Data-driven Machine Learning Models for Risk Stratification and Prediction of Emergence Delirium in Pediatric Patients Underwent Tonsillectomy/Adenotonsillectomy

Ann. Ital. Chir., 2024 95, 5: 944–955
https://doi.org/10.62713/aic.3485

Alessandro Simonini[1], Jeevitha Murugan[2], Alessandro Vittori[3], Roberta Pallotto[1], Elena Giovanna Bignami[4], Maria Grazia Calevo[5], Ornella Piazza[6], Marco Cascella[6]

[1]Department of Pediatric Anaesthesia and Intensive Care, S.C. SOD Anestesia e Rianimazione Pediatrica, Ospedale G. Salesi, 60123 Ancona, Italy
[2]BTech - Artificial Intelligence and Data Science, St Joseph's College of Engineering, 600119 Chennai, India
[3]Department of Anesthesia and Critical Care, ARCO Roma Ospedale Pediatrico Bambino Gesù IRCCS, 00165 Rome, Italy
[4]Anesthesiology, Critical Care and Pain Medicine Division, Department of Medicine and Surgery, University of Parma, 43126 Parma, Italy
[5]Epidemiology and Biostatistic Unit, Scientific Directorate, IRCCS Istituto Giannini Gaslini, 16147 Genoa, Italy
[6]Anesthesia and Pain Medicine, Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, 84081 Baronissi, Italy

AIM: In the pediatric surgical population, Emergence Delirium (ED) poses a significant challenge. This study aims to develop and validate machine learning (ML) models to identify key features associated with ED and predict its occurrence in children undergoing tonsillectomy or adenotonsillectomy.

METHODS: The analysis involved data cleaning, exploratory data analysis (EDA), supervised predictive modeling, and unsupervised learning on a medical dataset (n = 423). After preliminary data cleaning, EDA encompassed plotting histograms, boxplots, pairplots, and correlation heatmaps to understand variable distributions and relationships. Four predictive models were trained including logistic regression (LR), random forest (RF), Support Vector Machine (SVM), and Gradient Boosting (XGBoost). The models were evaluated and compared using Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), precision, recall, and feature importance. The RF model showed better performance and was used for the test (AUC-ROC 0.96, precision 1.00, and recall 0.92 on the validation set). K-means clustering was applied to find groups within the data. Elbow method and silhouette scores were used to determine the optimal number of clusters. The formed clusters were analyzed by aggregating features to understand the characteristics of each cluster.

RESULTS: EDA revealed significant positive correlations between age, weight, American Society of Anesthesiologists (ASA) health score, and surgery duration with the risk of developing ED. Among the ML models, RF achieved the highest performance. Key predictive variables, based on the model's feature importance, included delirium screening scales, extubation time, and time to regain consciousness. Unsupervised K-means clustering identified 2–3 optimal clusters, which represented distinct patient subgroups: younger, healthier, low-risk individuals (cluster 0), and older patients with increasing chronic disease burden, higher delirium screening scores, and consequently higher post-operative delirium risk (clusters 1 and 2).

CONCLUSIONS: ML techniques are valuable tools for extracting insights and making accurate predictions from healthcare data. High-performing algorithm-based models can be implemented for clinical decision support systems, facilitating early identification and intervention for ED in pediatric patients. By investigating various variables, it is possible to assess risk and implement preventive measures effectively. Furthermore, unsupervised clustering reveals distinct patient subgroups, enabling personalized perioperative management strategies and enhancing overall patient care.

Keywords: Emergence Delirium; machine learning; artificial intelligence; tonsillectomy/adenotonsillectomy; pediatric anesthesia

Introduction

Emergence Delirium (ED) in pediatric patients is a dissociative state that occurs during the recovery phase (emergence) from general anesthesia. This neurocognitive phenomenon is characterized by behavioral disturbances such as restlessness, agitation, hallucinations, crying, and disorientation [1]. The incidence of ED varies widely, ranging from 2% to 80%, depending on factors such as the grading method used and the type of anesthetic technique applied [2, 3]. Although this complication can affect all age groups, research indicates a higher prevalence among preschool-aged children [4]. Otolaryngology procedures are a significant risk factor for ED, likely due to factors such as the intensity of postoperative throat pain and the potential for airway obstruction or discomfort [4]. While ED includes hy-

Submitted: 6 June 2024 Revised: 19 August 2024 Accepted: 3 September 2024 Published: 20 October 2024
Correspondence to: Alessandro Vittori, Department of Anesthesia and Critical Care, ARCO Roma Ospedale Pediatrico Bambino Gesù IRCCS, 00165 Rome, Italy (e-mail: alexvittori82@gmail.com).

peractive, hypoactive, and mixed forms, hyperactive manifestations are most common in this surgical context [1, 2, 3, 4]. Despite extensive research on this important issue, the diagnosis of ED remains largely exclusionary, and its pathophysiology remains poorly understood [5].

Risk factors for Emergence Delirium (ED) can include various elements such as patient age, preoperative anxiety, pain levels, the type of surgery, and the chosen anesthesia technique [6]. For instance, evidence suggests that sevoflurane, compared to isoflurane, may be more closely associated with the development of ED [6].

The impact of ED can affect the postoperative course. While data does not support a longer hospital stay for children with ED, its onset can cause anxiety and stress for parents and may lead to patient accidents such as falls or removal of surgical dressings [7].

In this complex scenario, artificial intelligence (AI) strategies, including machine learning (ML) approaches, could be employed to identify risk factors or early signs of ED. Such tools could enable preemptive interventions to reduce the incidence and associated complications of ED. AI-driven predictive models could also enhance our understanding of ED's pathophysiological mechanisms, leading to more targeted research for effective prevention and management strategies. Despite several attempts to apply AI to delirium in adults, research specifically addressing ED in children is still limited [8, 9, 10].

This study aims to develop and validate various ML models to identify key factors associated with ED and predict its occurrence in children undergoing tonsillectomy or adenotonsillectomy procedures.

Materials and Methods

Dataset Implemented

The analyses were conducted on a dataset containing data on children who underwent tonsillectomy and/or adenoidectomy. The primary investigation was conducted at the IRCCS (Research Institute) Istituto Giannina Gaslini in Genoa, Italy, following approval from the competent Ethics Committee (protocol number 048/2018). The families provided informed consent for all aspects of the study.

The dataset comprises demographic variables such as age, sex, weight, the presence of neurocognitive issues, other comorbidities, presence and type of respiratory infections in the 7 days preceding the intervention, obstructive sleep apnea, anesthesiologic risk (American Society of Anesthesiologists (ASA) physical status classification), and medication use. Additional variables encompass anesthesia-related information, including pre-anesthesia (drug, dose), anesthesia induction (drug, dose), maintenance (technique, drugs, doses, fluid therapy), intraoperative events (bradycardia, tachycardia, hypotension, intraoperative movements), end of surgery/anesthesia emergence (extubation time in minutes, surgery duration, laryngospasm, desaturation, other adverse events). Bradycardia is defined as a decrease in heart rate (HR) by more than 20% compared to baseline. Tachycardia is characterized by an increase in HR by more than 20% compared to baseline. Hypotension is defined as a reduction in systolic blood pressure by more than 20% compared to baseline. Conversely, hypertension refers to an increase in mean arterial pressure by more than 20% compared to baseline. Desaturation is defined as a SpO_2 level dropping below 90% of the baseline value for more than 15 seconds.

