Archives of Anesthesiology and Critical Care (In Press); x(x): xx-xx.

Available online at https://aacc.tums.ac.ir/

Research Article

Machine Learning Model for Predicting Mortality in Pediatric Sepsis Patients Post-Laparotomy: A Retrospective Pilot Study

Retti Kartika Bakti[1,2], Arie Utariani[1,2], Nancy Margarita Rehatta[3], Adhi Dharma Wibawa[4], Kohar Hari Santoso[1,2], Lucky Andriyanto[1,2], Alief Wikarta[4]*

[1]*Department of Anesthesiology and Reanimation, Faculty of Medicine, Universitas Airlangga, Surabaya, East Java, Indonesia.*

[2]*Department of Anesthesiology and Reanimation, Dr. Soetomo General Academic Hospital, Surabaya, East Java, Indonesia.*

[3]*Universitas Kristen Petra, Surabaya, East Java, Indonesia.*

[4]*Institut Teknologi Sepuluh Nopember, Surabaya, East Java, Indonesia.*

ARTICLE INFO

Article history:
Received 18 July 2026
Revised 08 August 2026
Accepted 22 August 2026

Keywords:
Pediatric sepsis;
Laparotomy;
Mortality prediction;
Machine learning;
Ensemble methods

ABSTRACT

Background: This retrospective pilot study explored individual and ensemble machine learning models for predicting 30-day mortality in children with sepsis undergoing laparotomy.
Methods: A retrospective analysis was conducted for 60 children diagnosed with sepsis who received laparotomy at a hospital in Indonesia between 2020 and 2024. K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost) were evaluated before and after tuning. To address class imbalance inherent in mortality datasets, Synthetic Minority Oversampling Technique (SMOTE) was applied. Model performance was assessed using accuracy, class-specific F1-score, and area under the receiver operating characteristic curve (AUC). Ensemble strategies, including soft voting and stacking classifiers, were additionally employed to optimize predictive performance.
Results: The results showed that XGBoost and random forest demonstrated the highest accuracy among individual models. However, ensemble methods outperformed all single-model approaches. The stacking classifier achieved 100% accuracy on a small test set of 12 patients, while the soft voting ensemble demonstrated stable and consistent performance across metrics. The stacking classifier combining Random Forest and XGBoost, with logistic regression as the meta-learner and passthrough enabled, yielded an AUC of 0.70, indicating fair discriminative ability. Exploratory evaluation on an external dataset of six previously unused patients yielded a reported AUC of 0.62.
Conclusion: The findings are exploratory. The small sample size, few deaths, modest reported discrimination, and lack of independent validation prevent claims of clinical readiness. Larger multicenter and prospective validation is needed before clinical use.

Introduction

Children with sepsis are at risk of organ dysfunction and death. Sepsis is a serious condition that occurs when the body reacts to infection by attacking the inherent tissues. According to the Global Burden of Disease Study (2017), 48.9 million sepsis cases occur every year, with over 20 million affecting children, leading to more than 4.5 million deaths [1-2].

The authors declare no conflicts of interest.
*Corresponding author.
E-mail address: aliefwikarta@its.ac.id
DOI:

Copyright © 2027 Tehran University of Medical Sciences. Published by Tehran University of Medical Sciences.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

Routine laboratory measurements may help identify patients who need closer observation, but their predictive value must be assessed in the population in which they will be used. Laparotomy is a surgery often used to treat abdominal conditions in children, such as intestinal blockages or trauma [3]. However, children who develop sepsis after surgery are at higher risk of severe complications, including death [4]. Early identification of these high-risk patients is vital for improving outcomes. Candidate markers include immature granulocytes (IG), mean platelet volume (MPV), the MPV-to-platelet count ratio (MPV/P), and serum albumin [5–7]. Traditional methods for predicting outcomes often rely on clinical judgment and a few basic laboratory tests. However, these methods are less sophisticated for high-risk cases such as postoperative sepsis.

Children with sepsis undergoing laparotomy have clinical and laboratory characteristics that may interact in determining mortality risk. Machine learning methods can analyze these variables together and capture nonlinear relationships between organ dysfunction, nutritional status, and laboratory findings together with clinical characteristics [8–10]. However, investigations on the use of ML for predicting mortality in children with sepsis after laparotomy remain scarce. Therefore, this retrospective pilot study aimed to compare different ML models to predict 30-day mortality in patients.

