Journal of Anesthesia (2024) 38:747–755
https://doi.org/10.1007/s00540-024-03377-7

ORIGINAL ARTICLE

Pediatric cardiac surgery: machine learning models for postoperative complication prediction

Rémi Florquin[1,2] · Renaud Florquin[3] · Denis Schmartz[4] · Philippe Dony[1] · Giovanni Briganti[2]

Received: 8 December 2023 / Accepted: 4 July 2024 / Published online: 19 July 2024

Abstract

Purpose Managing children undergoing cardiac surgery with cardiopulmonary bypass (CPB) presents a significant challenge for anesthesiologists. Machine Learning (ML)-assisted tools have the potential to enhance the recognition of patients at risk of complications and predict potential issues, ultimately improving outcomes.

Methods We evaluated the prediction capacity of six models, ranging from logistic regression to support vector machine, using a dataset comprising 33 variables and 1364 subjects. The Area Under the Curve (AUC) and the F1 score served as the primary evaluation metrics. Our primary objectives were twofold: first, to develop an effective prediction model, and second, to create a user-friendly comprehensive model for identifying high-risk patients.

Results The logistic regression model demonstrated the highest effectiveness, achieving an AUC of 83.65%, and an F1 score of 0.7296, with balanced sensitivity and specificity of 77.94% and 76.47%, respectively. In comparison, the comprehensive three-layer decision tree model achieved an AUC of 72.84%, with sensitivity (79.41%) comparable to more complex models.

Conclusion Our machine learning-assisted tools provide an additional perspective and enhance the predictive capabilities of traditional scoring methods. These tools can assist anesthesiologists in making well-informed decisions. Furthermore, we have successfully demonstrated the feasibility of creating a practical white-box model. The next steps involve conducting clinical validation and multicenter cross-validation.

Trial registration NCT05537168

Keywords Anesthesiology · Artificial intelligence · Machine learning · Pediatric cardiac surgery

Introduction

Anesthesia is a pioneering medical specialty known for its focus on innovation [1] and prioritization of safety due to its high-risk nature [2]. Implementing anesthesia information management systems (AIMS) has also facilitated more personalized practice and seamless integration into the hospital's informatics environment [3]. This technological development has substantially increased the volume of data generated within the operating room, providing opportunities for research purposes [4] and enabling the measurement of the quality of care provided [5].

Traditional research underutilizes the potential of multivariate data, disregarding the intricate relationships between them. Machine learning (ML) can reveal patterns that may not be immediately evident in vast volumes of data. Moreover, it offers techniques for studying complex problems and constructing simpler models [6]. We can generate new hypotheses and gain deeper insights into understanding phenomena, such as the predictive factors influencing patient outcomes [7].

Managing pediatric patients undergoing cardiac surgery with cardiopulmonary bypass (CPB) poses a considerable challenge for anesthesiologists. Moreover, approximately 90% of the 1 million children born yearly with congenital heart disease (CHD) reside in low- and middle-income countries, where accessing quality care is challenging.

✉ Rémi Florquin
remi.florquin@gmail.com

1 Department of Anesthesiology, CHU Charleroi, Chaussée de Bruxelles 140, 6042 Lodelinsart, Belgium

2 Chair of Artificial Intelligence and Digital Medicine, Mons University, 7000 Mons, Belgium

3 Floconsult SPRL, 1480 Tubize, Belgium

4 Department of Anesthesiology, Hôpital Universitaire de Bruxelles (H.U.B), Université Libre de Bruxelles, 1070 Brussels, Belgium

Furthermore, patients in these countries often present with more advanced pathology and deteriorating physical conditions [8–10]. One vital aspect is the timely recognition of patients at risk of postoperative complications, particularly in settings with limited resources.

On the other hand, the abundance of data available through AIMS presents opportunities for developing predictability models in anesthesia that can assist in identifying patients at risk of complications and anticipating potential issues [11]. Jeffries et al. developed a model that achieved an impressive area under the curve (AUC) of 0.87 to predict mortality in intensive care [12]. Another recent example is a model by Kang et al. that predicts hypotension after the induction of anesthesia, achieving an AUC of 0.84 [13]. However, the development of machine learning in pediatric cardiac surgery faces obstacles due to the relatively small volume of cases and the unique characteristics of pediatric patients.

This study aims to harness the capabilities of machine learning to autonomously acquire knowledge by extracting patterns from the raw data [14] generated during daily operations in the operating room. Our objective is to develop models that accurately predict patients at risk of postoperative complications. We will compare a relevant set of machine learning models, initially focusing on identifying the best model for detecting patients at risk. Subsequently, we aim to create a white-box, user-friendly model.

