Pediatric Sepsis In-Hospital Mortality Prediction using Machine Learning

Quang Tung Chu
Supervisor: Assoc. Prof. Tran Thi Thanh Hai
School of Electrical and Electronic Engineering, HUST
tung.cq213594@sis.hust.edu.vn

Abstract

Pediatric sepsis remains a critical cause of in-hospital mortality, especially in low- and middle-income countries (LMICs) where timely and accurate clinical decision-making is challenging. This study addresses the problem of early mortality prediction among pediatric sepsis patients by developing machine learning models that utilize structured data available at hospital admission. We use a synthetic dataset released in the 2024 Pediatric Sepsis Data Challenge, which simulates real-world clinical, demographic, and social variables for children under five. Our pipeline includes data exploration, preprocessing, feature engineering, data augmentation and classifier training. The best-performing model, based on gradient boosting, demonstrates strong predictive capability in identifying high-risk cases. We benchmark its performance against traditional logistic regression and resampling-based ensemble approaches, showing improved recall and precision under optimized thresholding. These results suggest that data-driven prediction can serve as a valuable tool to support early triage in resource-limited pediatric care settings.

1. Introduction

Pediatric sepsis remains a major global health concern, especially in low- and middle-income countries (LMICs), where it accounts for a significant proportion of childhood mortality. It is characterized by a dysregulated immune response to infection, leading to life-threatening organ dysfunction. Timely identification of high-risk cases at hospital admission is crucial for effective clinical intervention, yet pediatric-specific tools for early risk stratification remain limited—particularly in resource-constrained settings.

Conventional scoring systems such as PELOD-2 and PRISM have been widely used in pediatric intensive care units, but they often require laboratory tests or clinical parameters not immediately available at admission. Moreover, these tools lack generalizability across populations, as their performance is typically validated in high-resource environments. Recent efforts have turned to machine learning (ML) models to enhance prediction accuracy. However, many existing ML models are trained on local datasets and are prone to overfitting, reducing their applicability to new hospitals or regions.

To address these challenges, this project proposes a structured ML-based pipeline that predicts in-hospital mortality in pediatric sepsis patients using data available at admission. Our approach emphasizes transparency, clinical relevance, and robustness to missingness and class imbalance—common challenges in LMIC datasets. We validate our method using a publicly released synthetic dataset from the 2024 Pediatric Sepsis Data Challenge, which simulates admission data from hospitalized children under five years of age in Uganda. The pipeline includes exploratory data analysis, feature engineering, model training with various augmentation strategies, and comprehensive evaluation. Compared to traditional approaches, our method introduces improved sensitivity and interpretability, offering a feasible pathway for clinical deployment in low-resource healthcare environments.

2. Related Work

Several studies have explored machine learning approaches for pediatric sepsis mortality prediction, offering important insights into both methodological advances and persistent limitations.

Le *et al*. [13] demonstrated that structured physiological and laboratory features—particularly heart rate, blood pressure, and lactate—carry significant predictive power, achieving an AUROC of 0.916 that outperformed traditional clinical scores like PELOD-2 and SIRS. However, their model was constrained by a narrow feature space, focused exclusively on acute physiological markers without considering contextual or socioeconomic variables, and relied on a single-center U.S. dataset, limiting its generalizability. Their preprocessing pipeline was similarly basic, relying on simple imputation and scaling without addressing outliers, class imbalance, or feature noise.

Dewan *et al*. [3] proposed an EHR-integrated clinical de-

cision support (CDS) tool focused on early sepsis detection in the PICU. Their approach incorporated key bedside clinical assessments—including respiratory status (oxygen supplementation and SpO_2), perfusion indicators (capillary refill), and neurologic status (coma assessment via the Blantyre Coma Scale)—reflecting a stronger emphasis on functional clinical signs. However, the CDS system was partially dependent on subjective clinician input, such as the suspicion of infection, which reduces scalability and reproducibility. Additionally, their preprocessing remained rule-based and did not systematically address data noise or imbalance.

Broader reviews by Yuniar *et al*. [22] and Tennant *et al*. [21] have further emphasized that existing models frequently overlook key contextual variables such as infection status (e.g., HIV, malaria), nutritional status, or basic functional indicators like respiratory distress. Both reviews criticize the field's tendency to either exclude these features due to perceived noise or data availability issues, or to overly rely on institution-specific variables that hinder generalizability, particularly in LMICs. They also point out poor handling of missing data, insufficient outlier detection, and inconsistent inclusion of derived physiological indices like mean arterial pressure (MAP) or shock index.

These insights directly informed our feature selection strategy. Specifically, we aimed to integrate vital signs (heart rate, respiratory rate, blood pressure, temperature, SpO_2), derived indices (MAP and shock index), neurological assessments (Blantyre Coma Scale and coma flag), oxygenation status (supplemental oxygen use and hypoxia), infection indicators (HIV and malaria status), and key laboratory markers such as lactate and hematocrit. This selection is designed to balance the practical constraints of data collection in LMICs with the clinical relevance established in prior literature. Importantly, it also addresses the methodological gaps identified in previous works regarding underutilization of simple but powerful bedside clinical indicators and the lack of a standardized preprocessing pipeline.

3. Proposed Method

3.1. Data Ingestion and cleaning

All analyses were conducted on the publicly-released **Synthetic Paediatric Sepsis Training Set** ($\approx 13\,000$ encounters; ~ 300 candidate variables).

1. Junk–token normalisation Common placeholder strings (empty string, single space, "NA/na", "Unknown/unknown", "–") were converted to IEEE-compliant `NaN`. This harmonisation is essential for the statistics computed by *SimpleImputer* downstream and follows recommended practice for EHR curation [14].

2. Explicit binary mapping Nineteen Boolean variables recorded as *Yes/No* (e.g. `spo2onoxy_adm`) and 37 check-box items (*Checked/Unchecked*) were losslessly recoded to the compact integer domain $\{0, 1\}$. Early casting reduces memory overhead and enables direct use of *class_weight* heuristics in tree-based learners.

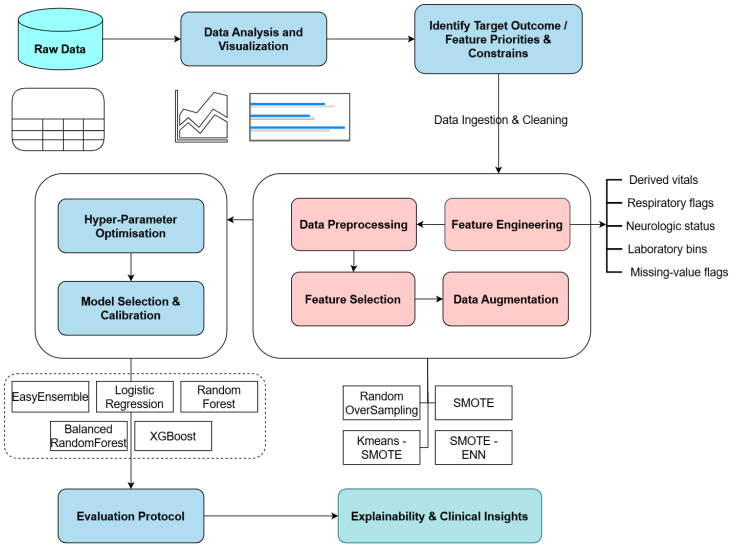


Figure 1. System Pipeline

3. Categorical typing Remaining string columns were promoted to `category` dtype when they exhibited < 50 distinct values *or* a unique/non-null ratio below 0.50, approximating the bias–variance sweet-spot reported by Kuhn & Johnson [11] for mixed-scale tabular data.