The aim was exploratory comparison and identification of issues requiring further validation, rather than development of a model ready for clinical implementation. We also evaluated whether ensemble learning could improve predictive performance.

Methods

This retrospective pilot study was carried out at Dr. Soetomo Hospital in Surabaya, Indonesia, from January 2020 to June 2024. The population included all children under 18 who had sepsis and received laparotomy during this period, provided the medical records were complete. However, neonates were excluded. This study received ethical clearance from the Health Research Ethics Committee (KEPK) of Dr. Soetomo Hospital under IRB approval number 1613/LOE/301.4.2/III/2024.

The model input variables included age category, sex, body weight, nutritional status, ventilator use, length of stay, and pediatric Sequential Organ Failure Assessment (pSOFA) score. The categorical variables were represented by predefined numerical codes.

Laboratory inputs recorded at H0, H1, and H2 included platelet count (PC), mean platelet volume (MPV), MPV-to-platelet count ratio (MPV/P), serum albumin, MPV-to-albumin ratio (MPV/A), hemoglobin, white blood cell count, immature granulocytes, neutrophils, and lymphocytes. Patient outcome was used as the target variable.

The analysis was performed in Python using Google Colab, NumPy, pandas, and scikit-learn. The data were explored to assess missing values and variable distributions. Of the 66 patients screened, six (9.1%) were excluded because of incomplete required predictor or outcome data, leaving 60 patients for analysis using a complete-case approach. Correlation analysis was exploratory, and no predictors were excluded on the basis of correlations.

Aftar data cleaning, the 60 patient dataset was split into a training (80%) and a testing set (20%). The mortality rate was 20%, which meant the data was imbalanced. SMOTE was applied to the training data for the subsequent ensemble evaluation to address the imbalance between deaths and survivors. This study developed and tuned four supervised ML models, including weighted K-Nearest Neighbors (KNN), Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel, Random Forest, and Extreme Gradient Boosting (XGBoost). For KNN and SVM, predictors were standardized using StandardScaler. The parameters were fine-tuned for each model, and performance was assessed using accuracy, precision, recall, and F1-score. Two ensemble methods, soft voting and stacking classifiers, were employed to check whether combining models would improve performance. The stacking classifier used either logistic regression or XGBoost as the meta-learner. Logistic regression was evaluated with and without passthrough of the original input features.

Results

This study included 60 children, with infants younger than one year forming the largest age group (23, 38.3%). Most patients were male (37, 61.7%), and nearly three-quarters required mechanical ventilation (44, 73.3%). Good nutritional status was the most common category (27, 45.0%), followed by poor nutritional status and overnutrition (13 each, 21.7%) and undernutrition (7, 11.7%). Body weight varied widely, ranging from 1.3 to 74.0 kg, with a mean of 19.69 kg (SD 19.15). The mean pSOFA score was 6.42 (SD 3.62), with scores ranging from 3 to 16. Length of stay ranged from 2 to 104 days, with a mean of 18.98 days (SD 20.53).

Data analysis showed that some variables were strongly correlated with the mortality condition. The pSOFA score demonstrated the strongest negative correlation with survival condition, while lower albumin levels were associated with higher mortality. As shown in (Figure 1), the heatmap visualizes the correlation between various clinical and laboratory variables. It shows the strong negative correlation between pSOFA score and survival condition, which emphasizes the significance of this variable in predicting mortality risk.

(Figure 2) shows a horizontal bar chart that further supports the correlation observed in (Figure 1),

illustrating how different features relate to mortality outcomes. This chart helps identify the most influential predictors of mortality.

Table 1 compares the performance of four individual ML models before and after hyperparameter tuning, without SMOTE or ensemble learning. This comparison informed the subsequent investigation of ensemble approaches, whose performance is presented in the following section. Initially, all models achieved an accuracy of approximately 83%, but their death-class F1-scores were 0.00. Following tuning, Random Forest and XGBoost showed improved mortality prediction, with death-class F1-scores increasing to 0.50 and macro-average F1-scores increasing from 0.45 to 0.70, while accuracy remained unchanged. SVM and KNN showed no improvement in the reported metrics. These findings supported further evaluation of Random Forest and XGBoost using SMOTE and ensemble methods.