Materials and methods

Study design

Our study entailed a retrospective analysis of a cohort of consecutive children aged 0–16 years who underwent cardiac surgery with CPB between 2008 and 2018 at Hôpital Universitaire des Enfants Reine Fabiola (HUDERF) in Brussels, which is a tertiary children's hospital. The study received approval from the HUDERF Ethics Committee, and the requirement for informed consent (CEH no 53/22) was waived due to the study's retrospective nature, and was registered on clinical trial (https://classic.clinicaltrials.gov/ct2/show/NCT05537168).

During the study period, our institution followed standardized intraoperative management practices; the specific details can be found in online resource 1. In cases of significant clinical bleeding, as determined by an algorithm developed by Despotis et al. [15], based on platelet count and standard coagulation tests (PT and aPTT), fresh-frozen plasma and platelet concentrates were administered after CPB.

Exclusion criteria

Due to their specific care requirements, we excluded children with an American Society of Anesthesiologists (ASA) score of 5 and Jehovah's Witnesses from the study.

Variables

A total of 33 variables were collected for this analysis, as outlined in Tables 1 and 2. The gathered data included demographic information such as age, sex, and pathology. Additionally, the type and volume of fluids administered in the operating room (OR) and the type and volume of fluids lost during the procedure were recorded. Cardiac pathology was categorized as cyanotic if the disease involved a right-to-left shunt. The preoperative physical status of the patients was described using the ASA score. The Risk Adjustment for Congenital Heart Surgery (RACHS) score was calculated to enable meaningful comparisons of hospital mortality among groups of children undergoing heart surgery for CHD [16]. At the end of the procedure, the Pediatric Index of Mortality and the Pediatric Risk of Mortality (PRISM) score were calculated.

Our primary outcome measure was the Modified Organ Dysfunction Score 2 (MODS2), which indicated severe postoperative morbidity and mortality. This composite outcome developed by Willems et al. MODS2 is a binary outcome, encompassing hospital death or the presence of at least two of the following: pulmonary failure, prolonged inotropic support, and renal failure, as defined in a previous publication [17]. Respiratory failure was determined by the need for mechanical ventilation for more than 90 h from pediatric intensive care unit admission until tracheal extubation, representing the 75th percentile of our study population [18]. Prolonged inotropic support was characterized by continuous infusion of vasoactive drugs for more than 48 h after surgery (excluding dopamine or dobutamine at doses ≤ 5 µg/kg/min). Renal failure was identified by the estimated postoperative creatinine clearance (eCCr) value displaying a reduction of at least 75% compared to the preoperative baseline eCCr [19]. This composite outcome provides a clear and efficient method to differentiate patients who encounter early severe complications. This method is simple and covers most of the severe postoperative complications. It not only simplifies our analysis but also ensures comprehensive coverage of important postoperative outcomes. Importantly, variables are collected by the end of the procedure, prior to the occurrence of postoperative complications (and so MODS2).

Table 1 Preoperative variables

Variable	MODS2 = no ($n = 1022$)	MODS2 = yes ($n = 342$)	P value
Age (months)	15.7 [5.0–54.9]	3.0 [0.6–8.0]	< 0.001
Male gender (%)	57	61	0.210
Preoperative weight (kg)	8.7 [5.3–16.2]	4.2 [3.3–5.9]	< 0.001
ASA physical status	3 [3–3]	3 [3–4]	< 0.001
Score 2	131 (13%)	3 (1%)	
Score 3	781 (77%)	209 (61%)	
Score 4	108 (10%)	129 (38%)	
Cyanotic heart disease (%)	401/1022 (39)	183/342 (54)	< 0.001
Previous heart surgeries (%)			< 0.001
0	70	87.0	
1	29.5	13.0	
2	0.5	0	
Elective surgery (%)	95%	87%	< 0.001
Hemoglobin (g dl^{-1})	13.0 [11.8–14.6]	13.0 [11.5–15.7]	0.447
Hematocrit (%)	38.7 [35.3–44.35]	40.3 [35.1–47.1]	0.040
Platelets ($\times 10^3$ μL^{-1})	319 [256–397]	319.5 [232.25–416.75]	0.738
INR	1.05 [0.99–1.12]	1.03 [0.97–1.12]	0.317
Fibrinogen (mg dL^{-1})	268 [229–320]	247 [195–309]	< 0.001
Creatinine (mg kg^{-1})	0.34 [0.27–0.44]	0.34 [0.25–0.45]	0.777

MODS2, composite outcome of severe postoperative complication;

Values represent the number (percentage) or median (25th and 75th quartile), as appropriate

ASA American Society of Anesthesiology, *RACHS* risk-adjusted classification for congenital heart disease

Data verification

To ensure data accuracy and mitigate potential misinterpretations, we conducted anomaly detection and feature analysis on our dataset. For this purpose, we utilized Python version 3.9.16 and employed the outlier detection function of scikit-learn [20, 21]. Subsequently, we thoroughly examined the clinical records corresponding to these outliers, which led to the discovery of 11 encoding errors (misplaced commas).