Post-anesthesia care unit (PACU) data include times for full awakening, the occurrence of ED, bradycardia, and desaturation. The pain was evaluated through the Face, Legs, Activity, Cry, Consolability scale (FLACC) or the Numeric Rating Scale (NRS). The FLACC Scale was implemented to assess pain in younger children (generally, in children up to 4 years of age) and those who are unable to communicate their pain verbally [11]. Previous investigations demonstrated the correlation between NRS and FLACC instruments with Pearson's correlation and Spearman's rho, up to $r = 0.97$ [12]. The FLACC scale includes five categories: Face, Legs, Activity, Cry, and Consolability. Each category is observed and scored from 0 to 2, with a total possible score of 0–10, indicating the severity of pain. The NRS measures the self-reported pain intensity. The child is asked to rate their pain on a scale from 0 to 10, where 0 means "no pain" and 10 means "the worst possible pain". The child's understanding of the scale is ensured by providing examples or using visual aids if necessary. The score is directly recorded as the child's reported number.

From PACU to discharge, we recorded postoperative nausea and vomiting, pain and its score, and discharge on time (24 hrs.). According to the primary investigation [7], the Pediatric Anesthesia Emergence Delirium (PAED) tool was utilized to assess ED. The tool assesses the presence and severity of ED in children post-anesthesia. The scale encompasses five items: eye contact, purposeful actions, awareness of surroundings, restlessness, and consolability. Each item is scored from 0 to 4 based on the observed behavior, with total scores ranging from 0 (no delirium) to 20 (severe delirium). In the postoperative monitoring within the PACU, the PAED scale was administered three times, at 10-minute intervals, by a dedicated nurse [13, 14]. For data collection purposes, the highest PAED score obtained among the three assessments was considered. The dataset is available at [15].

Preprocessing

The dataset was loaded into a Pandas DataFrame using the *pd.read_excel()* function. The *pd.DataFrame.info()* method was implemented to print information about the data frame including the column names, data types, and the number of non-null values. This preliminary step highlighted that the data consisted of a mix of numeric and object-type columns.

Summary statistics of the columns were printed using *df.describe()* to see basic metrics like mean, standard deviation (SD), min, max, etc. This method provided a high-level view of the distribution of key variables. The multi-indexed column names indicated multiple tables merged. Based on the initial inspection, several data-cleaning steps were applied. The columns were flattened to a single index using the *pd.DataFrame.droplevel()* method, simplifying the column names for analysis. Variables with greater than 80% missing values were removed from the dataset (*df.dropna*, thresh = 0.8). The column names were renamed to more readable names using a list of new names and *df.rename()*. Abbreviations were expanded for clarity. Object columns containing dates were converted to date time datatypes using *pd.to_datetime()*. This process enabled date-time operations like calculating patient age. A new 'Age' column was created by subtracting the "Surgery_Date" from "Birth_Date" using date time arithmetic. The variable Age was calculated in years. Numerical missing values were imputed using mean imputation with *df.fillna(df.mean())*. Categorical columns with missing values were imputed using mode imputation via *df.fillna(df.mode()[0])*. Non-numeric columns like identifiers, notes, etc. were dropped, keeping only useful numeric and date time data. Rows with remaining missing values were dropped using *df.dropna()* to keep only complete cases. This multi-step cleaning process handled the major data issues of missing values, non-standard formats, and unnecessary columns/rows. The output was a clean numeric data frame ready for analysis.

Additional preprocessing steps were applied to engineer new features. The engineered features process provided additional capabilities like handling categorical, normalizing data, transforming the target variable, and allowing for non-linear effects (feature engineering). Categorical columns like gender, ASA score, etc., were label encoded to numeric values using *LabelEncoder()*. Therefore, they were formatted for the subsequent modeling. Numeric columns were standardized using *StandardScaler()* to normalize the range of values. It subtracted the mean and scaled to unit variance. The target column "Emergence_DELIRIUM" was binarized to 0/1 based on a threshold of 4 on its original severity scale. Finally, the interaction columns were created by multiplying age with chronic conditions to test their combined effect. Polynomial terms were created from BMI and glucose to assess non-linear relationships.

Exploratory Data Analysis

Visualizations and summary statistics were used to understand relationships in the cleaned dataset. *df.describe()* generated statistics like mean, SD, and quartiles for key columns, highlighting the central tendency, spread, and shape of the distributions. Histogram plots were generated using the *df.hist()* function and pandas plotting capabilities. The histograms displayed the distribution of each key variable. Boxplots were generated using Seaborn's boxplot function to visualize the distribution of variables split by the delirium target. Side-by-side boxplots enabled comparing the distributions across the classes. Correlation heatmaps were created with Seaborn's heatmap function. These plotted the correlation matrix between all variables as a color-coded heatmap. Pairplots were generated using Seaborn's pairplot to show scatterplots between pairs of variables along with histogram subplots. This visualized the relationship and interactions between the variables. In total, over 20 visualizations were created to provide different perspectives on the relationships within the data based on summary statistics, variable distributions, correlations, interactions, and segmentation based on the target column. Key observations were that age, weight, ASA score, and surgery duration. Correlations with ED were tested.

Predictive Modeling

The data was split into train and test sets using *scikit-learn's train_test_split()* function. Four predictive models including logistic regression (LR) [16], random forest (RF), Gradient Boosting (XGBoost), and Support Vector Machine (SVM) [16, 17] were trained and evaluated. LR with default parameters was fit on the training data using *LogisticRegression()*. The test set predictions were generated using *.predict_proba()*. The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, precision, and recall were calculated using *sklearn.metrics*. The RF model was initialized with *RandomForestClassifier()*. Hyperparameter tuning was done using GridSearchCV [18] to find optimal parameters and included n_estimators 100 to 500, max_depth 5 to 20, and min_samples_split: 2 to 10. The tuned model was fit on the training set and predictions were generated on the test set. Model evaluation metrics were calculated using *.predict_proba()* and *sklearn.metrics*. These evaluation metrics were generated using probability predictions. The models were ultimately compared based on their ROC AUC, precision, recall, and sensitivity scores.

Unsupervised Learning

KMeans clustering was applied to find subgroups within the cleaned dataset [19]. Data was filtered to only numeric columns relevant for clustering. Values were standardized using StandardScaler to normalize ranges. The elbow method was used to determine the optimal number of clusters k. Inertia vs k was plotted and the "elbow" point was selected. Silhouette analysis was also performed to evaluate cluster coherence for different k values. KMeans model was initialized with "n_clusters=k" based on the above analysis. The model was fit on the scaled data using *.fit()*. Cluster labels were generated for each sample using *.predict()*. Aggregate statistics of features in each cluster were calculated using *.groupby()* and aggregations like *.mean()*.