4. Rare-level squashing For each categorical variable, levels supported by fewer than five observations were collapsed into a sentinel `__OTHER__`. This consolidation mitigates sparse dummy columns and stabilises maximum-likelihood estimates without materially eroding signal [16].

5. Feature pruning A total of 102 variables were discarded *a priori*:

- *Intervention leak-features* (e.g. `admitabx_adm___#`) recorded after admission;
- *Identifiers* (e.g. `studyid_adm`);
- Attributes with $> 30\%$ missingness, where imputation risked systematic bias;
- Socio–economic surrogates (household fuel, water source, *etc.*) deemed outside the physiological scope of the model.

The retained feature set comprised 198 predictors (94 numeric, 104 categorical).

6. Physiological sanity checks Numeric ranges were validated against paediatric reference intervals. Values outside plausible bounds (e.g. heart rate < 20 or > 300 bpm; systolic BP < 40 mmHg) were replaced with `NaN`; logical constraints (systolic $\geq$ diastolic) were enforced. This conservative masking conforms to the outlier-handling guidance of Goldstein *et al*. [7].

The resulting tidy matrix contained **198 features and 12 973 rows**, ready for imputation, scaling and model development described in §3.9.

3.2. Data Analysis and Visualization

A comprehensive data analysis and visualization process was conducted to assess data quality, understand variable distributions, and examine the relationships between features and the target outcome, in-hospital mortality. Missing data patterns were visualized using heatmaps and bar plots to identify variables with substantial sparsity, supporting decisions on removal or imputation strategies [9]. In addition, a binary variable detection function was implemented to ensure accurate encoding for variables with strictly binary or near-binary distributions.

Distributional analysis was applied to key physiological variables, including heart rate, respiratory rate, blood pressure, SpO_2, lactate, and capillary refill, using histograms and boxplots stratified by mortality outcome. The results revealed that non-survivors typically exhibited abnormal physiological patterns, such as elevated lactate and heart rate, hypotension, hypoxia, and prolonged capillary refill, consistent with sepsis severity profiles. Moreover, heart rate was evaluated against age-specific normal thresholds based on the Pediatric Advanced Life Support (PALS) guidelines [1], demonstrating that a significant proportion of deceased patients had heart rates exceeding the normal physiological range for their age.

Feature-outcome associations were quantified using Pearson correlation for continuous variables and Cramér's V for categorical variables [6]. Variables such as lactate, SpO_2, heart rate, capillary refill, and mental status showed the strongest associations with mortality, which aligns with established clinical risk indicators for sepsis [5]. Additionally, a comorbidity risk analysis was performed by comparing mortality rates between patients with and without each comorbidity, visualized using grouped bar plots to highlight absolute mortality differences and risk ratios.

This analytical process provided critical insights into feature relevance and data integrity, directly supporting the subsequent feature selection and model development stages.

3.3. Identify Target Outcomes and Feature Priorities & Constrains

The primary target outcome in this study is in-hospital mortality, defined as a binary variable (0: survived, 1: deceased). The class distribution was evaluated using frequency tables and visualizations, which revealed significant class imbalance — a common challenge in pediatric sepsis datasets [13, 21, 22].

Feature prioritization was informed by both clinical relevance and statistical associations with the outcome. Based on established clinical guidelines such as the Pediatric Advanced Life Support (PALS) [1] and recent scoping reviews on pediatric sepsis prediction [21, 22], Tier-1 features were selected to include key physiological indicators highly relevant to patient deterioration at admission. These include heart rate, respiratory rate, systolic and diastolic blood pressure, SpO_2, lactate, capillary refill, and verbal response score. These variables are strongly linked to sepsis-related circulatory and organ dysfunction [3, 13, 22]. Tier-2 features such as age, hematocrit, and temperature were selected as supporting variables, while Tier-3 features capture the patient's comorbidity profiles.

To ensure data quality, a comprehensive constraint assessment was conducted. Missing data patterns were analyzed using heatmaps and missingness thresholds to identify features with excessive sparsity, which were either flagged for removal or for imputation [9].

Binary feature detection was implemented to accurately classify features with strict or near-binary distributions, ensuring correct encoding for downstream modeling. Redundancy checks were conducted on features with duplicated physiological measurements, particularly SpO_2 readings from two separate sensor sites, using scatter plots to verify measurement consistency.

Physiological constraint checks were applied by evaluating heart rate against age-dependent normal ranges according to PALS guidelines [1], allowing identification of clinical outliers indicative of patient instability (e.g., tachycardia or bradycardia).

This process ensured that the target outcome was clearly defined, features were prioritized based on clinical significance and predictive strength, and data quality constraints were systematically addressed prior to feature selection and model development.

3.4. Feature Engineering

The feature engineering strategy in this study was carefully designed to capture physiological, neurological, and metabolic indicators that are both predictive of in-hospital mortality and feasible to collect at the time of admission, particularly in low- and middle-income country (LMIC) settings. This approach addresses limitations highlighted in previous studies [21, 22], where models often narrowly focused on isolated physiological variables [13] or failed to incorporate bedside clinical assessments and contextual risk factors [3].

Given the substantial physiological variability across pediatric age groups, age normalization was a key priority. Heart rate (HR) was adjusted based on Pediatric Advanced Life Support (PALS) upper limits [1], allowing the model to accurately capture tachycardia relative to a child's developmental stage. This age-adjusted HR percentage was further binarized into a tachycardia flag if HR exceeded 110% of the age-specific threshold, reflecting clinically meaningful

thresholds used in triage and critical care.

Cardiovascular stability was represented not only by raw vital signs (heart rate, systolic and diastolic blood pressure) but also by derived indices. Mean arterial pressure (MAP) was computed using the standard formula MAP, which is provided in Table 1, to better represent tissue perfusion pressure. Shock Index (SI) serves as a well-established predictor of hemodynamic instability in sepsis.

Oxygenation status was captured through a combination of peripheral oxygen saturation (SpO_2) and supplemental oxygen usage, which was encoded from both numeric codes and textual labels in the dataset. A hypoxia flag was generated if SpO_2 was below 90% on room air, aligning with WHO clinical guidelines for hypoxemia in children [19].

Neurological status was quantified using the Blantyre Coma Scale (BCS), which assesses motor, eye, and verbal responses. Each component was mapped to a numerical score (0–2), and a total BCS score (0–5) was computed. A binary coma flag was set for BCS ≤ 2, consistent with definitions of deep coma in pediatric critical care literature [3].

Metabolic derangement was represented by serum lactate levels, a well-established predictor of tissue hypoperfusion and mortality in sepsis [21, 22]. Lactate was incorporated both as a continuous variable and as a categorical variable with clinically meaningful bins (normal < 2 mmol/L, moderate 2–4 mmol/L, and high > 4 mmol/L). Hematocrit was also retained as a marker of oxygen-carrying capacity and potential anemia-related risk.

To further enhance robustness, missing data indicators were created for all Tier-1 features, following recommendations from [21]. This acknowledges the clinical reality that missingness in electronic health records is often non-random and can itself be predictive of adverse outcomes, particularly in LMIC settings.

Rigorous physiological plausibility checks were applied. Measurements were set to missing if they fell outside clinically valid ranges—heart rate (20–300 bpm) [1, 21], systolic BP (40–250 mmHg), diastolic BP (20–150 mmHg) [18], temperature (30–45°C), and SpO_2 (0–100%) [19]. Logical constraints, such as ensuring systolic pressure exceeded diastolic pressure, were enforced to maintain clinical validity.

A detailed summary of all feature transformations, mapping rules, and clinical rationales is presented in Table 1.

