

Original Article
Medical Informatics



Improving Delirium Prediction in Surgery: Enhanced Machine Learning Models vs. Traditional Risk Scores

Kyu Nam Kim ,¹ and Kounseok Lee ²

¹Department of Anesthesiology and Pain Medicine, College of Medicine, Hanyang University, Seoul, Korea

²Department of Psychiatry, College of Medicine, Hanyang University, Seoul, Korea



Received: Jul 1, 2025

Accepted: Oct 14, 2025

Published online: Apr 20, 2026

Author for Correspondence:

Kounseok Lee, MD, PhD

Department of Psychiatry, College of Medicine,
Hanyang University, 222 Wangsimni-ro,
Seongdong-gu, Seoul 04763, Korea.
Email: dual@hanyang.ac.kr

© 2026 The Korean Academy of Medical
Sciences.

This is an Open Access article distributed
under the terms of the Creative Commons
Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>)
which permits unrestricted non-commercial
use, distribution, and reproduction in any
medium, provided the original work is properly
cited.

ORCID iDs

Kyu Nam Kim

<https://orcid.org/0000-0003-2526-8348>

Kounseok Lee

<https://orcid.org/0000-0002-6084-5043>

Disclosure

The authors have no potential conflicts of
interest to disclose.

Data Availability Statement

The dataset used in this study is publicly
available from Dryad (<https://datadryad.org/dataset/doi:10.5061/dryad.51q03>). No
additional data were collected.

ABSTRACT

Background: Postoperative delirium (POD) is a prevalent complication among older surgical patients; however, accurately predicting its occurrence using traditional point-based risk scores remains challenging. This study aimed to develop a delirium prediction model using a contemporary dataset and to compare the performance of modern machine learning (ML) algorithms with that of the Inouye and Marcantonio risk indices.

Methods: A retrospective cohort of 283 surgical patients aged ≥ 50 years (POD incidence = 8.8%) was analyzed. Candidate predictors included demographic characteristics, frailty markers, comorbidities, surgical category, and planned intensive care unit (ICU) admission. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique. Five classifiers—ridge logistic regression (baseline), random forest, XGBoost, CatBoost, and a stacking ensemble—were trained using 5-fold stratified cross-validation. Model discrimination, assessed by the area under the receiver operating characteristic curve (AUC), along with F1 score, precision, and recall, was averaged across folds. For comparison, the Inouye 4-factor and Marcantonio 7-factor scoring systems were recalculated for each patient.

Results: The mean patient age was 65.7 ± 8.6 years, and 40.6% were female. The incidence of POD varied significantly across surgical specialties, ranging from 18.4% in hepatobiliary procedures to 3.7% in general and vascular surgeries. Univariate analysis identified older age, functional dependency, planned ICU admission, and high-risk surgery as significant risk factors. The AUCs of the ML models ranged from 0.57 (logistic regression) to 0.78 (stacking ensemble). Among individual models, CatBoost achieved the best overall balance (AUC = 0.77). In contrast, traditional risk scores demonstrated poor performance (Inouye AUC ≈ 0.50 ; Marcantonio ≈ 0.56).

Conclusion: Contemporary ML models outperformed conventional delirium risk scores in predicting POD, although their overall discrimination remained moderate. Key predictors included age, frailty markers, and surgical complexity. Integration of ML-based screening into electronic health records may enable earlier identification of high-risk patients and facilitate targeted delirium prevention strategies. External validation in larger, multi-center cohorts is warranted to confirm these findings.

Keywords: Emergence Delirium; Machine Learning; Risk Assessment

Author Contributions

Conceptualization: Lee K, Kim KN. Data curation: Lee K. Formal analysis: Lee K. Methodology: Lee K. Project administration: Lee K. Writing - original draft: Kim KN. Writing - review & editing: Lee K, Kim KN.

INTRODUCTION

Delirium is an acute neurocognitive disorder characterized by fluctuating attention and altered cognition, often precipitated by hospitalization, surgery, or critical illness. Postoperative delirium (POD) occurs in approximately 10–30% of older adults undergoing surgery and is associated with prolonged hospital stays, increased morbidity and mortality, and long-term cognitive decline.^{1–6} Early identification of individuals at high risk for POD is crucial, as delirium is frequently preventable through targeted interventions such as proactive geriatric consultation, minimizing deliriogenic medications, and optimizing pain control, sleep, and orientation.^{2,7}