Concerning missing values, we encountered several charts with incomplete variables, totaling 12 instances. As some machine learning models do not support missing values, we addressed this issue by replacing binary and ordinal variables with a value of 0, while continuous variables were replaced with the mean value of the population. While this method may not be the most sophisticated, it represents a conservative approach that guarantees data integrity [22].

Bivariate statistical analysis

Primary statistical analysis was conducted on the population, divided into two groups: those with complications (MODS = 1) and those without complications (MODS = 0). The homogeneity of variances for continuous variables was assessed using Bartlett's test, while the normality of their distribution was tested using Shapiro–Wilk's test. The results are reported as either median and interquartile range (IQR) or mean and standard deviation, depending on the assumption of normality. The comparison of these variables, based on their distribution, was conducted using either the T test or the Wilcoxon rank sum test. Categorical variables are presented as numbers and percentages (%) and were compared using Pearson's chi-square test. A p value less than 0.05 was considered statistically significant.

Statistical analyses were conducted using R software, version 3.6.2 [23]. In addition, we employed a form of unsupervised learning model, specifically kernel principal component analysis (PCA), utilizing Python version 3.9.16 along with the libraries pandas, numpy, matplotlib, scikit-learn, and Yellowbrick [20, 21, 24–27]. This analysis aimed to visualize the data distribution and reduce dimensionality, which refers to the number of input variables [6]. By employing this technique, we aimed to better understand the variables' impact on the outcomes and compare the results with traditional univariate analysis. This approach helped ensure that relevant variables were not excluded from our analysis.

Model development

Since our project focuses on binary classification, specifically identifying patients at high risk or not, we employed supervised machine learning methods.

Table 2 Per-operative and postoperative variables

Variable	MODS2 = no (n = 1022)	MODS2 = yes (n = 342)	P value
Intraoperative data			
Surgery time (min)	212 [179–251]	253 [208–308]	< 0.001
Antifibrinolytics (%)	98	99	0.789
Aortic cross clamp (%)	80	94	< 0.001
Aortic cross clamp (min)	46 [31–65]	61 [43–78]	< 0.001
Selective brain perfusion (%)	3.6	18.1	< 0.001
Selective brain perfusion (min)	34 [30–38]	33.5 [27.2–40.7]	0.612
Circulatory arrest (%)	4.7	13.7	< 0.001
Circulatory arrest time (min)	4 [2–14.25]	9 [3.5–27]	0.049
Minimal temperature (°C)	32 [31–33]	31 [27–32]	< 0.001
CBP			
CBP time (min)	99 [72–129]	133 [110–162]	< 0.001
CPB priming volume (ml kg^{-1})	38 [27–52]	55 [46–67]	< 0.001
CPB hemodilution	47 [38–61]	64 [54–76]	< 0.001
RBC transfusion (%)			
RBC transfusion during surgery	39	75	< 0.001
RBC in CPB priming	33	66	< 0.001
Blood loss			
Intraoperative blood loss (ml kg^{-1})	4.2 [2.5–7.0]	7.4 [4.4–12.7]	< 0.001
FFP transfusion during surgery (%)	8.7	40.0	< 0.001
Platelet transfusion during surgery (%)	1.8	11.4	< 0.001
Score at PICU arrival			
PIM score (%)	1.7 [1.1–2.4]	2.4 [1.6–4.5]	< 0.001
PRISM mortality (%)	2.2 [1.2–5.8]	7.3 [2.8–22.0]	< 0.001

Values represent the number (percentage) or median (25th and 75th quartile), as appropriate.

CPB cardiopulmonary bypass, *RBC* red blood cell, *FFP* fresh-frozen plasma, *PICU* pediatric intensive care unit, *PIM* pediatric index of mortality, *PRISM* pediatric risk of mortality

Both technical and clinical considerations influenced the selection of methods. Considering the relatively small number of dimensions (patient characteristics) and labeled samples (patients with severe postoperative morbidity and mortality), we excluded deep neural networks as viable options. Deep neural networks may not be optimal for modeling with such a limited training sample size [6]. Our goal was not to create the most intricate model but rather to develop a practical and easily interpretable model, adhering to the principles of eXplainable Artificial Intelligence (XAI) [28]. Simplicity served as a guiding principle in our approach.

The primary goal is to ensure that the developed model is accessible and usable by clinicians, even those unfamiliar with AI tools. From a technical perspective, the team implemented the model coding within the Python environment. This decision was motivated by the extensive availability of flexible and powerful libraries in Python version 3.9.16 [20] compared to other environments. Moreover, the team was familiar with the scikit-learn library in Python [21], and the Yellowbrick library was utilized to visualize the results [26].

Six mathematical models were chosen for this analysis: logistic regression, Gaussian naïve Bayes, decision tree, gradient-boosting classifier, random forest classifier, and support vector machine. These models utilized the same input features, underwent similar transformation steps (such as missing data imputation and data normalization), and employed grid search with k-fold cross-validation to identify the optimal hyperparameters for each model.