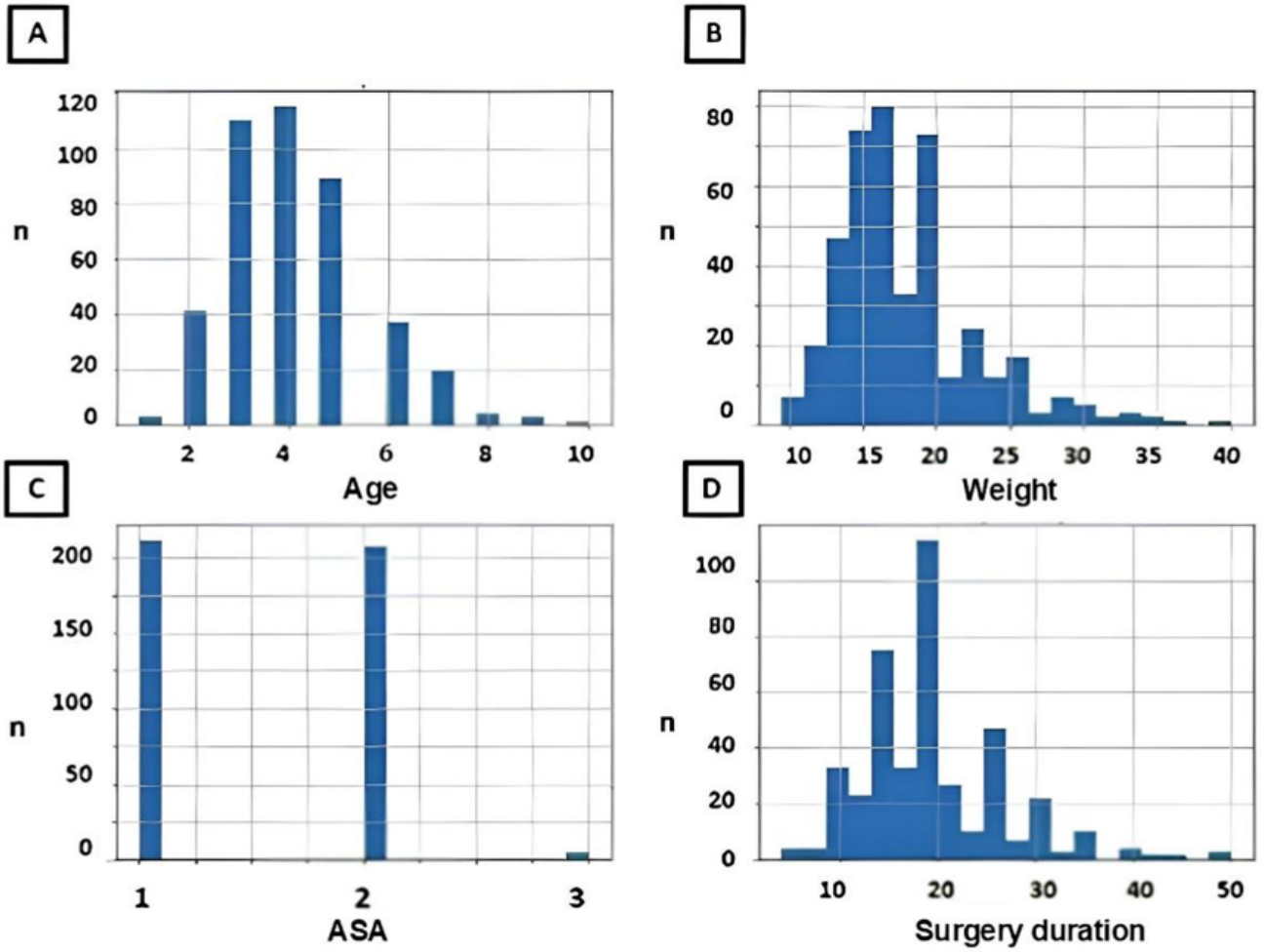

Fig. 1. Histograms of key variables including age (A), weight (kilograms) (B), ASA status (C), and surgery duration (minutes) (D). (n = 423). ASA, American Society of Anesthesiologists.

Software and Libraries Implemented

The software and libraries used for this project were primarily written in Python 3.6 (https://www.python.org/). For data manipulation and analysis, the Pandas library was employed, allowing for the seamless reading of CSV files into a DataFrame. Scikit-learn, a commonly used ML library for Python, provided various modules including *model_selection* for functions such as "train_test_split" and "GridSearchCV", for ensemble models like "RandomForestClassifier" and "GradientBoostingClassifier", *linear_model* for algorithms such as "LogisticRegression", metrics for evaluation metrics such as "roc_auc_score, precision_score", and "recall_score", and preprocessing for data preprocessing techniques such as "LabelEncoder" and "SimpleImputer".

Numerical operations were handled using Numpy, especially when converting the target variable, while data visualization and plotting were accomplished using *Matplotlib.pyplot* for charts like the elbow curve.

The code encompassed an end-to-end ML pipeline, which included data preprocessing, exploratory data analysis, predictive modeling using various techniques, and model evaluation and hyperparameter tuning with GridSearchCV. This combination of Pandas for data manipulation, scikit-learn for ML algorithms and utilities, Numpy for numerical operations, and Matplotlib for visualization provided a robust suite of tools for implementing ML models in Python.

Results

Exploratory Data Analysis Results

The exploratory data analysis provided insights into the variable distributions and relationships within the dataset. Given the original dataset [13], we analyzed data from 423 children (184 females, 239 males). The patients' ages range from 1.5 to 10.1 years old at the time of the procedure, based on their birth dates (mean 3.66 years; SD 1.33). Weights ranged from a minimum of 9.5 kg to 40 kg, with a mean weight of approximately 18 kg (SD 5). Most patients had an ASA score of 1 or 2, n = 211 (49.9%) and n = 207 (48.9%), respectively. The surgery duration spanned from short outpatient procedures of less than 10 minutes to cases lasting 50 minutes (mean 19.8 minutes, SD 5.08) (Fig. 1).