Over the past several decades, various tools have been developed to predict delirium risk.^{8–10} Inouye introduced a foundational model identifying four predisposing risk factors—cognitive impairment, visual impairment, severe illness, and dehydration (elevated blood urea nitrogen-to-creatinine ratio [BUN/Cr])—with the presence of three or more factors indicating high risk.¹¹ Marcantonio subsequently developed the 7-factor Delirium Risk Index for elective surgical patients, assigning points for age ≥ 70 years, history of alcohol use disorder, cognitive or functional impairment, abnormal laboratory findings (e.g., low albumin), and high-risk procedures such as aortic aneurysm repair or thoracic surgery.⁹ However, a systematic review reported that most delirium prediction models for older adults demonstrate only moderate discriminative ability and poor generalizability.¹² Traditional point-based risk scores also exhibit limited sensitivity when applied outside their original validation cohorts.^{13,14} In busy surgical settings, these tools often classify only a small proportion of patients as high risk, thereby failing to detect many individuals who ultimately develop POD.¹⁵

Machine learning (ML) has emerged as a powerful approach to improving clinical prediction by capturing complex, nonlinear relationships within high-dimensional data. Unlike conventional linear models, ML algorithms—such as random forest (RF) and gradient boosting—can model intricate interactions among diverse clinical variables. Recent studies have explored the application of ML for POD prediction with promising results.^{16–21} For instance, Song et al.¹⁸ compared logistic regression with several ML algorithms in older surgical patients and found that the RF model achieved an area under the receiver operating characteristic (ROC) curve (AUC) of approximately 0.80–0.81, outperforming logistic regression (AUC $\approx$ 0.68). Similarly, Zhang et al.¹⁹ developed an automated extreme gradient boosting (XGBoost) model for predicting delirium after spine surgery, achieving an AUC of $\sim$ 0.88 and sensitivity of $\sim$ 91%. Nagata et al.²⁰ applied ML methods in cardiac surgery patients and reported an AUC of $\sim$ 0.76—a modest but clinically meaningful improvement in this high-risk population. Collectively, these findings suggest that ML surpasses traditional regression approaches in identifying patients at risk, particularly when predictor relationships are complex or nonlinear.^{16,17} Delirium is a multifactorial syndrome influenced by baseline vulnerabilities (e.g., advanced age, frailty, cognitive impairment) and precipitating factors (e.g., surgical stress, anesthesia, intensive care unit [ICU] stay).^{5,11,22} ML models are well suited to capture these intricate dependencies.^{17–19,23}

Therefore, the present study aimed to evaluate whether incorporating advanced ML algorithms could enhance delirium risk prediction beyond traditional scoring systems. Using an augmented dataset of 283 surgical patients with an 8.8% incidence of POD,¹⁵ we hypothesized that advanced ML models would demonstrate superior discrimination (AUC)

and a more balanced sensitivity–specificity trade-off (F1 score) compared with logistic regression and conventional risk indices, thereby improving the identification of patients at risk for POD.

METHODS

Participants

A single-center retrospective dataset comprising 292 surgical patients aged ≥ 50 years who underwent major noncardiac procedures was analyzed, as originally reported by Jansen et al.¹⁵ After excluding nine patients without documented delirium assessments, 283 patients were included in the final analysis. POD was defined as either 1) a positive Confusion Assessment Method result—requiring acute onset and fluctuating course, inattention, and either disorganized thinking or altered level of consciousness—or 2) a physician-documented diagnosis in progress notes or discharge summaries consistent with the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria. International Classification of Diseases (ICD) codes were not used for case identification.⁸ Overall, 25 patients (8.8%) experienced POD. Predictor variables were extracted from preoperative assessments, intraoperative records, and immediate postoperative documentation. These included demographic factors (age and sex), comorbidities (Charlson comorbidity index, history of stroke, and dementia), baseline cognitive and functional status (memory impairment and need for assistance with activities of daily living [ADL]), substance use, surgical procedure type, anesthetic characteristics, and planned postoperative disposition. Planned ICU admission was recorded as a proxy for surgical complexity and coded as not applicable in emergency cases.²⁴ Frailty was assessed using an index derived from the Groningen Frailty Indicator (GFI), which consists of 15 binary items addressing functional limitations, sensory deficits, mood disturbances, and nutritional status. Laboratory abnormalities were recorded when available, although some data were missing.²⁵ Preoperative illness severity was quantified using the Acute Physiology and Chronic Health Evaluation II (APACHE II) score.²⁶

Two established delirium risk scores were calculated for each patient: the Inouye 4-factor score and the Marcantonio 7-factor score.^{9,27} The Inouye score assigns one point for each of the following: 1) cognitive impairment (preoperative memory or orientation deficits or diagnosed dementia), 2) visual impairment (requiring corrective lenses or low vision as defined by the GFI), 3) severe illness (APACHE II > 16), and 4) dehydration (BUN/Cr ≥ 18).²⁷ The total Inouye score (INOUE_TOT; range 0–4) was calculated, and patients with scores ≥ 3 were classified as high risk (INOUE_HIGH = 1). The Marcantonio score assigns points for age ≥ 70 years, history of alcohol use disorder, cognitive impairment, functional impairment (requiring assistance with ≥ 1 ADL), abnormal laboratory findings (albumin < 35 g/L or BUN/Cr > 18), and high-risk surgery (thoracic or complex abdominal procedures).⁹ The total Marcantonio score (range 0–7) was calculated, and patients with scores ≥ 2 were classified as high risk. For both scores, missing components were assigned 0 points under the assumption of normal values; the proportion of cases with missing data affecting score calculation was minimal (~4%) and was managed accordingly.