The processing chain was implemented using the scikit-learn pipeline, which served as a unified framework for evaluating various learning models and facilitated the seamless integration of new models. The code used to train and evaluate these models can be found in online resource 2.

Furthermore, the decision tree model is particularly user-friendly and easily understandable, making it an ideal option for developing a white-box, simple model, and we intentionally restricted the decision tree model to a maximum of four layers of nodes in the hyperparameters. Subsequently, we compared the performance of this simplified model with the more complex models.

Model evaluation

We initially divided the dataset into an 80% training subset and a 20% testing subset for model comparison. We employed a resampling technique in the training set to address the class imbalance issue caused by a higher number of "no complication" patients. This involved duplicating patients who experienced complications to ensure an equal number of patients with and without complications in the training set. However, it is worth noting that this technique may amplify the influence of patient characteristics associated with complications in our models. Nevertheless, given their higher prevalence, this approach effectively prevented imbalanced models from favoring the recognition of patients without complications. In contrast, during model evaluation using the test set, we maintained the original proportion of patients with and without complications. This allowed us to simulate a "real-world" scenario and assess model performance more realistically, while also mitigating the risk of overfitting.

We utilized the AUC plot to visualize the outcomes, with the F1 score as the primary endpoint. The F1 score combines the model's precision and recall, making it particularly advantageous for binary classifiers with comparable precision and recall [6]. In our study, the F1 score is better suited than other metrics, because it effectively handles imbalanced datasets where there is a smaller proportion of patients with complications. The objective is to detect these patients by minimizing false negatives and maximizing sensitivity [29]. We calculated the sensitivity (often referred to as recall in computer science), specificity, positive predictive value (PPV), and negative predictive value (NPV). The sum of the standard error of estimate (SEE) was determined to compare the disparities between the observed and predicted values, while the coefficient of determination (R-squared) was employed to assess the model's goodness of fit.

In the decision tree model, entropy serves as a critical parameter. Entropy quantifies the impurity of a split within the decision tree. The primary goal is to minimize entropy by carefully selecting the most relevant features for splitting [6]. We can see entropy as a quality measure, lower entropy corresponds to more homogeneous subsets of data, facilitating better classification.

Results

Subject characteristics

This retrospective analysis involved 1364 patients who underwent cardiac surgery with CPB between January 2008 and December 2018. The population's median age was 9.6 months, with a 25th percentile (P25) and 75th percentile (P75) of 3.1 and 43.0 months, respectively. Thirteen percent (185/1364) of the patients were younger than 1 month, and 57% (784/1364) were male. The median weight was 6.9 kg, with a P25 and P75 of 4.3 and 14.1 kg, respectively. Among our population, 342 patients (25%) developed MODS2 in the postoperative period.

Bivariate analysis

Out of the 33 variables in our dataset that were relevant to the study, 24 were found to have a significant association with MODS2 at the end of the surgery. The complete list of variables and their corresponding significance levels are presented in Tables 1 and 2.

The PCA results are illustrated in Fig. 1, with the length of the vectors representing their influence on the outcomes, whether positively or negatively, based on their orientation. When dealing with many dimensions, condensing them into a two- dimensional graph can compromise readability. Hence, we have selected the ten most significant vectors for presentation, which included the urgency of surgery, preoperative platelet count, the need for red blood cell priming, hemodilution level, surgical bleeding, duration of surgery, presence of cyanotic disease, preoperative hemoglobin level, history of previous surgery, preoperative weight, and minimum temperature. It is noticeable that patients with complications are primarily clustered on the left side, indicating that the vectors pointing toward the left contribute to an increased risk of complications, while the opposite holds for vectors pointing in the opposite direction.

Model capability

The results of the six models are shown in Table 3. The AUC values ranged from 0.7284 to 0.8365, while the precision ranged from 0.6434 to 0.7200.

The logistic regression model demonstrated the best AUC of 0.8365 (Fig. 2). However, all the models except our simplest model (decision tree) achieve an AUC above 0.8. The best sensitivity across our model was achieved by the decision tree and the random forest models. The best specificity was achieved by the Gaussian naïve Bayes. The best precision was achieved by the logistic regression model.

