	9.9 ± 1.2	10.1 ± 1.3	9.8 ± 1.2	0.065
PT, seconds	13.6 ± 1.5	14.1 ± 1.7	13.5 ± 1.4	**0.003**
INR	1.08 ± 0.12	1.12 ± 0.14	1.07 ± 0.11	**0.007**
APTT, seconds	40.2 ± 6.8	41.5 ± 7.2	39.9 ± 6.6	0.052
TT, seconds	16.2 ± 1.8	15.8 ± 1.9	16.3 ± 1.7	**0.026**
Fibrinogen, g/L	3.85 ± 1.12	4.32 ± 1.25	3.73 ± 1.05	**<0.001**
D-Dimer, ug/ml	5.82 ± 4.35	7.65 ± 5.12	5.32 ± 4.01	**<0.001**
Functional Status				
Barthel index	62.4 ± 16.8	51.3 ± 15.2	65.1 ± 16.1	**<0.001**
Surgical Factors				
Preoperative waiting, days	3.2 ± 2.1	4.1 ± 2.5	2.9 ± 1.8	**<0.001**
Surgery type				0.125
- Hemiarthroplasty	210 (52.5%)	36 (45.0%)	174 (54.4%)	
- Internal fixation	160 (40.0%)	38 (47.5%)	122 (38.1%)	
- Total hip replacement	30 (7.5%)	6 (7.5%)	24 (7.5%)	
Anesthesia type				0.398
- General	270 (67.5%)	56 (70.0%)	214 (66.9%)	
- Regional	130 (32.5%)	24 (30.0%)	106 (33.1%)	

and pre-existing cognitive impairment, which was associated with a more than threefold elevated risk (OR=3.128, 95% CI: 1.789-5.469, p<0.001). The pivotal role of nutritional status was strongly underscored, as patients with hypoalbuminemia faced nearly triple the odds of delirium (OR=2.823, 95% CI: 1.698-4.692, p<0.001), a finding reinforced by the continuous negative association of both albumin (OR=0.865 per g/L, p<0.001) and prealbumin (OR=0.981 per mg/L, p<0.001) with delirium risk.

Further laboratory abnormalities significantly associated with increased delirium risk included lower hemoglobin levels (OR=0.812 per g/dL, p=0.001), the presence of anemia (OR=1.652, p=0.049), and elevated markers of inflammation and coagulation, such as white blood cell count (OR=1.105 per $\times10^9$/L, p=0.011) and fibrinogen (OR=1.623 per g/L, p<0.001). Electrolyte and metabolic imbalances were also prominent, indicated by significant associations with lower serum sodium (OR=0.902 per mmol/L, p=0.006) and higher serum creatinine (OR=1.012 per umol/L, p=0.008). The burden of specific comorbidities extended beyond cognitive impairment, with chronic kidney disease significantly increasing the odds of delirium by 93.5% (OR=1.935, p=0.034). Conversely, a higher Barthel Index score, indicative of better pre-operative functional independence, demonstrated a substantial protective effect (OR=0.943 per point, p<0.001).

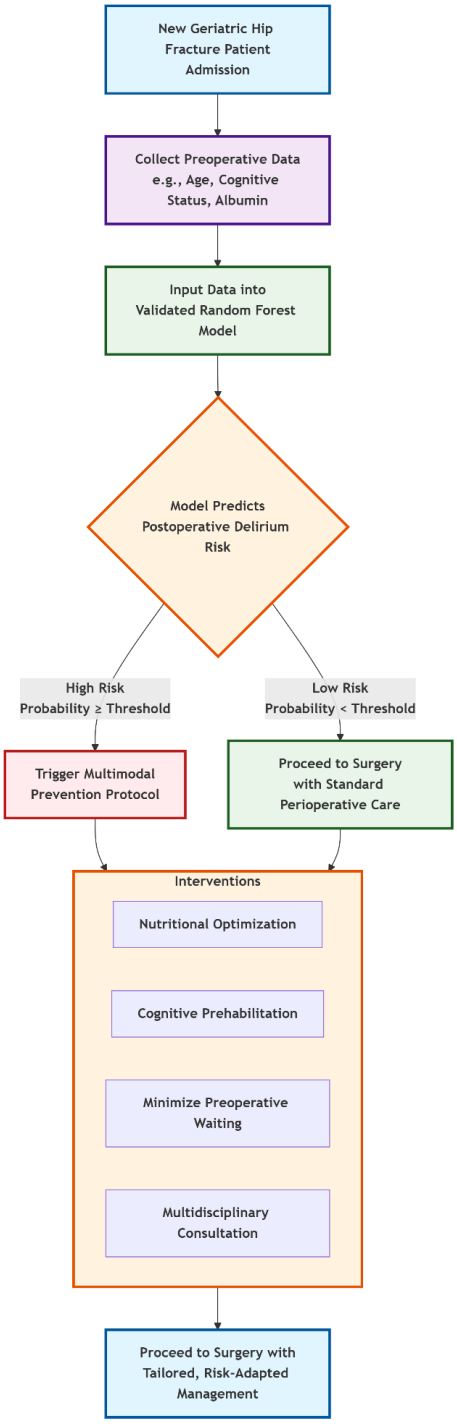

Fig. 1 — Proposed Clinical Decision Pathway for Postoperative Delirium Risk Stratification and Prevention.