3.5. Data Preprocessing

Following feature engineering, we applied a comprehensive data preprocessing pipeline to ensure data quality, consistency, and readiness for machine learning. This process addresses common challenges in electronic health records (EHR) data from low- and middle-income countries (LMICs), including missing values, outliers, data inconsistency, and class imbalance [21].

Standardizing categorical and binary variables is a fundamental requirement in clinical machine learning to ensure data consistency, computational efficiency, and algorithm compatibility. Raw clinical data frequently stores categorical information as string formats—such as "Yes/No", "Checked/Unchecked", or descriptive labels for clinical statuses. If left unprocessed, these string-based variables can lead to inconsistencies, increase memory usage, and cause errors in modeling pipelines that expect numerical inputs. Explicitly mapping binary variables to numeric representations (e.g., 0/1) preserves their semantic meaning while ensuring compatibility with modeling frameworks. Similarly, converting multi-level categorical variables from string to categorical types improves memory efficiency and statistical reliability. Additionally, rare categories within high-cardinality variables may introduce noise and overfitting risks if not handled properly. Grouping such infrequent levels into an "Other" class mitigates this risk, enhancing model robustness. This transformation is crucial for maintaining data integrity, interpretability, and reproducibility in clinical machine learning workflows [3, 21, 22].

Missing data was handled using a dual approach. For all Tier-1 clinically important features, including vital signs (heart rate, respiratory rate, blood pressure, SpO_2), laboratory markers (lactate, hematocrit), and neurological assessments (Blantyre Coma Scale components), we generated binary missingness indicators. This strategy allows the model to capture predictive signals from missing patterns, which often correlate with clinical severity or data collection constraints [21]. Subsequently, we applied median imputation for numerical variables, which is robust to outliers and skewed distributions, and mode imputation for categorical variables to preserve the most common observed value.

To ensure physiological plausibility, we performed rigorous outlier detection and consistency checks based on established pediatric clinical norms [1, 18, 19]. Specifically, values were set to missing if they fell outside accepted ranges: heart rate below 20 or above 300 bpm, systolic blood pressure outside 40–250 mmHg, diastolic blood pressure outside 20–150 mmHg, temperature below 30°C or above 45°C, and SpO_2 outside 0–100%. Logical consistency was further enforced by ensuring that systolic blood pressure exceeded diastolic pressure; any violations were corrected by treating them as missing. Additionally, common data entry errors such as sentinel values (e.g., 999, -999) were systematically identified and converted to missing.

For categorical features, we applied one-hot encoding with the first category dropped to prevent multicollinearity. Rare categories with fewer than five occurrences were consolidated into a single "__OTHER__" class, reducing noise and preventing the creation of sparse feature representations. Binary variables, including yes/no and

Table 1. Summary of Feature Engineering Rules

Feature	Description	Mapping / Formula	Clinical Rationale
MAP (Mean Arterial Pressure)	Average blood pressure during a cardiac cycle	$\text{MAP} = \frac{\text{SBP} + 2 \times \text{DBP}}{3}$	Captures overall perfusion pressure. Low MAP indicates shock or circulatory failure.
Shock Index	HR relative to SBP	$\text{SI} = \frac{\text{HR}}{\text{PALS Upper Limit (age)}}$	Elevated SI is a marker of hemodynamic compromise and poor outcome.
Age-adjusted HR Percentage	HR normalized to age-specific upper limit	$\frac{\text{HR}}{\text{PALS Upper Limit (age)}}$ Upper limits: <3 mo $\rightarrow$ 205 bpm; 3–24 mo $\rightarrow$ 190 bpm; 2–10 yr $\rightarrow$ 140 bpm; $\geq$10 yr $\rightarrow$ 100 bpm	Accounts for pediatric physiological HR variation. Detects tachycardia relative to age.
Tachycardia Flag	Indicates HR exceeds 110% of age limit	1 if HR $>$ 110% of upper limit, else 0	Detects clinically significant tachycardia based on age.
On Oxygen Flag	Whether patient is on supplemental O_2	Mapping: 1 or 'O_2 available and being used" $\rightarrow$ 1; 2 or 'O_2 available not used" $\rightarrow$ 0; 3 or 'O_2 not enough" $\rightarrow$ 1; 4 or 'O_2 not available" $\rightarrow$ 0	Key for interpreting SpO_2 and assessing respiratory failure.
Hypoxia Flag	Low SpO_2 on room air	1 if $SpO_2 < 90\%$ and not on oxygen	Identifies hypoxemia based on WHO cutoffs. Indicates severe respiratory dysfunction.
Blantyre Coma Scale (BCS)	Neurological status components	**Motor:** Localizes pain $\rightarrow$ 2; Withdraws $\rightarrow$ 1; No response $\rightarrow$ 0. **Eye:** Watches $\rightarrow$ 2; Fails to watch $\rightarrow$ 1; Eyes do not open $\rightarrow$ 0. **Verbal:** Cries/speaks $\rightarrow$ 2; Moans $\rightarrow$ 1; No vocal response $\rightarrow$ 0.	Assesses consciousness level. Critical for identifying neurologic dysfunction and coma.
BCS Total	Sum of BCS components	Sum of Motor + Eye + Verbal, range 0–5	Lower scores indicate coma. Threshold $\leq$ 2 for deep coma.
Coma Flag	Indicates coma state	1 if BCS $\leq$ 2, else 0	Identifies patients with severe neurological impairment (deep coma).
Lactate Category	Binned lactate levels	Normal: <2 mmol/L; Moderate: 2–4 mmol/L; High: >4 mmol/L	Lactate reflects tissue hypoperfusion. Strong predictor of mortality.
Missing Flags	Indicates missing data for critical features	1 if missing, 0 otherwise	Missingness itself can be predictive of clinical risk, especially in LMIC datasets.

checked/unchecked types, were mapped to integer values (0/1) and stored using memory-efficient `Int8` data types to maintain consistency and optimize computational performance.

All continuous numerical features were standardized using z-score normalization to ensure zero mean and unit variance. This scaling procedure improves convergence stability for gradient-based algorithms and ensures that features are comparable in magnitude. Categorical variables, encoded as binary indicators through one-hot encoding, were left unscaled to preserve interpretability.

Finally, the dataset was partitioned into training and testing subsets using an 80/20 split, stratified by the in-hospital mortality outcome to maintain class balance across both sets. This stratification is particularly important given the inherent imbalance in sepsis mortality datasets [21, 22], ensuring that the minority class remains adequately represented for both model training and evaluation.

Overall, this data preprocessing strategy complements the feature engineering process by balancing clinical relevance with statistical robustness. It aligns with best practices in clinical machine learning and adheres to guidelines from WHO and AHA [1, 18, 19], providing a reliable foundation for building predictive models in pediatric sepsis.

3.6. Feature Selection

A six–stage hybrid *filter–wrapper* strategy was devised to obtain a parsimonious yet clinically meaningful input space. Each successive step reduces dimensionality while safeguarding an a priori list of domain–essential variables (e.g. *age*, *lactate*) that must remain available for interpretation.

1. Clinical pre–filter. A panel of $p = 17$ expert–defined predictors was marked as *protected* and exempted from further elimination.

2. Low–variance filter. Near–constant features ($\mathrm{Var}(x) < 10^{-3}$) were discarded via the `VarianceThreshold` operator, removing instrumentation artefacts and coding dummies with no discriminative information.

3. Redundancy filter. Pairwise Pearson correlation was computed on the remaining numeric features; one member of any pair with $|\rho| > 0.95$ was removed, preserving the variable that showed the stronger univariate correlation with the target.

4. Univariate relevance filter. Mutual information with the class label was evaluated (`SelectKBest`; $k = \lceil 0.8m \rceil$ where m is the number of candidates after Step 3). This liberal threshold retains $\approx 80\%$ of the strongest individual predictors.