Simplest model

As outlined in the methodology chapter, we limited the decision tree model to a maximum of four layers. The resulting tree (Fig. 3) commenced with weight as the initial split, using a cut-off value of 6.895 kg. The second layer involved either the duration of CPB or surgery. The third layer encompassed variables, such as priming

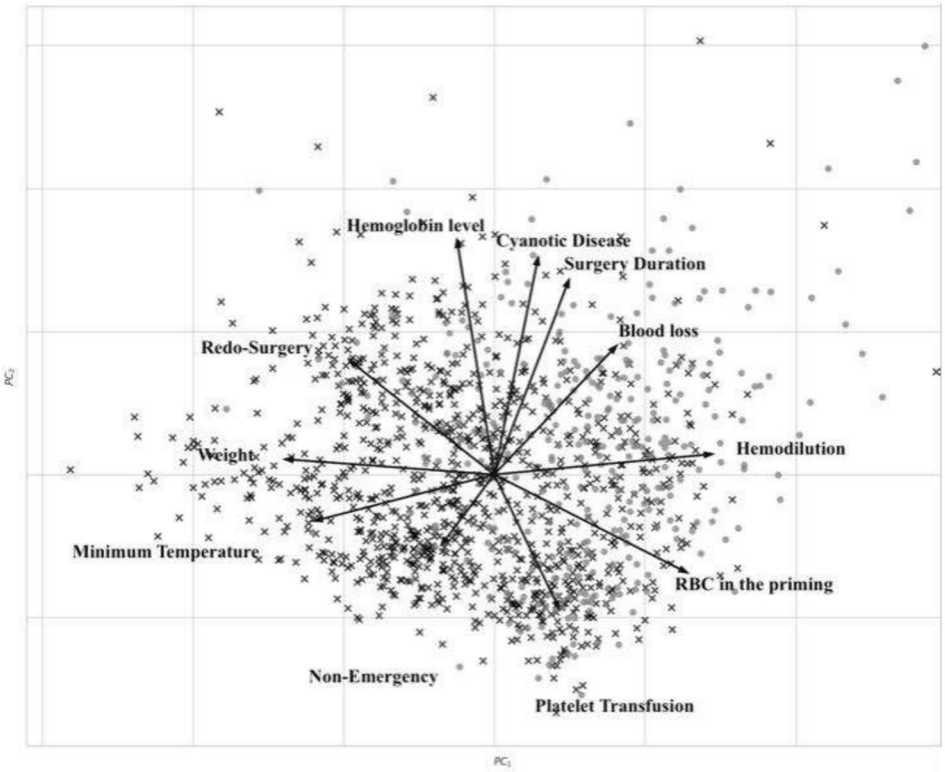

Fig. 1 Principal Component Analysis (PCA) plot with ten dimensions. The crosses represent patients without complications (MODS = 0) and the circles those with (MODS = 1). Most patients with complications are situated on the right side of the plot, indicating that dimensions aligned with the right axis might be detrimental, whereas dimensions aligned with the left axis might be protective

Table 3 Models' performances

	AUC	R^2	SEE	Sensitivity	Specificity	PPV	NPV	F1 score	Precision
Logistic regression	0.8365	0.7684	63	0.7794	0.7647	0.5247	0.9123	0.7296	0.7185
Gaussian naïve Bayes	0.8238	0.7574	66	0.6618	0.7892	0.5114	0.8750	0.7034	0.6931
Decision tree model	0.7284	0.6397	98	0.7941	0.5882	0.3913	0.8955	0.6171	0.6434
Gradient boosting	0.8036	0.7389	71	0.7059	0.7500	0.4848	0.8844	0.6932	0.6846
Random forest	0.8231	0.7353	72	0.7941	0.7157	0.4821	0.9125	0.7011	0.6973
Support vector machine	0.8206	0.7472	68	0.7353	0.7549	0.500	0.8953	0.7072	0.6977

There are variations in sensitivity and specificity among the models AUC, but they exhibited relatively similar performance in terms of AUC and accuracy

AUC area under the curve, *SEE* standard error of the estimate, *PPV* positive predictive value, *NPV* negative predictive value

volume, duration of surgery, preoperative hematocrit, and PRISM score. Finally, the last layer included variables, such as duration of surgery, PRISM score, RACHS score, presence of priming with red blood cells (RBCs), and level of hemodilution.

The more the box has a thick edge, the more the patient is at risk of postoperative complications; in contrast, the patient who ends up in a thin box is considered at low risk. The closer the entropy is close to zero, the more reliable the result is. As we can see, for example, in the fibrinogen box on the second layer, the following layer consists of two low-risk boxes, but one of them has a higher entropy.

Discussion

The relatively small difference between the results of the best and worst models in terms of AUC and precision suggests that the models have comparable predictive abilities. The random forest model demonstrated higher sensitivity, indicating its effectiveness in correctly identifying positive cases.

