DEVELOPMENT OF A MACHINE LEARNING ALGORITHM FOR MORTALITY PREDICTION IN PEDIATRIC PATIENTS HOSPITALIZED IN INTENSIVE CARE UNITS

MARIA ANGELICA BRAVO BRAVO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
MAESTRÍA EN INGENIERÍA ELECTRÓNICA
BUCARAMANGA
2026

DEVELOPMENT OF A MACHINE LEARNING ALGORITHM FOR MORTALITY PREDICTION IN PEDIATRIC PATIENTS HOSPITALIZED IN INTENSIVE CARE UNITS

MARIA ANGELICA BRAVO BRAVO

Thesis submitted in partial fulfillment of the requirements for the degree of Master in Electronic Engineering.

Advisor

Carlos A. Fajardo, Ph.D.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
MAESTRÍA EN INGENIERÍA ELECTRÓNICA
BUCARAMANGA

2026

CONTENT

LIST OF FIGURES

LIST OF TABLES

ANNEXES

RESUMEN

TÍTULO: DESARROLLO DE UN ALGORITMO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA MORTALIDAD EN PACIENTES PEDIÁTRICOS HOSPITALIZADOS EN UNIDADES DE CUIDADOS INTENSIVOS [*]

AUTOR: MARIA ANGELICA BRAVO BRAVO [**]

PALABRAS CLAVE: Desequilibrio de clases, variables clínicas, UCI, aprendizaje automático, odds ratio, mortalidad pediátrica, base de datos PIC, análisis SHAP.

DESCRIPCIÓN: La predicción de mortalidad en pacientes pediátricos críticos sigue siendo un reto en las Unidades de Cuidados Intensivos (UCI), donde una estimación precisa del riesgo es esencial para apoyar decisiones clínicas. Aunque sistemas como PRISM, PIM y PELOD se usan ampliamente, presentan limitaciones de generalización. Los modelos de aprendizaje automático ofrecen una alternativa flexible, pero su aplicación pediátrica está limitada por la escasez de bases públicas y la calidad de los datos; la base *"Pediatric Intensive Care"* (PIC) constituye el principal recurso abierto disponible.

Se realizó un análisis exploratorio y un OR univariado para evaluar la asociación de variables clínicas con la mortalidad, identificando al fosfato como el factor más asociado, junto con variables de coagulación, gases arteriales, equilibrio electrolítico y lactato. Luego, se entrenaron siete modelos en 4.272 pacientes de PIC, seleccionando LightGBM por su rendimiento, eficiencia y estabilidad. Para el desbalance de clases, se aplicaron estrategias de sobremuestreo, submuestreo y ajustes del modelo, destacando la pérdida focal por su mayor sensibilidad.

Finalmente, SHAP identificó como variables influyentes la estancia en UCI, lactato, hematocrito, pCO_2 máximos y frecuencia respiratoria media. Los resultados resaltan el potencial del aprendizaje automático, condicionado a datos más completos, representativos y estandarizados.

---

[*] Tesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Director: Carlos A. Fajardo, Ph.D.

ABSTRACT

TITLE: DEVELOPMENT OF A MACHINE LEARNING ALGORITHM FOR MORTALITY PREDICTION IN PEDIATRIC PATIENTS HOSPITALIZED IN INTENSIVE CARE UNITS [*]

AUTOR: MARIA ANGELICA BRAVO BRAVO [**]

KEYWORDS: Class imbalance, clinical variables, ICUs, machine learning, Odds Ratio, pediatric mortality, PIC database, SHAP analysis.

DESCRIPTION: Mortality prediction in critically ill pediatric patients remains a challenge in Intensive Care Units (ICUs), where accurate risk estimation is essential to support clinical decision-making. Although systems such as PRISM, PIM, and PELOD are widely used, they present limitations in generalizability. Machine learning models offer a flexible alternative, but their pediatric application is limited by the scarcity of public databases and data quality; the *"Pediatric Intensive Care"* (PIC) database constitutes the main open-access resource available.

An exploratory analysis and a univariate OR analysis were performed to assess the association between clinical variables and mortality, identifying phosphate as the most strongly associated factor, along with coagulation variables, arterial blood gases, electrolyte balance, and lactate. Then, seven models were trained on 4,272 PIC patients, selecting LightGBM for its performance, efficiency, and stability. To address class imbalance, oversampling, undersampling, and model-level adjustments were applied, with focal loss standing out for its higher sensitivity.

Finally, SHAP identified ICU length of stay, maximum lactate, maximum hematocrit, maximum pCO_2, and mean respiratory rate as influential variables. The results highlight the potential of machine learning, conditioned on more complete, representative, and standardized data.

---

[*] Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones. Advisor: Carlos A. Fajardo, Ph.D.

INTRODUCTION

Pediatric Mortality in Intensive Care Units

Pediatric mortality remains a major global health challenge, particularly among children admitted to Pediatric Intensive Care Units (PICUs). These specialized wards care for critically ill patients under 18 years of age who require continuous monitoring and advanced medical interventions to prevent death [1] [2]. Despite important advances in intensive care, mortality rates in PICUs remain substantial: approximately 41 % of pediatric ICU deaths occur in children under five years of age, and infant mortality continues to range between 5–20 % depending on region and clinical conditions [3] [4]. Children are admitted to PICUs for a wide range of critical conditions, including cardiovascular, renal, and respiratory diseases, as well as severe infections

---

[1] J. Qiu et al. «Comparison of the pediatric risk of mortality, pediatric index of mortality, and pediatric index of mortality 2 models in a pediatric intensive care unit in China: A validation study». En: *Medicine* 96.14 (2017), e6431. DOI: 10.1097/MD.0000000000006431.

[2] S. M. Altawalbeh et al. «Evaluating Intensive Care Unit Medication Charges in a Teaching Hospital in Jordan». En: *Expert Review of Pharmacoeconomics & Outcomes Research* (2019). DOI: 10.1080/14737167.2019.1571413.

[3] F Rosa-Mangeret et al. «2.5 million annual deaths—are neonates in low-and middle-income countries too small to be seen? A bottom-up overview on neonatal Morbi-mortality». En: *Trop. Med. Infect. Disease* 7.5 (2022), pág. 64.

[4] RM McAdams et al. «Predicting clinical outcomes using artificial intelligence and machine learning in neonatal intensive care units: a systematic review». En: *J Perinatol.* 42.12 (2022), págs. 1561-1575. DOI: 10.1038/s41372-022-01392-8.

such as sepsis [5] [6]. This persistent burden underscores the urgent need for accurate and timely risk prediction strategies to guide life-saving decisions in critically ill children. Mortality prediction tools play a central role in this context, providing rapid and objective risk assessments that support clinical decision-making. By enabling early identification of high-risk patients, these tools allow clinicians to prioritize interventions and optimize the allocation of scarce resources [7] [8].

Traditional Scoring Systems for Pediatric Mortality Prediction

In pediatrics, mortality prediction has traditionally relied on clinical scoring systems such as the Pediatric Risk of Mortality (PRISM-III/IV) [9], the Pediatric Index of Mortality (PIM-2/PIM-3) [10], and the Pediatric Logistic Organ Dysfunction score (PELOD-2) [11]. These scoring systems, based on physiological, laboratory, and clinical history

---

[5] M. Evans et al. «Development and validation of a pediatric model predicting trauma-related mortality». En: *BMC Pediatrics* 23 (2023). DOI: 10.1186/s12887-023-04437-9.

[6] P. Joshi et al. «Application of Pediatric Risk of Mortality (PRISM) III Score in Predicting Mortality Outcomes». En: *Journal of Nepal Health Research Council* 21 (2024), págs. 450-457.

[7] T. J. Pollard y L. A. Celi. «Enabling Machine Learning in Critical Care». En: *ICU management & practice* 17.3 (2017), págs. 198-199.

[8] C. Yue, C. Zhang y C. Ying. «A new nomogram for the individualized prediction of children's mortality risk in pediatric intensive care unit». En: *Am J Transl Res* 15.6 (2023), págs. 4172-4178.

[9] M. M. Pollack, K. M. Patel y U. E. Ruttimann. «PRISM III: an updated Pediatric Risk of Mortality score». En: *Critical Care Medicine* 24.5 (1996), págs. 743-752. DOI: 10.1097/00003246-199605000-00004.

[10] F Shann et al. «Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care». En: *Intensive Care Medicine* 23.2 (1997), págs. 201-7. DOI: 10.1007/s001340050317.

[11] S Leteurtre et al. «Development of a pediatric multiple organ dysfunction score: use of two strategies». En: *Medical Decision Making* 19.4 (1999), págs. 399-410. DOI: 10.1177/0272989X9901900408.

variables collected during the first hours of admission, have been widely adopted to classify patient severity and guide decision-making [12] [13].

Several studies have evaluated their performance across different settings. Shen and Jiang [12] reported pooled area under the curve (AUC) values of 0.82–0.84 and sensitivities around 0.75–0.78 for PRISM-III/IV, PIM-3, and PELOD-2 scores across more than 100,000 PICU admissions worldwide, and concluded that these scoring systems showed substantial heterogeneity and poor calibration across studies. Zhang et al. [14] compared PCIS, PRISM IV, and PELOD-2 scores in Chinese cohorts, obtaining AUCs of 0.74–0.80 and identifying determinants of mortality such as invasive ventilation and extremes of PaO_2 and pH. Similarly, L. Zhang et al. [15] found moderate discrimination for PRISM-III and PELOD-2 scores but variable calibration across Asian cohorts, emphasizing the need for recalibration and larger, more diverse datasets.

Despite their clinical utility, these scoring systems face important limitations. They rely on predefined sets of variables, require large amounts of baseline information, and often experience performance declines when applied outside their original de-

---

[12] Y Shen y J Jiang. «Meta-Analysis for the Prediction of Mortality Rates in a Pediatric Intensive Care Unit Using Different Scores: PRISM-III/IV, PIM-3, and PELOD-2». En: *Frontiers in Pediatrics* 9 (2021), pág. 712276. DOI: 10.3389/fped.2021.712276.

[13] L. J. Schlapbach et al. «Prognostic accuracy of age-adapted SOFA, SIRS, PELOD-2, and qSOFA for in-hospital mortality among children with suspected infection admitted to the intensive care unit». En: *Intensive Care Med* 44 (2018), págs. 179-188. DOI: 10.1007/s00134-017-5021-8.

[14] Z Zhang et al. «Performance of Three Mortality Prediction Scores and Evaluation of Important Determinants in Eight Pediatric Intensive Care Units in China». En: *Frontiers in Pediatrics* 8 (2020), pág. 522. DOI: 10.3389/fped.2020.00522.

[15] L Zhang et al. «Performance of PRISM III, PELOD-2, and P-MODS Scores in Two Pediatric Intensive Care Units in China». En: *Frontiers in Pediatrics* 9 (2021), pág. 626165. DOI: 10.3389/fped.2021.626165.

velopment cohorts [16] [17]. In addition, their rigid structure restricts the capacity to account for complex, non-linear relationships among predictors, which compromises generalizability and calibration across diverse populations [18] [19]. These challenges underscore the need for more flexible and adaptive approaches to pediatric mortality prediction.

Machine Learning Models for Predicting Mortality in Pediatric Patients

Machine learning (ML) has emerged as a powerful alternative to traditional scoring systems for pediatric mortality prediction. By leveraging the granularity of electronic health records (EHRs), ML models can analyze high-dimensional data, capture complex non-linear interactions among predictors, and adapt more flexibly to diverse clinical contexts [12] [14] [15] [20] [21]. These characteristics allow ML approaches to overcome many of the limitations of fixed-variable scoring systems, offering the potential for improved predictive power and generalizability.

---

[16] D. K. Richardson et al. «Score for neonatal acute physiology: A physiologic severity index for neonatal intensive care». En: *Pediatrics* 91.3 (1993), págs. 617-623.

[17] S Reid et al. «Comparing CRIB-II and SNAPPE-II as mortality predictors for very preterm infants». En: *Journal of Paediatrics Child Health* 51.5 (2015), págs. 524-528.

[18] H. Johnson et al. «Multiple Organ Dysfunction Syndrome and Pediatric Logistic Organ Dysfunction-2 Score in Pediatric Cerebral Malaria». En: *The American Journal of Tropical Medicine and Hygiene* 107.4 (2022), págs. 820-826. DOI: 10.4269/ajtmh.22-0140.

[19] S. Cameron et al. «Pediatric severe traumatic brain injury mortality prediction determined with machine learning-based modeling». En: *Injury* 53 (2022), págs. 992-998. DOI: 10.1016/j.injury.2022.01.008.

[20] S. Hong et al. «Predicting Risk of Mortality in Pediatric ICU Based on Ensemble Step-Wise Feature Selection». En: *Health Data Science* (2021), pág. 9365125. DOI: 10.34133/2021/9365125.

[21] J. Zhou et al. «Interpretable machine learning model for early prediction of disseminated intravascular coagulation in critically ill children». En: *Scientific Reports* 15 (2025), pág. 11217. DOI: 10.1038/s41598-025-91434-w.

