An Empirical Assessment of Machine Learning Models for Delirium Prediction Using the eICU-CRD v2.0 Database

Márcio Wellington de Souza[1,2], **Diego Minatel**[3]

[1]Department of Psychiatry, Hospital Israelita Albert Einstein,
São Paulo 05652-900, Brazil

[2]Department of Psychiatry , Hospital Beneficência Portuguesa,
São Paulo 01323-001, Brazil

[3]Institute of Mathematics and Computer Science, University of São Paulo,
São Carlos 13566-590, Brazil

marciowellsouza@gmail.com, dminatel@usp.br

Abstract. *Population aging intensifies the challenges faced by healthcare systems, particularly in managing conditions such as hospital-acquired delirium, a neuropsychiatric disorder characterized by acute onset and a fluctuating course, and associated with increased mortality and functional decline. Despite its clinical relevance, delirium frequently remains underdiagnosed during hospitalization. This paper investigates the use of machine learning to predict delirium early in intensive care units, using data from the eICU-CRD v2.0 database. A domain expert clinician selected the analyzed attributes, ensuring clinical coherence in the modeling process. We evaluated five classification algorithms and four data-balancing techniques using stratified cross-validation and metrics appropriate for imbalanced scenarios. Random Forest with Random Undersampling achieved the best overall performance. Our findings suggest that machine learning-based approaches can support earlier recognition of delirium and enhance clinical decision-making.*

1. Introduction

Population aging has become a structural reality of contemporary societies. It is no longer a distant projection but a condition that directly impacts healthcare systems and public policies. Estimates indicate that, by 2050, the global population aged 65 years or older will increase from 703 million in 2019 to 1.5 billion [United Nations 2019], creating unprecedented challenges for healthcare systems, which must manage chronic conditions and acute illnesses that predominantly affect older adults. One of the most critical conditions affecting this group is hospital-acquired delirium, a neuropsychiatric disorder characterized by acute onset and a fluctuating course, with profound consequences for patient morbidity, mortality, and functional independence [Inouye et al. 2014].

Although the medical literature extensively documents delirium, it remains underdiagnosed, particularly in its hypoactive forms, which clinicians often misinterpret as depression or normal aging-related changes [Ely et al. 2004]. Studies indicate that between 30% and 60% of delirium cases go unrecognized during hospitalization, resulting in significantly increased mortality, accelerated functional decline, a higher risk of institutionalization, and elevated healthcare costs [Witlox et al. 2010, Milbrandt et al. 2004].

This scenario underscores the urgent need for effective strategies to detect delirium early and manage it appropriately, particularly in light of population aging and the growing pressure on healthcare systems.

In recent years, Artificial Intelligence (AI) technologies, such as Machine Learning (ML) and Deep Learning, have demonstrated strong potential for addressing diagnostic and prognostic problems across various medical domains [Sarker 2021, Faceli et al. 2021]. In particular, classification algorithms, including neural networks and Random Forest models, have been successfully applied to identify patterns in clinical records, enabling disease prediction based on historical data. In this context, ML may represent an effective approach for diagnosing delirium in critically ill patients, provided that practitioners rigorously validate the models and incorporate explainability strategies to ensure that healthcare professionals can understand, trust, and safely apply their predictions in clinical practice.

Previous studies have explored delirium prediction primarily within specific patient subgroups, such as postoperative [Yang et al. 2024, Bishara et al. 2022, Davoudi et al. 2017] or COVID-19 [Viegas et al. 2025] populations, reporting promising results. However, predicting delirium for any patient admitted to the intensive care unit (ICU) using electronic health records (EHR) introduces additional challenges, including increased clinical heterogeneity and variability in risk factors. This paper addresses this broader scenario by conducting a comparative experimental study of machine learning techniques aimed at advancing the development of more generalizable and clinically applicable predictive models.