Conversely, the Gaussian naïve Bayes model showed higher specificity, indicating its ability to accurately identify negative cases. Notably, a simple decision tree

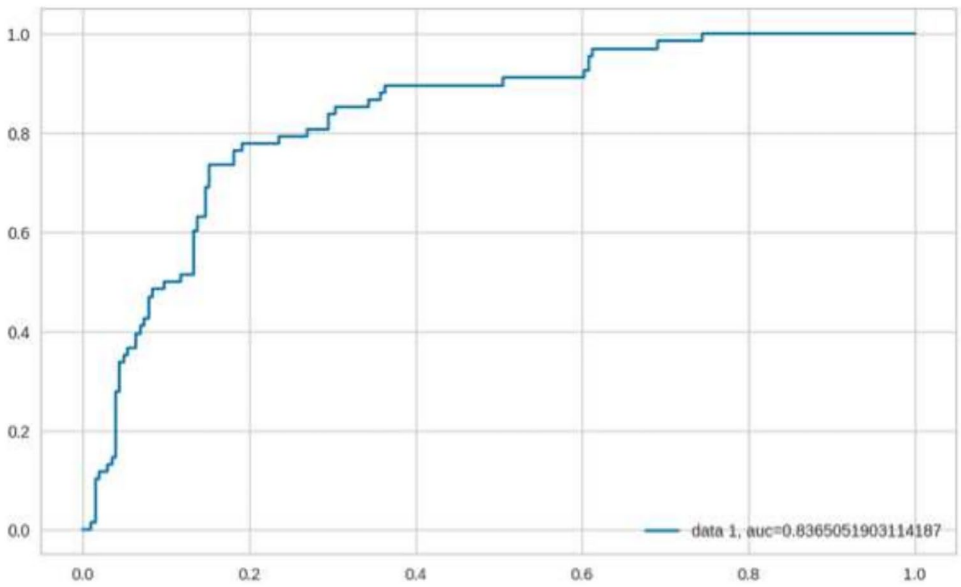

Fig. 2 The area under the curve of the logistic regression model

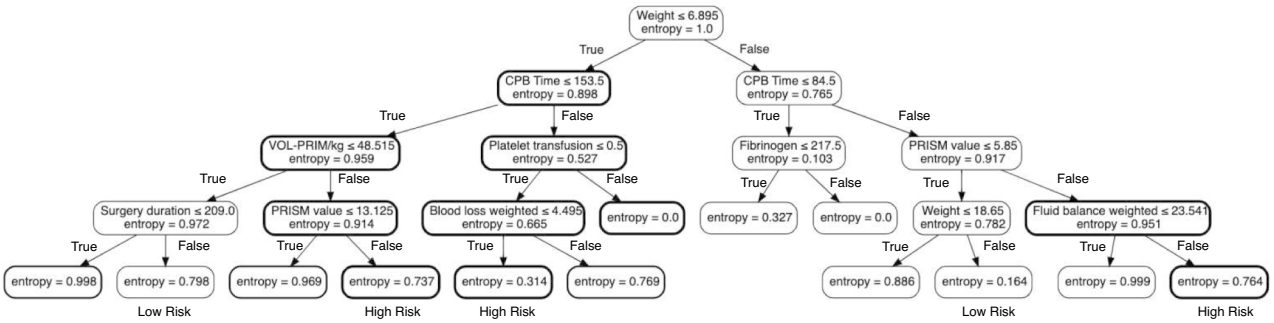

Fig. 3 Our decision tree model is limited to three levels. The boxes with thick edges are linked to a worse outcome than the thin boxes. For example, a patient with more than 6.9 kg and a CPB time of less than 84 min is already in a low-risk root

with only three layers produced results comparable to those of more complex models, achieving an AUC of 0.7284, a precision of 0.6434, and most importantly a sensitivity of 0.7941, which makes it effective in detecting cases at risk despite its simplicity. The development of simplified models with minimal loss of information compared to more intricate counterparts is significant. These streamlined models offer several advantages, including improved interpretability for end-users. The decision-making process behind these models is more transparent. Combining simplicity and comparable performance makes these models highly attractive for practical applications.

The logistic regression model stands out as the top-performing choice. Notably, its AUC establishes it as the best overall model, while its higher F1 score positions it as the optimal choice for detecting at-risk patients (emphasizing precision) while minimizing false negatives (prioritizing sensitivity). Furthermore, its low squared error of estimation (SEE) signifies minimal prediction errors compared to other models, and its higher *R*-squared value indicates a superior explanation of result variance. These results render it the most well-balanced model in consideration.

In recent years, several projects have emerged to develop predictive models in anesthesia. Zeng et al. [30] conducted a study focused on predicting postoperative complications in pediatric cardiac surgery. Their study, which involved 2,308 patients, utilized a gradient-boosting model and achieved an AUC of 0.82. It is worth noting that their patient population was considerably older than ours, with an average age of 22 months, while our population had a median age of 9.6 months. However, the rates of postoperative complications were similar, with 29.3% in their study and 25% in ours. Although the AUC and sensitivity were comparable, ours displayed higher specificity (75% vs. 80%).

Considering that the Zeng et al.'s study was conducted on a different population from ours, it would be intriguing

to pool our models and cross-validate the results on new populations with distinct characteristics. This would provide valuable insights into the robustness and generalizability of our models.

At the clinical level, the tools we have developed need to undergo prospective validation studies to evaluate their impact on patient outcomes.