Several studies have explored the application of ML models to pediatric mortality prediction. Lee et al. [22] developed a random forest model to predict PICU mortality within 72 hours of admission using private, multicenter data collected from four tertiary hospitals, achieving an AUC of 0.906 and outperforming the PIM-3 score (AUC of 0.845). Thorsen-Meyer et al. [23] developed a recurrent neural network for dynamic and explainable ICU mortality prediction, trained on high-frequency EHR data from four Danish hospitals and externally validated on data from a fifth institution. The authors reported AUCs ranging from 0.75 to 0.83 and concluded that their model consistently outperformed traditional risk scores.

More recently, Strutz et al. [24] trained ML models for early detection of critical events in hospitalized children using data from 135,621 admissions across three tertiary hospitals. Among the evaluated approaches, extreme gradient boosting (XGB) achieved the best performance, enabling a hospital-wide risk stratification framework applicable from general wards to ICUs.

Together, these studies highlight the growing potential of ML to enhance pediatric mortality prediction, not only by improving discriminative performance, but also by enabling dynamic, data-driven and individualized risk assessment. These approaches surpass static scoring systems by offering greater adaptability to diverse clinical populations, facilitating the development of more responsive and personalized decision-support tools in pediatric intensive care.

---

[22] B. Lee et al. «Development of a machine learning model for predicting pediatric mortality in the early stages of intensive care unit admission». En: *Sci Rep* 11.1 (2021), pág. 1263. DOI: 10.1038/s41598-020-80474-z.

[23] Hans-Christian Thorsen-Meyer y et al. «Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records». En: *The Lancet Digital Health* 2.4 (2020), e179-e191.

[24] S. Strutz et al. «Machine Learning for Predicting Critical Events Among Hospitalized Children». En: *JAMA Network Open* 8.5 (2025), e2513149. DOI: 10.1001/jamanetworkopen.2025.13149.

Pediatric Publicly Available ICU Databases

The development of robust ML models requires access to large, high-quality datasets. In adult critical care, public resources such as the Medical Information Mart for Intensive Care (MIMIC-III) [25] and the eICU Collaborative Research Database [26] have played a central role in advancing reproducible research. In pediatrics, however, publicly available ICU databases remain scarce, limiting opportunities for methodological innovation and large-scale validation.

To the best of our current knowledge, the Pediatric Intensive Care (PIC) database [27] represents the first openly accessible resource focused exclusively on critically ill children. Developed at the Children's Hospital of Zhejiang University, it includes detailed information for 13,449 ICU admissions from 12,881 patients collected between 2010 and 2018. The database covers a wide range of clinical domains, including vital signs, laboratory results, medications, interventions, fluid balance, diagnostic codes, and outcomes, thereby offering a rich foundation for both clinical and computational research.

Several studies have already demonstrated the utility of PIC database for advancing prognostic research. Morooka et al. [28] analyzed 1,505 PICU patients with acute kidney injury (AKI) and found that hyperlactatemia, low pH, and low bicarbonate were

---

[25] A. E. Johnson et al. «MIMIC-III, a freely accessible critical care database». En: *Scientific Data* 3 (2016), pág. 160035. DOI: 10.1038/sdata.2016.35.

[26] T. J. Pollard et al. «The eICU Collaborative Research Database, a freely available multi-center database for critical care research». En: *Scientific Data* 5 (2018), pág. 180178. DOI: 10.1038/sdata.2018.178.

[27] Xusheng Zeng et al. «PIC, a paediatric-specific intensive care database». En: *Scientific Data* 7.1 (2020), pág. 14. DOI: 10.1038/s41597-020-0355-4.

[28] H Morooka et al. «Prognostic Impact of Parameters of Metabolic Acidosis in Critically Ill Children with Acute Kidney Injury: A Retrospective Observational Analysis Using the PIC Database». En: *Diagnostics* 10.11 (2020), pág. 937. DOI: 10.3390/diagnostics10110937.

independently associated with 28-day mortality, underscoring the prognostic importance of metabolic acidosis parameters. Lu et al. [29] analyzed 5,114 PICU admissions to investigate the prognostic value of the Systemic Inflammatory Response Index (SIRI), a hematological marker derived from neutrophil, lymphocyte, and platelet counts. Their results showed that elevated SIRI levels were independently associated with both AKI and in-hospital mortality, suggesting that systemic inflammation plays a critical role in adverse outcomes among critically ill children. Based on these findings, the authors developed a nomogram that demonstrated good calibration and clinical utility for early risk stratification.

More recently, researchers have applied ML methods to the PIC database to enhance mortality prediction and feature discovery. Hong et al. [20] developed a logistic regression model to predict mortality in pediatric ICU patients, using 11 predictors selected from 397 variables, achieving an AUC of 0.753, compared to 0.690 for PRISM III reconstructed within the same cohort, while retaining interpretability. Prithula et al. [30] focused on respiratory admissions (1,188 patients), testing several ML algorithms. Their CatBoost model initially achieved an AUC of 0.722, but after implementing a novel data subdivision strategy to address class imbalance, performance improved to 0.852, highlighting the importance of data-level methods in pediatric mortality prediction.

---

[29] D. Lu et al. «Prognostic value of systemic inflammatory response index for acute kidney injury and the prognosis of pediatric patients in critical care units». En: *PLOS ONE* 19.8 (2024), e0306884. DOI: 10.1371/journal.pone.0306884.

[30] J. Prithula et al. «Improved pediatric ICU mortality prediction for respiratory diseases: machine learning and data subdivision insights». En: *Respiratory Research* 25.1 (2024), pág. 216. DOI: 10.1186/s12931-024-02753-x.

Class Imbalance Issue in Clinical Data

A central challenge when working with the PIC database, and with clinical data in general, is the issue of class imbalance [31]. Mortality, the outcome of greatest clinical relevance, occurs far less frequently than survival and therefore constitutes the minority class [32] [33], representing approximately 5 % of the total cohort. This imbalance causes most ML models to bias predictions towards the majority class, leading to poor sensitivity and an excess of false negatives. In medical contexts, this results in missing high-risk patients who could otherwise benefit from timely interventions [34]. Therefore, addressing class imbalance is not only a methodological concern, but also a clinical necessity for building reliable mortality prediction models in pediatric intensive care [34] [35].

Several strategies have been proposed to mitigate this issue. Data-level approaches, such as oversampling the minority class or undersampling the majority class,

---

[31] Vinod Kumar et al. «Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques». En: *Healthcare* 10.7 (2022). DOI: 10.3390/healthcare10071293.

[32] Nazim Uddin Niaz, K.M. Nadim Shahariar y Muhammed J. A. Patwary. «Class Imbalance Problems in Machine Learning: A Review of Methods And Future Challenges». En: *Proceedings of the 2nd International Conference on Computing Advancements*. ICCA '22. Dhaka, Bangladesh: Association for Computing Machinery, 2022, 485–490. DOI: 10.1145/3542954.3543024.

[33] Seifollah Gholampour. «Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable». En: *Machine Learning and Knowledge Extraction* 6.2 (2024), págs. 827-841. DOI: 10.3390/make6020039.

[34] Guo Haixiang et al. «Learning from class-imbalanced data: Review of methods and applications». En: *Expert systems with applications* 73 (2017), págs. 220-239.

[35] Tsung-Yi Lin et al. «Focal loss for dense object detection». En: *Proceedings of the IEEE international conference on computer vision*. 2017, págs. 2980-2988.

aim to rebalance the training distribution [36] [37]. Model-level strategies, including cost-sensitive learning and class weighting, adjust the learning process to penalize misclassifications of the minority class more heavily [38] [39] [40]. More recently, loss function adaptations—such as focal loss—have been introduced to down-weight easy-to-classify examples and focus the model on harder cases, demonstrating strong performance in medical applications with imbalanced data [35].

Among data-level techniques, the Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely used methods. Sáez et al. [41] applied it to clinical datasets, while Beckmann et al. [42] and Yu et al. [43] explored random undersampling in medical contexts. Hybrid strategies that combine both approaches have also been

---

[36] Nitesh V. Chawla et al. «SMOTE: Synthetic Minority Over-sampling Technique». En: *Journal of Artificial Intelligence Research* 16 (2002), págs. 321-357. DOI: 10.1613/jair.953.

[37] Cian Lin, Chih-Fong Tsai y Wei-Chao Lin. «Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study». En: *Artificial Intelligence Review* 56.2 (2023), págs. 845-863.

[38] Benjamin X Wang y Nathalie Japkowicz. «Boosting support vector machines for imbalanced data sets». En: *Knowledge and information systems* 25.1 (2010), págs. 1-20.

[39] H. He y E. A. Garcia. «Learning from Imbalanced Data». En: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), págs. 1263-1284. DOI: 10.1109/TKDE.2008.239.

[40] Sara Belarouci y Mohammed Amine Chikh. «Medical imbalanced data classification». En: *Advances in Science, Technology and Engineering Systems Journal* 2.3 (2017), págs. 116-124.

[41] J. A. Sáez, B. Krawczyk y M. Woźniak. «Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets». En: *Pattern Recognition* 57 (2016), págs. 164-178. DOI: 10.1016/j.patcog.2016.03.003.

[42] M. Beckmann, N.F.F. Ebecken y B.S.P. de Lima. «A KNN Undersampling Approach for Data Balancing». En: *Journal of Intelligent Learning Systems and Applications* 7.4 (2015), págs. 104-116. DOI: 10.4236/jilsa.2015.74010.

[43] H. Yu, J. Ni y J. Zhao. «ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data». En: *Neurocomputing* 101 (2013), págs. 309-318.

developed [44]. Beyond resampling, more advanced model-level adjustments—such as cost-sensitive learning, class weighting, and focal loss—have been shown to improve classifier performance by emphasizing hard-to-classify cases.

Further advances include hierarchical classification frameworks, such as the one proposed by Hosenie et al. [45], which combine SMOTE-based augmentation with multi-level classification. Other studies have experimented with hybrid approaches that integrate resampling methods with algorithmic modifications to address the challenges posed by imbalanced datasets [46] [47] [48] [49].

Despite these methodological advances, their application in pediatric critical care remains scarce. In particular, open databases such as PIC have seen limited exploration of imbalance-aware modeling for mortality prediction, underscoring the need for further research tailored to this domain.

---

[44] R. Alejo et al. «A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios». En: *Pattern Recognition Letters* 34.4 (2013), págs. 380-388. DOI: 10.1016/j.patrec.2012.11.002.

[45] Z. Hosenie et al. «Imbalance learning for variable star classification». En: *Monthly Notices of the Royal Astronomical Society* 493.4 (2020), págs. 6050-6059. DOI: 10.1093/mnras/staa776.

[46] D. Devi, S. K. Biswas y B. Purkayastha. «Correlation-based oversampling aided cost sensitive ensemble learning technique for treatment of class imbalance». En: *Journal of Experimental & Theoretical Artificial Intelligence* 34.2 (2022), págs. 143-174. DOI: 10.1080/0952813X.2020.1856110.

[47] M. Z. Abedin et al. «Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk». En: *Complex & Intelligent Systems* 9.5 (2023), págs. 3559-3579. DOI: 10.1007/s40747-023-01041-3.

[48] S. Kaisar y A. Chowdhury. «Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests». En: *ICT Express* 8.4 (2022), págs. 563-568. DOI: 10.1016/j.icte.2022.05.002.

[49] T. T. Khuat y M. H. Le. «Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems». En: *SN Computer Science* 1.2 (2020), pág. 108. DOI: 10.1007/s42979-020-00109-x.

1. OBJECTIVES

1.1. GENERAL OBJECTIVE

To develop a predictive algorithm based on machine learning techniques for mortality prediction in pediatric patients hospitalized in Intensive Care Units, using clinical variables.

1.2. SPECIFIC OBJECTIVES

To establish well-defined inclusion criteria for a pediatric ICU cohort, ensuring data relevance, minimizing bias, and avoiding cases that could skew the performance and clinical applicability of a machine learning model for mortality prediction.

To preprocess the selected medical database by cleaning the data, handling missing values, and transforming features to ensure its suitability for training the machine learning model.

To develop an optimized machine learning model capable of accurately predicting ICU mortality in pediatric patients using preprocessed clinical data.

To evaluate the performance of the developed machine learning model on pediatric patient data from the selected medical database using standard performance measures.

2. DATA AND VARIABLE DESCRIPTION, STATISTICAL AND ASSOCIATION ANALYSIS

2.1. DATA DESCRIPTION

The dataset analyzed in this study was obtained from the publicly available Pediatric Intensive Care (PIC) database [27], which contains de-identified clinical records from ICU admissions at the Children's Hospital of Zhejiang University School of Medicine between 2010 and 2018. It comprises 13,499 hospital admissions from 12,881 unique pediatric patients (aged 0–18 years) and includes more than 1,298 variables spanning vital signs, laboratory results, medications, fluid balance, diagnostic codes, and survival outcomes. All identifiers were anonymized according to the Health Insurance Portability and Accountability Act (HIPAA), and data access requires registration, research training with human subjects, and a signed data use agreement to ensure ethical use.

Cohort-level statistics reported in PIC database show that the mean patient age was 2.5 years (Q1–Q3: 0.1–3.3), 57.5 % were male, and overall in-hospital mortality was 7.1 %. The mean hospital stay was 17.6 days (Q1–Q3: 7.0–21.0), with a mean ICU stay of 9.3 days (Q1–Q3: 0.9–9.2). The most frequent discharge diagnoses were congenital malformations, perinatal conditions, and respiratory diseases. These distributions differ substantially from those observed in adult critical care, underscoring the value of a dedicated pediatric database.