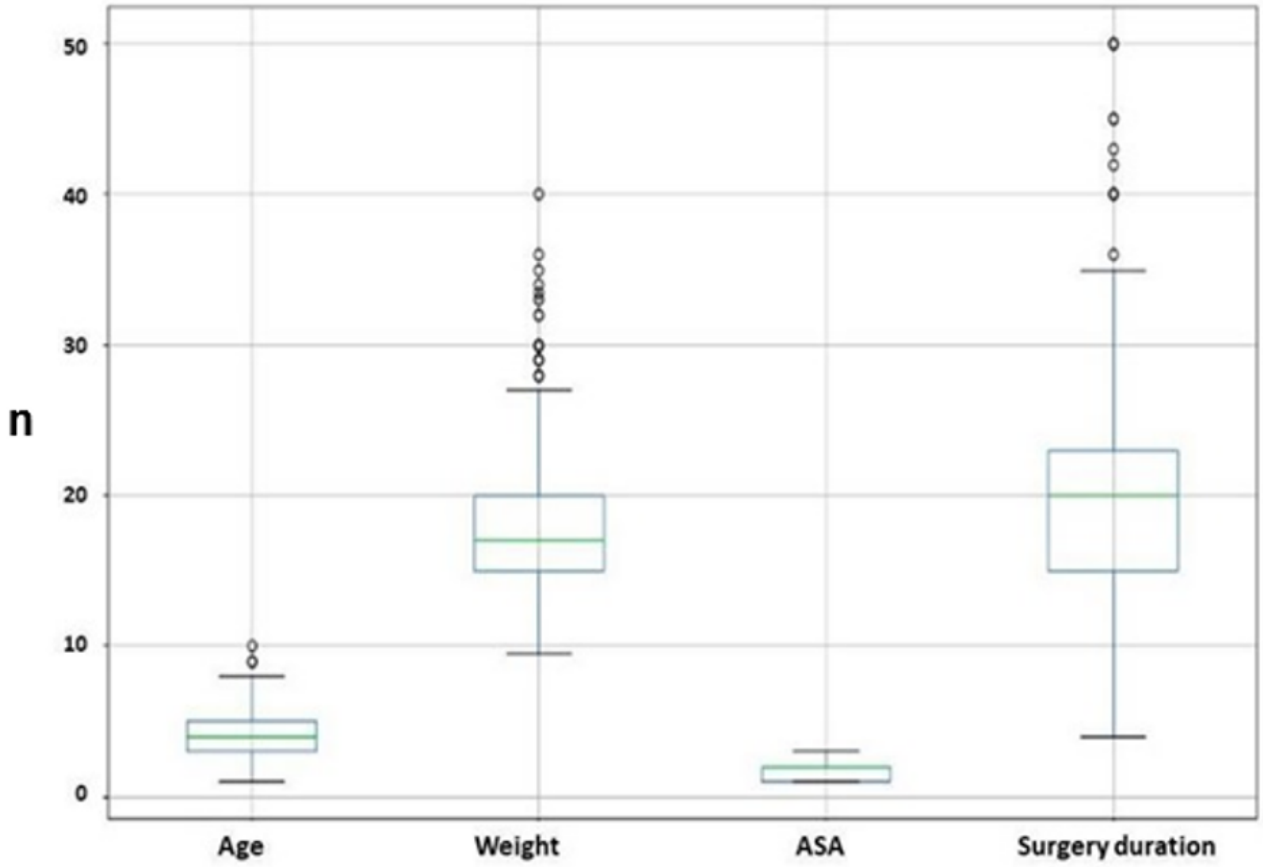

Fig. 2. Boxplots of key variables including age, weight (kilograms), ASA status, and surgery duration (minutes). The boxplots show each variable's distribution, central tendency (median), interquartile range (IQR), and potential outliers. The ages are relatively young with a few older outliers. There is a wider spread in weights, with significant outliers (up to about 40 kg) indicating a few patients with much higher weights. ASA scores are tightly clustered, reflecting less variability and suggesting most patients have similar health statuses. Surgery durations vary widely with a median duration of around 10 minutes and the IQR, indicating that most surgeries last between 5 and 20 units of time; the whiskers extend from about 2 to 35, with several outliers up to 50.

We included the boxplots of key variables to compare the distribution and spread of the four variables, providing insights into their central tendencies, variabilities, and the presence of any outliers. The variables Weight and Surgery Duration exhibit more outliers (Fig. 2).

The correlation heatmap, which helps identify the relationships between pairs of variables, showed weak or no correlations between the considered variables and ED. This indicates little to no linear relationship between them (Fig. 3).

Scatterplots in the pairplot clearly illustrated the positive relationship between age, weight, ASA score, delirium scales, and the target delirium variable (Fig. 4).

Predictive Modeling Results

Four ML models were employed including LG, XBoost, SVM, and RF. For train, validation, and test sets, we used the code:

• X_train, X_temp, y_train, y_temp = train_test_split (X, y, test_size = 0.3, random_state = 42)

• X_val, X_test, y_val, y_test = train_test_split (X_temp, y_temp, test_size = 0.5, random_state = 42)

A preliminary split of the dataset was performed to obtain 70% for training and 30% for testing/validation, followed by an equal split of the testing/validation set to achieve 15% each for testing and validation. Therefore, in this study, we conducted predictive modeling on a dataset split into a training set (70%), validation (15%), and a test set (15%).

Each model underwent hyperparameter tuning using GridSearchCV to identify the optimal parameters. The RF model, for instance, was tuned for n_estimators (100 to 500), max_depth (5 to 20), and min_samples_split (2 to 10). The models were evaluated using AUC-ROC, precision, recall, and log loss metrics.

The training set results showed excellent performance. The LR achieved an AUC-ROC of 0.9911, precision of 0.9733, recall of 0.9359, and log loss of 0.1099; RF and GBoosting both achieved perfect scores with an AUC-ROC of 1.0000 and precision and recall of 1.0000, with log losses of 0.0602 and 0.0003, respectively. The SVM model had an AUC-ROC of 0.9853, precision of 0.9552, recall of 0.8205, and log loss of 0.2836.