To this end, we used the eICU Collaborative Research Database (eICU-CRD v2.0), which, after preprocessing, presented a markedly imbalanced distribution, with only approximately 8% of cases corresponding to delirium. To mitigate this imbalance, we evaluated four data-balancing techniques—Random Undersampling, Random Oversampling, SMOTE, and ADASYN—alongside five classification algorithms. We assessed the resulting models using three evaluation metrics: accuracy, to provide an overall performance perspective; balanced accuracy, to account for class imbalance; and F1-score, to emphasize performance on the class of interest (*i.e.*, when delirium occurs). The findings indicate that the combination of Random Forest and Random Undersampling achieved the best overall performance under the adopted experimental protocol.

2. Related Work

Some previous studies have explored the use of machine learning techniques for delirium prediction. A relevant research direction has focused on predicting postoperative delirium, a specific form of delirium that occurs after surgical procedures. In [Bishara et al. 2022], the authors conducted a study to predict postoperative delirium by evaluating several classification algorithms, including Logistic Regression, Extreme Gradient Boosting, and Neural Networks. They used a dataset extracted from the Epic EHR system (Epic Systems Corporation, Verona, WI, USA), comprising data from 24,885 patients who underwent procedures requiring anesthesia, recovery in the post-anesthesia care unit, and at least one overnight hospital stay. The best-performing model achieved an AUC-ROC of up to 85%.

Along the same lines, [Yang et al. 2024] conducted a study on predicting post-

operative delirium after cardiac surgery. The dataset comprised 367 patients treated at the Affiliated Hospital of Southwest Medical University and the Affiliated Traditional Chinese Medicine Hospital of Southwest Medical University, of whom 28.6% developed delirium. The experimental protocol evaluated several machine learning algorithms, including Random Forest, Support Vector Machines, and k-Nearest Neighbors (k-NN). The best-performing model achieved an accuracy of up to 88%. In [Davoudi et al. 2017], the authors analyzed a preoperative EHR dataset comprising 51,457 adult patients, of whom only 3.12% developed postoperative delirium during hospitalization. The proposed models achieved an accuracy of up to 81%. Due to the pronounced class imbalance, the authors applied the SMOTE data-balancing technique in combination with Naïve Bayes, k-NN, and Random Forest. In the same study, the authors discussed that factors such as age, alcohol abuse, socioeconomic status, preexisting medical conditions, and the attending surgeon may influence the risk of delirium.

In contrast, [Zhang et al. 2023] investigated sepsis-associated delirium in ICU patients using the MIMIC-IV and eICU-CRD datasets. Sepsis, a severe infection-related organ dysfunction with high morbidity and mortality, represents a common critical condition in intensive care settings. The authors evaluated several machine learning algorithms, including Support Vector Machines, k-NN, Decision Trees, and Naïve Bayes. Their best-performing model achieved 69% accuracy and an AUC-ROC of 70%. [Viegas et al. 2025] applied a machine learning approach to predict delirium in patients with COVID-19. [Contreras et al. 2025] proposed a distinct methodological direction, introducing DeLLiriuM, an LLM-based framework for delirium prediction in ICU patients. Leveraging data from MIMIC-IV, eICU-CRD, and the University of Florida Integrated Data Repository, this method converts structured EHR data into textual representations to enable language model–driven prediction.

Our study differs from previous work primarily by focusing on delirium prediction in the general ICU population rather than on specific subgroups such as postoperative, COVID-19, or sepsis-associated patients. Predicting delirium at ICU admission for any critically ill patient introduces additional challenges, including greater clinical heterogeneity, a wider range of underlying conditions, and increased variability in risk factors. By addressing this broader population, our work seeks to advance the development of more generalizable and clinically applicable predictive models.

3. Methodology

As previously discussed, this paper applies classification algorithms to predict delirium based on patients' clinical and laboratory characteristics collected at ICU admission. In most hospitalizations, however, this condition does not manifest, leaving the dataset dominated by negative cases. Such imbalanced scenarios pose a challenge for classification algorithms, which often struggle to correctly identify the minority class (delirium in this case). For this reason, we employ data balancing techniques alongside the classifiers to develop more accurate predictive models for this task.