The PIC database is structured into 16 interlinked tables, with 'PATIENTS', 'ADMISSIONS', and 'ICUSTAYS' serving as the core sources for demographic, admission, and ICU stay information. Data are distributed as comma-separated value (CSV) files, facilitating integration into relational database systems. Mortality outcomes were derived from the "hospital_expire_flag" field, which indicates in-hospital death based

on EHR data.

2.2. VARIABLE SELECTION AND INCLUSION CRITERIA

Candidate predictors were first identified through an extensive literature review and consultation with a clinical expert. Subsequently, an exploratory analysis assessed their availability, completeness, measurement units, value ranges, and coding consistency within the dataset. The variable selection process was then refined through iterative clinical consultation to retain only those predictors that were both clinically meaningful and methodologically feasible.

The study population was defined according to three main inclusion criteria. First, patients aged between 1 month and 17 years were included, while neonates were excluded to ensure comparability with previous pediatric mortality studies and to avoid confounding clinical differences between neonatal and pediatric intensive care. Neonates typically present distinct physiological characteristics, disease patterns, and mortality risk factors that require specialized management and are often studied within separate neonatal ICU (NICU) frameworks [21] [50] [51]. Second, only the first ICU admission per patient was retained to prevent multiple records from the same individual being treated as independent observations [20] [21] [52]. Third, only data from the first 24 hours of ICU admission were considered, consistent with traditional scoring sys-

---

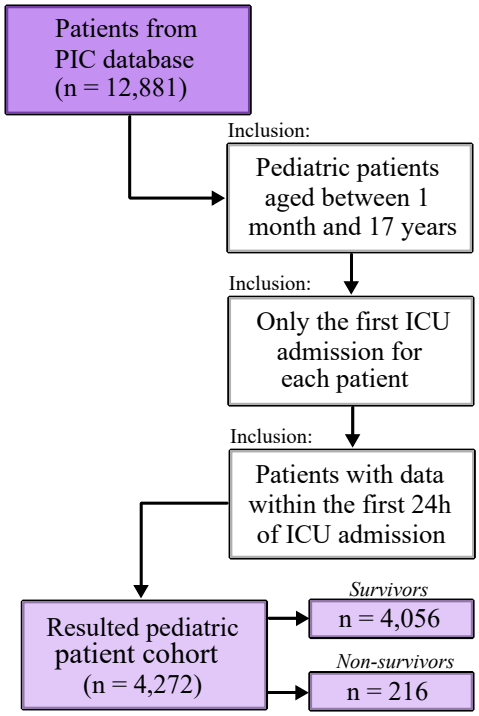
[50] A. Thukral et al. «Performance of Pediatric Risk of Mortality (PRISM), Pediatric Index of Mortality (PIM), and PIM2 in a pediatric intensive care unit in a developing country». En: *Pediatric Critical Care Medicine* 7.4 (2006), págs. 356-361. DOI: 10.1097/01.PCC.0000227105.20897.89.

[51] E. Botan et al. «Characteristics and timing of mortality in children dying in pediatric intensive care: a 5-year experience». En: *Acute and Critical Care* 37.4 (2022), págs. 644-653. DOI: 10.4266/acc.2022.00395.

[52] G. Kong, K. Lin e Y. Hu. «Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU». En: *BMC Medical Informatics and Decision Making* 20.251 (2020). DOI: 10.1186/s12911-020-01271-2.

tems such as PRISM III [9] and with recent ML studies that emphasize early prediction [20]. After applying these inclusion criteria, the final cohort consisted of 4,272 pediatric patients, of whom 4,056 survived and 216 died during their ICU stay (mortality rate: 5.06 %). Figure 1 illustrates the selection process of pediatric patients included in this study.

Figure 1. Inclusion criteria applied in this study to select pediatric patients extracted from the PIC database.



Subsequently, variable-level preprocessing was conducted. Clinical variables with more than 60 % missing values were excluded; however, this procedure affected only the feature set, not the patient cohort. In other words, all previously selected patients were retained, while only the variables exceeding the missingness threshold were removed. This criterion was defined based on exploratory data analysis and supported by previous studies, as imputing variables with over half of their observations missing can introduce substantial bias and reduce model reliability. The remaining

variables were imputed using *miceforest* [53], a LightGBM-based multiple imputation method capable of capturing nonlinear relationships among features.

The final dataset included 50 clinically meaningful candidate predictors distributed in 45 laboratory variables, 3 vital-sign variables, and 2 demographic variables. These variables were selected after excluding those with excessive missingness and by prioritizing features with established clinical relevance to pediatric critical care, as identified through literature review and expert consultation.

For vital signs, minimum, mean, and maximum values were computed; for laboratory variables, minimum and maximum values were retained; and for demographics, a single representative value was used. This feature engineering process expanded the dataset to 101 clinical variables, which served as inputs for subsequent analyses. Table 1 shows the feature engineering process applied to the 50 selected clinical predictors identified in this study. The complete list of all variables derived after feature engineering is presented in Annex A, as an extended version of Table 1.

---

[53] The miceforest Development Team. *miceforest: Fast, Memory Efficient Imputation with LightGBM*. Versión 6.0.3. 2024.

Table 1. Feature engineering process applied to clinical variables.

Variable category	Extracted value	Total of variables	Total of features extracted
Vital signs	Minimum, mean, and maximum values computed for each measurement.	3	9
Laboratory tests	Minimum and maximum values retained for each test.	45	90
Demographics	Single representative value retained.	2	2
Total			**101**

2.3. STATISTICAL AND ASSOCIATION ANALYSIS USING ODDS RATIO

The distribution of all clinical variables was first assessed using the Shapiro–Wilk test [54], which is well-suited for small to moderate sample sizes ($n < 5000$). Results indicated that none of the variables followed a normal distribution, leading us to employ non-parametric statistical tests. Specifically, the Mann–Whitney U test [55] was

[54] S. S. Shapiro y M. B. Wilk. «An Analysis of Variance Test for Normality (Complete Samples)». En: *Biometrika* 52.3-4 (1965), págs. 591-611. DOI: 10.2307/2333709.

[55] Henry B. Mann y Donald R. Whitney. «On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other». En: *The Annals of Mathematical Statistics* 18.1 (1947), págs. 50-60. DOI: 10.1214/aoms/1177730491.

applied to numerical variables, and the Chi-square test [56] to categorical variables, to identify differences between survivors and non-survivors.

Of the 101 clinical variables examined, 63 showed statistically significant differences between survivors and non-survivors, while 38 did not exhibit such differences in this analysis. Variables with significant group-level differences were considered potential candidates for mortality prediction in PICU patients. To further explore the strength and direction of these associations, and to complement the analysis, a univariate odds ratio (OR) analysis was subsequently conducted using logistic regression to identify variables independently associated with death risk.

Each model incorporated a single standardized predictor, and ORs were reported together with their 95 % confidence intervals. Standardization was necessary because the clinical variables were originally expressed in different units and showed markedly different numerical ranges. As a result, each OR represents the change in the odds of mortality associated with a one-standard-deviation increase in the predictor, which facilitates comparison of the relative strength of the associations across variables without altering their direction. This analysis provided an initial epidemiological perspective and a quantitative view of individual risk factors associated with PICU mortality. Moreover, it established a baseline for comparison with the interpretability-based ML analysis conducted in later stages of the study.

Results from the univariate OR analysis of 101 clinical variables suggest that approximately 20 variables show the strongest associations with pediatric mortality in the PIC dataset. Table 2 summarizes these variables, ranked in descending order according to their OR values.

Phosphate levels—particularly *max_phosphate* and *min_phosphate*—emerged as the variables with the strongest associations, underscoring their potential role as

---

[56] Kent State University Libraries. *SPSS Tutorials: Chi-Square Test of Independence*. Accessed: October 31, 2025. 2025. URL: https://libguides.library.kent.edu/spss/chisquare.

Table 2. Results of the 20 clinical variables most strongly associated with pediatric mortality, along with their respective OR values and 95 % CIs.

Clinical variable	OR value	95 % CI
max_phosphate	2.681	(2.099 - 3.424)
min_phospate	2.474	(1.938 - 3.159)
min_inr_pt	1.989	(1.492 - 2.652)
min_carboxyhemmoglobin	1.624	(1.212 - 2.175)
max_potassium	1.573	(1.340 - 1.846)
max_inr_pt	1.537	(1.318 - 1.792)
min_lactate	1.456	(1.341 - 1.580)
min_redbloodcells	1.332	(1.096 - 1.618)
max_redbloodcells	1.315	(1.074 - 1.610)
max_lactate	1.212	(1.171 - 1.255)
min_absolutelymphocytecount	1.127	(1.075 - 1.181)
max_absolutelymphocytecount	1.122	(1.075 - 1.171)
min_aniongap	1.100	(1.067 - 1.134)
min_pt	1.069	(1.042 - 1.097)
max_glucose	1.060	(1.036 - 1.084)
max_aniongap	1.056	(1.035 - 1.078)
max_hematocrit	1.044	(1.020 - 1.068)
resp_rate_mean	1.043	(1.030 - 1.057)
min_urea	1.043	(1.018 - 1.069)
max_urea	1.041	(1.017 - 1.064)

Note. OR: odds ratio; CI: confidence interval.

key prognostic markers. Coagulation function also proved important, with *min_inr_pt* (reflecting the combined activity of multiple clotting factors) showing a strong association with mortality. Similarly, *min_carboxyhemoglobin*, a marker derived from blood gas monitoring, was identified as relevant. Electrolyte balance stood out as well: *max_potassium*, corresponding to the peak serum potassium level, is critical for cardiac, neural, and muscular function, and its dysregulation is known to increase the risk of severe outcomes.

Lactate was another prominent marker, with both minimum and maximum values (*min_lactate* and *max_lactate*) strongly associated with mortality. Elevated lactate is a well-established early indicator of tissue hypoperfusion and sepsis. Hematolo-

gical variables also contributed significantly, as reflected in *min_redbloodcells* and *max_redbloodcells*, reinforcing the relevance of blood cell status in critical illness outcomes.

Other variables with smaller effect sizes but statistically consistent and clinically meaningful associations included lymphocyte counts (*min_absolutelymphocytecount* and *max_absolutelymphocytecount*), acid–base balance (*aniongap*), coagulation (*pt*), metabolic markers (*glucose* and *urea*), and hematological status (*hematocrit*). In particular, almost all of the top predictors were laboratory-based measurements, emphasizing the dominant role of biochemical and hematological markers. The only vital sign that appeared among the top 20 was the mean respiratory rate (*resp_rate_mean*), suggesting that while vital signs may contribute, their relative weight is smaller compared to laboratory features.

Overall, these findings indicate that within the PIC dataset, univariate OR analysis prioritizes laboratory biomarkers as the strongest mortality predictors, while vital signs play a comparatively minor role—partly reflecting the data composition, as the laboratory feature set ($\approx$ 90 variables) is substantially larger than the group of vital-sign variables (only 9 in total).

These results provided a preliminary understanding of the individual associations between clinical variables and pediatric mortality. However, given that univariate analyses cannot capture the combined or nonlinear effects among predictors, subsequent modeling with ML algorithms was conducted to explore multivariable interactions and enhance predictive performance.

3. MACHINE LEARNING MODELING AND CLASS IMBALANCE MITIGATION

3.1. EVALUATION OF MACHINE LEARNING MODELS FOR PEDIATRIC MORTALITY PREDICTION

Following the univariate association analysis based on ORs, predictive modeling using ML classifiers was performed to provide a complementary and more comprehensive evaluation of mortality risk. The models evaluated were Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Multi-Layer Perceptron (MLP), Gradient Boosting Machine (GBM), XGBoost (XGB), and LightGBM (LGBM).

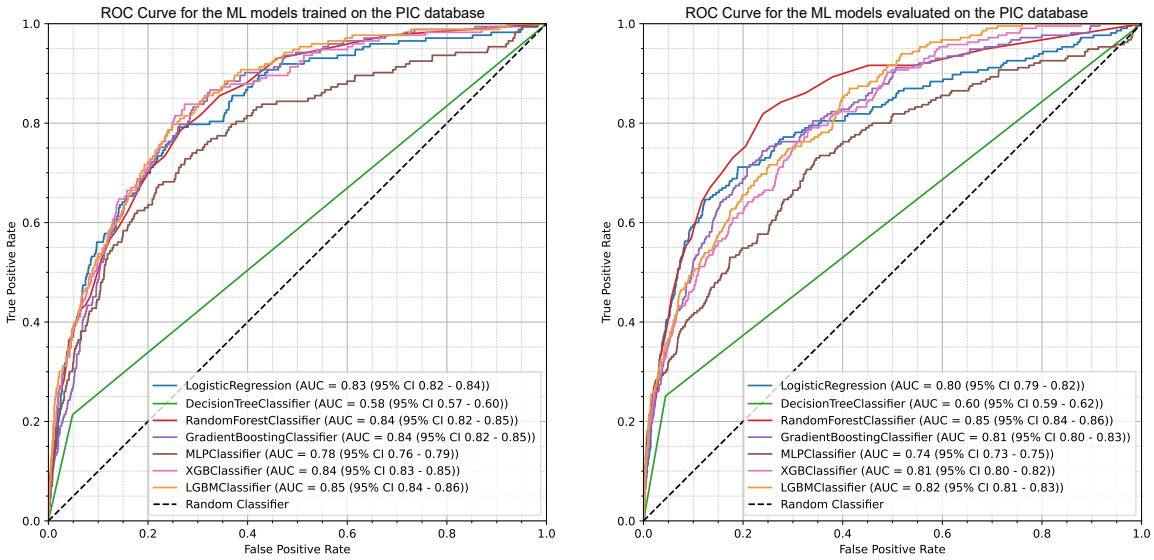
The PIC dataset was first divided into a training set (80 %) and an independent hold-out test set (20 %). Model development was conducted exclusively on the training set using 5-fold cross-validation, while the hold-out test set was reserved for final performance evaluation. Model performance was assessed using standard evaluation metrics, including sensitivity, specificity, F1-score, and area under the curve (AUC).

