Machine Learning Classifier Models Can Identify Delirium in Intensive Care Units

Anmin Hu (✉ anmin.edu@gmail.com)

Shenzhen People's Hospital <https://orcid.org/0000-0002-6507-6423>

Hui-Ping Li

Southern University of Science and Technology

Zhen Li

shenzhen research institute of big data

Zhongjun Zhang

Southern University of Science and Technology

Xiong-Xiong Zhong

Jinan University

Research

Keywords: Delirium Assessment, Intensive Care Unit Care, Predictive Modeling, Multiple Risk Factors, Machine learning.

Posted Date: August 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-798902/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Purpose: The aim of this study was to use machine learning to construct a model for the analysis of risk factors and prediction of delirium among ICU patients.

Methods: We developed a set of real-world data to enable the comparison of the reliability and accuracy of delirium prediction models from the MIMIC-III database, the MIMIC-IV database and the eICU Collaborative Research Database. Significance tests, correlation analysis, and factor analysis were used to individually screen 80 potential risk factors. The predictive algorithms were run using the following models: Logistic regression, naive Bayesian, K-nearest neighbors, support vector machine, random forest, and eXtreme Gradient Boosting. Conventional E-PRE-DELIRIC and eighteen models, including all-factor (AF) models with all potential variables, characteristic variable (CV) models with principal component factors, and rapid predictive (RP) models without laboratory test results, were used to construct the risk prediction model for delirium. The performance of these machine learning models was measured by the area under the receiver operating characteristic curve (AUC) of tenfold cross-validation. The VIMs and SHAP algorithms, feature interpretation and sample prediction interpretation algorithms of the machine learning black box model were implemented.

Results: A total of 78,365 patients were enrolled in this study, 22,159 of whom (28.28%) had positive delirium records. The E-PRE-DELIRIC model (AUC, 0.77), CV models (AUC, 0.77-0.93), CV models (AUC, 0.77-0.88) and RP models (AUC, 0.75-0.87) had discriminatory value. The random forest CV model found that the top five factors accounting for the weight of delirium were length of ICU stay, verbal response score, APACHE-III score, urine volume and hemoglobin. The SHAP values in the eXtreme Gradient Boosting CV model showed that the top three features that were negatively correlated with outcomes were verbal response score, urine volume, and hemoglobin; the top three characteristics that were positively correlated with outcomes were length of ICU stay, APACHE-III score, and alanine transaminase.

Conclusion: Even with a small number of variables, machine learning has a good ability to predict delirium in critically ill patients. Characteristic variables provide direction for early intervention to reduce the risk of delirium.

Introduction

Delirium is an acute fluctuating mental state change syndrome characterized by disturbances of consciousness, often accompanied by sleep-wake cycle disturbances, attention deficits, and cognitive and emotional disturbances [1]. In the intensive care unit (ICU), the incidence of delirium is 20%-50% when patients are not receiving mechanical ventilation and as high as 60%-80% when patients are receiving mechanical ventilation [2, 3]. Delirium can lead to a longer mechanical ventilation time, a longer stay in the ICU, increased medical expenses, decreased abilities of daily living after discharge, and poor prognosis [2-4]. It is particularly important to accurately identify and correct reversible causes of delirium in the early stage [5-7].

Existing clinical delirium risk prediction tools have achieved areas under the receiver operating characteristic curve (AUCs) of 0.68 to 0.89 [8-12]. The Prediction of Delirium in ICU Patients (PRE-DELIRIC) model and the early PRE-DELIRIC (E-PRE-DELIRIC) model can help medical staff identify groups at high risk of delirium [7-10]. However, the reliability of these models depends on whether the scorer has been professionally trained [13, 14]. Many factors that have been proven to be strongly related to delirium in recent years are missing from these models [7, 8, 15]. Recent work has highlighted the potential of machine learning (ML) algorithms for predicting delirium [11, 16, 17]. However, because data collection took place at a consistent time for all patients, they cannot accurately reflect the patient's pathological status and medical interventions performed during the delirium evaluation.