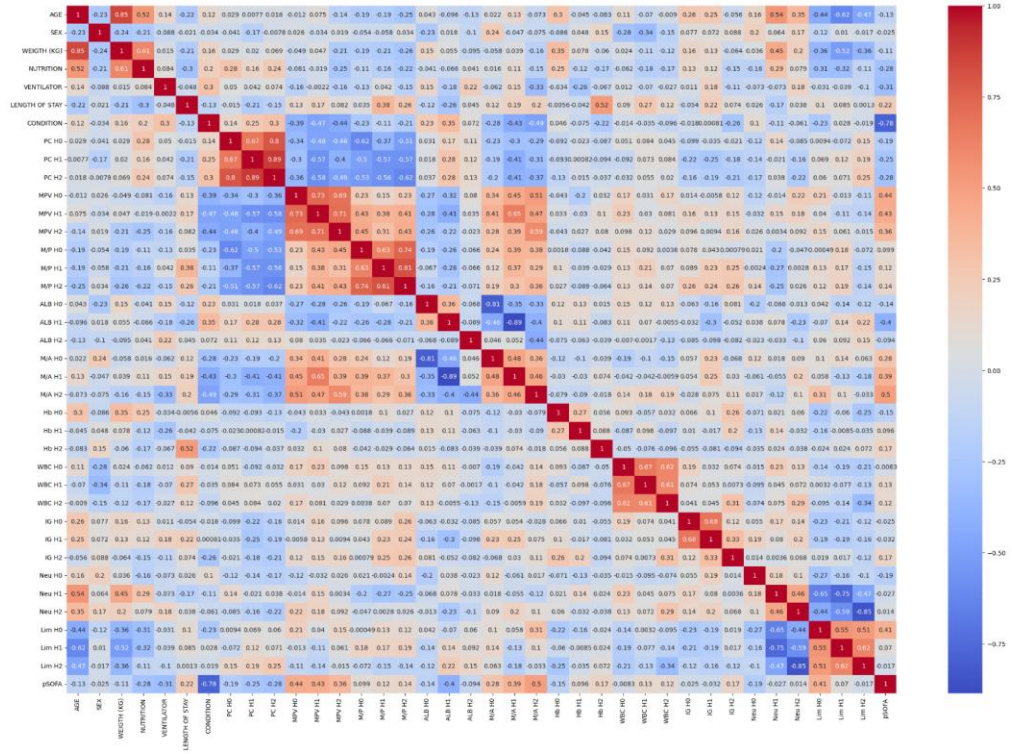


Figure 1- Heatmap correlation among variables

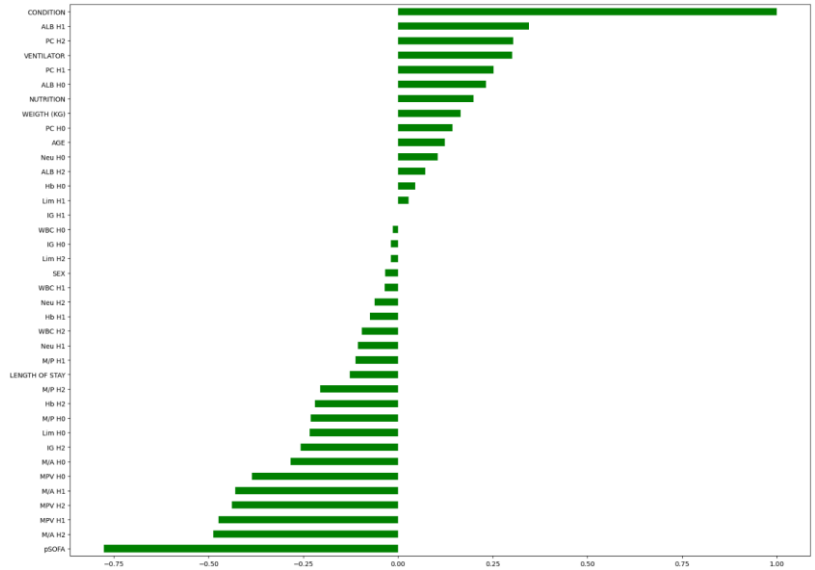


Figure 2- Horizontal bar chart of feature-outcome correlation

Moreover, Random Forest and XGBoost were combined using soft voting and stacking approaches. Random Forest used 100 trees and a maximum depth of 3, while XGBoost used 100 estimators, a maximum depth of 3, and a learning rate of 0.05, which both used a random seed of 42.