Data preprocessing

Data cleaning and preprocessing were performed before model development. Continuous variables were inspected for outliers and assessed for normality of distribution. Variables

that were markedly skewed and represented postoperative outcomes rather than preoperative predictors were excluded. Categorical variables were encoded as binary (0/1), and multilevel categorical features were transformed into one-hot encoded dummy variables, with infrequent categories consolidated when necessary.

Missing values were addressed using appropriate imputation methods. No patient records lacked age or demographic data. For missing binary risk factors, values were conservatively coded as “no,” which affected only one to two patients. Serum potassium, BUN/Cr ratio, and albumin exhibited variable degrees of missingness (4–40% of cases). Serum potassium was excluded because of its high rate of missingness and lack of association with delirium. For the Marcantonio score, missing albumin values were conservatively assumed to be within the normal range (≥ 35 g/L), as most missing values corresponded to elective surgical patients who were routinely screened and unlikely to have hypoalbuminemia. The Acute Physiology and Chronic Health Evaluation II (APACHE II) score was calculated using available data, and missing continuous variables were imputed using median values, a standard practice in validated severity scoring systems. Binary predictors with missing data were coded as “no.” To evaluate the robustness of these imputation assumptions, a sensitivity analysis was performed comparing alternative strategies (normal vs. median imputation for albumin; median vs. normal for APACHE II). As shown in **Supplementary Table 1**, the choice of imputation method did not materially alter model discrimination or calibration, supporting the validity of the primary approach.

ML models

Four ML models were developed to predict postoperative delirium: logistic regression with L2 regularization, RF, extreme gradient boosting (XGBoost), and categorical boosting (CatBoost). Logistic regression served as the baseline model, incorporating an L2-penalized solver ($C = 1$) to mitigate overfitting. Tree-based models were constructed using default hyperparameters: RF included 100 trees with Gini impurity as the split criterion, while both XGBoost and CatBoost employed 100 trees with a learning rate of 0.1.²⁸ For XGBoost, the maximum tree depth was set to six. Extensive hyperparameter tuning was not performed because of the limited dataset size; instead, model comparability under consistent default configurations was prioritized. All models were trained using identical preoperative predictors, including demographic characteristics (age, sex), comorbidities (e.g., Charlson comorbidity index, stroke history), frailty indicators, baseline cognitive and functional status, laboratory values (e.g., albumin, BUN/Cr), and surgical variables (procedure type and planned ICU admission). The Inouye and Marcantonio scores were recalculated separately according to their original published definitions. Model training employed 5-fold stratified cross-validation without repetition to ensure balanced class representation while prioritizing algorithmic comparison over hyperparameter optimization. All analyses were conducted in Python using the scikit-learn (v1.6.1), XGBoost (v3.0.0), and CatBoost (v1.2.8) libraries.

Addressing class imbalance

Given the low incidence of delirium (8.8%), the dataset exhibited substantial class imbalance (~1:11 ratio). To address this issue, oversampling was implemented within a 5-fold stratified cross-validation framework while maintaining consistent class distribution across folds. Specifically, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to training subsets to generate synthetic minority samples until class balance was achieved, whereas validation folds remained unaltered to prevent information leakage.²⁹ Logistic regression, RF, XGBoost, CatBoost, and stacking ensemble models were trained

on the resampled data. Although this design minimized the risk of leakage, oversampling may still alter the distribution of training data relative to real-world prevalence; therefore, results should be interpreted with caution. Model performance was evaluated on the original validation folds to assess discrimination (AUC) and threshold-based metrics (F1 score, precision, and recall) under realistic conditions.³⁰

Model evaluation metrics

Model performance was evaluated using the AUC to assess discrimination, along with F1 score, precision, and recall for the delirium class. Metrics were computed for each fold at a probability threshold of 0.5 and averaged across all folds. Given the class imbalance, emphasis was placed on minority-class performance metrics. DeLong's test was used to compare AUCs between models.³¹ For comparison, delirium risk was also calculated using the Inouye and Marcantonio scoring systems. The Inouye model assigned points for age ≥ 70 years, cognitive impairment, abnormal BUN/Cr or electrolyte imbalance, and functional impairment, with ≥ 3 points indicating high risk. The Marcantonio model included seven risk factors, with ≥ 2 points indicating high risk.⁹ Missing components were treated as normal values. Sensitivity, specificity, and AUC were evaluated for both traditional risk scores.