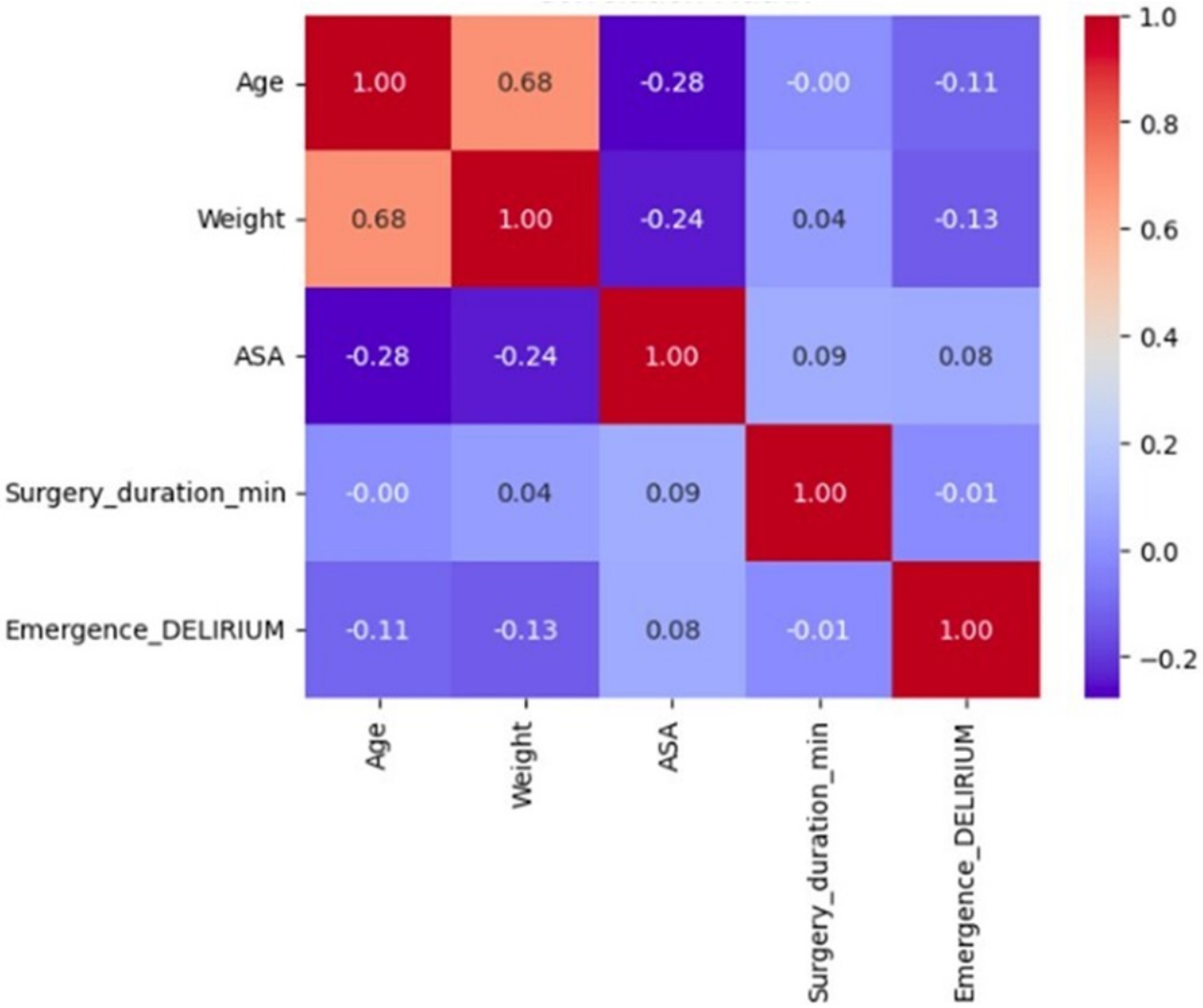

Fig. 3. Correlation matrix. The numbers in each cell represent the Pearson correlation coefficient between the row and column variables. This value ranges from –1 to 1, where 1 indicates a perfect positive correlation: as one variable increases, the other also increases; –1 indicates a perfect negative correlation: as one variable increases, the other decreases; and 0 indicates no linear correlation between the variables. Moreover, the color gradient helps visualize the strength and direction of the correlations; red shades represent positive correlations and blue shades express negative correlations. Moreover, the intensity of the color indicates the strength of the correlation (darker colors for stronger correlations). The correlation heatmap shows that there is a weak negative correlation between Age and Emergence Delirium (ED) (–0.11), indicating that older age is slightly associated with a lower incidence of that complication. There is a weak negative correlation between weight and ED (–0.13). Moreover, ASA has a moderate to strong positive correlation with ED. Regarding surgery duration, there is a very weak negative correlation (–0.01).

The training of the ML models is shown in Fig. 5.

The validation was performed to tune and select the best model and its hyperparameters. The four ML models showed high values of ROC AUC, and Recall. All models were featured by a precision of 100%, which means that every instance that the model predicted as positive is positive. The results of the validation process are reported in Table 1.

Despite XBoost's powerful modeling capabilities, RF's practical advantages—such as improved interpretability, greater robustness to overfitting and noisy data, and easier hyperparameter tuning—better align with our project goals and constraints. Therefore, we have selected RF as our preferred model for deployment. Moreover, the training demonstrated that RF was effective and reliable, providing good performance and generalization.

The RF model was evaluated on the test set. The performances were AUC-ROC 0.96, Precision 1.00, and Recall 0.92.

Key features of the RF model were the delirium screening scale PAED (importance 0.54), the time for full awake in minutes (0.040), weight (0.037), and extubation time (min) (0.037).

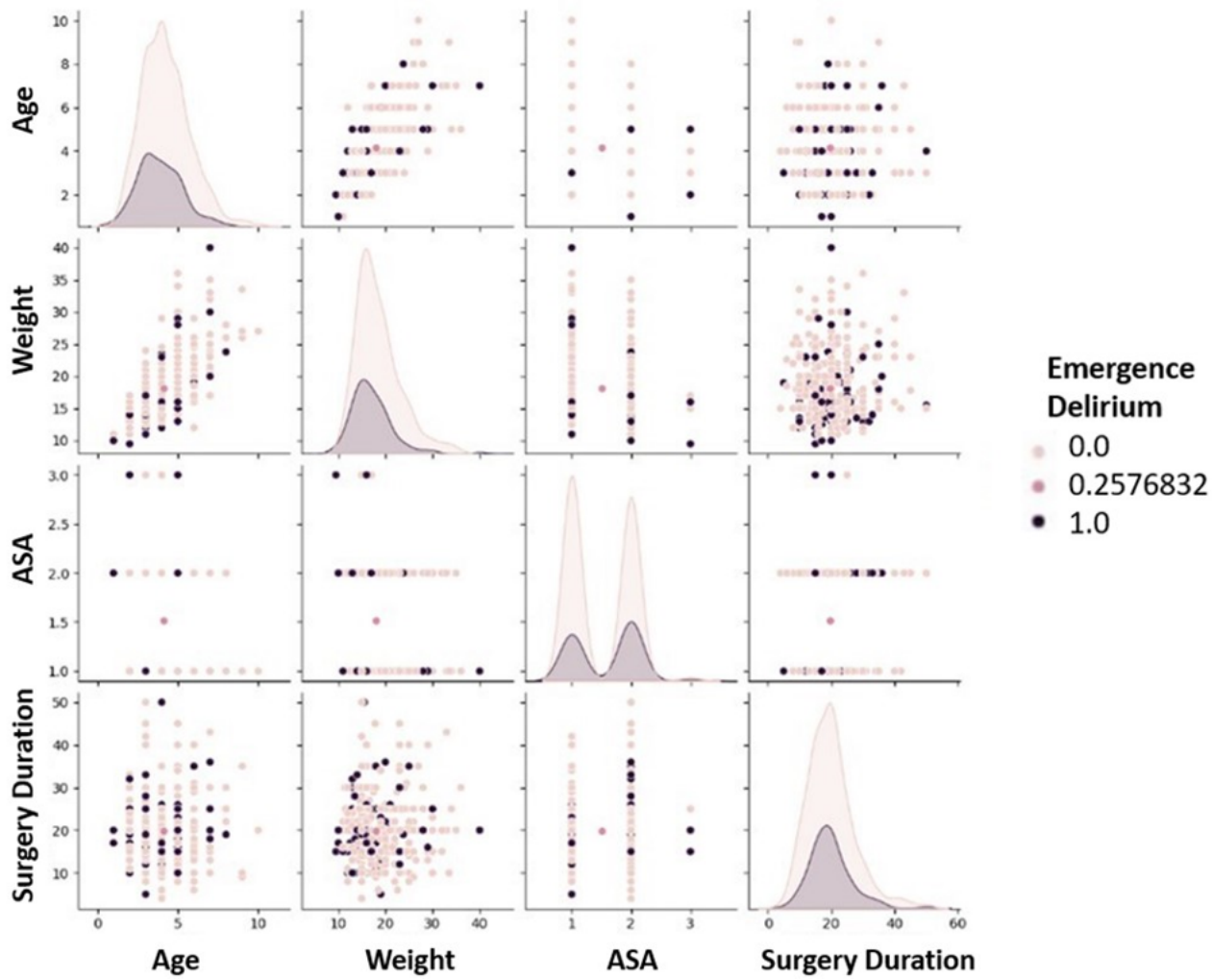

Fig. 4. Pairplot of the numeric variables by the target variable. The pairplots help assess the direction, strength, and nonlinearity of the relationships between the variables, informing which variables are related to the target and each other. Emergence Delirium (ED) is represented in the plot by the color intensity of the data points. The color gradient indicates the presence and severity of ED and helps identify patterns and correlations between variables and ED. Therefore, light pink represents no ED (0.0) and darker shades indicate increasing levels of ED (up to 1.0). About distributions, the diagonal plots show the distribution of each variable and the shaded areas show the density of data points, with darker areas representing higher densities. For pairwise relationships, off-diagonal plots show scatter plots for each pair of variables, colored by the ED severity. The scatter plots involving Age show that older patients tend to have higher instances of ED (darker points). On the other hand, there is no clear correlation between weight and ED, as indicated by the spread of color intensity. Concerning the correlation between the American Society of Anesthesiologists (ASA) score and ED, higher ASA scores (ASA 2 and 3) show some association with higher ED (more dark points). Finally, longer surgery durations (minutes) appear to correlate with higher ED.