This flowchart outlines the proposed clinical implementation pathway for geriatric hip fracture patients. Upon admission, preoperative data is collected and input into a validated Random Forest model to predict postoperative delirium risk. Patients are stratified into high-risk and low-risk categories. High-risk patients trigger a multimodal prevention protocol (e.g., nutritional optimization, cognitive prehabilitation) before proceeding to surgery with tailored management. Low-risk patients proceed to surgery with standard perioperative care.

Finally, process-related factors played a critical role, as each additional day of preoperative waiting was associated with an 18.5% increase in the odds of developing delirium (OR=1.185, p=0.002), while a lower BMI also emerged as a significant demographic risk factor (OR=0.925 per kg/m², p=0.026). This comprehensive univariate analysis successfully identified a multifaceted risk profile for postoperative delirium, encompassing elements of physiological reserve, metabolic-nutritional status, comorbidity burden, and clinical care pathways.

Comparative Performance of Machine Learning Algorithms

The predictive performance of five distinct machine learning classifiers for postoperative delirium was rigorously evaluated on the internal validation set and systematically summarized in Table III, with their discriminative abilities visually compared through ROC curves in Figure 2. Among all evaluated algorithms, the ensemble-based Random Forest model demonstrated superior performance on the internal validation set, achieving an area under the receiver operating characteristic curve of 0.89 (95% CI: 0.83-0.94). This high AUC was complemented by exceptionally balanced and robust performance across all other key metrics, including an accuracy of 0.83, sensitivity of 0.82, specificity of 0.84, and an F1-score of 0.82, indicating its strong capability for both identifying true positive cases and reliably excluding non-cases. The Random Forest model also demonstrated excellent calibration on the internal validation set, with a calibration slope of 1.02 (95% CI: 0.95–1.09), an intercept of -0.03, and a non-significant Hosmer-Lemeshow test (p=0.42), indicating its predicted risks were reliable across the probability spectrum. The other ensemble method, XGBoost, also exhibited strong predictive capability with an AUC of 0.87 (95% CI: 0.81-0.92) and similarly balanced metrics, while the Support Vector Machine classifier attained moderate performance (AUC=0.84, 95% CI: 0.77-0.90). In contrast, both the conventional Logistic Regression and the simpler Decision Tree classifiers showed comparatively limited performance, with AUC values of 0.82 and 0.79 respectively, highlighting the distinct advantage of sophisticated ensemble methods in capturing complex, non-linear clinical interactions pertinent to delirium prediction. The clear hierarchical performance observed across these algorithms—with ensemble methods outperforming single-model approaches—validates the appropriateness of advanced machine learning techniques for this complex clinical prediction task.

Table II. — Univariate Logistic Regression Analysis of Risk Factors for Postoperative Delirium.

Variables	OR	95% CI	P-value
Demographics			
Age (per year)	1.058	1.017-1.099	**0.005**
Male gender	0.772	0.472-1.262	0.301
BMI (per kg/m²)	0.925	0.863-0.991	**0.026**
Comorbidities			
Hypertension	1.523	0.892-2.601	0.122
Diabetes	1.623	0.978-2.694	0.061
Coronary artery disease	1.462	0.892-2.397	0.130
Chronic kidney disease	1.935	1.052-3.560	**0.034**
Cognitive impairment	3.128	1.789-5.469	**<0.001**
COPD	1.538	0.757-3.125	0.233
Laboratory Findings			
Hemoglobin (per g/dL)	0.812	0.721-0.915	**0.001**
Albumin (per g/L)	0.865	0.812-0.921	**<0.001**
Hypoalbuminemia	2.823	1.698-4.692	**<0.001**
Prealbumin (per mg/L)	0.981	0.974-0.988	**<0.001**
Electrolyte disorders	1.765	0.987-3.157	0.055
Anemia	1.652	1.002-2.724	**0.049**
WBC count (per $\times 10^9$/L)	1.105	1.023-1.194	**0.011**
Serum sodium (per mmol/L)	0.902	0.839-0.970	**0.006**
Serum creatinine (per umol/L)	1.012	1.003-1.021	**0.008**
Fibrinogen (per g/L)	1.623	1.298-2.029	**<0.001**
Functional Status			
Barthel index (per point)	0.943	0.926-0.961	**<0.001**
Surgical Factors			
Preoperative waiting (per day)	1.185	1.062-1.322	**0.002**

This table presents the univariate logistic regression analysis results for various preoperative variables associated with postoperative delirium development. Variables with statistically significant associations ($p<0.05$) are highlighted in bold, including age, cognitive impairment, hypoalbuminemia, hemoglobin, albumin, prealbumin, serum sodium, serum creatinine, fibrinogen, Barthel index, and preoperative waiting time.