Statistical analysis

Continuous variables were compared between the delirium and non-delirium groups using independent-samples *t*-tests or Mann–Whitney *U* tests, depending on data distribution. Categorical variables were analyzed using chi-square or Fisher's exact tests, as appropriate. Two-tailed *P* values were reported, with $P < 0.05$ considered statistically significant. Cross-validated AUCs were compared using DeLong's test for correlated ROC curves. All statistical analyses were performed using Python 3.12.0 and JASP 0.19.3.0.

Ethics statement

This study represents a secondary analysis of an open dataset and was approved by the Institutional Review Board of Hanyang University Hospital (approval No. HYUH 2025-05-011). The requirement for informed consent was waived by the board because the study utilized de-identified, publicly available data.

RESULTS

Baseline patient characteristics

A total of 283 patients were analyzed (mean age 65.7 ± 8.6 years; range 50–89 years), of whom 40.6% were female. POD occurred in 8.8% of patients (25/283). **Table 1** summarizes baseline cohort characteristics and univariate comparisons between patients with and without POD. Patients who developed delirium were older and exhibited greater baseline vulnerabilities than those who did not. Among patients with POD, 36% required preoperative assistance with ADLs compared with 14% of those without POD ($P = 0.008$), indicating that functional dependency was a significant predictor. Additionally, 68% of patients with POD had planned ICU admission postoperatively, compared with 28% of patients without POD ($P < 0.001$), suggesting higher delirium rates among individuals requiring ICU-level care. High-risk surgeries were also more common in the delirium group (32% vs. 11%, $P = 0.007$). Baseline cognitive impairment was more prevalent among delirious patients (28% vs. 12%, $P = 0.054$), although this difference did not reach statistical significance. Sex was not associated with POD occurrence ($P = 0.720$). Patients with delirium exhibited a greater comorbidity

Table 1. Baseline patient characteristics and univariate analysis by postoperative delirium status

Characteristics	All patients (N = 283)	Delirium (n = 25)	No delirium (n = 258)	P value
Age, mean $\pm$ SD, yr	65.7 $\pm$ 8.6	69.6 $\pm$ 7.9	65.3 $\pm$ 8.6	0.041*
Sex, female, No. (%)	115 (40.6%)	11 (44.0%)	104 (40.3%)	0.720
Comorbidity count, median [IQR]	2 [1–3]	3 [2–4]	2 [1–3]	0.021*
History of stroke (CVA), %	10.6%	20.0%	9.3%	0.100
Any baseline cognitive impairment (memory issues or confusion), %	13.8%	28.0%	12.4%	0.054
Dependent in $\geq$ 1 ADL, %	16.3%	36.0%	14.0%	0.008*
Planned ICU admission (elective cases), %	31.8%	68.0% (17/25)	28.3% (73/258)	< 0.001*
Emergency surgery (unplanned admission), %	9.9%	4.0%	10.5%	0.480
High-risk surgery (thoracic or complex abdominal), %	13.1%	32.0%	11.2%	0.007*
Duration of surgery, median [IQR], hr	3.2 [2.5–4.6]	3.5 [2.8–5.0]	3.2 [2.4–4.5]	0.270
APACHE II score, mean $\pm$ SD	6.2 $\pm$ 2.7	7.0 $\pm$ 3.1	6.1 $\pm$ 2.7	0.130
Frailty (GFI score $\geq$ 4), %	14.5%	28.0%	13.2%	0.053
Length of hospital stay, median [IQR], day	7 [5–10]	10 [7–16]	7 [5–9]	< 0.001*

SD = standard deviation, IQR = interquartile range, CVA = cerebrovascular accident, ADL = activities of daily living, ICU = intensive care unit, APACHE II = Acute Physiology and Chronic Health Evaluation II, GFI = Groningen Frailty Index.

*P < 0.05 (statistically significant).

burden (median of 3 vs. 2 diagnoses, $P = 0.021$). Emergency surgery was not more frequent among delirious patients, as most delirium cases followed elective but complex procedures. Collectively, these findings align with established risk factors for delirium, including advanced age, frailty, and surgical stress.^{5,6}

Incidence of postoperative delirium by surgical category

POD incidence varied significantly across surgical specialties (Table 2). Hepatobiliary procedures showed the highest rate (18.4%), followed by orthopedic/spine (16.7%) and oncologic surgeries (12.5%). In contrast, general and vascular surgeries demonstrated similarly low rates (~3.7%). These findings emphasize surgical complexity as a key precipitating factor for POD.