Unsupervised Learning Results

K-means clustering was applied to the dataset to discover patient subgroups. The Elbow method for determining the optimal number of clusters identified 3 as the most suitable number (Fig. 6A). The maximum mean silhouette score of 0.56 was achieved with 2 clusters (Fig. 6B).

Given the ambiguity, both 2 and 3-cluster solutions were analyzed:

Two Cluster Solution:

- Cluster 0 (475 points): Younger, lower ASA, delirium scales;
- Cluster 1 (822 points): Older, higher ASA, delirium scales.

Three Cluster Solution:

- Cluster 0 (402 points): Youngest, very low risks;
- Cluster 1 (574 points): Older, moderate chronic diseases;
- Cluster 2 (321 points): Older, highest chronic conditions.

In the 3-cluster solution, Clusters 1 and 2 mainly differ in having more extreme values for some features, such as chronic diseases. Overall, they include older and less healthy individuals. Consequently, the 3-cluster solution essentially splits the higher-risk cluster into two subgroups

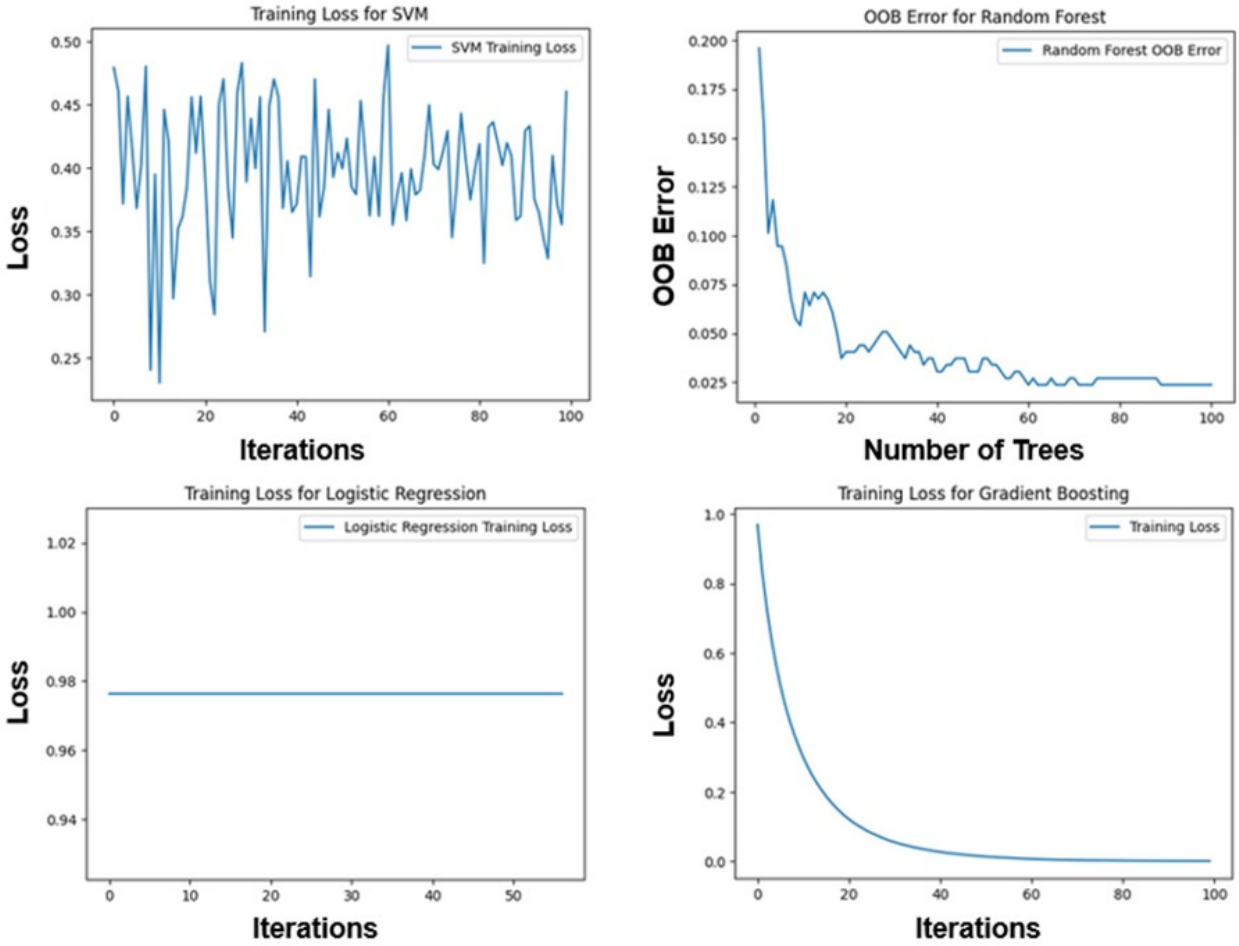

Fig. 5. Training of the considered model. Y-Axis (Loss) indicates the loss during training and X-Axis (Iterations) is the number of iterations. The training loss for SVM fluctuates significantly across iterations, indicating that the model's performance is quite variable and possibly sensitive to the chosen hyperparameters or training data. For RF, the Y-Axis (OOB error), is an estimate of the model's generalization error obtained during training. The OOB error decreases rapidly as the number of trees increases, stabilizing after around 40 trees. This indicates that adding more trees improves the model's performance up to a point, after which the improvement plateaus. The training loss for LR remains constant over iterations, suggesting that the model has reached its optimal solution quickly and does not improve further with more iterations. The training loss for Gradient Boosting decreases sharply at first and then more gradually, indicating continuous improvement in the model's performance as more iterations are added. This characteristic curve shows the model learning effectively over time. OOB, Out-of-Bag.

Table 1. Models' performances on the validation set (n = 63; 15% sample).

Model	ROC AUC	Precision	Recall
Logistic Regression	0.92	1.00	0.84
Random Forest	0.96	1.00	0.92
Gradient Boosting	0.96	1.00	0.92
Support Vector Machine	0.97	1.00	0.76

based on the severity of comorbidities. The clustering revealed distinct patient groups, ranging from younger, healthier individuals to older individuals with increasing levels of chronic disease burden and delirium risk (Fig. 7).

Discussion

In this study, we leveraged various ML techniques to construct predictive models aimed at identifying pivotal factors associated with ED and predicting its onset. Additionally, we utilized unsupervised learning to identify patient subgroups.

Overall, the exploratory data analysis (EDA) uncovered significant relationships between patient factors such as age, health status, surgery factors, and the risk of ED. For instance, the correlation heatmap, a visual representation of the relationships between variables, illustrated that several factors showed at least a weak correlation with the occurrence of delirium. Older age is slightly associated with a lower incidence of this complication, and there is a weak negative correlation between weight and ED.