In this study, we collected patient data based on the time point of delirium assessment. We aimed to validate the E-PRE-DELIRIC model and 18 prediction models for ICU patients. To this end, we used variable importance measures (VIMs) and Shapley additive explanations (SHAP) to explain the prediction models; therefore, the prediction models not only predict delirium but also provide a reasonable explanation for the prediction, which can greatly enhance users' trust in the models.

Methods

Study design

Multicenter cohort study.

Data source

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III (version 1.4) database, MIMIC-IV database and eICU Collaborative Research Database are maintained by the Laboratory for Computational Physiology at Massachusetts Institute of Technology [18-20]. The databases are accessible to researchers who have passed training courses on protecting human subjects.

Study population

The study population included all patients who underwent delirium assessment after admission to the ICU. For patients with multiple positive values, we included only the first episode. If all of the assessment results were negative, we also selected only the first episode. The relevant data for each patient were collected only once.

Dataset

We based the selection of predictor variables on a priori hypotheses guided by the literature and clinical knowledge and on ease of availability in the electronic health record [9, 10, 21, 22]. The study extracted 85 relevant variables, including baseline patient information, vital signs, laboratory test results, clinical diagnosis, and medical treatment, as candidate predictor variables. Considering the problem of predictors and the timing of the outcome measurement, we used the last time the patient was assessed for delirium

to calculate the value of vital signs, laboratory test results and clinical diagnostic candidate predictors. Medical treatment was provided within 24 hours before the delirium assessment, as described in supplemental file 1. Supplemental file 2 provides the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes used to define predictor variables.

Data analysis

Descriptive data are presented as the medians (25th to 75th percentiles) for continuous variables and as frequencies (%) for categorical variables. Categorical variables were compared between groups using the chi-square test. Unpaired t-tests or Kruskal-Wallis tests were used for continuous variables.

Missing data

The amount of missing data is detailed in supplemental file 3. All missing data were assumed to be missing at random and were imputed using MissForest [23, 24]. MissForest predicts missing values through a series of nonparametric random forest tree ensembles. Supplemental file 4 shows the frequency of missing data elements and the distribution of each parameter before and after imputation.

Feature selection

Statistically significant features were selected for further correlation analysis. For redundant features with strong correlations, factor analysis was used to confirm the collinearity of the variables, classify the collinearity as a latent factor, calculate the eigenvalues, visualize the gravel map of the eigenvalues, and select the feature root. Values >1 and the first few principal components for which the slope was decreased were used as principal component factors to eliminate redundant features and to identify more efficient, concise, and precise feature combinations, thereby improving the generalization and practical capabilities of the model. Factor analysis requires the data to be suitable according to the Bartlett sphere test and the Kaiser-Meyer-Olkin test.

ML models

We built different prediction models, including all-factor (AF) models, which included all potential variables; characteristic variable (CV) models, which included principal component factors; and rapid predictive (RP) models without laboratory test results.

The predictive algorithms were run using the following models: Logistic regression (LR), naive Bayesian (NB), K-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), and eXtreme Gradient Boosting (XGBoost). Categorical variables were converted to one-hot encoding, and data were centered to zero and scaled before the models were trained. The model built by the algorithm uses constant parameter optimization and model evaluation to compare the fitting effects of each model and to select the best model as the risk prediction model.

Validation

Model performance and evaluation

We used a tenfold cross-validation method on the training sets from each center. Each training set was divided randomly into ten parts, and the model was run ten times. In every such run, the model uses nine different parts of the training set for modeling and the tenth part for validation. Validation data were used to evaluate the prediction performance of the model and calculate the AUC, Youden index, relative risk (RR), positive predictive value (PPV), negative predictive value (NPV), accuracy, F1 score, positive likelihood ratio (PLR), and negative likelihood ratio (NLR) of the model for the test set.