Predictive performance of the ML model

ML models were trained to predict POD, and their performance metrics from 5-fold cross-validation are presented in Table 3. The stacking ensemble employed logistic regression as a meta-learner combining RF, XGBoost, and CatBoost models. Most ML models achieved AUC values ranging from 0.745 to 0.780, indicating moderate discrimination (Fig. 1). Logistic regression produced the lowest performance (AUC = 0.568; F1 = 0.189) with high recall (32%) but low precision (14%), suggesting frequent overprediction. RF achieved an AUC of 0.766 but an F1 score of only 0.04, detecting 4% of cases. XGBoost demonstrated balanced performance (AUC = 0.745; F1 = 0.216), while CatBoost achieved a comparable AUC (0.766) with a higher F1 score (0.287) and improved precision. The stacking ensemble model yielded the highest AUC (0.780) but a slightly lower F1 score (0.226) than CatBoost.

Table 2. Incidence of postoperative delirium by surgical category

Surgical category	No.	POD cases	Incidence, %
General surgery	53	2	3.8
Vascular surgery	82	3	3.7
Hepatobiliary surgery	38	7	18.4
Oncologic surgery	88	11	12.5
Orthopedic/Spine surgery	12	2	16.7
Total	273 ^a	25	8.8

POD = postoperative delirium.

^aIncludes 10 patients from other low-volume specialties not listed individually.

Table 3. Predictive performance of machine learning models compared to traditional risk scores

Model/Score	AUC	F1 score	Precision	Recall
Logistic regression	0.568	0.189	0.137	0.32
Random forest	0.766	0.050	0.067	0.04
XGBoost	0.745	0.216	0.233	0.24
CatBoost	0.766	0.287	0.355	0.28
Stacking ensemble	0.780	0.226	0.367	0.20
Inouye 4-factor score	0.500	0.000	0.000	0.00
Marcantonio 7-factor score	0.560	0.190	0.190	0.21

Machine Learning metrics represent mean performance across 5-fold cross-validation. Precision and recall indicate performance on the positive class (delirium). The stacking ensemble used a logistic regression meta-learner on outputs from random forest, XGBoost, and CatBoost.

AUC = area under the receiver operating characteristic curve.

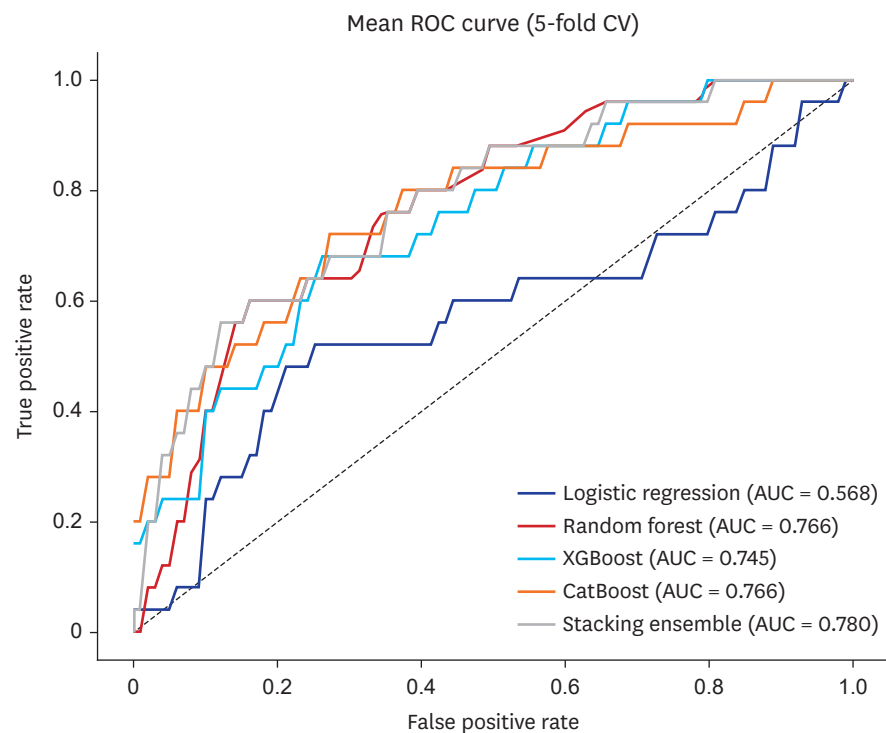


Fig. 1. ROC curves of machine learning models compared to chance. The ROC curve of the logistic regression model (blue) lies near the diagonal line, indicating limited discriminative ability. In contrast, the ROC curves of the tree-based models and the stacking ensemble (orange, green, red, and purple) curve upward toward the top-left corner, indicating better trade-offs between true-positive and false-positive rates and higher AUC values. Results for the Inouye and Marcantonio risk scores are not shown; their performance would correspond to points near the bottom of the plot, indicating very low sensitivity.