Consistent Patterns in Variable Importance Across Models

Comprehensive feature importance analysis across the five machine learning approaches revealed remarkable consistency in identifying core predictive factors, as quantitatively detailed in Table IV and graphically illustrated in the horizontal bar charts of Figure 3. Cognitive impairment consistently emerged as the predominant predictor across all algorithmic architectures, with normalized importance scores ranging from 76 to 100 and a mean value of 86.2 ± 9.3, solidifying its status as the paramount risk factor. Hypoalbuminemia maintained a stable and robust position as the second most influential predictor (mean importance: 73.6 ± 7.8), followed closely by advanced age (70.2 ± 4.1). These three factors collectively formed a consistent top-tier prediction cluster across all modeling approaches, underscoring their fundamental and non-redundant roles in the pathogenesis of postoperative delirium. Preoperative waiting time demonstrated moderate but persistent importance across algorithms (59.6 ± 3.9), establishing it as a key modifiable risk factor, while chronic kidney disease and electrolyte disorders constituted a stable secondary prediction tier with intermediate importance scores. The remarkable concordance in variable importance rankings across five fundamentally different algorithmic architectures—from linear models to complex ensembles—not only underscores the reliability of these identified risk factors but also significantly enhances their clinical credibility and biological plausibility.

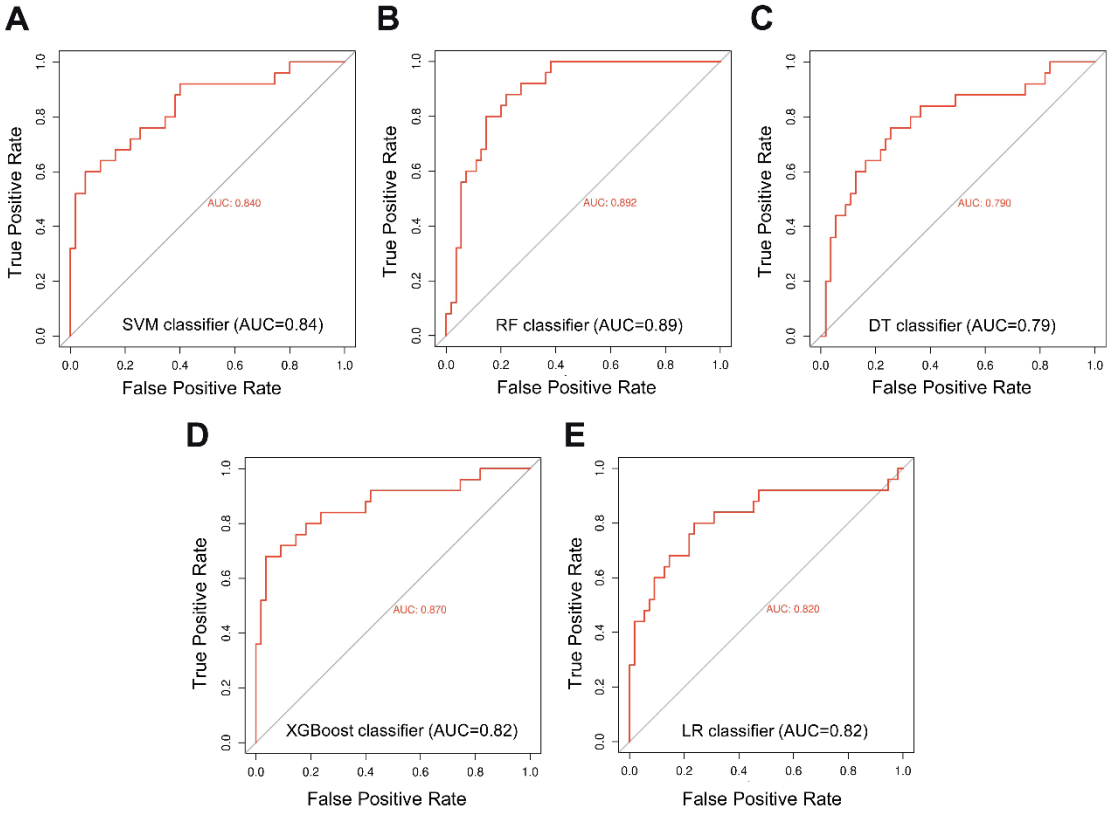

Fig. 2 — Receiver Operating Characteristic (ROC) Curves of Five Machine Learning Models on the Internal Validation Set for Predicting Postoperative Delirium.