In this section, we describe the methodology adopted in this study. We present the experimental protocol for assessing predictive modeling strategies for delirium detection in critically ill patients (Section 3.1). We also detail the selected clinical database (Section 3.2), the evaluated data-balancing techniques and their configurations (Section 3.3),

the classification algorithms and their hyperparameter settings (Section 3.4), the evaluation metrics (Section 3.5), and the rationale for the experimental design.

3.1. Proposed Experimental Protocol

Figure 1 presents an overview of the experimental protocol adopted in this study. First, we selected eICU-CRD v2.0[1], publicly released by PhysioNet, to perform delirium prediction in critically ill patients. Next, we selected relevant features from multiple relational tables. We integrated them into a single analytical dataset suitable for machine learning and then preprocessed it. Section 3.2 details the dataset and the preprocessing procedures.

Subsequently, we applied stratified cross-validation, partitioning the data into five folds while preserving, in each split, the original proportion of patients who developed delirium during hospitalization and those who did not. We adopted this configuration to reduce computational cost while maintaining a reliable and robust evaluation of the predictive models.

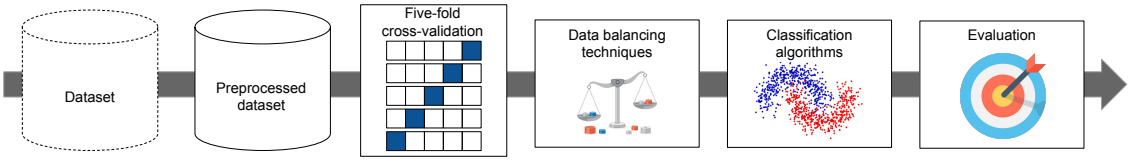

Figure 1. Experimental protocol for building classifiers aimed at predicting delirium in intensive care unit (ICU) patients.

In each cross-validation iteration, we applied eight data-balancing variants (described in Section 3.3) based on four distinct techniques. We also included a no-balancing setting, which we used as a baseline for comparison. For each problem representation generated by these techniques, we employed five classification algorithms, each evaluated under ten different configurations, as detailed in Section 3.4. After completing the five iterations, we computed the average metric scores for each classifier (Section 3.5) and report and discuss these results in Section 4. Overall, we evaluated the performance of 450 classifiers.

We developed the source code for this study in Python, using the `pandas` library for file reading, table concatenation, and result analysis; `imblearn` for data balancing techniques; and `scikit-learn` for data preprocessing, cross-validation, classification algorithms, and evaluation metrics. The source code is publicly available at: `https://github.com/diegominatel/delirium-prediction`.

3.2. Dataset

The eICU-CRD v2.0 is a database maintained by PhysioNet and developed by the MIT Lab for Computational Physiology in collaboration with Philips Healthcare. This repository aggregates anonymized clinical records from 200,859 patients admitted to 208 hospitals across the United States, collected between 2014 and 2015. The database includes patient information related to vital signs, laboratory tests, diagnoses, medications, and the APACHE[2] system, among other clinical variables [Pollard et al. 2018].

[1]Available at: https://physionet.org/content/eicu-crd/2.0/

[2]The Acute Physiology and Chronic Health Evaluation (APACHE) is a scoring system used to estimate the severity of illness in patients admitted to intensive care units.

Table 1 presents the attributes selected for delirium prediction, along with their descriptions and value types. A domain expert clinician, who is also the first author of this study, performed feature selection based on information available in the eICU-CRD v2.0 and the completeness of the corresponding columns. Accordingly, we extracted the attributes by combining the files `diagnosis.csv`, `apachePatientResult.csv`, and `lab.csv`, using the patient identifiers provided in `patient.csv`. For laboratory variables (Creatinine, Sodium, Albumin, Urea, and Lactate) obtained from `lab.csv`, we considered only the first recorded measurement. We defined the target variable (Delirium) based on the presence of the terms delirium, encephalopathy, or acute confusional state in the diagnosis records.