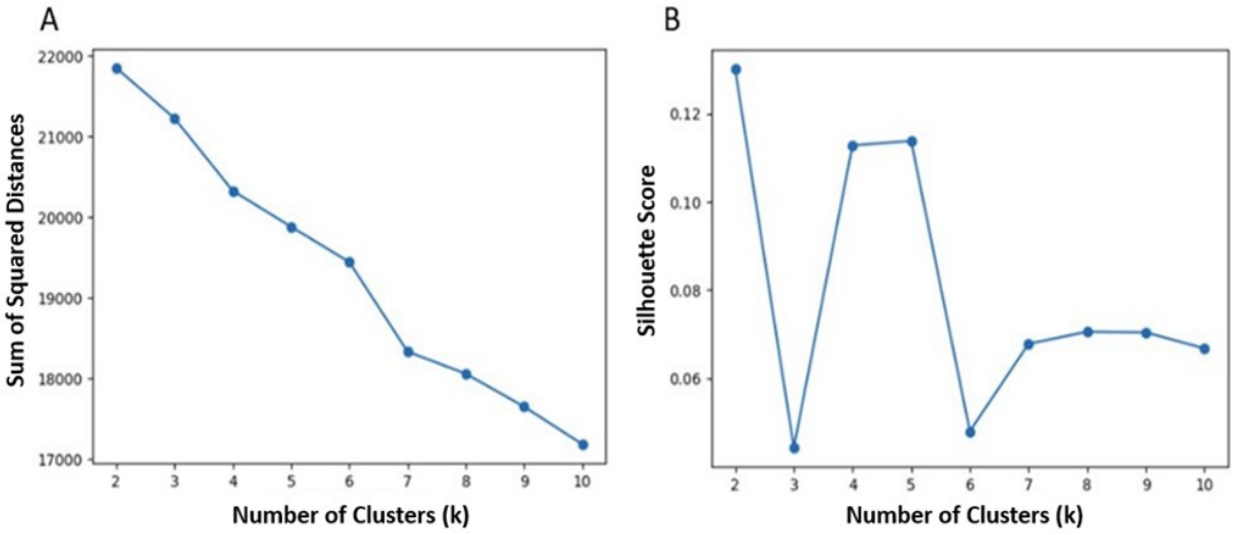

Fig. 6. The elbow method plots the sum of squared distances vs the number of clusters k. The "elbow" indicates optimal k where increasing k no longer significantly decreases the inertia. This k is chosen for the final clustering analysis. Silhouette scores measure how well samples fit within their clusters. Higher scores indicate better clustering. Therefore, k is chosen where the silhouette score is maximized. The Elbow method for optimal k indicated 3 as the optimal number of clusters (A). The Silhouette analysis score for optimal k indicated 2 clusters (0.56) (B).

Findings from other studies suggest the potential impact of demographic characteristics, pre-existing health conditions, and procedural factors on the likelihood of experiencing delirium in the postoperative period [20, 21]. However, this topic is highly debated, and precise correlations among these factors remain undetermined. Furthermore, study on pediatric ED have mostly identified associated factors using relatively small sample sizes, with only a limited number focusing on the development of predictive models [22]. This highlights the need for research efforts to develop more robust predictive models in this area.

In a recent study, Yu *et al*. [13] developed different predictive ML models for ED using a large dataset (n = 43,830). However, the authors acknowledged that the models they tested exhibited only moderate predictive performance with ROC AUC ranging from 0.74 to 0.75. In our study, the results of the ML investigations showed that while all models performed well, the RF demonstrated superior metrics compared to the other models. Notably, according to the model, in the supervised analysis (output delirium prediction), the PAED scale, time to wake up, weight, and extubation time were among the most important for predicting postoperative delirium. This finding underscores the critical role that these parameters play in the postoperative period and the reliability of the tool used for the assessment of the complication. Time to wake up, which measures the time required for a patient to regain consciousness after anesthesia, can offer insights into both the recovery process and the effectiveness of anesthetic management. Similarly, extubation time, the period from the end of surgery to the removal of the endotracheal tube, is a crucial indicator of a patient's ability to maintain airway patency and resume adequate spontaneous breathing. Therefore, optimizing these parameters through careful anesthetic management and monitoring, and targeted interventions could potentially reduce the incidence of ED, leading to better outcomes for patients. Strategies could include the use of fast-acting anesthetic agents, meticulous intraoperative monitoring, and protocols for early and safe extubation.

Remarkably, in contrast with the EDA analysis, we failed to find an association between the variable "age" and the risk of ED in the RF model, suggesting that age may not be a significant factor in predicting or understanding ED. On the contrary, previous study has noted a correlation between advancing age and reduced occurrence of ED [23]. Age can serve as a proxy for various physiological, developmental, and procedural factors that contribute to the risk of this complication, although the underlying biological mechanisms of this phenomenon should be better explained [24]. In contrast to other study, variables like midazolam premedication were not found to have an association [13].

The unsupervised clustering pipeline identified distinct groups of patients based on their clinical profile characterized by the set of features. The 3-cluster solution essentially split the older, higher-risk cluster into two subgroups based on the severity of comorbidities. The clustering uncovered distinct patient groups ranging from younger, healthier individuals to older with increasing levels of chronic disease burden and delirium risk. Consequently, this subgroup identification confirms the results of previous investigations [4, 5] and could facilitate personalized care approaches and targeted interventions tailored to specific patient profiles.

Cluster	Gender	Weight	ASA	Intellectual_disability	\
0	0.565056	17.778016	1.507483	0.014931	
1	0.574194	18.621290	1.522581	0.019355	

Cluster	Cardiovascular_drugs	Neuropsychiatric_therapy	Chronic_diseases	\
0	0.003735	0.003735	0.048512	
1	0.006452	0.006452	0.051613	

Cluster	Infections_7_days_preoperatively	Type_of_infection	OSAS_dichotomy	\
0	0.100750	3.862454	0.261392	
1	0.103226	3.787097	0.406452	

Cluster	...	Delirium_scale_PAED	Pain_dichotomy	Pain_scale_FLACC_NRS	\
0	...	9.884758	0.048556	5.981413	
1	...	10.161290	0.083871	6.032258	

Cluster	BRADICARDIA_PACU	Management_PACU	Oxygen_desaturation_PACU	\
0	0.070843	1.996283	0.029959	
1	0.032258	1.987097	0.109677	

Cluster	PONV_Dichotomy	Pain_PACU	Pain_value_FLACC_NRS	\
0	0.071010	0.115857	5.576208	
1	0.154839	0.251613	5.109677	

Cluster	Discharge_on_time_24_hrs
0	0.962623
1	0.916129

[2 rows x 58 columns]

Fig. 7. Cluster solutions summary and performances.

Clinical Implications

The findings have the potential to inform perioperative care practices and enhance patient outcomes. In particular:

- The high-performing predictive models hold promise for clinical decision support systems aimed at early identification and intervention for ED in pediatric patients. Preoperative and intra-operative variables can proactively be used to assess the risk of ED and implement preventive measures to optimize patient outcomes. Furthermore, the development of a user-friendly risk calculator based on these variables could facilitate clinicians in quantifying the risk of ED and tailoring interventions accordingly. Therefore, these models could be adopted to design information and communication technologies devices useful for different aims [24].

- The identification of distinct patient subgroups through unsupervised clustering offers opportunities for personalized medicine approaches. Clinicians can tailor perioperative management strategies based on individual patient profiles, optimizing resource allocation, and improving overall patient care and satisfaction.