Stacking employed five-fold internal cross-validation and either logistic regression with default settings, with or without passthrough, or an XGBoost meta-learner with 10 estimators, a maximum depth of 1, a learning rate of 0.1, and a random seed of 42. The voting classifier used soft voting.

As shown in Table 2, soft voting achieved a mean cross-validation accuracy of 88.3% (SD 4.1%). Stacking with logistic regression, an XGBoost meta-learner, and logistic regression with passthrough achieved mean accuracies of 85.0% (SD 8.2%), 80.0% (SD 4.1%), and 90.0% (SD 8.2%), respectively. Stacking with logistic regression and passthrough therefore had the highest mean cross-validation accuracy among the evaluated ensemble configurations.

In a separate evaluation on the 12-patient test set, the stacking classifier with logistic regression and passthrough correctly classified both deaths and all ten survivors.

With death defined as the positive outcome, its confusion matrix comprised two true positives (TP), ten true negatives (TN), zero false positives (FP), and zero false negatives (FN). Although this yielded good test accuracy, the result should be interpreted with extreme caution because the test set included only two deaths.

Discussion

The best-performing model had a reported ROC AUC of 0.6979, approximately 0.70 (Figure 3). The reported AUC was lower, at 0.62, in the external dataset. These estimates indicate modest discrimination and give a more cautious picture of performance than the perfect classification observed in the small test set. The lower AUC in the additional cohort also raises concerns about how well the model performs on other patients. The XGBoost feature-importance analysis ranked pSOFA as the leading contributor, followed by MPV and platelet count measurements (Figure 4).

The current findings do not establish clinical utility, even for triage. No universal minimum AUC or F1-score defines clinical acceptability. This depends on the intended use and the consequences of false-negative and false-positive predictions. Prospective validation, calibration assessment, and evaluation of clinically acceptable sensitivity are required before implementation.

These findings suggest that organ dysfunction and platelet-related measurements contributed to the fitted model. However, gain-based importance does not establish independent prognostic significance or causal relationships.

Platelet activation and sepsis-associated coagulopathy provide a biological rationale for investigating platelet-related predictors [11-12].

Pediatric studies have evaluated MPV changes [13] and MPV/P [14], although their populations and measurement schedules differ from those of this postoperative cohort. Evidence concerning IG suggests that its incremental value beyond other blood count measurements may be limited [15].

Albumin has also been investigated in children undergoing abdominal surgery [7]. These studies support the selection of candidate predictors but do not validate the present model or establish the clinical usefulness of individual markers.

Clinical deployment may introduce alert fatigue, algorithmic bias, limited interpretability, and excessive reliance on predictions. Prospective evaluation is needed before these models can support clinical judgment in pediatric intensive care.

This study has several limitations. The small cohort of 60 patients, including only 12 patients in the test set, limits the precision of performance estimates and increases the risk of overfitting and model-selection optimism. Cross-validation accuracy also varied across folds (SD 0.082 for stacking with logistic regression and passthrough), indicating unstable performance estimates. No formal sample size calculation was performed, and 95% confidence intervals for AUC and F1-score were not calculated, and ROC analysis was limited to the selected model, leaving uncertainty in these estimates unquantified. Calibration and fairness across age, sex, and nutritional-status subgroups were not assessed.

The retrospective single-center design also limits generalizability. Although laboratory measurements were collected at discrete time points, continuous time-series data were unavailable, limiting the assessment of evolving clinical trajectories and real-time prediction.

The inclusion of patients with complete medical records may introduce selection bias, and the extent of missing data and the procedures used to address missingness require clearer documentation. Prospective validation has not been performed. Larger multicenter cohorts and prospective evaluation are needed to assess the reproducibility and clinical utility of the models.

Table 1- Classification performance of individual ML model.