We propose a two-step process, starting with an initial evaluation using a simple and efficient model, followed by a subsequent evaluation utilizing a more complex model to assist clinicians in making resource allocation decisions. The simplicity of our decision tree model has the benefit of not demanding extensive resources; a simple paper printout would be sufficient. This characteristic is particularly valuable in areas with limited resources, especially considering that 90% of newborns with heart defects are born in low- or middle-income countries [8, 9]. Similar to other AI solutions, this particular product will require validation and may potentially need to be retrained for these specific populations.

Conducting rigorous prospective studies to validate our models would offer clinicians reliable and evidence-based tools to support decision-making. Implementing these predictive tools has the potential to enhance coordinated patient prioritization, thereby leading to improved patient outcomes.

Our study has several limitations, primarily due to its retrospective nature, which introduces the potential for uncontrolled bias. Furthermore, certain variables relied on clinician reports rather than calculated values. Another limitation is the relatively small sample size in our database. Given the restricted size of our patient cohort, we faced technical limitations that hindered our ability to stratify the data based on age. This stratification would have ideally enhanced the accuracy of our analysis. Machine learning models typically require a large sample size for effective training and validation, and our specific population size presents challenges when attempting to study the models on a broader scale. The employment of a composite outcome can present interpretative challenges, primarily due to the heterogeneity inherent in its components. Furthermore, its dichotomous nature may potentially reduce the precision of the analyses. This study concentrates on the intraoperative phase, with no consideration given to factors from the postoperative period. It is important to note that our study was conducted at a single center, potentially limiting the generalizability of our findings.

Additionally, our primary outcome focuses solely on in-hospital mortality and morbidity, providing insights into short-term outcomes only. Another limitation is the duration of the study period. Although clinical practice remained standardized, the 10-year time frame is relatively long, with the upper limit of the period already being 5 years old. This raises the risk of practice variation, emphasizing the need for prospective evaluation of our models, especially considering that new technologies such as ROTEM and TEG may have potentially altered the way platelet or plasma transfusion is assessed.

Our study boasts several strengths. The internal validity of our findings is enhanced by the homogeneity of the clinical practice and the population under study. The data's high quality, with minimal outliers and missing data, is a notable feature of our study. Rather than aiming to develop the most complex model, our study focused on efficiency and user-friendliness by exploring various types of models. Furthermore, including an engineer on our team with expertise in machine learning models provided valuable and diverse perspectives that complemented the medical expertise within the team.

To promote AI, collaboration is crucial. This involves creating inter-center data sharing, collaborative models, and validation across diverse populations. Clinical validation is pivotal, requiring prospective studies to evaluate AI model impact on patient outcomes. Rigorous validation assesses effectiveness, safety, and ethical implications. Proper training and comprehension are essential for ethical AI tool use in medicine. In summary, successful integration relies on collaboration, clinical validation, and ethical awareness.

In anesthesia, timely identification of high-risk cases can significantly improve patient outcomes, reduce morbidity, and optimize resource allocation. Furthermore, integrating AI tools into clinical practice provides anesthesiologists with valuable decision support, allowing for more precise risk assessment and personalized interventions.

Conclusion

We have effectively showcased the potential of AI models for predicting patients at high risk of major postoperative complications. These models exhibited noteworthy performances, with AUCs of over 80%. Additionally, we have developed a reliable decision tree model that, while slightly less powerful than other techniques, offers the advantages of simplicity, interpretability, and user-friendliness.

This study emphasizes the importance of leveraging the abundant data generated in the OR daily to develop tools that can optimize resource allocation and enhance the prediction of complications, ultimately resulting in improved patient outcomes. By harnessing data utilization, safety can be enhanced, leading to an improved quality of care.

A prospective clinical study is necessary to assess the impact of implementing these new tools on patient outcomes. Establishing collaborations among multiple healthcare centers is crucial to create comprehensive

databases that would facilitate the development of robust and validated tools across diverse populations.

Data availability The dataset used for this study is not available online due to the decision of the ethics committee.

Declaration

Conflit of interest The authors have no conflict of interest to declare.