Figure 2 summarizes the predictive performance of the seven models in the training (Figure 3(a)) and testing (Figure 3(b)) sets in terms of AUC. The results suggest that RF, and ensemble boosting methods (GBM, XGB, LGBM) achieved the highest AUC values. Notably, LR, despite its simplicity, demonstrated performance comparable to that of more complex ML algorithms.

When comparing performance on both training and test sets, RF and LGBM consistently outperformed the other ML models. RF achieved the highest AUC with a value of 0.8459 (95 % CI: 0.8351 - 0.8568), but with the lowest sensitivity, whereas LGBM reached an AUC of 0.8211 (95 % CI: 0.8096 - 0.8326) with a more balanced trade-off.

Table 3 summarizes the results of all seven models across multiple evaluation metrics. These results highlight that GBM and XGB also performed strongly. Although GBM showed slightly higher sensitivity and F1-score, LGBM was selected for further

Figure 2. Predictive performance (AUC) of seven machine learning models for mortality prediction in PICU patients using the PIC database.



(a) Results of predictive performance in terms of AUC of the seven ML models for the training set.

(b) Results of predictive performance in terms of AUC of the seven ML models for the test set.

analysis because it achieved the highest AUC among the evaluated boosting-based models, indicating a slightly better overall discriminative capacity. In addition, LGBM offered lower computational cost and greater flexibility for hyperparameter optimization and for incorporating additional strategies in subsequent stages.

The LGBM model was subsequently optimized using Bayesian hyperparameter tuning with Optuna [57], adjusting parameters such as *number of iterations*, *learning rate*, *number of leaves*, *depth*, *minimum data in leaf*, *bagging fraction*, *lambda_l1*, and *lambda_l2*, following search ranges commonly recommended in the ML literature. After optimization, the decision threshold was recalibrated to maximize the F1-score, improving the balance between sensitivity and precision.

---

[57] Takuya Akiba et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Available at https://optuna.org/. 2019.

Table 3. Performance metrics results of the seven ML models evaluated on PIC dataset.

Model	Sens	Spec	F1	AUC (95 % CI)
LR	11 %	99 %	17 %	0.8038 95 % CI (0.7918 - 0.8157)
DT	25 %	95 %	24 %	0.6034 95 % CI (0.5887 - 0.6181)
RF	4 %	99 %	7 %	0.8459 95 % CI (0.8351 - 0.8568)
GBM	10 %	99 %	17 %	0.8145 95 % CI (0.8028 - 0.8261)
MLP	22 %	99 %	29 %	0.7399 95 % CI (0.7267 - 0.7530)
XGB	9 %	99 %	15 %	0.8069 95 % CI (0.7951 - 0.8188)
LGBM	6 %	99 %	11 %	0.8211 95 % CI (0.8096 - 0.8326)

Note. Sens: sensitivity; Spec: specificity; F1: F1-score.

Table 4 shows the results obtained for the performance metrics evaluated using the new threshold by the optimized LGBM model.

Table 4. Performance metrics results of the optimized LGBM model evaluated on PIC dataset with a decision threshold that maximizes the F1-score.

Model	Sens	Spec	F1	AUC (95 % CI)
LGBM	25 %	99 %	33 %	0.8371 95 % CI (0.826 - 0.8482)

Note. Sens: sensitivity; Spec: specificity; F1: F1-score.

Although this adjustment improved overall model performance, sensitivity remained relatively low. This limitation highlights the persistent challenge of reducing false negatives in mortality prediction and motivated the implementation of imbalance-handling strategies described in the following sections.

3.2. CLASS IMBALANCE HANDLING

As previously mentioned, the PIC dataset exhibits a significant class imbalance, with approximately 95 % of pediatric patients being survivors and just 5 % being non-survivors. This imbalance has an adverse effect on model performance, particularly with regard to sensitivity and the F1-score. To address this limitation, imbalance-handling strategies were systematically evaluated at both the data and model levels.

Experiments were conducted within a five-fold cross-validation framework to ensure robustness and minimize bias.

At the data level, oversampling and undersampling techniques were explored to rebalance the dataset. Oversampling was performed using SMOTE [36], which generates synthetic minority-class samples through interpolation, and Borderline-SMOTE [58], a variant that focuses on minority cases located near the decision boundary, where misclassifications are more frequent. In parallel, Random Under-Sampling [39] was tested to reduce the dominance of the majority class by randomly discarding instances.

At the model level, class weighting schemes were implemented to increase the penalty associated with misclassifying minority cases. Three weighting strategies were compared: (i) a “raw” scheme, which assigns proportional penalties based on class frequency; (ii) a “balanced” scheme, which normalizes these penalties across classes; and (iii) “LGBM_balanced”, which leverages LightGBM’s internal class weighting. The terms “raw”, “balanced”, and “LGBM_balanced” are descriptive labels adopted in this study for differentiation purposes.

Additionally, focal loss [35] was evaluated as an alternative loss function that dynamically down-weights easy examples to emphasize harder, clinically relevant cases.

The optimized LGBM model served as the reference model for all evaluations, allowing the comparison of data-level and model-level imbalance-handling approaches under a consistent learning framework.

3.2.1. SMOTE and Borderline-SMOTE Techniques for Oversampling Given the extreme imbalance in the PIC dataset (approximately one non-survivor for every

---

[58] H. Han, W. Y. Wang y B. H. Mao. «Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning». En: *Advances in Intelligent Computing (ICIC 2005)*. Ed. por D. S. Huang, X. P. Zhang y G. B. Huang. Vol. 3644. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, págs. 878-887. DOI: 10.1007/11538059_91.

19 survivors), oversampling techniques were applied only to the training subsets within each fold of the five-fold cross-validation procedure, ensuring that synthetic data generation did not influence model evaluation. Both SMOTE and Borderline-SMOTE were evaluated with different resampling ratios (1:9, 1:7, 1:5, 1:3, and 1:1). Since correctly identifying non-survivors is clinically more relevant, sensitivity was prioritized as the main evaluation metric. Tables 5 and 6 present the detailed results of these experiments.

Table 5 summarizes the results for SMOTE. The best performance was achieved with a 1:9 ratio, improving sensitivity to 32 % (vs. 25 % in the baseline optimized LGBM model) while maintaining high specificity (96 %). The 1:5 ratio produced the highest AUC (0.8224, 95 % CI: 0.8109–0.8338), whereas a fully balanced scenario (1:1) increased sensitivity but reduced specificity and F1-score.

Table 5. Performance metrics of the optimized LGBM model applying SMOTE oversampling technique with different balance ratios, compared to the baseline model (without oversampling) evaluated on the PIC dataset.

Ratio	Sens (%)	Spec (%)	F1 (%)	AUC (95 % CI)
Baseline (no SMOTE)	25	99	33	0.8371 (0.8260 - 0.8482)
1:9	**32**	**96**	**31**	**0.8214 (0.8099 - 0.8329)**
1:7	27	97	29	0.8172 (0.8056 - 0.8288)
1:5	24	97	26	0.8224 (0.8109 - 0.8338)
1:3	27	96	27	0.8161 (0.8045 - 0.8278)
1:1	32	93	24	0.8109 (0.7991 - 0.8226)

Note. “Baseline” refers to the optimized LGBM model without any data resampling. Sens: sensitivity; Spec: specificity; F1: F1-score.

Table 6 summarizes the results for Borderline-SMOTE. These results shows that Borderline-SMOTE achieved its best sensitivity with the 1:7 ratio, reaching 33 %, a slight improvement over standard SMOTE (32 %). This gain is consistent with the method’s design, which emphasizes generating synthetic samples near the decision boundary where misclassifications are more likely. In terms of overall discrimination, the 1:9 ratio obtained the highest AUC (0.8435, 95 % CI: 0.8326–0.8544), confirming

that Borderline-SMOTE can enhance sensitivity without compromising AUC.

Table 6. Performance metrics of the optimized LGBM model applying Borderline-SMOTE oversampling technique with different balance ratios, compared to the baseline model (without oversampling) evaluated on the PIC dataset.

Ratio	Sens (%)	Spec (%)	F1 (%)	AUC (95% CI)
Baseline (no Borderline-SMOTE)	25	99	33	0.8371 (0.8260 - 0.8482)
1:9	27	97	31	0.8435 (0.8326 - 0.8544)
1:7	**33**	**95**	**29**	**0.8386 (0.8276 - 0.8496)**
1:5	25	98	29	0.8369 (0.8259 - 0.8480)
1:3	26	96	27	0.8387 (0.8277 - 0.8497)
1:1	29	95	26	0.8269 (0.8155 - 0.8382)

Note. "Baseline" refers to the optimized LGBM model without any data resampling. Sens: sensitivity; Spec: specificity; F1: F1-score.

Overall, both methods showed that moderate ratios (1:7 and 1:9) provided a more stable trade-off between sensitivity, specificity, and AUC compared to more aggressive balancing.

3.2.2. Random Under-Sampling In contrast to oversampling, Random Under-Sampling balances the dataset by reducing majority-class samples, which can, however, lead to information loss. The same resampling ratios were applied as in oversampling. Table 7 presents the results obtained with the RandomUnderSampler technique for this subsampling approach.

Table 7. Performance metrics of the optimized LGBM model applying RandomUnderSampler technique with different balance ratios, compared to the baseline model (without undersampling) evaluated on the PIC dataset.

Ratio	Sens (%)	Spec (%)	F1 (%)	AUC (95% CI)
Baseline (no UnderSampler)	25	99	33	0.8371 (0.8260 - 0.8482)
1:9	33	97	34	0.8326 (0.8214 - 0.8437)
1:7	31	97	32	0.8314 (0.8202 - 0.8426)
1:5	33	96	33	0.8340 (0.8229 - 0.8452)
1:3	**37**	**94**	**31**	**0.8303 (0.8191 - 0.8416)**
1:1	32	95	29	0.7934 (0.7813 - 0.8055)

Note. "Baseline" refers to the optimized LGBM model without any data resampling. Sens: sensitivity; Spec: specificity; F1: F1-score.

As shown in Table 7, the 1:3 ratio achieved the best sensitivity (37 %), representing a substantial improvement over the baseline optimized LGBM model and outperforming the oversampling approaches (see Table 5 and Table 6).

In terms of overall discrimination, the highest AUC was observed with the 1:5 ratio (0.8340, 95 % CI: 0.8229–0.8452), while performance declined under the fully balanced scenario (1:1).

To enable a direct comparison among the three data-level imbalance handling strategies — SMOTE, Borderline-SMOTE, and Random UnderSampling — Table 8 summarizes their best-performing configurations alongside the baseline optimized LGBM model. This consolidated view highlights how each method impacts sensitivity, specificity, and overall discrimination (AUC).

Table 8. Comparison of the best-performing configurations of data-level imbalance handling techniques (SMOTE, Borderline-SMOTE, and Random UnderSampling) against the baseline optimized LGBM model without resampling, evaluated on the PIC dataset.

Method	Best Ratio	Sens (%)	Spec (%)	F1 (%)	AUC (95 % CI)
Baseline (no resampling)	—	25	99	33	0.8371 (0.8260 – 0.8482)
SMOTE	1:9	32	96	31	0.8214 (0.8099 – 0.8329)
Borderline-SMOTE	1:7	33	95	29	0.8386 (0.8276 – 0.8496)
Random UnderSampling	1:3	37	94	31	0.8303 (0.8191 – 0.8416)

Note. Sens: sensitivity; Spec: specificity; F1: F1-score. "Best ratio" refers to the class distribution yielding the optimal sensitivity for each method.

3.2.3. Model-Level Class Weighting In addition to data-level resampling, class weighting was explored as a model-level alternative to address class imbalance. This technique assigns higher penalties to minority-class misclassifications during training, guiding the model to better recognize misclassified non-survivors. These experiments were conducted using the baseline optimized LGBM model without any prior data-level resampling, ensuring that the observed effects could be attributed solely to the weighting strategies.

Three weighting schemes were tested, formally defined as follows:

1. Raw weighting ("raw"):

$$w_i = \frac{n_{\text{total}}}{n_{\text{samples, class } i}} \tag{1}$$

This formulation directly scales weights by class frequency, producing stronger penalties for the minority class.

2. Balanced weighting ("balanced"):

$$w_i = \frac{n_{\text{total}}}{n_{\text{classes}} \cdot n_{\text{samples, class } i}} \tag{2}$$

Here, weights are normalized by the number of classes, resulting in more moderate penalties.