ROC = receiver operating characteristic, CV = cross-validation, AUC = area under the receiver operating characteristic curve.

Traditional delirium risk scores performed poorly. The Inouye model classified only four patients (1.4% of 283) as high risk (≥ 3 factors), none of whom developed delirium. All 25 POD patients were categorized as low risk (1–2 factors), yielding 0% sensitivity and precision, with an AUC of 0.50. These findings suggest that the Inouye criteria, originally developed for older medical inpatients, did not generalize to this surgical cohort. Most patients underwent elective procedures and were relatively healthy, failing to meet “high-risk” thresholds despite developing POD. Furthermore, some predictors—such as dehydration (BUN/Cr abnormalities)—were corrected perioperatively, diminishing their predictive value. The Marcantonio index demonstrated slightly better but still limited performance. It classified

27 patients (9.5%) as high risk (score ≥ 2), correctly identifying 5 of 24 delirium cases with complete data (sensitivity $\approx 20.8\%$). The AUC difference between the ML models and traditional risk scores was statistically significant ($P < 0.001$, DeLong's test).

DISCUSSION

In this study, we developed and evaluated ML models for predicting POD in an older surgical population and compared their performance with established clinical risk indices. The ML models, particularly tree-based algorithms and ensemble methods, demonstrated superior discriminative performance compared with the traditional scoring systems developed by Inouye and Marcantonio. The top-performing ML model achieved an AUC of 0.78, significantly exceeding that of conventional tools (AUCs = 0.50–0.56) and exhibiting greater capability in stratifying patients by delirium risk.

These findings are consistent with previous studies applying ML approaches to delirium prediction, which similarly demonstrated enhanced performance over traditional risk scores that often show low sensitivity in external validation settings.^{13,32} ML algorithms can integrate a greater number of variables and their complex interactions, thereby identifying additional predictors. In the present dataset, planned ICU admission and surgical procedure type emerged as strong predictors—factors not fully represented in traditional indices such as the Inouye and Marcantonio models.²⁴ Feature importance analysis from the CatBoost model further identified age and baseline functional or cognitive impairment as key predictors, in alignment with established clinical evidence regarding frailty and aging.^{6,22,24} Collectively, these results support the robustness of ICU admission and surgical complexity as predictors and suggest that ML models may detect nuanced risk patterns overlooked by conventional point-based tools. However, the performance differences among ML algorithms were modest. Both RF and XGBoost achieved AUCs of approximately 0.75, while boosting and stacking methods yielded only marginal gains (AUC ≈ 0.78) without notable improvements in F1 score. This observation indicates that simpler tree-based models captured most of the predictive signal within the available dataset.

The limited performance of the Inouye and Marcantonio indices was anticipated, given that these models were derived from distinct populations and include a restricted set of risk factors. Many patients who developed delirium in this cohort did not meet the high-risk thresholds defined by these tools, underscoring the need for recalibration or the adoption of more adaptive, data-driven ML approaches. Although rule-based algorithms were insufficient, traditional predictors such as age and cognitive impairment retained their predictive value within ML models. The enhanced discriminative performance of ML approaches has important clinical implications, enabling targeted prevention strategies such as orientation protocols and judicious sedative use. In the best-performing CatBoost model, 36% of high-risk patients developed POD, compared with 0% and 19% identified by the Inouye and Marcantonio systems, respectively. While the overall precision and F1 scores were modest, these models may still serve as valuable screening tools when prioritizing sensitivity and early detection, particularly in contexts where preventive interventions are low risk. Conversely, for resource-intensive strategies, models with higher precision would be required. Incorporating biomarkers or additional perioperative variables may further enhance precision and improve risk stratification.

This study demonstrates that ML modestly enhances POD prediction through the integration of preoperative data. These models could be incorporated into electronic health records to support clinical decision-making, using automated dashboards to estimate delirium risk following surgical scheduling and to identify high-risk patients for preventive interventions. Model performance may improve with larger datasets, while recalibration could enhance external validity and generalizability. Incorporating intraoperative and postoperative variables may further strengthen predictive accuracy, although doing so would require more complex data inputs.