Limitations

A significant limitation of this study is its generalizability. The findings may not apply to populations outside the specific context of pediatric patients undergoing tonsillectomy or adenotonsillectomy procedures. Additionally, the generalizability is constrained by data collection from a single center. To enhance clinical relevance and generalizability, the model should be trained on a larger dataset and validated

on both external datasets and/or a larger internal dataset.

Furthermore, several variables have not been verified. For example, laboratory values such as the neutrophil-lymphocyte ratio, which appears to be closely related to ED [20], were not included. Another significant limitation is the sample size, which may be insufficient for a robust ML approach. With a 70:15:15 train-validation-test split, the test set consists of only 63 patients. Such a small dataset increases the risk of overfitting, particularly given the high AUCs observed. To address this, it is essential to include more cases, train the model on a larger dataset, and validate it on external datasets and/or a larger internal dataset to improve its clinical relevance and generalizability.

Conclusions

In summary, this study identified key risk factors for post-operative delirium in pediatric patients using predictive modeling, and unsupervised learning. The RF model achieved the highest accuracy in forecasting delirium, highlighting the importance of variables such as delirium scales, extubation time, and time to regain consciousness. Unsupervised learning revealed distinct patient subgroups with varying risk profiles, providing insights for risk stratification and personalized intervention strategies. These findings underscore the potential of data-driven approaches to enhance the prediction and management of this discomforting complication in pediatric patients. Further research and clinical validation are encouraged to refine and implement these findings in real-world healthcare settings.

Availability of Data and Materials

All experimental data included in this study can be obtained by contacting the first author if needed.

Author Contributions

AS, and MC, designed the research study. AV, RP, EGB, and OP performed the research. JM and MC analyzed the data. MGC interpreted data for the study. MC wrote the original draft. MC and JM performed text and analysis revisions according to the editorial and reviewers' comments. All authors revised the manuscript critically for important intellectual content. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Ethical approval for this study (protocol number 048/2018) was provided by the Regional Ethics Committee of Istituto Giannina Gaslini in Genoa, Italy. The families provided informed consent for all aspects of the study. This study adhered to the principles outlined in the Helsinki Declaration.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Urits I, Peck J, Giacomazzi S, Patel R, Wolf J, Mathew D, *et al*. Emergence Delirium in Perioperative Pediatric Care: A Review of Current Evidence and New Directions. Advances in Therapy. 2020; 37: 1897–1909.

[2] Dahmani S, Delivet H, Hilly J. Emergence delirium in children: an update. Current Opinion in Anaesthesiology. 2014; 27: 309–315.

[3] Modi D, Goyal S, Kothari N, Sharma A, Kumar R, Chhabra S, *et al*. Comparison of incidence of emergence delirium in pediatric patients with three different techniques of general anesthesia using sevoflurane and propofol: a randomized controlled trial. Brazilian Journal of Anesthesiology (Elsevier). 2022; 72: 841–842.

[4] Kanaya A. Emergence agitation in children: risk factors, prevention, and treatment. Journal of Anesthesia. 2016; 30: 261–267.

[5] Liu T, Luo F. The Topics and Publication Trends in Emergence Deliri-Um: A Bibliometric Analysis from 2002 to 2022. Journal of Pain Research. 2023; 16: 2729–2745.

[6] Farag RS, Spicer AC, Iyer G, Stevens JP, King A, Bain PA, *et al*. Incidence of emergence agitation in children undergoing sevoflurane anesthesia compared to isoflurane anesthesia: An updated systematic review and meta-analysis. Paediatric Anaesthesia. 2024; 34: 304–317.

[8] Neto PCS, Rodrigues AL, Stahlschmidt A, Helal L, Stefani LC. Developing and validating a machine learning ensemble model to predict postoperative delirium in a cohort of high-risk surgical patients: A secondary cohort analysis. European Journal of Anaesthesiology. 2023; 40: 356–364.

[9] 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 Anesthesiology. 2022; 22: 8.

[10] Chen D, Wang W, Wang S, Tan M, Su S, Wu J, *et al*. Predicting postoperative delirium after hip arthroplasty for elderly patients using machine learning. Aging Clinical and Experimental Research. 2023; 35: 1241–1251.

[11] Marchetti G, Vittori A, Tortora V, Bishop M, Lofino G, Pardi V, *et al*. Prevalence of pain in the departments of surgery and oncohematology of a paediatric hospital that has joined the project "Towards pain free hospital". La Clinica Terapeutica. 2016; 167: 156–160.

[12] Skov ST, Bünger C, Li H, Vigh-Larsen M, Rölfing JD. Lengthening of magnetically controlled growing rods

caused minimal pain in 25 children: pain assessment with FPS-R, NRS, and r-FLACC. Spine Deformity. 2020; 8: 763–770.

[13] Yu H, Simpao AF, Ruiz VM, Nelson O, Muhly WT, Sutherland TN, *et al*. Predicting pediatric emergence delirium using data-driven machine learning applied to electronic health record dataset at a quaternary care pediatric hospital. JAMIA Open. 2023; 6: ooad106.

[14] Blankespoor RJ, Janssen NJJF, Wolters AMH, Van Os J, Schieveld JNM. Post-hoc revision of the pediatric anesthesia emergence delirium rating scale: clinical improvement of a bedside-tool? Minerva Anestesiologica. 2012; 78: 896–900.

[15] Simonini A. Dataset_Pediatric_Delirium 2024. Available at: https://doi.org/10.5281/zenodo.10471867 (Accessed: 23 July 2024).

[16] Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. Journal of Thoracic Disease. 2019; 11: S574–S584.

[17] Salditt M, Humberg S, Nestler S. Gradient Tree Boosting for Hierarchical Data. Multivariate Behavioral Research. 2023; 58: 911–937.

[18] Sarıdemir M, Topçu İ, Özcan F, Severcan M. Prediction of long-term effects of GGBFS on compressive strength of concrete by artificial neural networks and fuzzy logic. Construction and Building Materials. 2009; 23: 1279–1286.

[19] Pei S, Chen H, Nie F, Wang R, Li X. Centerless Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023; 45: 167–181.

[20] Feng B, Guo Y, Tang S, Zhang T, Gao Y, Ni X. Association of preoperative neutrophil-lymphocyte ratios with the emergence delirium in pediatric patients after tonsillectomy and adenoidectomy: an observational prospective study. Journal of Anesthesia. 2024; 38: 206–214.

[21] Chen W. Regarding the risk factors for the emergence delirium in pediatric patients after tonsillectomy and adenoidectomy. Journal of Anesthesia. 2024; 38: 575.

[22] Petre MA, Saha B, Kasuya S, Englesakis M, Gai N, Peliowski A, *et al*. Risk prediction models for emergence delirium in paediatric general anaesthesia: a systematic review. BMJ Open. 2021; 11: e043968.

[23] Sikich N, Lerman J. Development and psychometric evaluation of the pediatric anesthesia emergence delirium scale. Anesthesiology. 2004; 100: 1138–1145.

[24] Dantas C, Machado N, Ortet S, Leandro F, Burnard M, Grünloh C, *et al*. The Iterative Model of Ethical Analysis for Large-Scale Implementation Of ICT Solutions. Translational Medicine @ UniSa. 2020; 23: 1–9.