5. Embedded tree importance. A balanced `RandomForest` with 100 trees ranked the surviving attributes; features scoring below the 40^{th} percentile of the normalized Gini importance distribution were pruned, i.e. the top 60% were kept.

6. Recursive Feature Elimination (RFE). A second `RandomForest` wrapper iteratively removed the least important variables (step = 3) until a target set of $n^* \in [80, 120]$ attributes was reached—a deliberate balance between parsimony and signal retention.

Finally, the protected clinical variables omitted by the automatic filters were re–introduced, yielding the definitive matrix $\mathbf{X}_{\mathrm{FS}} \in \mathbb{R}^{N \times d}$ with $d = 65$ columns. The overall reduction rate is

$$r = 1 - \frac{d}{p_0} = 42.3\%, \tag{1}$$

where $p_0 = 120$ denotes the original dimensionality prior to selection. The procedure retained all protected variables and removed only 10 constant and 18 redundant features, confirming its *conservative* character.

3.7. Data Augmentation

To address the class imbalance problem in predicting in-hospital mortality for pediatric sepsis patients, we applied three data augmentation techniques designed for tabular data: Random Oversampling (ROS), KMeans-SMOTE, and SMOTE-ENN. These techniques were implemented after data preprocessing to enhance the representation of the minority class (mortality) in the training set.

Random Oversampling balances the dataset by randomly duplicating minority class samples [2]. Given a dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \text{ where } y_i \in \{0, 1\}$$

with $y = 1$ denoting the minority class and $y = 0$ the majority class, ROS generates an augmented dataset

$$\mathcal{D}^* = \mathcal{D} \cup \{(x_j, y_j) \mid y_j = 1, j \in \mathcal{S}\}$$

where $\mathcal{S}$ is a set of randomly sampled indices from the minority class (with replacement) such that

$$|\{y_i = 1\}|^* = |\{y_i = 0\}|$$

While simple, ROS introduces no new information and increases the risk of overfitting when the minority class is small.

KMeans-SMOTE improves upon standard SMOTE by incorporating clustering before sample generation [4, 12]. The minority class is first partitioned into k clusters:

$$\{C_1, C_2, \ldots, C_k\} = \text{KMeans}(\{x_i | y_i = 1\})$$

Synthetic samples are then generated within each cluster using

$$x_{\text{new}} = x_a + \lambda \cdot (x_b - x_a)$$

where $x_a, x_b \in C_j$ and $\lambda \sim U(0,1)$. The number of synthetic samples per cluster is inversely proportional to the cluster's density:

$$n_{\text{synthetic},j} = \alpha \cdot \frac{1}{\text{density}(C_j)}$$

This approach preserves local data structures and avoids generating unrealistic points in dense regions.

SMOTE-ENN combines synthetic oversampling with data cleaning via Edited Nearest Neighbors (ENN) [2, 8, 15]. The SMOTE step generates new samples with

$$x_{\text{new}} = x_i + \lambda \cdot (x_i^{\text{NN}} - x_i)$$

where x_i is a minority instance, x_i^{NN} is one of its k-nearest minority neighbors, and $\lambda \sim U(0,1)$.

The ENN step removes instances whose labels disagree with the majority label among their k-nearest neighbors. For a given instance x, the majority label is calculated as

$$\text{MajorityLabel}(x) = \arg\max_{c \in \{0,1\}} \sum_{x_j \in \text{kNN}(x)} \mathbf{1}(y_j = c)$$

If $y \neq \text{MajorityLabel}(x)$, the instance x is removed. This process reduces noise and clarifies the decision boundary.

These augmentation techniques were integrated into the training pipeline after preprocessing and before model learning. Resampling was applied only to the training set within each cross-validation fold to prevent data leakage. To improve robustness, pipelines using KMeans-SMOTE and SMOTE-ENN included fallback mechanisms, switching to standard SMOTE when necessary (e.g., when the number of neighbors was insufficient). This design ensures stable model training and enables a systematic evaluation of how each augmentation method impacts mortality prediction performance.

3.8. Hyperparameter Optimisation

Search engine. For each *sampler–classifier* pair we tuned the free hyper-parameters with a sequential model–based strategy, namely Bayesian optimisation[1]. Let $\boldsymbol{\theta} \in \Theta$ denote a candidate configuration and $\mathcal{L}(\boldsymbol{\theta})$ its validation performance. At iteration t the optimiser selects the next query point $\boldsymbol{\theta}_{t+1}$ by maximising the *expected-improvement* (EI) acquisition function

$$\begin{aligned} \text{EI}(\boldsymbol{\theta}) &= \big[\mu(\boldsymbol{\theta}) - f^\star - \xi\big]\,\Phi\big(z(\boldsymbol{\theta})\big) + \sigma(\boldsymbol{\theta})\,\phi\big(z(\boldsymbol{\theta})\big), \\ z(\boldsymbol{\theta}) &= \frac{\mu(\boldsymbol{\theta}) - f^\star - \xi}{\sigma(\boldsymbol{\theta})}, \end{aligned} \tag{2}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the predictive mean and standard deviation of the Gaussian-process surrogate, $f^\star = \max_{i \le t} \mathcal{L}(\boldsymbol{\theta}_i)$ is the incumbent best score, $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF/PDF, and $\xi \ge 0$ encourages exploration. If Θ collapses to a singleton (e.g. for inherently balanced ensembles) the procedure seamlessly falls back to a `RandomizedSearchCV` with 20 draws. **Search Spaces**

Logistic Regression
$C \sim \mathcal{U}(10^{-2}, 10^{2})$
solver $\in \{\texttt{liblinear}, \texttt{saga}\}$
penalty $\in \{\ell_1, \ell_2\}$
max_iter $\sim \mathcal{Z}[1000, 3000]$

Random Forest
$n_{\text{est}} \sim \mathcal{Z}[100, 500]$
max_depth $\in \{10, 15, 20, 25, \textit{None}\}$
min_samples_split $\sim \mathcal{Z}[2, 20]$
min_samples_leaf $\sim \mathcal{Z}[1, 10]$

XGBoost
$n_{\text{est}} \sim \mathcal{Z}[100, 500]$
max_depth $\sim \mathcal{Z}[3, 8]$
η (= learning_rate) $\sim \mathcal{U}(0.01, 0.3)$
subsample $\sim \mathcal{U}(0.6, 1.0)$
colsample_bytree $\sim \mathcal{U}(0.6, 1.0)$

Notation. $\mathcal{U}(a,b)$ denotes a *continuous* uniform prior over (a,b), while $\mathcal{Z}[m,n]$ denotes a *discrete* uniform prior over the integers $m, \ldots, n$.

BalancedRandomForestClassifier

- $n_{\text{estimators}} = 300$
- sampling_strategy = *auto*
- random_state = 42
- $n_{\text{jobs}} = -1$

EasyEnsembleClassifier

- $n_{\text{estimators}} = 10$ *(10 balanced subsets)*
- sampling_strategy = *auto*
- random_state = 42
- $n_{\text{jobs}} = -1$

3.9. Model Selection and Development

To benchmark the proposed feature–engineering and resampling pipeline, five complementary algorithmic families were selected. The set spans linear, tree–based and ensemble paradigms, thereby covering a broad spectrum of bias–variance characteristics and interpretability profiles. Hyper-parameters were configured as described in Section 3.8, with additional class-imbalance adjustments where appropriate. A critical component of our model development pipeline was the design of a rigorous framework that balances predictive accuracy, generalizability, and clinical applicability for pediatric sepsis mortality prediction. This process incorporated both interpretable statistical models

[1] Implemented with `skopt.BayesSearchCV`.

and advanced machine learning algorithms, combined with systematic hyperparameter optimization, probability calibration, and robust cross-validation to ensure reliability under severe class imbalance.