These results are consistent with previous investigations reporting the benefits of ML for delirium prediction.^{17-19,21,32} Nagata et al.²⁰ achieved an AUC of approximately 0.80 using gradient boosting to predict POD in cardiovascular surgery patients, while other studies in orthopedic and ICU populations demonstrated that tree-based models outperform conventional approaches. In our study, ML models exhibited modest predictive performance, likely constrained by limited sample size and low event rates; nevertheless, they substantially outperformed traditional risk scores, which achieved AUCs of approximately 0.5. This observation underscores the limitations of generalized point-based algorithms and highlights the advantages of data-driven modeling tailored to specific clinical contexts. Similarly, Benovic et al.³³ reported an AUC of 0.81 using a simple linear support vector machine in 878 surgical patients, comparable to our ensemble models. Our findings further confirm that increasing model complexity yielded only marginal gains over simpler RF and XGBoost models within comparable datasets.

This study has several limitations. First, the small sample size constrained the models' ability to capture complex patterns and increased the risk of overfitting, despite the use of cross-validation and regularization. Additionally, cross-validation was not repeated multiple times, potentially limiting the stability of performance estimates. Second, as a single-center study with a relatively small sample, the patient population and delirium assessment methods may limit generalizability to other settings. The absence of external validation further necessitates cautious interpretation of these results and underscores the need for replication in larger, multicenter cohorts. Third, missing data for certain predictors and the reliance on surrogate or imputed values may have introduced bias. Fourth, although SMOTE oversampling was confined to training folds, it may have altered the original class distribution, thereby limiting real-world applicability. These limitations collectively emphasize the importance of validating model performance in larger, diverse populations without oversampling. Finally, although the present models were restricted to preoperative and intraoperative factors, excluding postoperative variables may have constrained predictive accuracy but preserved clinical utility for preemptive applications.

Despite these limitations, our findings indicate that advanced ML algorithms enhance the identification of surgical patients at risk for POD compared with traditional risk scoring systems. Ensemble and boosting models effectively integrated multiple predictors and captured nonlinear relationships that conventional models could not represent. From a clinical standpoint, this advancement improves preoperative risk stratification and enables earlier identification of patients who may benefit from targeted delirium prevention strategies. Although prediction was the primary focus of this study, model-derived insights into feature importance may assist clinicians in recognizing modifiable risk factors. Ultimately, the goal is to develop a predictive tool that is both accurate and clinically actionable, facilitating targeted prevention and improving postoperative outcomes.

Advanced ML models outperformed conventional delirium risk scores in predicting POD among older surgical patients. Key predictive factors included age, frailty indicators, surgical procedure type, and planned ICU admission—variables often underrepresented or underweighted in traditional scoring systems. These findings underscore the advantages of data-driven modeling approaches in perioperative care. Integrating ML-based risk stratification into clinical workflows may facilitate the early identification of high-risk patients, support targeted prevention strategies, and ultimately improve postoperative outcomes.

SUPPLEMENTARY MATERIAL

Supplementary Table 1

Predictive performance of CatBoost model by different missing data handling strategies

REFERENCES

1. Schenning KJ, Deiner SG. Postoperative delirium in the geriatric patient. *Anesthesiol Clin* 2015;33(3):505-16. [PUBMED](#) | [CROSSREF](#)
2. Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383(9920):911-22. [PUBMED](#) | [CROSSREF](#)
3. Aoki Y, Kurita T, Nakajima M, Imai R, Suzuki Y, Makino H, et al. Association between remimazolam and postoperative delirium in older adults undergoing elective cardiovascular surgery: a prospective cohort study. *J Anesth* 2023;37(1):13-22. [PUBMED](#) | [CROSSREF](#)
4. Choi E, Seo YE, Yoo HS, Kim DY. Incidence and risk factors of postoperative recovery room delirium in elderly patients after surgery under general anesthesia. *Korean J Adult Nurs* 2022;34(2):215-24. [CROSSREF](#)
5. Oh ST, Park JY. Postoperative delirium. *Korean J Anesthesiol* 2019;72(1):4-12. [PUBMED](#) | [CROSSREF](#)
6. Kwak MJ. Delirium in frail older adults. *Ann Geriatr Med Res* 2021;25(3):150-9. [PUBMED](#) | [CROSSREF](#)
7. Siddiqi N, Harrison JK, Clegg A, Teale EA, Young J, Taylor J, et al. Interventions for preventing delirium in hospitalised non-ICU patients. *Cochrane Database Syst Rev* 2016;3(3):CD005563. [PUBMED](#) | [CROSSREF](#)
8. Inouye SK, van Dyck CH, Alessi CA, Balkin S, Siegel AP, Horwitz RI. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med* 1990;113(12):941-8. [PUBMED](#) | [CROSSREF](#)
9. Marcantonio ER, Goldman L, Mangione CM, Ludwig LE, Muraca B, Haslauer CM, et al. A clinical prediction rule for delirium after elective noncardiac surgery. *JAMA* 1994;271(2):134-9. [PUBMED](#) | [CROSSREF](#)
10. Dworkin A, Lee DS, An AR, Goodlin SJ. A simple tool to predict development of delirium after elective surgery. *J Am Geriatr Soc* 2016;64(11):e149-53. [PUBMED](#) | [CROSSREF](#)
11. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons. Predictive model and interrelationship with baseline vulnerability. *JAMA* 1996;275(11):852-7. [PUBMED](#) | [CROSSREF](#)
12. Lindroth H, Bratzke L, Purvis S, Brown R, Coburn M, Mrkobrada M, et al. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ Open* 2018;8(4):e019223. [PUBMED](#) | [CROSSREF](#)
13. Ruppert MM, Lipori J, Patel S, Ingersent E, Cupka J, Ozrazgat-Baslanti T, et al. ICU delirium-prediction models: a systematic review. *Crit Care Explor* 2020;2(12):e0296. [PUBMED](#) | [CROSSREF](#)
14. Ely EW, Inouye SK, Bernard GR, Gordon S, Francis J, May L, et al. Delirium in mechanically ventilated patients: validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). *JAMA* 2001;286(21):2703-10. [PUBMED](#) | [CROSSREF](#)
15. Jansen CJ, Absalom AR, de Bock GH, van Leeuwen BL, Izaks GJ. Performance and agreement of risk stratification instruments for postoperative delirium in persons aged 50 years or older. *PLoS One* 2014;9(12):e113946. [PUBMED](#) | [CROSSREF](#)
16. Bishara A, Chiu C, Whitlock EL, Douglas VC, Lee S, Butte AJ, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol* 2022;22(1):8. [PUBMED](#) | [CROSSREF](#)