References

1. Kwon AH, Marshall ZJ, Nabzdyk CS. Why anesthesiologists could and should become the next leaders in innovative medical entrepreneurism. Anesth Analg. 2017;124(3):998–1004.
2. Lienhart A, Auroy Y, Péquignot F, Benhamou D, Warszawski J, Bovet M, Jougla E. Survey of anesthesia-related mortality in France. Anesthesiology. 2006. https://doi.org/10.1097/00000542-200612000-00008.
3. Kadry B, Feaster WW, Macario A, Ehrenfeld JM. Anesthesia information management systems: past, present, and future of anesthesia records. Mt Sinai J Med. 2012;79(1):154–65.
4. Deng F, Hickey JV. Anesthesia information management systems: an underutilized tool for outcomes research. AANA J. 2015;83(3):189–95.
5. Murphy PJ. Measuring and recording outcome. Br J Anaesth. 2012;109(1):92–8.
6. Al G. Hands-on machine learning with scikit-learn, keras, and tensorflow concepts, tools, and techniques to build intelligent systems. 2nd ed. Sebastopoled: O'Reilly Media, Inc.; 2019.
7. Briganti G. On the use of bayesian artificial intelligence for hypothesis generation in psychiatry. Psychiatr Danub. 2022;34(8):201–6.
8. Iyer PU, Iyer KS. Research in pediatric cardiac anesthesia and intensive care in low- and middle- income countries and low resource settings: challenges and opportunities. Ann Pediatr Cardiol. 2021;14(3):356–8.
9. Cvetkovic M. Challenges in pediatric cardiac anesthesia in developing countries. Front Pediatr. 2018. https://doi.org/10.3389/fped.2018.00254.
10. Li B, Zhang R, Zhang M, Zheng J. Current anesthesia practices of pediatric cardiac surgeries in tertiary maternity and children's hospitals in China: a national survey. J Cardiothorac Vasc Anesth. 2023;37(7):1213–22.
11. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology. 2020;132(2):379–94.
12. Jeffries HE, Soto-Campos G, Katch A, Gall C, Rice TB, Wetzel R. Pediatric index of cardiac surgical intensive care mortality risk score for pediatric cardiac critical care*. Pediatr Crit Care Med. 2015. https://doi.org/10.1097/PCC.0000000000000489.
13. Kang AR, Lee J, Jung W, Lee M, Park SY, Woo J. Development of a prediction model for hypotension after induction of anesthesia using machine learning. PLoS ONE. 2020;15(4):e0231172.
14. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT press; 2016.
15. Despotis G, Avidan M, Eby C. Prediction and management of bleeding in cardiac surgery. J Thromb Haemost. 2009;7(1):111–7.
16. Jenkins KJ, Gauvreau K, Newburger JW, Spray TL, Moller JH, Iezzoni LI. Consensus-based method for risk adjustment for surgery for congenital heart disease. J Thorac Cardiovasc Surg. 2002;123(1):110–8.
17. Willems A, Van Lerberghe C, Gonsette K, De Villé A, Melot C, Hardy JF. The indication for perioperative red blood cell transfusions is a predictive risk factor for severe postoperative morbidity and mortality in children undergoing cardiac surgery. Eur J Cardiothorac Surg. 2014;45(6):1050–7.
18. Székely A, Sápi E, Király L, Szatmári A, Dinya E. Intraoperative and postoperative risk factors for prolonged mechanical ventilation after pediatric cardiac surgery. Paediatr Anaesth. 2006;16(11):1166–75.
19. Zappitelli M, Washburn KK, Arikan AA, Loftis L, Ma Q, Devarajan P, Parikh CR, Goldstein SL. Urine neutrophil gelatinase-associated lipocalin is an early marker of acute kidney injury in critically ill children: a prospective cohort study. Crit Care. 2007. https://doi.org/10.1186/cc6089.
20. GaD VR, Fred L. Python 3 reference manual. California: CreateSpace; 2009.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(85):2825–30.
22. Kubben P, Dumontier M, Dekker A. Fundamentals of clinical data science. Berlin: Springer Nature; 2019.
23. R Development Core Team. R: a language and environment for statistical computing. 362nd ed. Vienna: R Foundation for Statistical Computing; 2023.
24. McKinney W (2010) Data structures for statistical computing in python. In: McKinney W (ed) Proceedings of the 9th Python in Science Conference. Austin, TX, 2010
25. Price-Whelan AM, Sipőcz B, Günther H, Lim P, Crawford S, Conseil S. The astropy project: building an open-science project and status of the v2. 0 core package. Astron J. 2018. https://doi.org/10.3847/1538-3881/aabc4f.
26. Bengfort B, Bilbro R. Yellowbrick: visualizing the scikit-learn model selection process. J Open Sour Softw. 2019;4(35):1075.
27. Charles R, Harris K, Jarrod M, van der Walt SJ, Ralf G, Pauli V, David C, Eric W, Julian T, Sebastian B, Nathaniel JS, Robert K, Matti P, Stephan H, van Kerkwijk MH, Matthew B, Allan H, Mark W, Pearu P, Gérard-Marchant P, Kevin S, Tyler R, Warren W, Hameer A, Christoph G, Travis EO. Array programming with NumPy. Nature. 2020. https://doi.org/10.1038/s41586-020-2649-2.
28. Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA. Four principles of explainable artificial intelligence. Maryland: Gaithersburg; 2020.
29. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.
30. Zeng X, An J, Lin R, Dong C, Zheng A, Li J. Prediction of complications after paediatric cardiac surgery. Eur J Cardiothorac Surg. 2020;57(2):350–8.