3. LGBM-balanced ("lgb_balanced"):
An internal implementation in LightGBM, which automatically applies the balanced strategy using the training data distribution.

These weighting schemes mainly differ in the intensity of penalization applied to the minority class. While raw weighting may lead to very high penalties in severely imbalanced datasets, the balanced and lgb_balanced strategies offer more stable adjustments. Table 9 summarizes the performance metrics obtained under each scheme. As shown in Table 9, the *balanced* weighting scheme achieved the highest sensitivity (34 %), substantially improving the model's ability to identify non-survivors while also delivering the best F1-score (34 %) and maintaining high specificity (96 %). This gain came without compromising overall discrimination, with an AUC of 0.8364 (95 % CI: 0.8253–0.8475). In contrast, the *raw* and *lgb_balanced* schemes produced lower

Table 9. Performance metrics of the optimized LGBM model applying different types of class weights evaluated on PIC dataset.

Type	Sens	Spec	F1	AUC (95 % CI)
"raw"	31 %	97 %	32 %	0.8343 (0.8231 - 0.8454)
"balanced"	**34 %**	**96 %**	**34 %**	**0.8364 (0.8253 - 0.8475)**
"lgb_balanced"	30 %	97 %	32 %	0.8352 (0.8241 - 0.8463)

Note. Sens: sensitivity; Spec: specificity; F1: F1-score.

sensitivities (31 % and 30 %, respectively), reflecting a reduced capacity to detect high-risk patients despite preserving specificity.

3.2.4. Evaluation of Focal Loss In this section, the use of *focal loss* was evaluated as the cost function in the optimized LGBM model. This approach assigns greater weight to hard-to-classify cases while reducing the influence of easy ones. The loss is defined in Equation 3, where the parameter $\alpha \in [0, 1]$ balances the relative importance of the positive and negative classes, and γ controls the strength of the focus on misclassified examples.

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{3}$$

Here, p_t denotes the predicted probability of the true class for each sample. When a sample is correctly classified with high confidence (i.e., p_t close to 1), the term $(1-p_t)^{\gamma}$ approaches zero, effectively reducing its contribution to the overall loss. Conversely, for misclassified or low-confidence samples (small p_t), this term increases, amplifying their influence during optimization. This mechanism enables the model to focus its learning on minority and hard-to-predict cases—such as non-survivors in imbalanced clinical datasets—mitigating the dominance of easy majority-class examples.

To assess its effect, three values of α were tested, and for each, three values of γ were applied. Table 10 summarizes the results obtained from these experiments.

Table 10. Performance metrics of the optimized LGBM model with focal loss implementation evaluated on PIC dataset.

Alpha	Gamma	Sens	Spec	F1	AUC (95 % CI)
0.25	1	40 %	94 %	31 %	0.8245 (0.8131 - 0.8359)
	3	30 %	97 %	32 %	0.8171 (0.8055 - 0.8287)
	5	25 %	98 %	31 %	0.8137 (0.8020 - 0.8253)
0.5	1	29 %	98 %	33 %	0.8220 (0.8105 - 0.8335)
	3	33 %	96 %	31 %	0.8099 (0.7981 - 0.8216)
	5	34 %	96 %	32 %	0.8119 (0.8002 - 0.8236)
0.75	1	33 %	97 %	34 %	0.8257 (0.8144 - 0.8371)
	3	**43 %**	**93 %**	**32 %**	**0.8147 (0.8030 - 0.8263)**
	5	31 %	96 %	31 %	0.8119 (0.8002 - 0.8236)

Note. Sens: sensitivity; Spec: specificity; F1: F1-score.

As shown in Table 10, the performance of the optimized LGBM model varied with the choice of α and γ. The best configuration was $\alpha = 0.75$ and $\gamma = 3$, which achieved the highest sensitivity (43 %) while maintaining competitive AUC (0.8147, 95 % CI: 0.8030–0.8263), F1-score (32 %), and a reasonable specificity (93 %). This configuration outperformed the baseline optimized LGBM model as well as the data-level and class-weight strategies previously evaluated. Overall, focal loss proved effective in enhancing sensitivity, though its performance was strongly parameter-dependent, with more extreme values of γ reducing the balance across metrics.

4. MODEL INTERPRETABILITY AND COMPARATIVE INSIGHTS

4.1. INTERPRETABILITY THROUGH SHAP ANALYSIS

While the ML models demonstrated reasonable predictive performance, their clinical usefulness ultimately depends on understanding why certain predictions are made. In contrast to the univariate OR analysis—which quantified the strength of individual associations with mortality—model interpretability allows exploring how these variables collectively contribute to prediction outcomes within a multivariable context.

Interpretability is a fundamental requirement for ML models in medicine, where understanding the contribution of each clinical variable is as important as achieving high predictive performance [59] [60]. In intensive care, clinicians not only require accurate predictions of mortality risk but also transparent explanations of why patients are classified as high risk, in order to support timely and informed interventions.

To address this need, we applied SHapley Additive exPlanations (SHAP) [61], one of the most widely used methods for interpreting complex ML models. SHAP is based on an idea from cooperative game theory: imagine that each input variable is like a player on a team, and together they work to produce the model's prediction. SHAP calculates how much each "player" contributes by checking what happens

---

[59] Talal A. A. Abdullah, Mohd Soperi Mohd Zahid y Waleed Ali. «A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions». En: *Symmetry* 13.12 (2021). DOI: 10.3390/sym13122439.

[60] Chang Ho Yoon, Robert Torrance y Naomi Scheinerman. «Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned?» En: *Journal of Medical Ethics* 48.9 (2022), págs. 581-585. DOI: 10.1136/medethics-2020-107102.

[61] Scott M Lundberg y Su-In Lee. «A Unified Approach to Interpreting Model Predictions». En: *Advances in Neural Information Processing Systems 30*. Ed. por I. Guyon et al. Curran Associates, Inc., 2017, págs. 4765-4774.

when that player joins or leaves different combinations of teammates. In this way, SHAP provides a clear estimate of how much each variable influences the prediction—showing both the direction of the effect (whether it increases or decreases the predicted risk) and the magnitude (how strong that influence is). By providing complementary perspectives—global interpretability, through feature ranking across the cohort, and local interpretability, by explaining individual patient outcomes—SHAP facilitates clinical understanding, enhances trust, and supports the integration of ML models as decision-support tools in ICU settings.

In this study, SHAP analysis was applied to the optimized LGBM model trained with the focal loss function (α = 0.75, γ = 3), which achieved the highest sensitivity among all evaluated configurations. This model was selected for interpretability analysis because it provided the most clinically relevant trade-off between sensitivity and overall discrimination, making it particularly suitable for understanding how the algorithm identifies high-risk patients. By focusing on the best-performing configuration, the SHAP analysis aimed to reveal the underlying predictive logic of a model that not only performs well but also aligns with clinical priorities, such as early and accurate identification of non-survivors.

Global interpretability was explored through two complementary visualizations: (i) a bar plot ranking predictors by their mean absolute SHAP values, and (ii) a beeswarm plot showing how variations in feature values influenced predicted mortality risk across individual patients. These visualizations not only identify the most influential predictors but also clarify the directionality of their effects—indicating whether higher or lower values of a given variable tend to increase or reduce the predicted risk of mortality. This distinction provides clinically meaningful insight into which features act as risk-enhancing versus protective factors within the model's decision-making process.

Figure 3 presents the global SHAP feature ranking for the focal loss–optimized LGBM

model, with variables ordered by their relative influence on mortality prediction. The length of stay (*los*) emerged as the most influential predictor, followed by laboratory markers such as *maximum lactate*, *maximum gamma-glutamyltransferase*, *maximum hematocrit*, and *maximum* pCO_2. Other biochemical indicators, including *maximum creatine kinase* (CK) and *minimum hemoglobin*, also showed strong contributions. Beyond laboratory features, the model highlighted a few physiological variables—particularly *mean respiratory rate*, *maximum respiratory rate*, and *mean temperature*—indicating that both laboratory biomarkers and vital signs play complementary roles in shaping mortality predictions.

Figure 3. Bar plot for SHAP analysis applied to the optimized LGBM model with focal loss (α = 0.75, γ = 3) for feature importance in pediatric mortality prediction using PIC dataset.

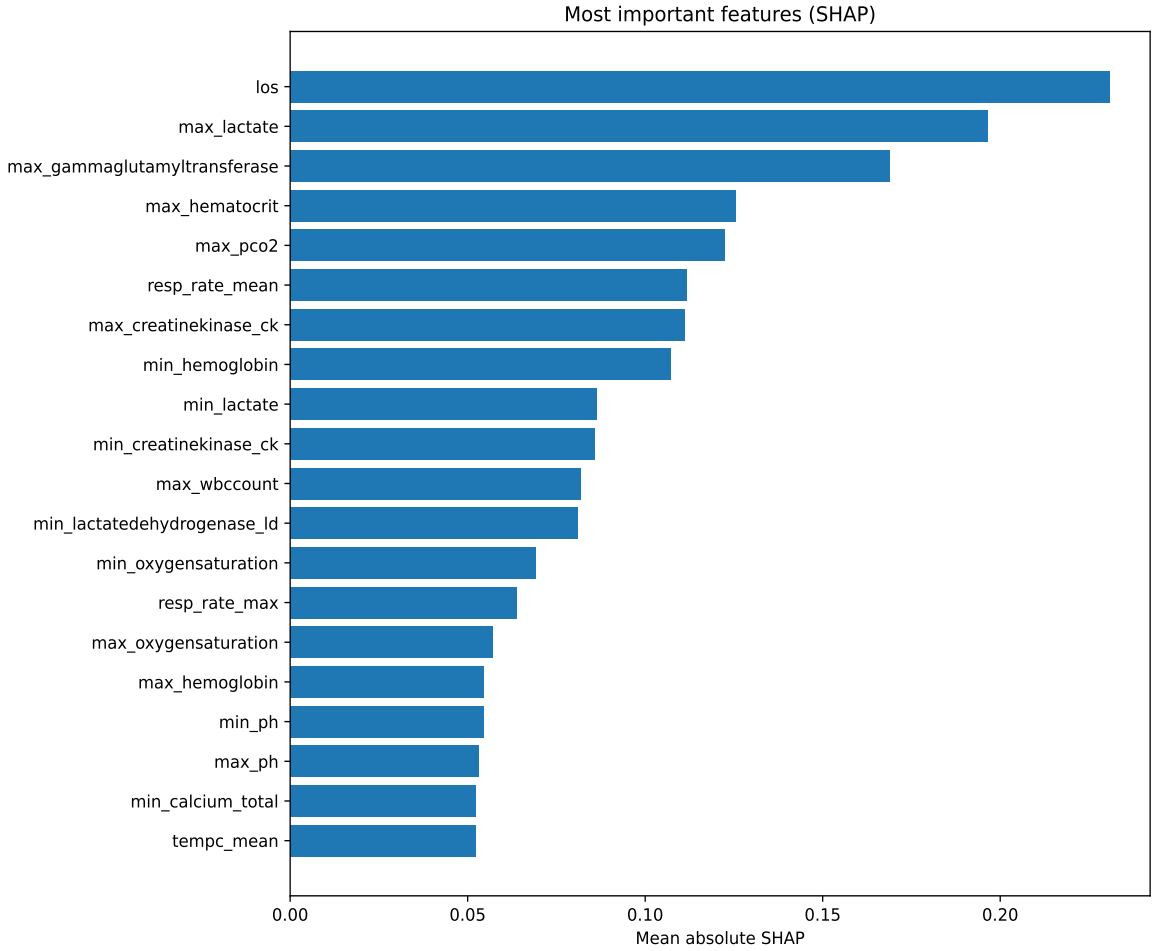
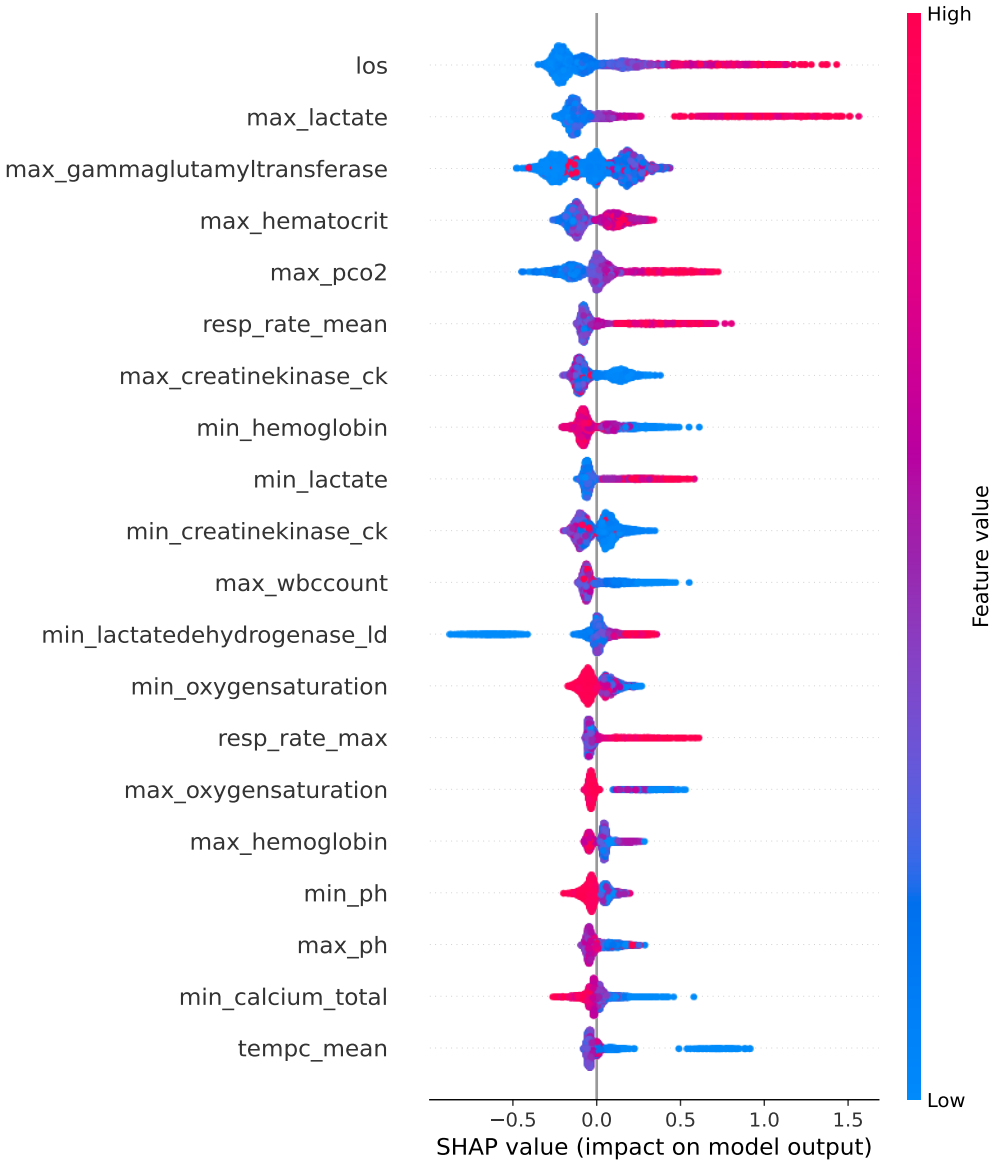