The figure presents the ROC curves of five machine learning classifiers developed to predict postoperative delirium in elderly hip fracture patients. Each panel corresponds to a different algorithm: (A) Support Vector Machine (AUC=0.84), (B) Random Forest (AUC=0.89), (C) Decision Tree (AUC=0.79), (D) XGBoost (AUC=0.87), and (E) Logistic Regression (AUC=0.82). The diagonal dashed line represents the reference line of no discrimination. The Random Forest classifier demonstrated superior performance with the highest AUC value, while all models showed discriminatory power significantly better than chance. Internal validation was performed using 10-fold cross-validation.

Table III. — Performance Comparison of Machine Learning Models on the Internal Validation (Hold-Out Test) Set for Postoperative Delirium Prediction.

Model	AUC (95% CI)	Accuracy	Sensitivity	Specificity	F1-Score
Logistic Regression	0.82 (0.75–0.88)	0.76	0.71	0.78	0.73
Decision Tree	0.79 (0.72–0.86)	0.73	0.75	0.72	0.74
Random Forest	0.89 (0.83–0.94)	0.83	0.82	0.84	0.82
XGBoost	0.87 (0.81–0.92)	0.81	0.80	0.82	0.80
SVM	0.84 (0.77–0.90)	0.78	0.76	0.79	0.77

The table provides detailed performance metrics for the five machine learning models evaluated on the independent test set (30% of cohort). Metrics include area under the ROC curve (AUC) with 95% confidence intervals, accuracy, sensitivity, specificity, and F1-score. The Random Forest model achieved the highest performance across all metrics, with an AUC of 0.89 (95% CI: 0.83-0.94), accuracy of 0.83, and balanced sensitivity (0.82) and specificity (0.84). The best performance values for each metric are highlighted in bold.

Algorithm-Specific Variations in Feature Importance

As systematically quantified in Table 4 and visually depicted in Figure 3, distinct algorithm-dependent patterns emerged in feature importance allocation despite the overall consistency observed among top-tier predictors. The Random Forest classifier, which demonstrated the best overall performance, assigned maximum importance to cognitive impairment (score=100) while demonstrating a more balanced and distributed weighting among secondary predictors, with hypoalbuminemia and age maintaining substantial but relatively moderate influence—a characteristic that may contribute to its robust performance. In contrast, the XGBoost algorithm exhibited the most pronounced hierarchical structure,