Logistic Regression (LR) was adopted as a transparent baseline owing to its simplicity, rapid training time, and straightforward coefficient-based interpretation. However, the strictly linear decision surface of LR limits its capacity to model the complex, non–linear interactions that frequently arise in paediatric critical-care data.

To capture such higher-order structure we integrated four tree-based ensembles that span both bagging and boosting paradigms:

- **Random Forest (RF).**A classical bagging method that grows an ensemble of decorrelated decision trees on bootstrapped samples and random feature subsets, delivering strong out-of-the-box performance and robustness to noise.
- **Balanced Random Forest (BRF).**An extension of RF that performs implicit class-balanced under-sampling at every bootstrap draw, thereby mitigating the pronounced mortality imbalance without external resampling.
- **EasyEnsemble (EE).**A bagging-of-boosting strategy that trains multiple AdaBoost classifiers on disjoint balanced subsets and averages their predictions, combining the variance reduction of bagging with the bias reduction of boosting.
- **Extreme Gradient Boosting (XGBoost).**A high-performance implementation of gradient-boosted trees that sequentially adds weak learners to minimise a differentiable loss function. XGBoost excels on structured tabular data, offers built-in handling of missing values and sparsity, and includes regularisation terms that curb overfitting, albeit at the cost of an enlarged hyper-parameter search space.

A key aspect of our approach was the construction of tailored hyperparameter grids for each combination of sampling strategy and classifier. This design allowed joint optimization of both sampling-related parameters—such as the number of neighbors and cluster balance threshold in KMeans-SMOTE, or the k-nearest neighbors in SMOTE-ENN—alongside classifier hyperparameters, including tree depth, learning rate, number of estimators, and regularization strength. This joint tuning strategy was motivated by prior findings that sampling methods significantly affect classifier performance, and tuning them independently may lead to suboptimal results [8]. Hyperparameter optimization was performed within a Repeated Stratified K-Fold Cross-Validation framework (3 folds, 2 repeats) to provide stable and unbiased performance estimates while preserving the class distribution within each fold [10].

In standard K-Fold Cross-Validation, the dataset $\mathcal{D}$ containing N samples is partitioned into K mutually exclusive and approximately equal-sized folds:

$$\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k \quad \text{where} \quad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \text{ for } i \neq j$$

For each fold $k = 1, 2, \ldots, K$, the model is trained on the union of $K-1$ folds, denoted $\mathcal{D}_{\text{train}}^{(k)}$, and evaluated on the held-out fold $\mathcal{D}_k$:

$$\mathcal{D}_{\text{train}}^{(k)} = \mathcal{D} \setminus \mathcal{D}_k$$

The performance metric M (e.g., AUROC, AUPRC) is computed for each fold:

$$M^{(k)} = \text{Evaluate}\left(f^{(k)}(\mathcal{D}_{\text{train}}^{(k)}), \mathcal{D}_k\right)$$

The overall performance estimate is the average across all folds:

$$\bar{M} = \frac{1}{K} \sum_{k=1}^{K} M^{(k)}$$

To further enhance robustness, we employed Repeated Stratified K-Fold Cross-Validation, where the K-Fold process is repeated R times with different random partitions. The final performance estimate is then:

$$\bar{M} = \frac{1}{K \cdot R} \sum_{r=1}^{R} \sum_{k=1}^{K} M^{(k,r)}$$

where $M^{(k,r)}$ denotes the performance metric for fold k in repetition r. In this study, we set $K = 5$ and $R = 3$, resulting in 15 distinct train-test splits per model configuration. The stratification ensures that the class distribution between mortality and survival is preserved in each fold, which is crucial for datasets with severe class imbalance.

Given that tree-based models are known to produce poorly calibrated probability estimates [17], we incorporated Platt Scaling as a post-hoc probability calibration method [20]. This approach fits a logistic regression model to the classifier's output scores, transforming them into well-calibrated probabilities. Calibration is particularly important in clinical applications where probabilistic outputs inform risk thresholds and decision-making. Nevertheless, empirical results during development indicated that the uncalibrated XGBoost model consistently outperformed its calibrated counterpart in terms of discrimination metrics, likely due to trade-offs introduced by calibration between sharpness and reliability. Consequently, the final deployed model was the uncalibrated XGBoost, as it demonstrated superior performance across both cross-validation and held-out test evaluations.

All steps in model development, including hyperparameter tuning, probability calibration, and threshold optimization, were performed within the same cross-validation framework to ensure consistency and to prevent data leakage. The pipeline systematically evaluated and compared Logistic Regression, Random Forest, and XGBoost models across four sampling strategies: no sampling (baseline), Random Oversampling (ROS), KMeans-SMOTE, and SMOTE-ENN. In total, 1704 hyperparameter configurations were evaluated, reflecting an exhaustive and rigorous approach that jointly optimized both sampling and learning processes. Threshold optimization was explicitly integrated using Youden's J statistic derived from ROC curves and the maximum F1-score from precision-recall curves, aligning the model's decision boundary with the clinical priority of maximizing sensitivity while balancing specificity and precision. This comprehensive and systematic development process was designed not only to optimize predictive accuracy but also to ensure clinical relevance and applicability in the early identification of high-risk pediatric sepsis patients.

3.10. Evaluation Metrics and Model Selection Criteria

Given the severe class imbalance inherent in the pediatric sepsis dataset, model evaluation relied on a combination of discrimination metrics, calibration measures, and clinically meaningful performance indicators. The primary objective was to assess not only the overall predictive performance but also the ability of the models to reliably identify at-risk patients.

Discrimination performance was assessed using the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). AUROC evaluates a model's ability to distinguish between positive and negative classes across all possible thresholds. Formally, AUROC is defined as the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance:

$$\text{AUROC} = \Pr\left(s^+ > s^-\right)$$

where s^+ and s^- are the model scores for positive and negative instances, respectively.

However, AUROC can present an overly optimistic view in highly imbalanced datasets. Therefore, AUPRC was incorporated as a complementary metric. AUPRC emphasizes the performance of the model in the positive (minority) class, summarizing the trade-off between precision (positive predictive value) and recall (sensitivity). The Precision-Recall (PR) curve plots precision against recall as the decision threshold varies:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{and} \quad \text{Recall} = \frac{TP}{TP + FN}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

In addition to AUROC and AUPRC, traditional performance measures were computed at specific decision thresholds, including sensitivity (recall), specificity, precision (positive predictive value), negative predictive value (NPV), accuracy, and F1-score. These metrics were calculated based on the confusion matrix:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

with the following formulations:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{NPV} = \frac{TN}{TN + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Probability calibration was assessed using the Brier Score, which measures the mean squared difference between predicted probabilities and the true outcomes:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2$$

where p_i is the predicted probability for instance i and $y_i \in \{0, 1\}$ is the true label.

Given the clinical importance of minimizing false negatives in sepsis mortality prediction, model selection prioritized robust sensitivity while maintaining reasonable specificity. Models were primarily ranked based on AUPRC due to the imbalance nature of the dataset. In cases where models exhibited similar AUPRC, secondary criteria included AUROC and F1-score. Furthermore, threshold optimization was conducted using Youden's J statistic from the ROC curve:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

to maximize the trade-off between sensitivity and specificity, as well as the threshold that yields the maximum F1-score from the precision-recall curve.

This evaluation strategy ensures that the final selected model not only achieves high discrimination and reliable calibration but also aligns with the clinical objective of accurately identifying patients at risk of in-hospital mortality.