Figure 4 complements this ranking with the beeswarm plot, illustrating how variations in these key features (e.g., high vs. low values) influence individual mortality predictions. Each dot represents a patient, with color encoding the original feature value (fuchsia for higher values and blue for lower ones). The position along the x-axis (SHAP value) reflects both the direction and magnitude of the feature's effect on mortality risk—positive SHAP values correspond to an increased predicted risk, while negative values indicate a protective influence.

For example, higher *maximum lactate* and longer lengths of stay (*los*) appear predominantly on the right side in fuchsia, showing that elevated values in these variables are associated with an increased risk of mortality. Conversely, for features such as minimum hemoglobin (*min_hemoglobin*), lower values shift towards the right, indicating a higher predicted risk, whereas higher hemoglobin values appear on the left, suggesting a protective effect. Together, these visualizations provide an interpretable overview of how the most influential clinical variables shape the model's output across individual patients, offering clinically meaningful insights into both the direction and strength of each variable's contribution.

Figure 4. Beeswarm plot for SHAP analysis applied to the optimized LGBM model with focal loss (α = 0.75, γ = 3) for feature relationship with pediatric mortality prediction using PIC dataset.



Together, these analyses emphasize the central role of laboratory biomarkers and selected physiological indicators in shaping model predictions. They also reveal non-linear patterns and interactions that traditional statistical methods may overlook. Overall, SHAP analysis offered valuable insights into both the relative importance

and functional role of clinical predictors, enhancing transparency and providing a clinically meaningful interpretation of the focal loss–optimized LGBM model's predictions.

4.2. COMPARISON BETWEEN ASSOCIATION ANALYSIS AND MACHINE LEARNING APPROACHES

After conducting the initial statistical analysis through OR and subsequently training multiple ML models, the next step was to compare both analytical perspectives. This comparison sought to examine how the variables identified as significant in the traditional association analysis aligned with those highlighted as most influential by the interpretable ML model. Such integration allows assessing the consistency between conventional statistical associations and model-derived feature importance, providing a broader understanding of which clinical variables are most strongly linked to pediatric ICU mortality.

After identifying the 20 variables most strongly associated with mortality through the univariate OR analysis (see Table 2), these findings were compared with the top 20 predictors obtained from the global SHAP interpretability analysis of the focal loss–optimized LGBM model (see Figure 3). The objective was to evaluate the degree of convergence between traditional statistical associations and model-derived feature importance obtained from the best-performing configuration after addressing class imbalance.

Variables highlighted by both methods can be regarded as robust candidates, as they demonstrate consistency across epidemiological association and predictive modeling. In contrast, features with high SHAP values but low ORs may not represent clear independent risk factors, yet they contribute meaningfully to predictive performance through non-linear effects or interactions. Conversely, features with strong ORs but limited SHAP relevance indicate statistical associations that, while clinically

relevant, offer limited utility within a multivariable ML framework. Features with low importance in both analyses are unlikely to represent meaningful risk factors or reliable predictors.

The comparison revealed four clinical variables consistently identified by both approaches: *min_lactate*, *max_lactate*, *max_hematocrit*, and *resp_rate_mean*. These overlapping features stand out as strong candidates for pediatric ICU mortality prediction, as they are simultaneously recognized as statistically significant and as influential contributors in the focal loss–optimized model. Lactate, in both its minimum and maximum values, emerged as a particularly robust predictor across methods, consistent with its well-established clinical relevance in critical care.

Closer inspection highlights nuanced divergences. While *min_lactate* and*max_lactate* proved robust in both analyses, *max_hematocrit* and *resp_rate_mean* exhibited weaker statistical associations but ranked highly in the SHAP analysis. This suggests that their predictive value is better captured through complex, multivariable interactions rather than through isolated univariate effects. Importantly, their high SHAP contributions indicate that, despite modest OR values, these features substantially strengthen the model's predictive capacity—underscoring the value of interpretability methods for uncovering clinically meaningful predictors.

On the other hand, several variables with strong univariate associations, such as *max_phosphate* and *min_phosphate*, were not prioritized by SHAP, suggesting that their predictive value diminishes once multivariable and non-linear relationships are considered. Conversely, features highly ranked by SHAP but absent from the OR top 20 (e.g., length of stay, *min_pH*, and *max_creatinekinase_ck*) likely capture complex patterns that are overlooked by conventional statistical methods.

Overall, the overlap reinforces the robustness of certain predictors, while the divergences highlight the complementary value of combining traditional statistical analysis with interpretable machine learning. Together, these approaches provide a more

comprehensive understanding of clinically relevant predictors for pediatric ICU mortality, especially when analyzing models optimized under effective imbalance mitigation strategies such as focal loss.

5. DISCUSSION

This study investigated the development and evaluation of ML models for predicting pediatric intensive care mortality using the PIC database. Overall, the models demonstrated reasonable discriminative capacity; however, their sensitivity identifying non-survivor patients remained limited. This finding suggests that, despite methodological refinements—including model optimization and class imbalance handling—accurately detecting high-risk patients continues to pose a challenge. This challenge highlights a broader issue: the reduced sensitivity is probably not due to the modelling approach itself, but instead to the nature and completeness of the available data.

Indeed, advances in pediatric critical care modeling remain largely constrained by data scarcity and heterogeneity. Unlike the adult domain—where open databases such as MIMIC-III and eICU have enabled reproducible research and methodological innovation—pediatric intensive care lacks equivalent multicenter resources. To the best of current knowledge, the PIC database stands as a crucial step forward, being the only openly accessible dataset dedicated to critically ill children. While it offers a valuable foundation for data-driven research, it also highlights the limitations of current data availability and standardization in pediatrics.

An exploratory assessment of the PIC database, supported by a literature review and consultation with a pediatric critical care expert, revealed both advantages and restrictions. The dataset includes valuable laboratory biomarkers such as lactate and creatinine, which are well-established predictors of mortality. However, it lacks key clinical variables—particularly vital signs and neurological indicators such as blood pressure, Glasgow Coma Score, pupillary reactions, urinary output, and capillary refill—that are routinely integrated into conventional mortality risk scores. Several studies have emphasized the importance of these parameters in pediatric critical

care and their inclusion in established severity scores such as PRISM, PIM, and PELOD [5] [6] [18] [19] [62] [63] [64]. According to the consulted pediatric intensive care expert, the absence of these variables in the PIC database significantly limits the model's ability to capture hemodynamic and neurological dimensions of critical illness, thereby reducing sensitivity to early physiological deterioration. Moreover, since these same variables are fundamental components of conventional mortality scores, their omission prevents direct comparisons between ML–based models trained on PIC data and traditional scoring systems. These limitations extend beyond variable count to encompass data quality and completeness: more coherent and comprehensive information would likely improve model performance and bring sensitivity values closer to clinically meaningful thresholds.

The comparison between traditional statistical association and ML interpretability further illustrates these constraints. The univariate OR analysis identified *phosphate*, *potassium*, *carboxyhemoglobin*, and *lactate* as variables strongly associated with mortality. In contrast, SHAP analysis applied to the optimized LGBM model with focal loss (α = 0.75, γ = 3)—the best-performing configuration after addressing class imbalance—highlighted length of stay (*los*) as the most influential predictor, followed by *maximum lactate*, *maximum gamma-glutamyltransferase*, *maximum hematocrit*, and *maximum arterial pCO_2*. Among physiological indicators, *mean respiratory rate* and *mean temperature* also emerged as relevant contributors, underscoring their

---

[62] L. Straney et al. «Paediatric index of mortality 3: an updated model for predicting mortality in pediatric intensive care». En: *Pediatric Critical Care Medicine* 14.7 (2013), págs. 673-681. DOI: 10.1097/PCC.0b013e31829760cf.

[63] S. Y. Kim et al. «A deep learning model for real-time mortality prediction in critically ill children». En: *Critical Care* 23.1 (2019), pág. 279. DOI: 10.1186/s13054-019-2561-z.

[64] A. L. Graciano et al. «The Pediatric Multiple Organ Dysfunction Score (P-MODS): development and validation of an objective scale to measure the severity of multiple organ dysfunction in critically ill children». En: *Critical Care Medicine* 33.7 (2005), págs. 1484-1491. DOI: 10.1097/01.ccm.0000170943.23633.47.

relevance in the model's decision-making process.

The limited overlap between OR and SHAP results highlights their complementary strengths and may suggest potential structural constraints within the PIC dataset. Specifically, the database is heavily populated with laboratory variables, whereas vital signs and neurological indicators are underrepresented or inconsistently recorded. Consequently, both analyses likely operate on a partial representation of the patient's physiological state, which may explain why certain variables appear significant in one approach but not the other. These discrepancies likely reflect data composition rather than methodological inconsistency. The consistent prominence of lactate across both analyses supports its robustness as a prognostic biomarker, while divergences in other variables may result from incomplete data and highlight the added value of ML interpretability in revealing non-linear and interaction effects that traditional statistics cannot capture.

Beyond variable selection, class imbalance emerged as another key challenge. In the study cohort, approximately 95 % of patients were survivors and only 5 % non-survivors, creating a highly skewed distribution that biased models toward the majority class. Systematic evaluation of imbalance-handling strategies showed that over-sampling methods such as SMOTE and Borderline-SMOTE improved sensitivity under moderate resampling ratios, whereas Random Under-Sampling achieved the highest sensitivity (37 %) but at the cost of reducing data diversity. Model-level techniques such as class weighting provided modest yet stable improvements, while focal loss achieved the best trade-off between sensitivity (43 %) and overall discrimination without artificially altering the data distribution. These results underscore the importance of addressing imbalance at both data and model levels to achieve clinically meaningful performance.

Taken together, these findings highlight the following central insights. First, while the PIC database provides an invaluable open resource for pediatric mortality predic-

tion, it lacks several key physiological and neurological variables, limiting the capacity of ML models to reach the sensitivity required for clinical deployment. Second, methodological strategies such as class imbalance mitigation can partially compensate for these data limitations but cannot fully overcome them. Ultimately, the convergence between traditional statistical and ML interpretability analyses will depend on the availability of richer, higher-quality, and more consistent pediatric datasets. Strengthening these data foundations would not only enhance model sensitivity and robustness but also increase transparency and trust in ML-based decision-support tools within pediatric intensive care.

6. CONCLUSIONS

This study developed and evaluated several ML models for predicting pediatric ICU mortality using the PIC database. By integrating univariate OR analysis with SHAP-based interpretability applied to the best-performing model—an optimized LGBM with focal loss (α = 0.75, γ = 3)—it was possible to examine variable relevance from both epidemiological and computational perspectives.

The comparative analysis revealed partial overlap between both approaches. *Minimum* and *maximum lactate*, *maximum hematocrit*, and *mean respiratory rate* were consistently identified as relevant predictors, with *lactate* standing out as a robust and clinically coherent biomarker. Variables such as *hematocrit* and *respiratory rate*, though less significant in univariate analyses, gained importance in the SHAP model, suggesting complex, multivariable interactions. These findings underscore the complementary value of combining traditional statistics with ML interpretability and the strong influence of data completeness on model stability and understanding.