Model	Accuracy		Class 0 F1		Class 1 F1		Macro Avg F1		Weighted Avg F1	
	Initial	Tuning	Initial	Tuning	Initial	Tuning	Initial	Tuning	Initial	Tuning
SVM	0.83	0.83	0.00	0.00	0.91	0.91	0.45	0.45	0.76	0.76
KNN	0.83	0.83	0.00	0.00	0.91	0.91	0.45	0.45	0.76	0.76
Random Forest	0.83	0.83	0.00	0.50	0.91	0.90	0.45	0.70	0.76	0.83
XGBoost	0.83	0.83	0.00	0.50	0.91	0.90	0.45	0.70	0.76	0.83

Table 2- All scores from cross-validation of the ensemble model

Model	Mean Accuracy	Std Dev	All Scores
Voting Classifier	0.8833	0.0408	[0.9167, 0.9167, 0.9167, 0.8333, 0.8333]
Stacking: LogisticRegression	0.8500	0.0816	[0.9167, 0.9167, 0.9167, 0.75, 0.75]
Stacking: XGBoost meta-learner	0.8000	0.0408	[0.8333, 0.8333, 0.8333, 0.75, 0.75]
Stacking: LogisticRegression + passthrough	0.9000	0.0816	[1.0, 1.0, 0.8333, 0.8333, 0.8333]

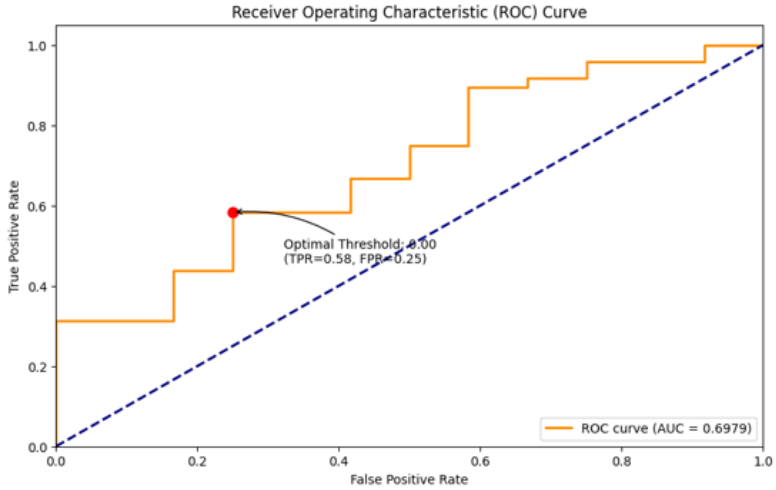


Figure 3- ROC Curve for the Best-Performing Model

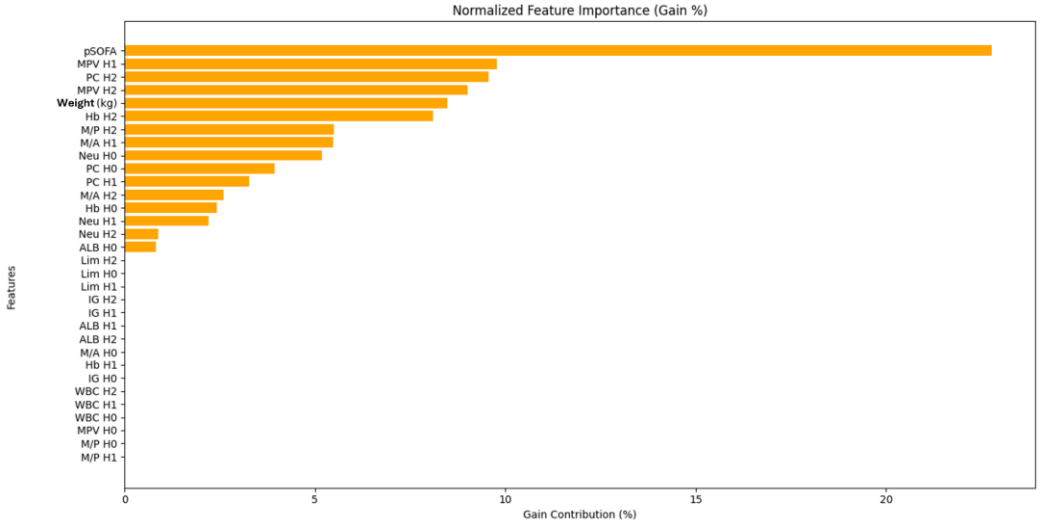


Figure 4- Feature contribution from the XGBoost model

Conclusion

In this retrospective pilot study, stacking with logistic regression and passthrough achieved the highest mean cross-validation accuracy among the ensemble models. Its perfect test accuracy should not be overinterpreted, as the test set contained only 12 patients, including two deaths. These exploratory findings support further investigation but do not establish clinical utility. Larger independent cohorts and prospective validation are required before clinical use.