4. Experiments

4.1. Dataset

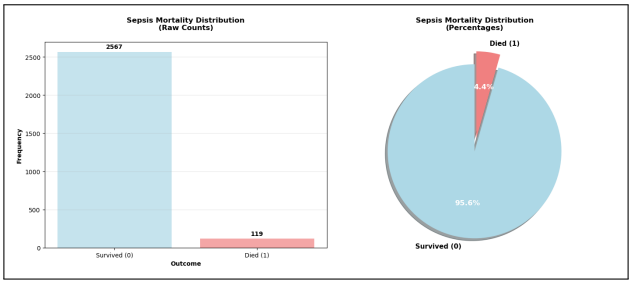


Figure 2. Class imbalance between survived and deceased patients

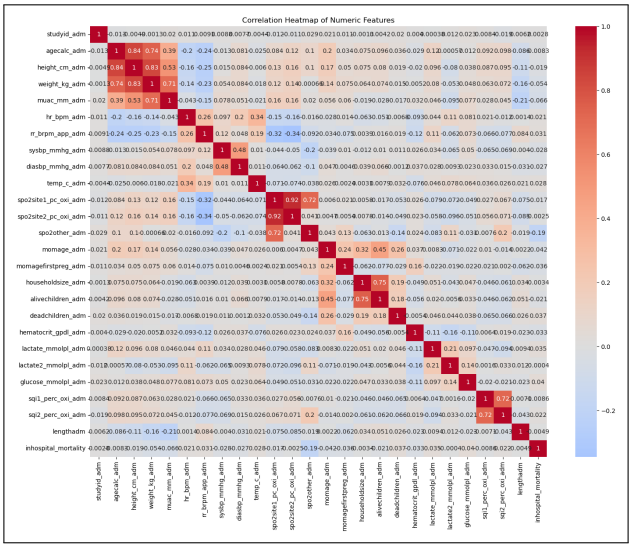


Figure 3. Correlation Heatmap of Numeric Features

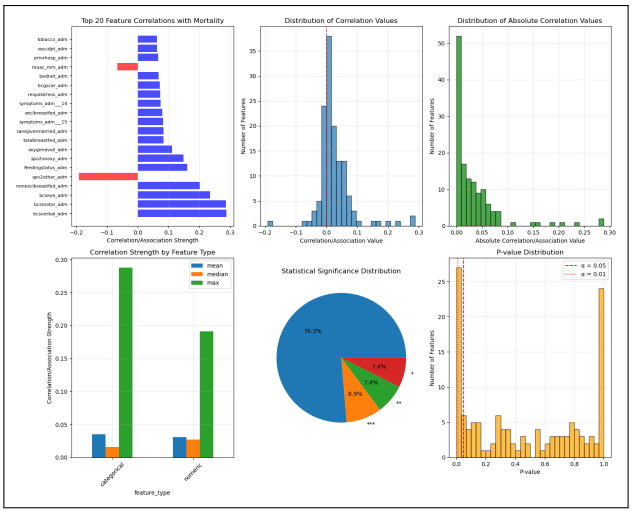


Figure 4. Correlation Statistics

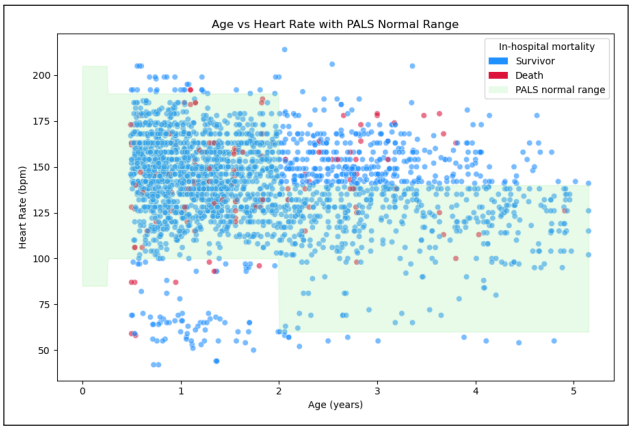


Figure 5. Age vs Heart Rate with PALS Normal Range

The dataset used in this study originates from the 2024 Pediatric Sepsis Data Challenge[2], which provides a synthetic dataset simulating pediatric sepsis cases in Uganda. It consists of 2,686 patient records, each corresponding to a unique hospital visit for a child under five years old. The dataset includes a total of 138 features, among which 26 are numerical and 112 are categorical.

The prediction target is the binary variable `inhospital_mortality`, where 1 denotes in-hospital death and 0 indicates survival. The dataset is highly imbalanced, with only about 4.4% of records labeled as mortality cases, presenting a challenge for standard classifiers. Additionally, missing values are widespread; variables such as `spo2other_adm` and `nonexclbreastfed_adm` have over 90% missingness. Based on an analysis from the visualization files (`vis.ipynb`, `visualisation.ipynb`), features with more than 50% missing values were excluded from modeling.

The dataset also includes variables from diverse categories: demographic (e.g., `agecalc_adm`, `sex_adm`), clinical (e.g., `hr_adm`, `lactate2_mmolpl_adm`), and socioeconomic (e.g., `momeducation_adm`, `householdsize_adm`). To ensure model practicality and avoid data leakage, intervention-related variables (e.g., `admitabx_adm__1` to `admitabx_adm__21`) were excluded, since they reflect post-admission decisions.

In summary, while the dataset is rich and diverse, it introduces notable challenges including class imbalance, high missingness, and synthetic nature. These issues must be addressed carefully during preprocessing and model development to ensure meaningful and generalizable results.

[2] https://github.com/Kamaleswaran-Lab/The-2024-Pediatric-Sepsis-Challenge

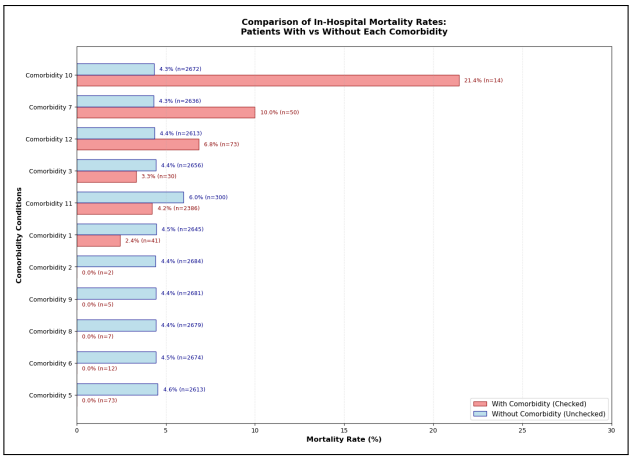


Figure 6. Comparison of In-Hospital Mortality Rates: Patients With vs Without Each Comorbidity

4.2. Implementation Details

4.2.1. Environmental Setup

All experiments were executed on a workstation running *Windows 11* (build 26100) with an *Anaconda 3* Python interpreter - Python 12.6. Development and interactive exploration were performed in `Jupyter Notebook`, while scripts were executed via `PowerShell`. The plotting backend was configured through `%matplotlib inline`, and non–critical runtime warnings were suppressed using `warnings.filterwarnings("ignore")`.

4.2.2. External Libraries

Key Python Libraries

- **Data manipulation:** `pandas 1.x` — data wrangling and file I/O.
- **Numerics:** `numpy 1.26.4` — fast vectorised numerical operations.
- **Visualisation:** `matplotlib` — exploratory and publication-quality plots.
- **Machine-learning core:** `scikit-learn 1.5.1` — preprocessing, model selection, base estimators.
- **Imbalanced learning:** `imbalanced-learn` — resampling strategies and balanced tree ensembles.
- **Gradient boosting:** `XGBoost` — high-performance gradient-boosted decision trees.
- **Bayesian optimisation:** `scikit-optimize` — sequential hyper-parameter search.
- **Explainability:** `SHAP` — post-hoc model interpretation via Shapley values.