A good model explanation must be presented for the black box model. VIMs based on the random forest technique were measured by the randomForest package [25]. The marginal contribution of a feature based on the SHAP algorithm was calculated by the SHAPforxgboost package, which explains the influence of characteristic factors included in the CV model for model prediction and distinguishes the attributes of the factors (risk factors and protective factors) [26].

The modeling and statistical analyses were performed using R version 3.62.

Results

In addition to the 46,428 ICU patients and 61,051 ICU admissions in the MIMIC-III database, 50,048 ICU patients and 69,619 ICU admissions in the MIMIC-IV database and 177,863 ICU patients and 626,858 ICU admissions in the eICU Collaborative Research Database were available. A total of 621,189 sequential delirium assessment records were available. The final 78,365 patients were included in this study. We identified 22,159 (28.28%) patients with positive records. The demographic characteristics and pertinent outcomes of the cohort are described in supplemental file 4. Briefly, the group of patients with delirium had statistically significant differences from the group without delirium in terms of age, ethnicity, admission type, ICU type, vital signs, laboratory tests, hospital characteristics, and time to delirium assessment.

No candidate variables were missing in approximately 5% of patients. The proportion of missing values in the overall data was approximately 20.0%. We eliminated variables for which less than 40% of the real data were available, including bands (9.1%), bicarbonate (26.0%), base deficit (23.7%), FiO₂ (19.1%), and sedation score (15.0%). The amounts of missing data are detailed in supplemental file 3. The proportion of missing values for the remaining covariate data was approximately 20.0%.

Predictor selection

Using the abovementioned statistically significant factors for the correlation analysis, the correlation coefficient matrix heat map of the features shows that the top ten features that were negatively correlated with the outcomes were the verbal response score, pain score, albumin, eye opening score, motor response score, hemoglobin, hematocrit, Glasgow Coma Scale score, region, and platelets; the top ten characteristics that were positively correlated with outcomes were the APACHE-III score, mechanical

ventilation, length of ICU stay, benzodiazepines, opioid analgesics, respiratory failure, alpha2-adrenergic receptor agonist, loop diuretics, blood urea nitrogen, and teaching (supplemental file 5). In addition, strong correlations were found between many features. For example, the correlation coefficient between alanine transaminase (ALT) and glutamic oxaloacetic transaminase (AST) reached 0.85; therefore, it was necessary to reduce redundant features.

The Kaiser-Meyer-Olkin test gave a value of 0.7, and Bartlett's test of sphericity showed a significance level of $P < 0.001$, indicating that the factor analysis was effective. Factor analysis and visualization of the characteristic root gravel map and load matrix revealed that the seventeen principal components and eight factors were the most predictive (supplemental files 6-7); for example, the correlation between hemoglobin and the second main factor reached 0.84. Considering the accuracy and practicability of using the CV models, clinical experience and actual comparisons were combined to select seventeen features representing the eight principal component factors, namely, age, APACHE-III score, Diastolic blood pressure (DBP), hemoglobin, urine volume, AST, BUN, verbal response score, hypertension, diabetes, mechanical ventilation, opioid analgesics use, alpha2-adrenergic receptor agonists use, adrenergic agonists use, number of beds, and length of ICU stay.

In addition, we selected risk factors that can be used to build RP models that do not require laboratory test results. These risk factors included age, admission type, DBP, urine volume, verbal response score, mechanical ventilation, opioid analgesics use, alpha 2-adrenergic receptor agonists use, adrenergic agonists use, number of beds, and length of ICU stay. We believe that these variables provide a good representation of delirium risk factors. In addition, these variables are easy to obtain in clinical settings. They also reduce the incidence of missing values.