Despite the methodological refinements—including hyperparameter optimization and imbalance handling —the overall sensitivity of the models remained limited. This suggests that the main constraints are structural rather than algorithmic, stemming from the restricted scope and quality of available pediatric data. The PIC database, while valuable for its laboratory information, lacks critical hemodynamic and neurological variables—such as blood pressure, Glasgow Coma Score, and capillary refill—integral to established mortality scores. Their absence restricts both physiological representation and comparability with conventional methods, highlighting the need for richer, multicenter, and publicly available pediatric datasets that integrate vital signs and neurological parameters to support the development of robust, generalizable, and clinically actionable predictive models.

Furthermore, this study demonstrates the critical role of explicitly addressing class

imbalance in pediatric mortality prediction. Among the strategies evaluated, focal loss achieved the highest sensitivity, making it particularly suitable when identifying high-risk patients is clinically crucial. These results suggest that class imbalance should be treated as an inherent modeling challenge rather than merely a data limitation.

Overall, integrating traditional statistical analysis with model interpretability methods provides a richer understanding of how different clinical variables contribute to mortality prediction, revealing both methodological complementarities and data-driven limitations. Additionally, advancing ML-based mortality prediction in pediatric critical care requires not only algorithmic and methodological innovation but also a concerted effort to expand and standardize high-quality, multicenter pediatric databases. Addressing these structural data gaps is essential for developing predictive models that achieve high predictive power, sensitivity, and transparency necessary for meaningful clinical application.

BIBLIOGRAPHY

Abdullah, Talal A. A., Mohd Soperi Mohd Zahid y Waleed Ali. «A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions». En: *Symmetry* 13.12 (2021). DOI: 10.3390/sym13122439 (vid. pág. 40).

Abedin, M. Z. et al. «Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk». En: *Complex & Intelligent Systems* 9.5 (2023), págs. 3559-3579. DOI: 10.1007/s40747-023-01041-3 (vid. pág. 20).

Akiba, Takuya et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Available at https://optuna.org/. 2019 (vid. pág. 31).

Alejo, R. et al. «A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios». En: *Pattern Recognition Letters* 34.4 (2013), págs. 380-388. DOI: 10.1016/j.patrec.2012.11.002 (vid. pág. 20).

Altawalbeh, S. M. et al. «Evaluating Intensive Care Unit Medication Charges in a Teaching Hospital in Jordan». En: *Expert Review of Pharmacoeconomics & Outcomes Research* (2019). DOI: 10.1080/14737167.2019.1571413 (vid. pág. 11).

Beckmann, M., N.F.F. Ebecken y B.S.P. de Lima. «A KNN Undersampling Approach for Data Balancing». En: *Journal of Intelligent Learning Systems and Applications* 7.4 (2015), págs. 104-116. DOI: 10.4236/jilsa.2015.74010 (vid. pág. 19).

Belarouci, Sara y Mohammed Amine Chikh. «Medical imbalanced data classification». En: *Advances in Science, Technology and Engineering Systems Journal* 2.3 (2017), págs. 116-124 (vid. pág. 19).

Botan, E. et al. «Characteristics and timing of mortality in children dying in pediatric intensive care: a 5-year experience». En: *Acute and Critical Care* 37.4 (2022), págs. 644-653. DOI: 10.4266/acc.2022.00395 (vid. pág. 23).

Cameron, S. et al. «Pediatric severe traumatic brain injury mortality prediction determined with machine learning-based modeling». En: *Injury* 53 (2022), págs. 992-998. DOI: 10.1016/j.injury.2022.01.008 (vid. págs. 14, 50).

Chawla, Nitesh V. et al. «SMOTE: Synthetic Minority Over-sampling Technique». En: *Journal of Artificial Intelligence Research* 16 (2002), págs. 321-357. DOI: 10.1613/jair.953 (vid. págs. 19, 33).

Development Team, The miceforest. *miceforest: Fast, Memory Efficient Imputation with LightGBM*. Versión 6.0.3. 2024 (vid. pág. 25).

Devi, D., S. K. Biswas y B. Purkayastha. «Correlation-based oversampling aided cost sensitive ensemble learning technique for treatment of class imbalance». En: *Journal of Experimental & Theoretical Artificial Intelligence* 34.2 (2022), págs. 143-174. DOI: 10.1080/0952813X.2020.1856110 (vid. pág. 20).

Evans, M. et al. «Development and validation of a pediatric model predicting trauma-related mortality». En: *BMC Pediatrics* 23 (2023). DOI: 10.1186/s12887-023-04437-9 (vid. págs. 12, 50).

Gholampour, Seifollah. «Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable».

En: *Machine Learning and Knowledge Extraction* 6.2 (2024), págs. 827-841. DOI: 10.3390/make6020039 (vid. pág. 18).

Graciano, A. L. et al. «The Pediatric Multiple Organ Dysfunction Score (P-MODS): development and validation of an objective scale to measure the severity of multiple organ dysfunction in critically ill children». En: *Critical Care Medicine* 33.7 (2005), págs. 1484-1491. DOI: 10.1097/01.ccm.0000170943.23633.47 (vid. pág. 50).

Haixiang, Guo et al. «Learning from class-imbalanced data: Review of methods and applications». En: *Expert systems with applications* 73 (2017), págs. 220-239 (vid. pág. 18).

Han, H., W. Y. Wang y B. H. Mao. «Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning». En: *Advances in Intelligent Computing (ICIC 2005)*. Ed. por D. S. Huang, X. P. Zhang y G. B. Huang. Vol. 3644. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, págs. 878-887. DOI: 10.1007/11538059_91 (vid. pág. 33).

He, H. y E. A. Garcia. «Learning from Imbalanced Data». En: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), págs. 1263-1284. DOI: 10.1109/TKDE.2008.239 (vid. págs. 19, 33).

Hong, S. et al. «Predicting Risk of Mortality in Pediatric ICU Based on Ensemble Step-Wise Feature Selection». En: *Health Data Science* (2021), pág. 9365125. DOI: 10.34133/2021/9365125 (vid. págs. 14, 17, 23, 24).

Hosenie, Z. et al. «Imbalance learning for variable star classification». En: *Monthly Notices of the Royal Astronomical Society* 493.4 (2020), págs. 6050-6059. DOI: 10.1093/mnras/staa776 (vid. pág. 20).

Johnson, A. E. et al. «MIMIC-III, a freely accessible critical care database». En: *Scientific Data* 3 (2016), pág. 160035. DOI: 10.1038/sdata.2016.35 (vid. pág. 16).

Johnson, H. et al. «Multiple Organ Dysfunction Syndrome and Pediatric Logistic Organ Dysfunction-2 Score in Pediatric Cerebral Malaria». En: *The American Journal of Tropical Medicine and Hygiene* 107.4 (2022), págs. 820-826. DOI: 10.4269/ajtmh.22-0140 (vid. págs. 14, 50).

Joshi, P. et al. «Application of Pediatric Risk of Mortality (PRISM) III Score in Predicting Mortality Outcomes». En: *Journal of Nepal Health Research Council* 21 (2024), págs. 450-457 (vid. págs. 12, 50).

Kaisar, S. y A. Chowdhury. «Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests». En: *ICT Express* 8.4 (2022), págs. 563-568. DOI: 10.1016/j.icte.2022.05.002 (vid. pág. 20).

Khuat, T. T. y M. H. Le. «Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems». En: *SN Computer Science* 1.2 (2020), pág. 108. DOI: 10.1007/s42979-020-00109-x (vid. pág. 20).

Kim, S. Y. et al. «A deep learning model for real-time mortality prediction in critically ill children». En: *Critical Care* 23.1 (2019), pág. 279. DOI: 10.1186/s13054-019-2561-z (vid. pág. 50).

Kong, G., K. Lin e Y. Hu. «Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU». En: *BMC Medical Informatics and Decision Making* 20.251 (2020). DOI: 10.1186/s12911-020-01271-2 (vid. pág. 23).

Kumar, Vinod et al. «Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques». En: *Healthcare* 10.7 (2022). DOI: 10.3390/healthcare10071293 (vid. pág. 18).

Lee, B. et al. «Development of a machine learning model for predicting pediatric mortality in the early stages of intensive care unit admission». En: *Sci Rep* 11.1 (2021), pág. 1263. DOI: 10.1038/s41598-020-80474-z (vid. pág. 15).

Leteurtre, S et al. «Development of a pediatric multiple organ dysfunction score: use of two strategies». En: *Medical Decision Making* 19.4 (1999), págs. 399-410. DOI: 10.1177/0272989X9901900408 (vid. pág. 12).

Libraries, Kent State University. *SPSS Tutorials: Chi-Square Test of Independence*. Accessed: October 31, 2025. 2025. URL: https://libguides.library.kent.edu/spss/chisquare (vid. pág. 27).

Lin, Cian, Chih-Fong Tsai y Wei-Chao Lin. «Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study». En: *Artificial Intelligence Review* 56.2 (2023), págs. 845-863 (vid. pág. 19).

Lin, Tsung-Yi et al. «Focal loss for dense object detection». En: *Proceedings of the IEEE international conference on computer vision*. 2017, págs. 2980-2988 (vid. págs. 18, 19, 33).

Lu, D. et al. «Prognostic value of systemic inflammatory response index for acute kidney injury and the prognosis of pediatric patients in critical care units». En:

PLOS ONE 19.8 (2024), e0306884. DOI: 10.1371/journal.pone.0306884 (vid. pág. 17).

Lundberg, Scott M y Su-In Lee. «A Unified Approach to Interpreting Model Predictions». En: *Advances in Neural Information Processing Systems 30*. Ed. por I. Guyon et al. Curran Associates, Inc., 2017, págs. 4765-4774 (vid. pág. 40).

Mann, Henry B. y Donald R. Whitney. «On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other». En: *The Annals of Mathematical Statistics* 18.1 (1947), págs. 50-60. DOI: 10.1214/aoms/1177730491 (vid. pág. 26).

McAdams, RM et al. «Predicting clinical outcomes using artificial intelligence and machine learning in neonatal intensive care units: a systematic review». En: *J Perinatol.* 42.12 (2022), págs. 1561-1575. DOI: 10.1038/s41372-022-01392-8 (vid. pág. 11).

Morooka, H et al. «Prognostic Impact of Parameters of Metabolic Acidosis in Critically Ill Children with Acute Kidney Injury: A Retrospective Observational Analysis Using the PIC Database». En: *Diagnostics* 10.11 (2020), pág. 937. DOI: 10.3390/diagnostics10110937 (vid. pág. 16).

Niaz, Nazim Uddin, K.M. Nadim Shahariar y Muhammed J. A. Patwary. «Class Imbalance Problems in Machine Learning: A Review of Methods And Future Challenges». En: *Proceedings of the 2nd International Conference on Computing Advancements*. ICCA '22. Dhaka, Bangladesh: Association for Computing Machinery, 2022, 485–490. DOI: 10.1145/3542954.3543024 (vid. pág. 18).

Pollack, M. M., K. M. Patel y U. E. Ruttimann. «PRISM III: an updated Pediatric Risk of Mortality score». En: *Critical Care Medicine* 24.5 (1996), págs. 743-752. DOI: 10.1097/00003246-199605000-00004 (vid. págs. 12, 24).

Pollard, T. J. y L. A. Celi. «Enabling Machine Learning in Critical Care». En: *ICU management & practice* 17.3 (2017), págs. 198-199 (vid. pág. 12).

Pollard, T. J. et al. «The eICU Collaborative Research Database, a freely available multi-center database for critical care research». En: *Scientific Data* 5 (2018), pág. 180178. DOI: 10.1038/sdata.2018.178 (vid. pág. 16).

Prince, R. D. et al. «A Machine Learning Classifier Improves Mortality Prediction Compared With Pediatric Logistic Organ Dysfunction-2 Score: Model Development and Validation». En: *Critical Care Explorations* 3.5 (2021), e0426. DOI: 10.1097/CCE.0000000000000426.

Prithula, J. et al. «Improved pediatric ICU mortality prediction for respiratory diseases: machine learning and data subdivision insights». En: *Respiratory Research* 25.1 (2024), pág. 216. DOI: 10.1186/s12931-024-02753-x (vid. pág. 17).

Qiu, J. et al. «Comparison of the pediatric risk of mortality, pediatric index of mortality, and pediatric index of mortality 2 models in a pediatric intensive care unit in China: A validation study». En: *Medicine* 96.14 (2017), e6431. DOI: 10.1097/MD.0000000000006431 (vid. pág. 11).

Reid, S et al. «Comparing CRIB-II and SNAPPE-II as mortality predictors for very preterm infants». En: *Journal of Paediatrics Child Health* 51.5 (2015), págs. 524-528 (vid. pág. 14).

Richardson, D. K. et al. «Score for neonatal acute physiology: A physiologic severity index for neonatal intensive care». En: *Pediatrics* 91.3 (1993), págs. 617-623 (vid. pág. 14).

Rosa-Mangeret, F et al. «2.5 million annual deaths—are neonates in low-and middle-income countries too small to be seen? A bottom-up overview on neonatal Morbi-mortality». En: *Trop. Med. Infect. Disease* 7.5 (2022), pág. 64 (vid. pág. 11).

Sáez, J. A., B. Krawczyk y M. Woźniak. «Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets». En: *Pattern Recognition* 57 (2016), págs. 164-178. DOI: 10.1016/j.patcog.2016.03.003 (vid. pág. 19).