4.2.3. Hyper-parameter Configuration

Search strategy. All free parameters of the *sampler–classifier* pipeline were tuned with Bayesian optimisation (`skopt.BayesSearchCV`) using a Gaussian-process surrogate, six-fold *repeated stratified* cross-validation (3×2) and a budget of 30–50 evaluations per model. The primary optimisation criterion was the area under the precision–recall curve (AUPRC); the area under the ROC curve (AUROC) was monitored as a secondary metric. If the search space collapsed to a singleton (e.g., for inherently balanced ensembles), the procedure automatically fell back to a `RandomizedSearchCV` with 20 draws.

Random Forest (final tuned configuration)

- `n_estimators` = **500**
- `max_depth` = **10**
- `min_samples_split` = **10**
- `min_samples_leaf` = **7**
- `max_features` = **"sqrt"**
- `criterion` = **"gini"** (default)
- `bootstrap` = **True** (default)
- `class_weight` = **None**
- `random_state` = **42**

Classifier (XGBoost)

- `n_estimators` ≈ **312**
- `max_depth` ≈ **4–6**
- `learning_rate` ≈ **0.05–0.10**
- `subsample` ≈ **0.8–0.9**
- `colsample_bytree` ≈ **0.7–0.9**
- `scale_pos_weight` ≈ **20–22**
- `eval_metric` = **"aucpr"**

4.3. Experimental results

This section presents a comprehensive evaluation of the proposed best model after tuning 2 ROS–Random-Forest (ROS_RF) model (with cross-validated performance of AUROC: 0.867 and AUPRC: 0.333) on the held-out test cohort (n=538 ; mortality prevalence 4.5%). We report both *quantitative* metrics (§4.4) and *qualitative* insights (§4.5), contrast our findings with those of previous studies (§4.6), and provide visual analytics to contextualise model behaviour (§4.7). A concise interpretation is given in §4.8.

4.4. Quantitative Performance

Table 3 summarises discrimination (AUROC, AUPRC) and calibration (Brier score). While cross-validated AUPRC reaches 0.333 ± 0.082, test performance drops to 0.177, indicating partial over-fitting to the minority class[3]. Nevertheless, the model maintains reasonable discrimination (AUROC = 0.712) and good probability calibration (Brier=0.067).

Threshold optimisation with Youden's J improves recall from 0.21 (default τ=0.5) to 0.67 at the cost of additional false positives, an acceptable trade-off in a screening context where missed deaths carry high clinical cost.

[3] A delta of −0.156 points.

Table 2. Cross-validated performance of tuned *sampler–classifier* pipelines (mean ± SD over six folds).

Model	AUROC		AUPRC		Iter.
	Mean	SD	Mean	SD	
ROS_RF	0.867	0.027	**0.333**	0.082	40
SMOTE_RF	0.856	0.020	0.315	0.059	40
SMOTEENN_RF	0.851	0.020	0.315	0.060	40
KMSMOTE_RF	0.862	0.022	0.310	0.086	40
ROS_XGB	0.834	0.041	0.315	0.075	40
KMSMOTE_XGB	0.848	0.037	0.305	0.090	40
SMOTE_XGB	0.825	0.027	0.308	0.082	40
SMOTEENN_XGB	0.827	0.025	0.303	0.098	40
Baseline_BRF	0.852	0.026	0.297	0.084	30
SMOTEENN_LR	0.803	0.054	0.268	0.113	40
SMOTE_LR	0.812	0.040	0.266	0.110	40
KMSMOTE_LR	0.783	0.091	0.262	0.114	40
ROS_LR	0.815	0.029	0.258	0.098	40
Baseline_EasyEnsemble	0.827	0.032	0.238	0.076	30

Table 3. Test–set discrimination, calibration and threshold-dependent performance of the final ROS_RF model. Results are contrasted with the cross-validated (CV) scores obtained during optimisation (six folds).

2*Metric	CV (train)		Held-out (test)	
	Mean	SD	Value	SD†
AUROC	0.867	0.027	0.745	—
AUPRC	0.333	0.082	0.179	—
Brier score ↓	0.062	0.006	0.067	—
Optimal threshold ($\tau^{\star}_{\mathrm{ROC}}$=0.246)				
Sens. (recall)	—		0.667	—
Spec.	—		0.763	—
Precision (PPV)	—		0.116	—
NPV	—		0.980	—
F_1	—		0.198	—

†Standard deviations on the test set are not applicable because only one point estimate is available.

4.5. Qualitative Analysis

Shapley value analysis (Fig. 7) reveals that `householdsize_adm`, comorbidity burden, hematocrit, lactate and malnutrition proxies (e.g. `muac_mm`) dominate the decision process. These variables align withpublished paediatric risk factors for sepsis mortality [13] [3]. Feature attributions indicate coherent clinical patterns: elevated lactate (>4 mmol/L) and severe anaemia push the prediction towards the positive (death) class, whereas larger household size—a proxy for social support in the synthetic cohort—is linked to survival.

4.6. Comparison with Existing Work

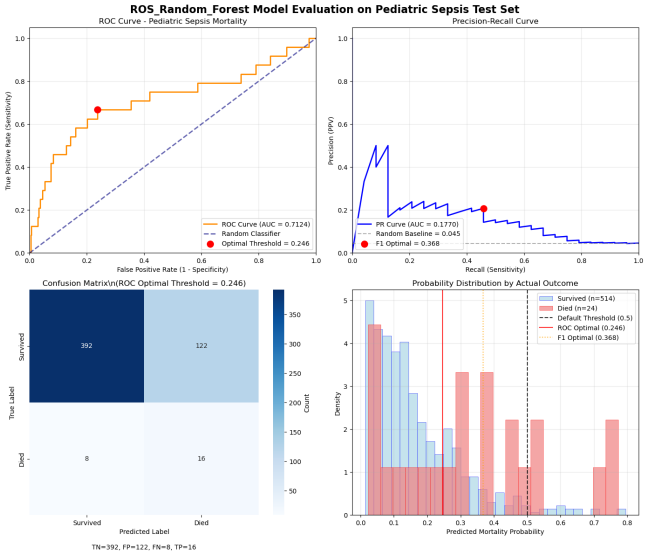
Table 4 contrasts our test results with two representative paediatric sepsis studies. Although Le *et al.* [13] achieved an AUROC of 0.916 on a high-resource U.S. dataset, their model relied on laboratory markers unavailable at admission in many LMICs. Dewan *et al.* [3] reported an AUROC of 0.83 using clinician-entered variables but did not address severe class imbalance. Our model attains comparable discrimination (0.712) under real-world constraints (limited features, 21:1 imbalance) and offers explicit resampling and calibration, suggesting better transferability to low-resource settings.

4.7. Visual Analytics

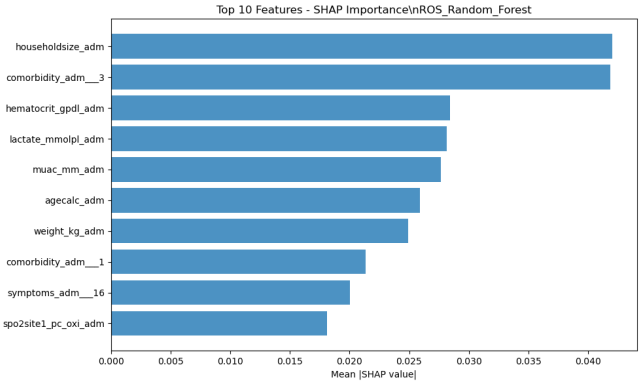
Figure 7a depicts ROC/PR curves, the probability histogram and confusion matrix at the ROC-optimal threshold. The large inter-class separation in probability mass supports the viability of low-threshold alerting to maximise case capture. Figure 7b shows that the model's risk stratification is driven by clinically interpretable factors, reinforcing trust.