Comparison of models

The output of the E-PRE-DELIRIC model and the 18 prediction models is presented in Table 1. The E-PRE-DELIRIC model had an AUC of 0.77, a Youden index of 0.38, an RR of 6.42, a PPV of 0.85, an NPV of 0.54, an accuracy of 0.76, an F1 score of 0.56, a PLR of 1.82, and an NLR of 0.28.

The AUC was higher for the all features set conditions of the AF models (range of AUC, 0.77–0.93 across algorithms) than for the selected features set conditions of the CV models (AUC range, 0.77–0.88) and the fast features set conditions of the RP models (AUC range, 0.75–0.87).

The RF AF model had the highest AUC (0.93), Youden Index (70) and RR (32) and was considered a more comprehensive evaluation approach. The XGBoost AF model had a slightly lower AUC of 0.92, Youden index of 66, and RR of 31. Similarly, the best-performing CV models were the RF model and the XGBoost model, which had AUCs of 0.88 and 0.86, respectively. Among the RP models, the KNN model had the highest AUC (0.87) and RR (14.52); the NB model had the lowest AUC (0.75) and RR (5.82).

Interpretation and evaluation of the machine learning model

Random forest models can provide measures of the importance of variables, thus providing some insight into the factors with the greatest influence on the predictions. The ten most highly ranked variables in this model were length of ICU stay, verbal response score, APACHE-III score, urine volume, hemoglobin, alanine transaminase, blood urea nitrogen, diastolic blood pressure, age, and mechanical ventilation (Figure 1A). Based on the SHAP algorithm, the characteristics of length of ICU stay, APACHE-III score, alanine transaminase, hypertension, and blood urea nitrogen correlated positively with the outcomes and were the top five risk factors; additionally, verbal response score, urine volume, hemoglobin, diastolic blood pressure, and alpha2 adrenergic receptor agonists use correlated negatively with the outcomes and were the top five protective factors (Figure 1B).

Discussion

Prediction models have the potential to provide improved diagnosis, risk stratification, and treatment of delirium in the ICU. The results of the present study show that our models are more reliable and more accurate for predicting delirium for critically ill patients in the ICU than the conventional E-PRE-DELIRIC model. As expected, the slight decrease in the AUC of the CV and RP models based on selected features demonstrated that other variables could be excluded without a marked negative effect on model performance. In particular, correlation analysis and characteristic analysis showed that a longer ICU stay, lower verbal response score and higher APACHE-III score had strong correlations with delirium. Physicians should be alert to possible delirium when encountering such patients.

In our study, the predictive model developed with combined RF had marginally better discrimination than the AF models and CV models. In addition to identifying risk and protective factors, the results suggested that the RF algorithm with 80 predictors achieved the best prediction effect, with an AUC of 0.93. In the RP models with 13 predictors, NB analysis showed the lowest AUC of 0.75. In contrast to our study, Racine et al developed ML prediction models for postoperative delirium in older adults undergoing major elective noncardiac surgery [27]. The highest AUC (0.71) was observed for the neural network algorithm when they used the 71 predictors in the full feature set. Finally, Coombes et al used 31 clinical actions based on a logistic regression model to predict delirium and found an AUC of 83% [12]. With our CV and RP models, we were able to predict delirium with good accuracy based on variables that are readily available in the clinical setting, thus offering improved convenience and efficiency.

This study explored the model-agnostic interpretation technique for describing potential factors that contribute to delirium in the ICU. Leveraging VIMs and SHAP analysis, our study generated a visualization format for interpreting patient-associated risks based on the clinical variables. The top three risk factors were length of ICU stay, verbal response score, and APACHE-III score in both the RF model and XGBoost model. Longer ICU stay, lower verbal response score and higher APACHE-III score had strong correlations with delirium. Furthermore, by highlighting significant clinical variables that contribute to risk prediction, such visualization can assist practitioners with the preemptive and early identification of key factors, including modifiable ones (e.g., longer ICU stay, lower urine volume, lower hemoglobin, lower diastolic blood pressure), that contribute to patients' risk of developing delirium. Physicians can use such insights

to quickly identify factors that may contribute to delirium risk and select evidence-based treatment protocols to mitigate such risks.