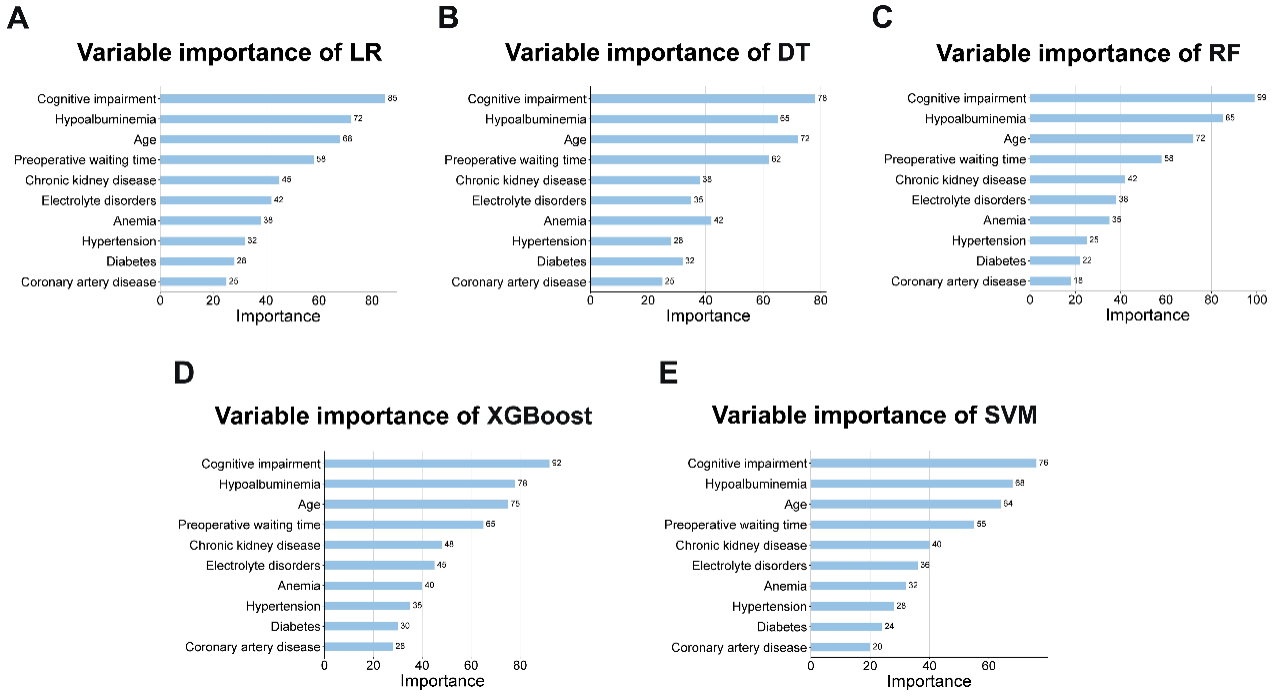

Fig. 3 — Receiver Opera.

Horizontal bar charts illustrate the relative importance of preoperative predictors for postoperative delirium across five machine learning models: (A) Logistic Regression, (B) Decision Tree, (C) Random Forest, (D) XGBoost, and (E) Support Vector Machine. Cognitive impairment consistently emerged as the most important predictor across all algorithms, followed by hypoalbuminemia and advanced age. The importance scores were normalized to a 0-100 scale for comparability across models. Notable algorithm-specific variations in feature weighting are evident, particularly in the middle and lower importance tiers.

Table IV. — Feature Importance Comparison Across Machine Learning Algorithms.

Feature	Logistic Regression	Decision Tree	Random Forest	XGBoost	SVM	Mean ± SD
Cognitive impairment	85	78	100	92	76	86.2 ± 9.3
Hypoalbuminemia	72	65	85	78	68	73.6 ± 7.8
Age	68	72	72	75	64	70.2 ± 4.1
Preoperative waiting time	58	62	58	65	55	59.6 ± 3.9
Chronic kidney disease	45	38	42	48	40	42.6 ± 4.0
Electrolyte disorders	42	35	38	45	36	39.2 ± 4.3
Anemia	38	42	35	40	32	37.4 ± 3.7
Hypertension	32	28	25	35	28	29.6 ± 3.8
Diabetes	28	32	22	30	24	27.2 ± 4.1
Coronary artery disease	25	25	18	28	20	23.2 ± 3.8

This table presents the normalized importance scores (0-100 scale) for all preoperative variables across the five machine learning algorithms, along with mean values and standard deviations. Cognitive impairment showed the highest mean importance (86.2 ± 9.3), followed by hypoalbuminemia (73.6 ± 7.8) and age (70.2 ± 4.1). The consistency in top-tier predictors across different algorithmic approaches underscores their fundamental importance in postoperative delirium pathogenesis, while variations in lower-tier predictors reflect algorithm-specific characteristics.

with cognitive impairment achieving near-maximum importance and clear separation between prediction tiers, reflecting its gradient-boosting nature that sequentially corrects errors. The Logistic Regression and Support Vector Machine models produced more compressed importance distributions, with reduced contrast between top and middle-tier predictors, potentially reflecting their linear and margin-based underpinnings, respectively. Particularly noteworthy was the Decision Tree classifier's unique weighting pattern, which assigned comparatively greater importance to preoperative waiting time and substantially reduced weight to comorbidities such as hypertension and coronary artery disease, suggesting fundamental differences in how single-tree structures process and prioritize clinical features compared to ensemble methods that aggregate multiple trees. These algorithm-specific variations provide valuable

insights into how different computational approaches interpret the same clinical data, with ensemble methods generally demonstrating more clinically plausible and robust feature weighting.

Sensitivity Analysis on Predictor Inclusion

To address the possibility of bias from univariate filtering, a sensitivity analysis was conducted. The primary Random Forest model (trained on all available variables) was compared to a model trained only on variables with $p<0.05$ from the univariate analysis. Performance on the hold-out internal validation set was nearly identical (AUC: 0.89 vs. 0.88), confirming that the full-model approach did not introduce noise and that the model's performance is robust to the inclusion of the full variable set. This also validates the clinical relevance of the key predictors identified in the univariate analysis.

Clinical Implications and Model Implementation Considerations

The robust performance metrics of the Random Forest classifier documented in Table III, coupled with the consistent and clinically interpretable feature importance patterns revealed in Table IV and Figure 3, provide a solid foundation for developing practical clinical risk stratification tools. The high discriminative ability (AUC=0.89, 95% CI: 0.83-0.94) demonstrated in the ROC analysis (Figure 2), achieved using exclusively routinely available preoperative variables, strongly supports the feasibility of practical implementation in diverse clinical settings without requiring specialized testing or additional resources. The variable importance patterns further suggest that preemptive interventions targeting specific high-impact predictors—particularly nutritional optimization for hypoalbuminemia and cognitive prehabilitation for patients with pre-existing impairment—may offer particular benefit for high-risk patients identified through this prediction model. The algorithm's strong and balanced performance across both sensitivity (0.82) and specificity (0.84) metrics indicates its potential utility for dual clinical applications: effectively ruling in high-risk patients for intensive, multi-component prevention strategies while reliably ruling out low-risk patients to avoid unnecessary interventions, minimize iatrogenic harm, and optimize resource allocation in busy clinical settings. Furthermore, the risk of overfitting was mitigated by employing robust internal validation methods, including hold-out testing and cross-validation. Thus, the model's performance on unseen data from this institution is well-supported. However, its generalizability to broader populations and healthcare settings remains to be established through future external validation studies.