17. Song YX, Yang XD, Luo YG, Ouyang CL, Yu Y, Ma YL, et al. Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: a retrospective study. *CNS Neurosci Ther* 2023;29(1):158-67. [PUBMED](#) | [CROSSREF](#)
18. Song Y, Zhang D, Wang Q, Liu Y, Chen K, Sun J, et al. Prediction models for postoperative delirium in elderly patients with machine-learning algorithms and SHapley Additive exPlanations. *Transl Psychiatry* 2024;14(1):57. [PUBMED](#) | [CROSSREF](#)
19. Zhang Y, Wan DH, Chen M, Li YL, Ying H, Yao GL, et al. Automated machine learning-based model for the prediction of delirium in patients after surgery for degenerative spinal disease. *CNS Neurosci Ther* 2023;29(1):282-95. [PUBMED](#) | [CROSSREF](#)
20. 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. [PUBMED](#) | [CROSSREF](#)
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. [PUBMED](#) | [CROSSREF](#)
22. Partridge JS, Harari D, Dhesei JK. Frailty in the older surgical patient: a review. *Age Ageing* 2012;41(2):142-7. [PUBMED](#) | [CROSSREF](#)
23. Ren Y, Loftus TJ, Datta S, Ruppert MM, Guan Z, Miao S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. *JAMA Netw Open* 2022;5(5):e2211973. [PUBMED](#) | [CROSSREF](#)
24. Onwochei DN, Fabes J, Walker D, Kumar G, Moonesinghe SR. Critical care after major surgery: a systematic review of risk factors for unplanned admission. *Anaesthesia* 2020;75 Suppl 1:e62-74. [PUBMED](#) | [CROSSREF](#)
25. Pol RA, van Leeuwen BL, Visser L, Izaks GJ, van den Dungen JJ, Tielliu IF, et al. Standardised frailty indicator as predictor for postoperative delirium after vascular surgery: a prospective cohort study. *Eur J Vasc Endovasc Surg* 2011;42(6):824-30. [PUBMED](#) | [CROSSREF](#)
26. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985;13(10):818-29. [PUBMED](#) | [CROSSREF](#)
27. Inouye SK, Zhang Y, Jones RN, Kiely DK, Yang F, Marcantonio ER. Risk factors for delirium at discharge: development and validation of a predictive model. *Arch Intern Med* 2007;167(13):1406-13. [PUBMED](#) | [CROSSREF](#)
28. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 2009;10(1):213. [PUBMED](#) | [CROSSREF](#)
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321-57. [CROSSREF](#)
30. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep* 2024;14(1):6086. [PUBMED](#) | [CROSSREF](#)
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837-45. [PUBMED](#) | [CROSSREF](#)
32. Xie Q, Wang X, Pei J, Wu Y, Guo Q, Su Y, et al. Machine learning-based prediction models for delirium: a systematic review and meta-analysis. *J Am Med Dir Assoc* 2022;23(10):1655-1668.e6. [PUBMED](#) | [CROSSREF](#)
33. 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. [PUBMED](#) | [CROSSREF](#)