Our study has four strengths. First, the time point of data inclusion was not fixed in our study. In previous studies of predictive models, data collection time points for different patients were often fixed, for example, taking place 24 hours after admission and during the postoperative period [16, 28]. However, we collected data dynamically according to the time point of the delirium assessment. Our data could better reflect disease status and treatment scenarios. Second, our study built prediction models based on three variable sets: All potential variables, principal component factors, and rapid predictive variables. Our prediction models can be used to effectively predict delirium in ICU patients, as evidenced by the AUCs of 0.93 for the AF RF model, 0.88 for the CV RF model, and 0.87 for the KNN RP model. We were able to accurately predict delirium based on variables that are available in different clinical settings. Third, uniform data collection, a large sample size and a population-based design that covered all known delirium events minimized potential sources of bias. These features contribute to the representativeness of the predictive models developed in the present study.

Our study had three limitations. First, due to the number of cases in our study, we could not perform further subdivision according to some variables, such as the route of administration, drug dosage, and severity of disease. Additionally, some drugs, such as clonidine, fluoxetine, and quetiapine, are negatively related to the occurrence of delirium [29], and there are no records of the use of these medications in the MIMIC databases and eICU databases. Future research may aim to analyze a greater variety of variables. Second, to facilitate the practical application of the model, the variables in our CV model were representative. We used the VIMs algorithm and the SHAO algorithm to show the risk factors that explain the CV model itself. Since the CV model does not include all risk factors for delirium, the variables in this model cannot represent or explain all risk factors for delirium. Therefore, the conclusions of the VIM algorithm and the SHAP algorithm cannot fully explain all the risk factors for delirium. Third, our study was limited to patients in the United States who underwent a delirium assessment, most of whom were aged 45 years and older. The inclusion of a large, racially and geographically diverse patient population may enhance generalizability, although future studies are needed to examine the external validity of these models for predicting delirium in patients from different countries before their widespread deployment.

Conclusions

With a small number of predictive variables, ML algorithms can be established to predict the occurrence of delirium in ICU patients. Physicians can use such insights to quickly identify factors that may contribute to delirium risk and select evidence-based treatment protocols to mitigate such risks.

Declarations

Ethics approval and consent to participate

Consent was obtained for the original data collection and the institutional review boards of the Massachusetts Institute of Technology (Cambridge, MA, USA) approved the establishment of the database. Therefore, the ethical approval statement and informed consent were waived for this manuscript.

Consent for publication

Not applicable.

Availability of data and materials

Analyses of the data in this study are still ongoing. We shall make fully anonymized data available on the website <https://figshare.com/s/81bc9cb206af0095d5b7> in an estimated five years from the publication of this manuscript.

Declaration of interest

The authors declare that they have no competing interests.

Author' contributions

According to the guidelines of the International Committee of Medical Journal Editors (ICMJE), all authors contributed to the four criteria. AMH, HPL, and ZJZ conceived and designed the study. AMH and ZL acquired the data. AMH and XXZ analyzed and interpreted the data. AMH and HPL drafted the manuscript. AMH and ZJZ critically revised the manuscript for valuable intellectual content. AMH, ZL, and XXZ performed statistical analysis. All authors read and approved the final manuscript.

Funding

The current study is supported by a grant from Guangdong Medical Science and Technology Research Fund Project (A2021058).

Acknowledgments

The authors thank the patients for their participation.

References

1. Association ED, Society AD. The DSM-5 criteria, level of arousal and delirium diagnosis: inclusiveness is safer. *BMC Med.* 2014;12:141.
2. Salluh JI, Wang H, Schneider EB, Nagaraja N, Yenokyan G, Damluji A, Serafim RB, Stevens RD. Outcome of delirium in critically ill patients: systematic review and meta-analysis. *Bmj.* 2015;350:h2538.