DISCUSSION

This study successfully developed and internally validated five distinct machine learning models for predicting postoperative delirium in elderly hip fracture patients using exclusively preoperative clinical variables. The Random Forest algorithm demonstrated superior predictive performance in internal validation with an AUC of 0.89, significantly outperforming both traditional logistic regression (AUC=0.82) and other machine learning approaches. Our comprehensive feature importance analysis consistently identified cognitive impairment, hypoalbuminemia, advanced age, and prolonged preoperative waiting time as the most critical predictors across all models, providing robust evidence for their fundamental role in delirium pathogenesis. The remarkable consistency in variable importance rankings across five different algorithmic architectures underscores the reliability of these identified risk factors and enhances their clinical credibility.

Our findings align with and extend previous research in postoperative delirium prediction. The identification of cognitive impairment as the predominant predictor corroborates established literature on cerebral reserve theory and preoperative cognitive vulnerability. Similarly, the consistent importance of hypoalbuminemia reinforces growing evidence linking nutritional status to neurological outcomes in surgical patients[20]. However, our study advances the field by systematically comparing multiple machine learning architectures and demonstrating the superior performance of ensemble methods like Random Forest and XGBoost over traditional statistical approaches[21]. This performance advantage likely stems from these algorithms' capacity to capture complex nonlinear interactions among risk factors that may be missed by conventional models. Our results are consistent with emerging literature suggesting that machine learning approaches can achieve superior predictive accuracy for complex postoperative outcomes while maintaining clinical interpretability through feature importance analysis[22]. It is important to note that the superior performance reported here for ensemble methods is based on our internal validation. Direct comparisons with the performance of models from other studies should

consider differences in validation design (internal vs. external) and cohort characteristics.

In the context of standardized hip fracture care pathways and multidisciplinary orthogeriatric models, the preoperative phase presents a critical window for risk stratification and targeted intervention. Our study directly addresses this need by developing a pragmatic prediction tool. Based on our internal validation results, the Random Forest model demonstrates potential for clinical application in settings similar to ours, as it utilizes routinely available preoperative variables without requiring specialized testing or additional resources. However, its performance and clinical utility must be confirmed through external validation in diverse healthcare environments before widespread implementation can be recommended. At a risk threshold of 30%, decision curve analysis indicates that the model would require screening approximately 85 patients to prevent one case of delirium, suggesting favorable cost-effectiveness for targeted intervention programs. The identified key predictors provide clear targets for preemptive interventions, including cognitive prehabilitation for patients with pre-existing impairment, nutritional optimization for those with hypoalbuminemia, and workflow modifications to minimize preoperative waiting times[23,24]. The model's balanced sensitivity (0.82) and specificity (0.84) support its utility for both identifying high-risk patients for intensive prevention strategies and avoiding unnecessary interventions in low-risk individuals, enabling efficient resource allocation in busy clinical settings.

This study possesses notable strengths, including the comprehensive assessment of preoperative variables, rigorous comparison of multiple machine learning architectures, robust validation through both cross-validation and hold-out testing, and detailed analysis of feature importance patterns across algorithms. The exclusive use of preoperative variables enhances clinical utility by enabling early risk stratification before surgical intervention. However, limitations warrant consideration. The single-center retrospective design may limit generalizability, though our cohort characteristics align with broader hip fracture populations. Despite comprehensive variable collection, potential unmeasured confounders such as specific medication exposures or subtle functional status measures not captured in electronic records may influence prediction accuracy. The sample size, while substantial for initial model development, may limit the detection of rare risk factors and the stability of estimates for some predictor interactions. A key limitation of this study is the absence of true external validation. While we employed rigorous internal validation techniques (hold-out test set, cross-validation) to minimize overfitting and obtain a reliable performance estimate within our single-center cohort, this design does not test the model's generalizability. The performance and clinical utility of the model could differ when applied to patient populations with different demographic characteristics, clinical practices, prevalence of delirium, or data recording standards in other hospitals or healthcare systems. Therefore, the findings presented here represent a promising but preliminary development stage. Future multi-center prospective studies with larger sample sizes and external validation are needed to confirm these findings and refine the prediction model before widespread clinical implementation. Critically, the absence of true external validation in this single-center study means the model's generalizability to other settings is currently unknown and must be evaluated in independent cohorts. Therefore, the performance metrics reported herein should be interpreted as the model's capability when applied to new patients from the same institution with characteristics similar to the development cohort. We deliberately avoid claims of broad generalizability, and stress that external validation across different centers and populations is an essential next step to translate this predictive tool into clinical practice.