Table 1. Description of the features selected for delirium prediction.

Feature	Description	Type
Age	Patient age in years	Numerical
Sex	Patient sex	Categorical
Ethnicity	Self-reported ethnicity or ethnicity recorded in the medical record	Categorical
Unittype	Type of ICU in which the patient was admitted (*e.g.*, Med-Surg ICU, Neuro ICU, Cardiac ICU)	Categorical
Unitadmitsource	Patient admission source to the ICU (*e.g.*, emergency department, operating room, another hospital unit)	Categorical
Unitstaytype	Type of ICU stay, such as direct admission or internal transfer	Categorical
Admissionweight	Patient weight at the time of admission (kg)	Numerical
Acutephysiologyscore	Acute physiology score from the APACHE system. Sum of points based on abnormalities in vital signs and laboratory tests during the first hours after admission	Numerical
Apachescore	Overall APACHE score, which incorporates the Acutephysiologyscore, age, and the presence of chronic diseases	Numerical
Predictedicumortality	Estimated probability (between 0 and 1) of ICU mortality, based on APACHE parameters	Numerical
Predictediculos	Estimated length of ICU stay (in days), based on the initial clinical score	Numerical
Chloride	Blood chloride level (mEq/L)	Numerical
Creatinine	Blood creatinine level (mg/dL)	Numerical
Sodium	Blood sodium concentration (mEq/L)	Numerical
Albumin	Blood albumin level (g/dL)	Numerical
Urea	Blood urea level (mg/dL)	Numerical
Lactate	Blood lactate level (mmol/L)	Numerical
Hypertension	Indicates whether the patient has a diagnosis of hypertension	Categorical

After defining the features, we removed instances with missing values in any of these attributes. As a result, the final dataset contained 103,052 instances, of which 7.66% belong to the positive class (delirium diagnosis) and 92.34% to the negative class (absence

of a delirium diagnosis). Figure 2 illustrates the distribution of instances by age group and sex after this filtering step.

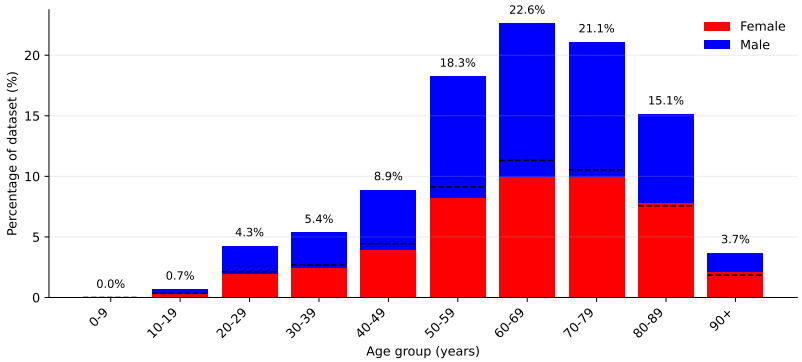

Figure 2. Distribution of instances by age group after preprocessing. For each bar, colors represent the proportion of male (blue) and female (red) instances. The dashed line indicates half of the total height of each bar.

Next, we performed data preprocessing, which included one-hot encoding of categorical attributes and standardization of numerical attributes. After this procedure, the total number of attributes increased from 18 to 48.

3.3. Data-balancing techniques

We selected Random Undersampling, Random Oversampling, SMOTE, and ADASYN as the data balancing techniques for this study. Table 2 summarizes the evaluated parameters and their values for each technique. We retained the default settings defined in the `imblearn` library for parameters not explicitly listed in the table.

Table 2. Configurations adopted for each data-balancing technique.