3. Milbrandt EB, Deppen S, Harrison PL, Shintani AK, Speroff T, Stiles RA, Truman B, Bernard GR, Dittus RS, Ely EW. Costs associated with delirium in mechanically ventilated patients. *Crit Care Med*. 2004;32(4):955–62.
4. Shi Z, Mei X, Li C, Chen Y, Zheng H, Wu Y, Zheng H, Liu L, Marcantonio ER, Xie Z, et al. Postoperative Delirium Is Associated with Long-term Decline in Activities of Daily Living. *Anesthesiology*. 2019;131(3):492–500.
5. Oh ES, Fong TG, Hsieh TT, Inouye SK. Delirium in Older Persons: Advances in Diagnosis and Treatment. *Jama*. 2017;318(12):1161–74.
6. Delaney A, Hammond N, Litton E. Preventing Delirium in the Intensive Care Unit. *Jama*. 2018;319(7):659–60.
7. Herling SF, Greve IE, Vasilevskis EE, Egerod I, Bekker Mortensen C, Møller AM, Svenningsen H, Thomsen T. Interventions for preventing intensive care unit delirium in adults. *Cochrane Database Syst Rev*. 2018;11(11):Cd009783.
8. Wassenaar A, Schoonhoven L, Devlin JW, van Haren FMP, Slooter AJC, Jorens PG, van der Jagt M, Simons KS, Egerod I, Burry LD, et al. Delirium prediction in the intensive care unit: comparison of two delirium prediction models. *Crit Care (London England)*. 2018;22(1):114.
9. van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, van der Hoeven JG, Donders R, van Achterberg T, Schoonhoven L. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. *Bmj*. 2012;344:e420.
10. Wassenaar A, van den Boogaard M, van Achterberg T, Slooter AJ, Kuiper MA, Hoogendoorn ME, Simons KS, Maseda E, Pinto N, Jones C, et al. Multinational development and validation of an early prediction model for delirium in ICU patients. *Intensive care medicine*. 2015;41(6):1048–56.
11. Oh J, Cho D, Park J, Na SH, Kim J, Heo J, Shin CS, Kim JJ, Park JY, Lee B. Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning. *Physiol Meas*. 2018;39(3):035004.
12. Coombes CE, Coombes KR, Fareed N. A novel model to label delirium in an intensive care unit from clinician actions. *BMC Med Inform Decis Mak*. 2021;21(1):97.
13. Neto AS, Nassar AP Jr, Cardoso SO, Manetta JA, Pereira VG, Espósito DC, Damasceno MC, Slooter AJ. Delirium screening in critically ill patients: a systematic review and meta-analysis. *Crit Care Med*. 2012;40(6):1946–51.
14. van Eijk MM, van den Boogaard M, van Marum RJ, Benner P, Eikelenboom P, Honing ML, van der Hoven B, Horn J, Izaks GJ, Kalf A, et al. Routine use of the confusion assessment method for the intensive care unit: a multicenter study. *Am J Respir Crit Care Med*. 2011;184(3):340–4.
15. Zaal IJ, Devlin JW, Peelen LM, Slooter AJ. A systematic review of risk factors for delirium in the ICU. *Crit Care Med*. 2015;43(1):40–7.
16. Wang Y, Lei L, Ji M, Tong J, Zhou CM, Yang JJ. Predicting postoperative delirium after microvascular decompression surgery with machine learning. *J Clin Anesth*. 2020;66:109896.

17. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of Incident Delirium Using a Random Forest classifier. *J Med Syst*. 2018;42(12):261.
18. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
19. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. 2018;5:180178.
20. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R: **MIMIC-IV (version 0.4)**. *PhysioNet* 2020.
21. Halladay CW, Sillner AY, Rudolph JL. Performance of Electronic Prediction Rules for Prevalent Delirium at Hospital Admission. *JAMA Netw Open*. 2018;1(4):e181405.
22. Burry L, Hutton B, Williamson DR, Mehta S, Adhikari NK, Cheng W, Ely EW, Egerod I, Fergusson DA, Rose L. Pharmacological interventions for the treatment of delirium in critically ill adults. *Cochrane Database Syst Rev*. 2019;9(9):Cd011749.
23. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–8.
24. Kanwal F, Taylor TJ, Kramer JR, Cao Y, Smith D, Gifford AL, El-Serag HB, Naik AD, Asch SM. Development, Validation, and Evaluation of a Simple Machine Learning Model to Predict Cirrhosis Mortality. *JAMA Netw Open*. 2020;3(11):e2023780.
25. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.
26. Yang L, Allan J. **SHAPforxgboost: SHAP Plots for 'XGBoost'**. *R package version 010* 2020, **3**.
27. Racine AM, Tommet D, D'Aquila ML, Fong TG, Gou Y, Tabloski PA, Metzger ED, Hsieh TT, Schmitt EM, Vasunilashorn SM, et al. Machine Learning to Develop and Internally Validate a Predictive Model for Post-operative Delirium in a Prospective, Observational Clinical Cohort Study of Older Surgical Patients. *J Gen Intern Med*. 2021;36(2):265–73.
28. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and Validation of an Electronic Health Record-Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. *JAMA Netw Open*. 2018;1(4):e181018.
29. Guan X, Zhang B, Fu M, Li M, Yuan X, Zhu Y, Peng J, Guo H, Lu Y. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. *Ann Med*. 2021;53(1):257–66.

Table

Table 1. Model performance in the internal ten-fold cross-validation validation sets.

Model	AUC	Youden index (%)	RR	PPV (%)	NPV (%)	Accuracy (%)	F1 score	PLR	NLR
E-PRE-DELIRIC model	0.77 (0.77-0.78)	38 (35-40)	6.42 (6.25-6.52)	85 (83-88)	54 (47-57)	76 (76-76)	0.56 (0.53-0.57)	1.82 (1.65-1.93)	0.28 (0.26-0.3)
AF models									
LR	0.84 (0.83-0.84)	57 (56-57)	14.1 (13.38-14.57)	84 (84 -84)	73 (72-73)	82 (81-82)	0.63 (0.62-0.64)	3.09 (3.02-3.16)	0.22 (0.22-0.23)
NB	0.77 (0.77-0.78)	39 (37-40)	6.3 (6.06-6.75)	84 (83-84)	55 (53-57)	76 (75-76)	0.56 (0.55-0.57)	1.86 (1.79-1.91)	0.29 (0.28-0.3)
KNN	0.89 (0.88-0.89)	58 (57-60)	30.12 (29.08-30.99)	93 (92-94)	65 (64-66)	86 (85-86)	0.72 (0.71-0.73)	2.67 (2.61-2.74)	0.10 (0.09-0.11)
SVM	0.84 (0.84-0.85)	54 (53-55)	12.73 (12.2-13.25)	85 (85-86)	68 (67-70)	81 (80-81)	0.67 (0.66-0.67)	2.71 (2.61-2.81)	0.21 (0.21-0.21)
RF	0.93 (0.93-0.93)	70 (69-70)	32.53 (30.69-33.06)	86 (85-86)	84 (83-85)	85 (85-86)	0.76 (0.76-0.77)	5.47 (5.16-5.58)	0.17 (0.16-0.17)
XGBoost	0.92 (0.92-0.93)	66 (65-67)	30.62 (26.31-32.33)	91 (90-91)	75 (75-76)	86 (86-87)	0.76 (0.74-0.76)	3.71 (3.57-3.81)	0.12 (0.11-0.14)
CV models									
LR	0.81 (0.81-0.81)	52 (52-52)	10.49 (10.26-10.84)	82 (82-82)	70 (70-70)	79 (79-80)	0.57 (0.57-0.58)	2.72 (2.69-2.75)	0.26 (0.25-0.26)
NB	0.77 (0.77-0.78)	36 (35-37)	6.72 (6.58-7.06)	88 (88-88)	48 (47-49)	77 (76-77)	0.54 (0.53-0.55)	1.69 (1.67-1.73)	0.25 (0.24-0.26)
KNN	0.80 (0.80-0.80)	41 (40-41)	10.32 (9.80-11.15)	92 (91-92)	47 (46-48)	80 (79-80)	0.56 (0.55-0.57)	1.75 (1.71-1.79)	0.16 (0.15-0.16)
SVM	0.82 (0.82-0.83)	45 (45-46)	12.56 (12.04-12.98)	91 (91-92)	55 (54-56)	81 (81-82)	0.61 (0.61-0.62)	1.98 (1.75-2.26)	0.17 (0.16-0.17)
RF	0.88 (0.88-0.88)	59 (58-59)	17.34 (16.68-17.88)	88 (87-88)	71 (70-72)	83 (82-83)	0.7 (0.7-0.71)	3.06 (2.94-3.17)	0.18 (0.17-0.18)
XGBoost	0.86 (0.86-0.87)	52 (51-53)	14.21 (13.35-14.97)	89 (88-90)	63 (61-64)	82 (81-82)	0.66 (0.65-0.67)	2.39 (2.33-2.49)	0.17 (0.15-0.18)