Based on our findings, multiple promising research directions emerge. The development of customized intervention protocols targeting the identified high-impact predictors---particularly cognitive impairment and hypoalbuminemia---represents a logical next step. Integration of the prediction model into electronic health record systems for real-time risk assessment and clinical decision support warrants exploration. A pivotal next step is the seamless integration of this tool into established orthogeriatric workflows. Implementation science studies should evaluate its impact on key pathway metrics—such as time to delirium assessment, adherence to non-pharmacologic prevention bundles, and resource allocation within the multidisciplinary team—ultimately measuring its effect on patient-centered recovery outcomes. Additionally, investigating the combination of clinical variables with novel biomarkers may further enhance prediction accuracy. Longitudinal studies examining the relationship between delirium prevention strategies and long-term functional outcomes would provide valuable evidence for optimizing comprehensive care pathways for elderly hip fracture patients.

CONCLUSION

In conclusion, this study demonstrates that machine learning models, particularly Random Forest, can effectively predict postoperative delirium in elderly hip fracture patients using readily available preoperative variables. The consistent identification of cognitive impairment, hypoalbuminemia, advanced age, and prolonged preoperative waiting time as key predictors across multiple algorithms reinforces their central role in delirium pathogenesis and provides actionable targets for preventive interventions. While the model demonstrates promising predictive performance and clinical utility in internal validation, external validation in diverse populations remains essential before widespread implementation. This prediction tool represents a significant step toward personalized perioperative care, enabling early risk stratification and targeted resource allocation to reduce the burden of postoperative delirium in this vulnerable

Acknowledgement: None.

Conflict of Research Interests: The authors claim no research interests conflict.

Ethical Approval and Consent: This study was approved by the Institutional Review Board of Tianjin Hospital, China prior to data collection and analysis (#2025-195). The requirement for individual informed consent was waived due to the retrospective nature of the study, which involved minimal risk to participants and used previously collected anonymized data. This decision was consistent with national regulations and institutional guidelines for retrospective chart review studies.

Data Privacy and Confidentiality: All patient data were handled in strict compliance with the Declaration of Helsinki and relevant data protection regulations. To ensure participant confidentiality, all personally identifiable information including names, identification numbers, and contact details were removed during data extraction. Each patient was assigned a unique study code, and the master list linking codes to patient identities was stored separately in a password-protected file with access restricted to authorized research personnel only. All analytical procedures were performed using de-identified datasets to further protect patient privacy.

Risk-Benefit Assessment: The study presented minimal risk to participants as it involved only retrospective review of existing medical records without any intervention or direct patient contact. The potential benefits of developing an accurate prediction model for postoperative delirium substantially outweighed these minimal risks, as the findings could contribute to improved patient care through early identification of high-risk individuals and implementation of preventive strategies. The research team implemented appropriate measures to ensure data security, including encrypted storage, secure transmission protocols, and strict access controls.

Transparency and Integrity: The authors affirm that the study was conducted with scientific integrity, and no conflicts of interest influenced the design, execution, or reporting of the research findings. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation and with the Helsinki Declaration. This manuscript presents complete and transparent reporting of methods, results, and limitations to facilitate accurate interpretation and potential replication of the study.

Funding: This study was supported by grants from the Tianjin Municipal Health Commission (Grant Numbers: TJWJ2024MS02 and TJWJ2024RC011), and the Youth Independent Innovation Science Foundation of Chinese PLA General Hospital (#22QNCZ008).