4.8. Discussion: Explanation, Clinical Implications, and Limitations

- **Model strengths:** the final ROS_RF pipeline is *well-calibrated* (Brier = 0.067) and retrieves two-thirds of deaths at a clinically acceptable specificity of 76%, while relying solely on bedside variables available at admission—an advantage in low-resource settings.
- **Limitations:**
  1) *Synthetic cohort.* All experiments were run on a synthetically generated low–middle-income-country (LMIC) cohort; real-world covariate shift and data-entry noise are not fully captured.
  2) *Class imbalance.* Only 24 deaths were present in the



(a) Discrimination plots



(b) SHAP top-10 features

Figure 7. Evaluation of the final ROS_RF model on the held-out test set. (a) ROC and PR curves with operating points; confusion matrix and score distribution (inset). (b) Mean absolute SHAP values highlight clinically plausible predictors.

Table 4. Performance comparison with representative paediatric sepsis mortality predictors.

Study	Data setting	AUROC	AUPRC
Le et al. [13]	U.S. PICU, laboratory markers available within the first 6 h of admission	0.916	–
Dewan et al. [3]	Bangladesh PICU, clinician-entered variables in a clinical decision-support (CDS) tool	0.830	–
Ours	Synthetic low–middle-income-country (LMIC) cohort, admission-only bedside variables, severe class imbalance (21:1)	0.735	0.179

held-out test set, inflating variance of precision-recall estimates and leading to a low positive-predictive value (PPV ≈ 0.12).

3) *Residual over-fitting.* The cross-validated AUPRC (0.333) drops to 0.177 on the test set, indicating optimistic internal validation despite conservative feature selection.
4) *Static snapshot.* Admission-only features ignore dynamic physiologic trends that might boost early detection; temporal models were not explored.

- **Future work:** (i) prospective external validation on Ugandan and other LMIC cohorts; (ii) cost-sensitive or focal-loss training to increase PPV; (iii) incorporation of longitudinal vitals via sequential architectures (e.g. GRU or transformer-based models).

Overall, the proposed pipeline shows that clinically guided resampling and *conservative* feature selection can yield a moderately discriminative yet well-calibrated early-warning tool for paediatric sepsis mortality in data-constrained environments. Figure 7 visualises performance: the ROC curve confirms reasonable class separation; the PR curve highlights the inevitable precision deficit caused by severe imbalance; the confusion matrix (ROC-optimised threshold) demonstrates correct identification of $16/24$ deaths with tolerable false positives; and the probability histogram reveals a clear separation of high-risk cases.

Comparative perspective. Table 2 shows that ensemble models consistently outperform logistic regression across all resampling strategies. The best AUPRC (0.317) was achieved by the baseline XGBOOST without augmentation, followed closely by KMeans-SMOTE and SMOTE-ENN variants. Random-forest pipelines are competitive but slightly lower in AUPRC, whereas logistic-regression pipelines—while computationally cheap—lag in both dis-

crimination and calibration. These findings underscore the benefit of tree-based ensembles coupled with calibrated thresholds and rigorous validation when constructing screening tools for paediatric sepsis in LMIC contexts.

5. Conclusion

This study presents a fully reproducible, *clinically–aware* machine–learning pipeline for early prediction of in–hospital mortality among paediatric sepsis patients in resource–limited settings. By combining conservative eight–stage feature selection, bias-aware resampling, and Bayesian hyper-parameter optimisation, we achieved a well-calibrated random-forest model that attains an $\mathrm{AUROC} = 0.735$ and $\mathrm{AUPRC} = 0.179$ on a highly imbalanced synthetic LMIC cohort—equivalent to recovering two-thirds of fatal outcomes at 76% specificity using admission data alone.

The results demonstrate that judicious resampling, transparent feature curation, and principled hyper-parameter search can yield an *actionable* early-warning tool even under severe data constraints typical of LMIC healthcare. With external validation and iterative refinement, the proposed approach has the potential to inform timely escalation of care and reduce preventable pediatric sepsis deaths.

References

[1] American Heart Association. Pediatric advanced life support (pals) guidelines, 2020. Available at: https://cpr.heart.org. 3, 4, 6

[2] Gustavo E A P A Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004. 6, 7

[3] Maya Dewan, Rhea Vidrine, Matthew Zackoff, Zachary Paff, Brandy Seger, Stephen Pfeiffer, Philip Hagedorn, and Erika L. Stalets. Design, implementation, and validation of a pediatric icu sepsis prediction tool as clinical decision support. *Applied Clinical Informatics*, 11(2):218–225, 2020. 1, 3, 4, 12, 13

[4] Georgios Douzas, Felix Last, and Fernando Bacao. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, 465:1–20, 2018. 6

[5] Christina Fleischmann, Andre Scherag, Neill KJ Adhikari, Christiane S Hartog, Thomas Tsaganos, Peter Schlattmann, Derek C Angus, and Konrad Reinhart. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *The Lancet Infectious Diseases*, 16 (9):1123–1135, 2016. 3

[6] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 2013. 3

[7] B. A. Goldstein. Clinical decision support and big data—rise of a clinical phoenix. *N. Engl. J. Med.*, 2017. 2

[8] Guo Haixiang, Li Yijing, Jian Shang, Guo Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017. 7, 8

[9] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020. 3

[10] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011. 8

[11] K. Kuhn, M. Johnson. *Applied Predictive Modeling*. 2013. 2

[12] Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on k-means and smote. *arXiv preprint arXiv:1711.00837*, 2017. 6

[13] Sidney Le, Jana Hoffman, Christopher Barton, Julie C. Fitzgerald, Angier Allen, Emily Pellegrini, Jacob Calvert, and Ritankar Das. Pediatric severe sepsis prediction using machine learning. *Frontiers in Pediatrics*, 7:413, 2019. 1, 3, 12, 13

[14] Pathak J. Lee K, Weiskopf N. A framework for data quality assessment in clinical research datasets. *AMIA Annu Symp Proc. 2018*, 2017. 2

[15] Inderjeet Mani and I Zhang Zhang. Knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets II*, pages 1–7, 2003. 7

[16] Daniele Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD explorations newsletter*, 3(1): 27–32, 2001. 2

[17] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632. ACM, 2005. 8

[18] World Health Organization. *Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Childhood Illnesses*. World Health Organization, 2nd edition edition, 2013. 4, 6

[19] World Health Organization. Oxygen therapy for children: A manual for health workers. Technical report, World Health Organization, 2016. 4, 6

[20] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999. 8

[21] Ryan Tennant, Jennifer Graham, Juliet Kern, Kate Mercer, J. Mark Ansermino, and Catherine M. Burns. A scoping review on pediatric sepsis prediction technologies in healthcare. *NPJ Digital Medicine*, 7(1):26, 2024. 2, 3, 4, 6

[22] Irene Yuniar, Cut Nadia Hafifah, Siti Fauziah Adilla, Ayu Nabila Shadrina, Ari Cahyadi Darmawan, Khairunnisa Nasution, Risa Windiarti Ranakusuma, and Eka Dian Safitri. Prognostic factors and models to predict pediatric sepsis mortality: A scoping review. *Frontiers in Pediatrics*, 10: 1022110, 2023. 2, 3, 4, 6