Schlapbach, L. J. et al. «Prognostic accuracy of age-adapted SOFA, SIRS, PELOD-2, and qSOFA for in-hospital mortality among children with suspected infection admitted to the intensive care unit». En: *Intensive Care Med* 44 (2018), págs. 179-188. DOI: 10.1007/s00134-017-5021-8 (vid. pág. 13).

Shann, F et al. «Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care». En: *Intensive Care Medicine* 23.2 (1997), págs. 201-7. DOI: 10.1007/s001340050317 (vid. pág. 12).

Shapiro, S. S. y M. B. Wilk. «An Analysis of Variance Test for Normality (Complete Samples)». En: *Biometrika* 52.3-4 (1965), págs. 591-611. DOI: 10.2307/2333709 (vid. pág. 26).

Shen, Y y J Jiang. «Meta-Analysis for the Prediction of Mortality Rates in a Pediatric Intensive Care Unit Using Different Scores: PRISM-III/IV, PIM-3, and PELOD-2».

En: *Frontiers in Pediatrics* 9 (2021), pág. 712276. DOI: 10.3389/fped.2021.712276 (vid. págs. 13, 14).

Straney, L. et al. «Paediatric index of mortality 3: an updated model for predicting mortality in pediatric intensive care». En: *Pediatric Critical Care Medicine* 14.7 (2013), págs. 673-681. DOI: 10.1097/PCC.0b013e31829760cf (vid. pág. 50).

Strutz, S. et al. «Machine Learning for Predicting Critical Events Among Hospitalized Children». En: *JAMA Network Open* 8.5 (2025), e2513149. DOI: 10.1001/jamanetworkopen.2025.13149 (vid. pág. 15).

Thorsen-Meyer, Hans-Christian y et al. «Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records». En: *The Lancet Digital Health* 2.4 (2020), e179-e191 (vid. pág. 15).

Thukral, A. et al. «Performance of Pediatric Risk of Mortality (PRISM), Pediatric Index of Mortality (PIM), and PIM2 in a pediatric intensive care unit in a developing country». En: *Pediatric Critical Care Medicine* 7.4 (2006), págs. 356-361. DOI: 10.1097/01.PCC.0000227105.20897.89 (vid. pág. 23).

Viloria, A., O. B. P. Lezama y N. Mercado-Caruzo. «Unbalanced data processing using oversampling: Machine learning». En: *Procedia Computer Science* 175 (2020), págs. 108-113. DOI: 10.1016/j.procs.2020.07.016.

Wang, Benjamin X y Nathalie Japkowicz. «Boosting support vector machines for imbalanced data sets». En: *Knowledge and information systems* 25.1 (2010), págs. 1-20 (vid. pág. 19).

Yoon, Chang Ho, Robert Torrance y Naomi Scheinerman. «Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned?» En: *Journal of Medical Ethics* 48.9 (2022), págs. 581-585. DOI: 10.1136/medethics-2020-107102 (vid. pág. 40).

Yu, H., J. Ni y J. Zhao. «ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data». En: *Neurocomputing* 101 (2013), págs. 309-318 (vid. pág. 19).

Yue, C., C. Zhang y C. Ying. «A new nomogram for the individualized prediction of children's mortality risk in pediatric intensive care unit». En: *Am J Transl Res* 15.6 (2023), págs. 4172-4178 (vid. pág. 12).

Zeng, Xusheng et al. «PIC, a paediatric-specific intensive care database». En: *Scientific Data* 7.1 (2020), pág. 14. DOI: 10.1038/s41597-020-0355-4 (vid. págs. 16, 22).

Zhang, L et al. «Performance of PRISM III, PELOD-2, and P-MODS Scores in Two Pediatric Intensive Care Units in China». En: *Frontiers in Pediatrics* 9 (2021), pág. 626165. DOI: 10.3389/fped.2021.626165 (vid. págs. 13, 14).

Zhang, Z et al. «Performance of Three Mortality Prediction Scores and Evaluation of Important Determinants in Eight Pediatric Intensive Care Units in China». En: *Frontiers in Pediatrics* 8 (2020), pág. 522. DOI: 10.3389/fped.2020.00522 (vid. págs. 13, 14).

Zhou, J. et al. «Interpretable machine learning model for early prediction of disseminated intravascular coagulation in critically ill children». En: *Scientific Reports* 15 (2025), pág. 11217. DOI: 10.1038/s41598-025-91434-w (vid. págs. 14, 23).

ANNEXES

Annex A. Complete list of variables included in the study

This annex presents the complete list of engineered variables included in the final dataset used for model development. As described in Table 1, vital signs were summarized using minimum, mean, and maximum values; laboratory variables were represented by minimum and maximum values; and demographic variables were retained as a single representative value.

Table 11. Complete list of variables included in the study.

Variable name	Category	Extracted value	Description
heartrate_min	Vital sign	Minimum	Minimum heart rate during the first 24 h of ICU stay
heartrate_mean	Vital sign	Mean	Mean heart rate during the first 24 h of ICU stay
heartrate_max	Vital sign	Maximum	Maximum heart rate during the first 24 h of ICU stay
tempc_min	Vital sign	Minimum	Minimum body temperature during the first 24 h of ICU stay
tempc_mean	Vital sign	Mean	Mean body temperature during the first 24 h of ICU stay

Variable name	Category	Extracted value	Description
tempc_max	Vital sign	Maximum	Maximum body temperature during the first 24 h of ICU stay
resp_rate_min	Vital sign	Minimum	Minimum respiratory rate during the first 24 h of ICU stay
resp_rate_mean	Vital sign	Mean	Mean respiratory rate during the first 24 h of ICU stay
resp_rate_max	Vital sign	Maximum	Maximum respiratory rate during the first 24 h of ICU stay
max_absolutelymphocyte count	Laboratory	Maximum	Maximum absolute lymphocyte count
min_absolutelymphocyte count	Laboratory	Minimum	Minimum absolute lymphocyte count
max_carboxyhemoglobin	Laboratory	Maximum	Maximum carboxyhemoglobin level
min_carboxyhemoglobin	Laboratory	Minimum	Minimum carboxyhemoglobin level
max_alanineaminotransferase_alt	Laboratory	Maximum	Maximum alanine aminotransferase (ALT)
min_alanineaminotransferase_alt	Laboratory	Minimum	Minimum alanine aminotransferase (ALT)

Variable name	Category	Extracted value	Description
max_creatinekinase_ck	Laboratory	Maximum	Maximum creatine kinase (CK)
min_creatinekinase_ck	Laboratory	Minimum	Minimum creatine kinase (CK)
max_creatinekinase _mbiso-enzyme	Laboratory	Maximum	Maximum creatine kinase MB isoenzyme
min_creatinekinase _mbiso-enzyme	Laboratory	Minimum	Minimum creatine kinase MB isoenzyme
max_eosinophils	Laboratory	Maximum	Maximum eosinophil count
min_eosinophils	Laboratory	Minimum	Minimum eosinophil count
max_gammaglutamyl transfe-rase	Laboratory	Maximum	Maximum gamma-glutamyl transferase (GGT)
min_gammaglutamyl transfe-rase	Laboratory	Minimum	Minimum gamma-glutamyl transferase (GGT)
max_globulin	Laboratory	Maximum	Maximum globulin level
min_globulin	Laboratory	Minimum	Minimum globulin level
max_inr_pt	Laboratory	Maximum	Maximum international normalized ratio / proth-rombin time ratio
min_inr_pt	Laboratory	Minimum	Minimum international nor-malized ratio / prothrombin time ratio

Variable name	Category	Extracted value	Description
max_lactatedehydrogenase_ld	Laboratory	Maximum	Maximum lactate dehydrogenase (LDH)
min_lactatedehydrogenase_ld	Laboratory	Minimum	Minimum lactate dehydrogenase (LDH)
max_methemoglobin	Laboratory	Maximum	Maximum methemoglobin level
min_methemoglobin	Laboratory	Minimum	Minimum methemoglobin level
max_phosphate	Laboratory	Maximum	Maximum phosphate level
min_phosphate	Laboratory	Minimum	Minimum phosphate level
max_po2	Laboratory	Maximum	Maximum partial pressure of oxygen (pO_2)
min_po2	Laboratory	Minimum	Minimum partial pressure of oxygen (pO_2)
max_pt	Laboratory	Maximum	Maximum prothrombin time (PT)
min_pt	Laboratory	Minimum	Minimum prothrombin time (PT)
max_ppt	Laboratory	Maximum	Maximum partial thromboplastin time (PTT/PPT)
min_ppt	Laboratory	Minimum	Minimum partial thromboplastin time (PTT/PPT)
max_redbloodcells	Laboratory	Maximum	Maximum red blood cell count

Variable name	Category	Extracted value	Description
min_redbloodcells	Laboratory	Minimum	Minimum red blood cell count
max_sodium_wholeblood	Laboratory	Maximum	Maximum whole blood sodium level
min_sodium_wholeblood	Laboratory	Minimum	Minimum whole blood sodium level
max_totalbileacid	Laboratory	Maximum	Maximum total bile acid level
min_totalbileacid	Laboratory	Minimum	Minimum total bile acid level
max_alb_glb	Laboratory	Maximum	Maximum albumin/globulin ratio
min_alb_glb	Laboratory	Minimum	Minimum albumin/globulin ratio
max_albumin	Laboratory	Maximum	Maximum albumin level
min_albumin	Laboratory	Minimum	Minimum albumin level
max_amylase	Laboratory	Maximum	Maximum amylase level
min_amylase	Laboratory	Minimum	Minimum amylase level
max_aniongap	Laboratory	Maximum	Maximum anion gap
min_aniongap	Laboratory	Minimum	Minimum anion gap
max_asparateaminotransferase_ast	Laboratory	Maximum	Maximum aspartate aminotransferase (AST)
min_asparateaminotransferase_ast	Laboratory	Minimum	Minimum aspartate aminotransferase (AST)

Variable name	Category	Extracted value	Description
max_basophilscount	Laboratory	Maximum	Maximum basophil count
min_basophilscount	Laboratory	Minimum	Minimum basophil count
max_bicarbonate	Laboratory	Maximum	Maximum bicarbonate level
min_bicarbonate	Laboratory	Minimum	Minimum bicarbonate level
max_bilirubin_direct	Laboratory	Maximum	Maximum direct bilirubin level
min_bilirubin_direct	Laboratory	Minimum	Minimum direct bilirubin level
max_bilirubin_indirect	Laboratory	Maximum	Maximum indirect bilirubin level
min_bilirubin_indirect	Laboratory	Minimum	Minimum indirect bilirubin level
max_bilirubin_total	Laboratory	Maximum	Maximum total bilirubin level
min_bilirubin_total	Laboratory	Minimum	Minimum total bilirubin level
max_c_reactiveprotein	Laboratory	Maximum	Maximum C-reactive protein level
min_c_reactiveprotein	Laboratory	Minimum	Minimum C-reactive protein level
max_calcium_total	Laboratory	Maximum	Maximum total calcium level

Variable name	Category	Extracted value	Description
min_calcium_total	Laboratory	Minimum	Minimum total calcium level
max_creatinine	Laboratory	Maximum	Maximum creatinine level
min_creatinine	Laboratory	Minimum	Minimum creatinine level
max_glucose	Laboratory	Maximum	Maximum glucose level
min_glucose	Laboratory	Minimum	Minimum glucose level
max_hematocrit	Laboratory	Maximum	Maximum hematocrit
min_hematocrit	Laboratory	Minimum	Minimum hematocrit
max_hemoglobin	Laboratory	Maximum	Maximum hemoglobin
min_hemoglobin	Laboratory	Minimum	Minimum hemoglobin
max_lactate	Laboratory	Maximum	Maximum lactate level
min_lactate	Laboratory	Minimum	Minimum lactate level
max_lymphocytes_percent	Laboratory	Maximum	Maximum lymphocyte percentage
min_lymphocytes_percent	Laboratory	Minimum	Minimum lymphocyte percentage
max_neutrophils	Laboratory	Maximum	Maximum neutrophil count
min_neutrophils	Laboratory	Minimum	Minimum neutrophil count
max_oxygensaturation	Laboratory	Maximum	Maximum oxygen saturation
min_oxygensaturation	Laboratory	Minimum	Minimum oxygen saturation
max_pco2	Laboratory	Maximum	Maximum partial pressure of carbon dioxide (pCO_2)

Variable name	Category	Extracted value	Description
min_pco2	Laboratory	Minimum	Minimum partial pressure of carbon dioxide (pCO_2)
max_ph	Laboratory	Maximum	Maximum blood pH
min_ph	Laboratory	Minimum	Minimum blood pH
max_plateletcount	Laboratory	Maximum	Maximum platelet count
min_plateletcount	Laboratory	Minimum	Minimum platelet count
max_potassium	Laboratory	Maximum	Maximum potassium level
min_potassium	Laboratory	Minimum	Minimum potassium level
max_protein_total	Laboratory	Maximum	Maximum total protein level
min_protein_total	Laboratory	Minimum	Minimum total protein level
max_urea	Laboratory	Maximum	Maximum urea level
min_urea	Laboratory	Minimum	Minimum urea level
max_wbccount	Laboratory	Maximum	Maximum white blood cell count
min_wbccount	Laboratory	Minimum	Minimum white blood cell count
gender	Demographic	Single value	Biological sex recorded for the patient
los	Demographic / Clinical stay	Single value	Length of ICU stay