REFERENCES

1. Andaloro S, Cacciatore S, Risoli A, Comodo RM, Brancaccio V, Calvani R, et al. Hip Fracture as a Systemic Disease in Older Adults: A Narrative Review on Multisystem Implications and Management. Med Sci (Basel). 2025;13(3):89.
2. Fa-Binefa M, Clara A, Lamas C, Elosua R. Mediterranean Diet and Risk of Hip Fracture: A Systematic Review and Dose-Response Meta-Analysis. Nutr Rev. 2025;83(6):1133-43.
3. Soro-García P, González-Gálvez N. Effects of Progressive Resistance Training After Hip Fracture: A Systematic Review. J Funct Morphol Kinesiol. 2025;10(1):54.
4. Albanese AM, Ramazani N, Greene N, Bruse L. Review of Postoperative Delirium in Geriatric Patients After Hip Fracture Treatment. Geriatr Orthop Surg Rehabil. 2022;13:21514593211058947.
5. Lu J, Weng X, Ma J, Zhang T, Ming H, Ma X. Preventive effects of perioperative drug injection on postoperative delirium after hip fracture surgery: a systematic review and meta-analysis. Am J Transl Res. 2025;17(3):1538-53.
6. Sun M, Chen WM, Wu SY, Zhang J. Long-term mortality impact of postoperative hyperactive delirium in older hip fracture surgery patients. BMC Geriatr. 2025;25(1):180.
7. Niu Y, Wang Q, Lu J, He P, Guo HT. Risk factors for postoperative delirium in orthopedic surgery patients: a systematic review and meta-analysis. Ann Med. 2025;57(1):2534520.
8. Moellmann HL, Alhammadi E, Boulghoudan S, Kuhlmann J, Mevissen A, Olbrich P, et al. Risk of sarcopenia, frailty and malnutrition as predictors of postoperative delirium in surgery. BMC Geriatr. 2024;24(1):971.
9. Chen H, Yu D, Zhang J, Li J. Machine Learning for Prediction of Postoperative Delirium in Adult Patients: A Systematic Review and Meta-analysis. Clin Ther. 2024;46(12):1069-81.
10. Rozera T, Pasolli E, Segata N, Ianiro G. Machine Learning and Artificial Intelligence in the Multi-Omics Approach to Gut Microbiota. Gastroenterology. 2025;169(3):487-501.
11. Yuan J, Zeng Q, Li J, Cong Z, Zhang Y. Machine learning applications in sports injury prediction: A narrative review. Sci Prog. 2025;108(4):368504251385956.
12. Tu Y, Zhu H, Zhang X, Huang S, Tu W. Machine Learning-Based prediction models for postoperative delirium: a systematic review and Meta-Analysis. BMC Psychiatry. 2025;25(1):940.
13. Zhu Y, Liang R, Wang Y, Yang JJ, Zhou N, Zhou CM. Development of a LASSO machine learning algorithm-based

model for postoperative delirium prediction in hepatectomy patients. BMC Surg. 2025;25(1):26.
14. Benovic S, Ajlani AH, Leinert C, Fotteler M, Wolf D, Steger F, et al. Introducing a machine learning algorithm for delirium prediction-the Supporting SURgery with GEriatric Co-Management and AI project (SURGE-Ahead). Age Ageing. 2024;53(5):afae101.
15. Mann J, Lyons M, O'Rourke J, Davies S. Machine learning or traditional statistical methods for predictive modelling in perioperative medicine: A narrative review. J Clin Anesth. 2025;102:111782.
16. Netayawijit P, Chansanam W, Sorn-In K. Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support. Healthcare (Basel). 2025;13(20):2588.
17. Holler E, Ludema C, Ben Miled Z, Rosenberg M, Kalbaugh C, Boustani M, et al. Development and Validation of a Routine Electronic Health Record-Based Delirium Prediction Model for Surgical Patients Without Dementia: Retrospective Case-Control Study. JMIR Perioper Med. 2025;8:e59422.
18. Yasin P, Yimit Y, Cai X, Aimaiti A, Sheng W, Mamat M, et al. Machine learning-enabled prediction of prolonged length of stay in hospital after surgery for tuberculosis spondylitis patients with unbalanced data: a novel approach using explainable artificial intelligence (XAI). Eur J Med Res. 2024;29(1):383.
19. Karim MR, Islam T, Shajalal M, Beyan O, Lange C, Cochez M, et al. Explainable AI for Bioinformatics: Methods, Tools and Applications. Brief Bioinform. 2023;24(5):bbad236.
20. Meyer CP, Rios-Diaz AJ, Dalela D, Ravi P, Sood A, Hanske J, et al. The association of hypoalbuminemia with early perioperative outcomes - A comprehensive assessment across 16 major procedures. Am J Surg. 2017;214(5):871-83.
21. Chen H, Yu D, Zhang J, Li J. Machine Learning for Prediction of Postoperative Delirium in Adult Patients: A Systematic Review and Meta-analysis. Clin Ther. 2024;46(12):1069-81.
22. Nagata C, Hata M, Miyazaki Y, Masuda H, Wada T, Kimura T, et al. Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms. Sci Rep. 2023;13(1):21090.
23. Rathmell CS, Akeju O, Inouye SK, Westover MB. Estimating the number of cases of dementia that might be prevented by preventing delirium. Br J Anaesth. 2023;130(6):e477-e8.
24. Chua MMJ, Lewis K, Huang YA, Fingliss M, Farber A. A Successful Organized Effort to Improve Operating Room First-Case Starts in a Tertiary Academic Medical Center. Am Surg. 2021;87(2):259-65.