Technique	Parameter	Values
Random Undersampling	—	—
Random Oversampling	—	—
SMOTE	Number of nearest neighbors for synthetic sample generation	[5, 10, 15]
ADASYN	Number of nearest neighbors for synthetic sample generation	[5, 10, 15]

As mentioned in Section 3.1, we also included a no-balancing baseline for comparison, aiming to assess whether data balancing techniques provide performance gains for this task. Consequently, at the end of the balancing step, we generated nine distinct dataset representations: one without data balancing (the original imbalanced dataset), one using Random Undersampling, one using Random Oversampling, three SMOTE configurations, and three ADASYN settings.

3.4. Classification algorithms

We selected five classification algorithms for evaluation: Decision Tree (DT), k-NN, Logistic Regression (LR), Multilayer Perceptron (MLP), and Random Forest (RF). For each algorithm, we tested ten distinct configurations. Table 3 reports the hyperparameters used to generate these configurations, along with their corresponding value ranges. We retained the remaining hyperparameters at their default values as defined in the `scikit-learn` library.

Table 3. Configurations adopted for each classification algorithm. Numerical variations in the format $(i : f : p)$ in the *Value Range* column indicate that i and f correspond to the initial and final values, respectively, and that p denotes the increment step. The increment r indicates that the step advances to the next odd prime number.

Alg.	Hyperparameter	Value Range
DT	Minimum number of samples required to be at a leaf node	(1 : 19 : 2)
k-NN	Number of nearest neighbors	(1 : 29 : r)
LR	C	(0.80 : 1.25 : 0.05)
MLP	Number of neurons in the hidden layer (a single hidden layer is used)	(10 : 55 : 5)
RF	Number of estimators	(100 : 550 : 50)

3.5. Metrics

To evaluate the classifiers' predictive performance, we used three metrics: accuracy, balanced accuracy, and F1-score. Accuracy provides an overall assessment of model performance. However, because the dataset is imbalanced (as discussed in Section 3.2), this metric alone may be insufficient. For this reason, we included balanced accuracy, which computes the average recall across the two analyzed classes (delirium and non-delirium cases), thereby assigning equal importance to both classes regardless of their prevalence.

Furthermore, since the primary objective of this work is to correctly identify patients with delirium—that is, the positive class—we used the F1-score to emphasize performance on this class. The F1-score corresponds to the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives.

4. Results and Discussion

This section presents and discusses the results derived from the empirical protocol described in Section 3.

Table 4 presents the average predictive performance scores for each classification algorithm. For each algorithm, we computed the reported results as the average performance across 90 classifiers (10 hyperparameter settings $\times$ 9 data-balancing techniques). Therefore, the values shown in this table reflect the overall behavior of each algorithm across different hyperparameter configurations and data-balancing strategies.

Table 4. Average scores (in %) per classification algorithm. Values in parentheses indicate the standard deviation. Bold values denote the best result for each metric, while underlined values represent the second-best result.

Algorithm	Accuracy	Balanced Accuracy	F1-score
DT	<u>79.89% (6.62%)</u>	53.23 (1.22%)	13.68% (1.65%)
k-NN	69.14% (10.73%)	55.75% (2.45%)	15.23% (4.37%)
LR	65.48% (9.61%)	**59.81% (3.51%)**	**17.29% (6.10%)**
MLP	73.65% (7.85%)	<u>56.84% (2.88%)</u>	<u>16.43% (4.48%)</u>
RF	**89.32% (5.65%)**	51.89% (3.47%)	6.20% (5.91%)

As observed in Table 4, RF achieved the highest average accuracy, whereas LR obtained the best average performance in terms of balanced accuracy and F1-score. DT

achieved the second-best accuracy, and MLP recorded the second-highest averages for balanced accuracy and F1-score. A relevant aspect of the results presented in Table 4 is that the two algorithms with the highest accuracy (RF and DT) exhibited the poorest performance on the metrics more suitable for this problem, namely balanced accuracy and F1-score. This contrast highlights that, in imbalanced scenarios such as this one, accuracy can be misleading. In this context, a high accuracy value primarily reflects the model's ability to correctly classify the majority class, which does not align with the task's central objective—accurately identifying delirium cases.