RP models									
LR	0.77 (0.77- 0.77)	47 (46- 49)	8.38 (7.93- 9.12)	80 (80- 81)	67 (66- 69)	78 (78- 78)	0.53 (0.51- 0.53)	2.44 (2.37- 2.58)	0.29 (0.29- 0.3)
NB	0.75 (0.74- 0.75)	33 (32- 33)	5.82 (5.48- 5.97)	88 (87- 88)	45 (45- 45)	76 (75- 76)	0.51 (0.50- 0.51)	1.59 (1.57- 1.6)	0.27 (0.27- 0.29)
KNN	0.87 (0.86- 0.87)	53 (52- 54)	14.52 (13.98- 15.74)	89 (88- 90)	64 (63- 64)	82 (81- 83)	0.65 (0.64- 0.67)	2.41 (2.36- 2.49)	0.16 (0.15- 0.16)
SVM	0.76 (0.76- 0.77)	45 (45- 46)	9.57 (9.13- 10.08)	88 (88- 89)	57 (56- 57)	79 (78- 80)	0.61 (0.61- 0.2)	2.18 (2.05- 2.35)	0.22 (0.22- 0.23)
RF	0.86 (0.86- 0.86)	54 (53- 55)	14.37 (13.82- 14.75)	88 (87- 89)	66 (65- 67)	82 (81- 82)	0.67 (0.67- 0.68)	2.58 (2.53- 2.67)	0.18 (0.17- 0.19)
XGBoost	0.83 (0.83- 0.84)	46 (45- 47)	10.93 (10.55- 11.44)	89 (88- 90)	57 (55- 59)	80 (80- 81)	0.62 (0.61- 0.62)	2.06 (2.01- 2.13)	0.19 (0.18- 0.2)

Abbreviation: LR, Logistic Regression; NB, Naive Bayesian; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; RF, Random Forest; XGBoost, eXtreme Gradient Boosting; AUC, Area Under the Receiver Operating Characteristic; RR, Relative Risk; PPV, Positive Predictive Value; NPV, Negative Predictive Value; PLR, Positive Likelihood Ratio; NLR, Negative Likelihood Ratio.

*Data shown as mean (95% confidence interval).

Figures

Figure 1

Visualization of characteristic variables in the prediction models. A, Relative importance measures of the variables associated with delirium in the random forest model; B, mean SHAP in the XGBoost model.

Supplementary Files

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

- [Supplementfile.docx](#)