References

[1] Markwart R, Saito H, Harder T, Tomczyk S, Cassini A, Fleischmann-Struzek C, et al. Epidemiology and burden of sepsis acquired in hospitals and intensive care units: a systematic review and meta-analysis. Intensive Care Med. 2020; 46(8):1536-1551.

[2] Miranda M, Nadel S. Pediatric sepsis: a summary of current definitions and management recommendations. Curr Pediatr Rep. 2023; 11(2):29-39.

[3] Ten Broek RPG, Krielen P, Di Saverio S, Coccolini F, Biffl WL, Ansaloni L, et al. Bologna guidelines for diagnosis and management of adhesive small bowel obstruction (ASBO): 2017 update of the evidence-based guidelines from the world society of emergency surgery ASBO working group. World J Emerg Surg. 2018; 13:24.

[4] Ylimartimo AT, Nurkkala J, Koskela M, Lahtinen S, Kaakinen T, Vakkala M, et al. Postoperative complications and outcome after emergency laparotomy: a retrospective study. World J Surg. 2023; 47(1):119-129.

[5] Labgaa I, Joliat GR, Kefleyesus A, Mantziari S, Schäfer M, Demartines N, et al. Is postoperative decrease of serum albumin an early predictor of complications after major abdominal surgery? A prospective cohort study in a European centre. BMJ Open. 2017; 7(4):e013966.

[6] Sinaga RH, Utariani A, Wardhani P, Hardiono H. Immature granulocyte and mean platelet volume as a predictor of 30-day postoperative mortality in patients with sepsis caused by peritonitis. Bali J Anesth. 2020; 4(4):166-71.

[7] Dewi R, Silitonga FGM, Mangunatmadja I. Impact of albumin levels on clinical outcomes in children underwent abdominal surgery. Paediatr Indones. 2020; 60(3):149-53.

[8] Ranjbarzadeh R, Bagherian Kasgari A, Jafarzadeh Ghoushchi S, Anari S, Naseri M, Bendechache M. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Sci Rep. 2021; 11(1):10930.

[9] Poncette AS, Mosch L, Spies C, Schmieding M, Schiefenhövel F, Krampe H, et al. Improvements in patient monitoring in the intensive care unit: survey study. J Med Internet Res. 2020; 22(6):e19091.

[10] Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology. 2020; 132(2):379-394.

[11] Vélez-Páez JL, Legua P, Vélez-Páez P, Irigoyen E, Andrade H, Jara A, et al. Mean platelet volume and mean platelet volume to platelet count ratio as predictors of severity and mortality in sepsis. PLoS One. 2022; 17(1):e0262356.

[12] Tsantes AG, Parastatidou S, Tsantes EA, Bonova E, Tsante KA, Mantzios PG, et al. Sepsis-induced coagulopathy: an update on pathophysiology, biomarkers, and current guidelines. Life (Basel). 2023; 13(2):350.

[13] Lestari RP, Sumadiono S, Arguni E. An increase in mean platelet volume (MPV) as a predictor of mortality in children with sepsis. J Med Sci. 2022; 54(1).

[14] Bafna T, Sarangi BU, Walimbe A, Shankar GH. Neutrophil-to-lymphocyte ratio (NLR) and Mean platelet volume (MPV) to platelet count ratio in the prognosis of neonatal and pediatric sepsis. Indian J Pediatr. 2026; 93(1):30-5.

[15] Nguyen VT, Pham NH, Le H, Vo QB, Manh TT, Phan HD, et al. Predicting septic shock and mortality in pediatric sepsis using CBC indices with and without immature granulocyte: a prospective study. Egypt Pediatr Assoc Gaz. 2025; 73(1):134.