We now focus the analysis on the average results obtained for each data-balancing technique. Table 5 presents these results, computed as the average performance across 50 classifiers (5 classification algorithms $\times$ 10 hyperparameter configurations) for Random Undersampling, Random Oversampling, and the no-balancing setting. For SMOTE and ADASYN, which each include three variations, the reported results correspond to the average performance across 150 classifiers (5 classification algorithms $\times$ 10 hyperparameter configurations $\times$ 3 variations). The symbol "—" denotes the absence of a balancing technique, meaning that the classifiers were trained using the original imbalanced dataset.

Table 5. Average scores (in %) per data-balancing technique. Values in parentheses indicate the standard deviation. Bold values denote the best result for each metric, while underlined values represent the second-best result.

Technique	Accuracy	Balanced Accuracy	F1-score
—	**90.95% (2.01%)**	50.68% (0.84%)	3.86% (4.36%)
Random Undersampling	64.70% (4.95%)	**59.74% (2.25%)**	**19.11% (1.80%)**
Random Oversampling	<u>75.60% (11.39%)</u>	<u>55.76% (4.33%)</u>	13.62% (7.19%)
SMOTE	75.08% (10.78%)	55.66% (3.58%)	<u>14.71% (4.86%)</u>
ADASYN	74.32% (11.44%)	55.45% (3.55%)	14.40% (4.90%)

The no-balancing setting resulted in a higher average accuracy than the other techniques, with a difference of at least 15%. However, it yielded the worst performance in terms of balanced accuracy and, in particular, F1-score, with a decrease of approximately 10% compared to the lowest average obtained among the balancing techniques (Random Oversampling) for this metric. These results demonstrate that, for this problem, data balancing is essential, particularly to improve predictive performance for the positive class.

Techniques that generate synthetic samples, such as SMOTE and ADASYN, achieved intermediate performance in terms of balanced accuracy and F1-score. Random Undersampling delivered the best results across these metrics, outperforming the second-best approach by at least 4 percentage points, making it the most effective technique for this scenario. This finding highlights an important insight: increasing the amount of training data does not necessarily lead to better performance. In this case, reducing the size of the training set produced classifiers that were more accurate across both classes.

Table 6 presents the best result achieved by each classification algorithm, along with the corresponding data-balancing technique that produced this performance. We selected these configurations based on the highest F1-score obtained for each algorithm. Notably, selecting the best models based on balanced accuracy would have yielded the same classifiers, reinforcing the consistency of the model ranking across both metrics.

Table 6. Best-performing classifier for each classification algorithm and its corresponding data-balancing technique. Model selection was performed based on the highest F1-score. Bold values denote the best result for each metric, while underlined values represent the second-best result.

Algorithm	Technique	Accuracy	Balanced Accuracy	F1-score
DT	Undersampling	62.15%	56.93%	17.06%
k-NN	Undersampling	61.66%	60.94%	19.38%
LR	Undersamplng	<u>63.87%</u>	<u>61.62%</u>	<u>20.00%</u>
MLP	Oversampling	61.66%	60.94%	19.38%
RF	Undersampling	**73.55%**	**61.82%**	**21.77%**

The combination of Random Undersampling and the Random Forest algorithm, with `n_estimators` set to 550, produced the best classifier for predicting delirium presence/absence, achieving the highest scores across all three evaluated metrics. The second-best performance was obtained with the combination of Random Undersampling and Logistic Regression, although its accuracy was approximately 10 percentage points lower than that of the best configuration. Finally, it is worth noting that MLP was the only algorithm whose best performance did not occur with Random Undersampling; instead, it achieved its highest performance when combined with Random Oversampling.

5. Concluding Remarks

This paper presents a comparative analysis of classification algorithms combined with data-balancing techniques for binary classification of delirium presence/absence in ICU patients, using the eICU-CRD v2.0 database. A domain expert clinician carefully selected the analyzed attributes, ensuring their clinical relevance and strengthening the reliability of the experimental design. The experimental results showed that the Random Forest algorithm with 550 estimators and Random Undersampling achieved the best overall performance in predicting both classes. An additional key finding is that, across classification algorithms, Random Undersampling consistently improved balanced accuracy and F1-score, establishing it as the most effective balancing technique for this task.

As future work, we plan to expand the experimental protocol by including more hyperparameter configurations and additional classification algorithms, particularly those based on deep learning. We also intend to investigate alternative data-balancing techniques, as well as data augmentation approaches, with an emphasis on adversarial learning methods that have demonstrated strong performance in similar tasks. Furthermore, we aim to validate the findings of this study using other, more comprehensive and diverse datasets, including unstructured data such as textual diagnostic descriptions, by applying Natural Language Processing (NLP) techniques.

Another important direction is to assess the explainability of the generated models using methods such as SHAP and LIME. Adopting these strategies may increase healthcare professionals' trust and facilitate the clinical implementation of the proposed model, ultimately contributing to reducing delirium underdiagnosis in Brazilian hospitals.

References

Bishara, A., Chiu, C., Whitlock, E. L., Douglas, V. C., Lee, S., Butte, A. J., Leung, J. M., and Donovan, A. L. (2022). Postoperative delirium prediction using machine learning

models and preoperative electronic health record data. *BMC anesthesiology*, 22(1):8.

Contreras, M., Kapoor, S., Zhang, J., Davidson, A., Ren, Y., Guan, Z., Ozrazgat-Baslanti, T., Sena, J., Nerella, S., Bihorac, A., et al. (2025). A large language model for delirium prediction in the intensive care unit using structured electronic health records. *Scientific Reports*, 15(1):38890.

Davoudi, A., Ozrazgat-Baslanti, T., Ebadi, A., Bursian, A. C., Bihorac, A., and Rashidi, P. (2017). Delirium prediction using machine learning models on predictive electronic health records data. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 568–573. IEEE.

Ely, E. W., Shintani, A., Truman, B., Speroff, T., Gordon, S. M., Harrell Jr, F. E., Inouye, S. K., Bernard, G. R., and Dittus, R. S. (2004). Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. *Jama*, 291(14):1753–1762.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, Rio de Janeiro, 2nd edition.

Inouye, S. K., Westendorp, R. G., and Saczynski, J. S. (2014). Delirium in elderly people. *The lancet*, 383(9920):911–922.

Milbrandt, E. B., Deppen, S., Harrison, P. L., Shintani, A. K., Speroff, T., Stiles, R. A., Truman, B., Bernard, G. R., Dittus, R. S., and Ely, E. W. (2004). Costs associated with delirium in mechanically ventilated patients. *Critical care medicine*, 32(4):955–962.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

United Nations (2019). World population ageing 2019: Highlights.

Viegas, A., Von Rekowski, C. P., Araújo, R., Viana-Baptista, M., Macedo, M. P., and Bento, L. (2025). Predicting icu delirium in critically ill covid-19 patients using demographic, clinical, and laboratory admission data: A machine learning approach. *Life*, 15(7):1045.

Witlox, J., Eurelings, L. S., de Jonghe, J. F., Kalisvaart, K. J., Eikelenboom, P., and Van Gool, W. A. (2010). Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *Jama*, 304(4):443–451.

Yang, T., Yang, H., Liu, Y., Liu, X., Ding, Y.-J., Li, R., Mao, A.-Q., Huang, Y., Li, X.-L., Zhang, Y., et al. (2024). Postoperative delirium prediction after cardiac surgery using machine learning models. *Computers in Biology and Medicine*, 169:107818.

Zhang, Y., Hu, J., Hua, T., Zhang, J., Zhang, Z., and Yang, M. (2023). Development of a machine learning-based prediction model for sepsis-associated delirium in the intensive care unit. *Scientific reports*, 13(1):12697.