Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PICTOR versus PRISM-3: a comparative evaluation of mortality prediction models in a Pakistani paediatric PICU: a retrospective cohort study

Muhammad Amin Ghias
Institute of Business Administration Karachi
Saadia Humayun
saadiah@iba.edu.pk

Institute of Business Administration Karachi
Naveed ur Rehman Siddiqui
Aga Khan University Hospital
Hassan Naqvi
Aga Khan University Hospital
Humaira Shafaq
Aga Khan University Hospital
Rabiya Owais
Aga Khan University Hospital
Qalab Abbas
Aga Khan University Hospital
Fyezah Jehan
Aga Khan University Hospital
Tariq Mahmood
Institute of Business Administration Karachi

Research Article

Keywords: deep learning, ensemble learning, machine learning, mortality prediction, paediatric intensive care, PICU

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10992486/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PICTOR versus PRISM-3: a comparative evaluation of mortality prediction models in a Pakistani paediatric PICU: a retrospective cohort study

Muhammad Amin Ghias[1], Saadia Humayun[1*], Naveed ur Rehman Siddiqui[2], Hassan Naqvi[2], Humaira Shafaq[2], Rabiya Owais[2], Qalab Abbas[2], Fyezah Jehan[2], Tariq Mahmood[1]

[1]Institute of Business Administration, Karachi, Pakistan.
[2]Aga Khan University Hospital, Karachi, Pakistan.

*Corresponding author(s). E-mail(s): saadiah@iba.edu.pk;
Contributing authors: m.ghias.25366@khi.iba.edu.pk; naveed.siddiqui@aku.edu; hassan.naqvi@aku.edu; humaira.shafaq@aku.edu; rabiya.owais@aku.edu; qalab.abbas@aku.edu; fyezah.jehan@aku.edu; tmahmood@iba.edu.pk;

Abstract

Background: PRISM-3 and PIM-3 are the paediatric severity scores widely used to estimate mortality risk in the paediatric intensive care unit (PICU), yet both were derived from high-income country populations and have never been formally validated in Pakistani hospitals. Whether their predicted probabilities remain clinically trustworthy in lower-middle-income country (LMIC) settings is unknown. This study formally assessed PRISM-3 calibration in a Pakistani PICU, identified clinical predictors of in-hospital mortality, and developed PICTOR (Pediatric Intensive Care Outcome Risk Predictor), a locally trained, age-stratified ensemble framework providing continuously updated mortality risk predictions at per-minute, 30-minute, and hourly horizons from routinely collected monitoring data.

Methods: A retrospective cohort study of 1,586 patients (1,851 visits; 191 deaths) admitted to the PICU at Aga Khan University Hospital, Pakistan (September 2018–September 2021). PRISM-3 calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test. Predictors of mortality were identified by Mann-Whitney U test and univariate logistic regression. PICTOR was trained

within five paediatric age groups using clinician-guided variable selection, cluster-based oversampling within the training partition only, and ensemble stacking of machine learning and deep learning models. All hyperparameter optimisation was performed within the training set only.
Results: PRISM-3 was miscalibrated (Hosmer-Lemeshow HL = 272.57, $\boldsymbol{p < 0.001}$; confirmed in all five age groups individually), systematically underestimating observed mortality. The strongest predictors of death were hypercapnia (OR 1.87, 95% CI: 1.64–2.14, $\boldsymbol{p < 0.001}$) and overall physiological severity (partial PRISM-3 score OR 1.97, 95% CI: 1.69–2.29, $\boldsymbol{p < 0.001}$). PICTOR outperformed partial PRISM-3 on both primary metrics in every age group (F1: 0.367–0.580 vs. 0.238–0.360; AUPRC: 0.302–0.496 vs. 0.169–0.327), achieving an AUCROC of **0.871** (95% CI: 0.870–0.872) and F1-score of **0.441** (95% CI: 0.439–0.442) for minute-by-minute prediction, improving to **0.649** at one hour.
Conclusions: PRISM-3 is miscalibrated in this Pakistani PICU cohort and should not be used as a standalone mortality estimator without local recalibration. PICTOR, trained on local data and delivering multi-horizon continuous risk prediction from routine monitoring, provides a clinically meaningful alternative with direct implications for other LMIC settings where internationally derived severity tools may systematically underestimate mortality risk.

Keywords: deep learning, ensemble learning, machine learning, mortality prediction, paediatric intensive care, PICU

1 Background

Severity scoring systems are the cornerstone of risk stratification in the paediatric intensive care unit (PICU). PIM-3 and PRISM-3, the two most widely used scores, provide clinicians with an estimated probability of death based on physiological parameters collected in the first hour or 24 hours of admission, respectively [1]. These estimates drive decisions about escalation of care, resource allocation, and family communication across hundreds of PICUs worldwide. Yet both scoring systems were developed and validated in high-income countries—primarily from North American, European, and Australian cohorts—and have never been formally assessed for calibration in Pakistani hospitals [2, 3]. If these scores systematically mis-estimate mortality risk in a lower-middle-income country (LMIC) setting, clinicians relying on them are receiving misleading prognostic information for every high-risk patient they assess.

This matters beyond Pakistan. Approximately two-thirds of the world's children live in LMICs, where paediatric critical care infrastructure is growing rapidly but evidence-based risk stratification tools calibrated to local populations are virtually absent. The assumption that severity scores derived from high-income populations transfer directly to LMIC settings has rarely been tested with formal calibration methods, and the consequences of miscalibration— systematic underestimation of risk—can lead to delayed escalation, inadequate resource allocation, and missed opportunities for preventive intervention [4, 5].

A further limitation of PRISM-3 and PIM-3 is structural: both are static scores computed once at admission, discarding all monitoring data collected thereafter [2, 3]. A child's clinical trajectory can change dramatically over hours, yet the admission score provides no mechanism for updating risk estimates as new physiological information becomes available. Machine learning models trained on continuous vital-sign and laboratory time series have consistently outperformed admission-based scoring in PICU mortality prediction [6–8], and ensemble approaches combining multiple models have shown particular promise [6, 9, 10].

This study makes three contributions. First, it provides the first formal calibration assessment of PRISM-3 in a Pakistani PICU using the Hosmer-Lemeshow goodness-of-fit test. Second, it identifies the clinical predictors of mortality in this population through systematic statistical analysis. Third, it introduces **PICTOR** (Pediatric Intensive Care Outcome Risk Predictor), a locally trained, age-stratified ensemble framework that delivers continuously updated mortality risk estimates at per-minute, 30-minute, and hourly prediction horizons from routinely collected monitoring data—a capability unavailable from any current clinical severity score.

1.1 Related work

Machine learning models have consistently outperformed conventional severity scores for ICU mortality prediction, with ensemble and time-tracking approaches showing the strongest results. Random Forest achieved an AUCROC of 0.81 on adult ICU time-series data [11], and deep-learning models—including LSTM, GRU, and BiLSTM architectures—have since reached AUROCs of 0.83–0.92 across a range of ICU populations [12–15]. Transformer-based models extended these gains to multimodal EHR data, reaching up to 0.92 AUROC [16, 17]. Stacking ensembles that combine multiple base models have consistently improved over any single model alone, with reported AUROCs of 0.89–0.96 in stroke, haemorrhagic stroke, and heart failure mortality tasks [10, 18–20]. In Pakistani paediatric populations, available work is limited to standard models on non-EHR data [21, 22]; no study has combined ensemble deep learning with continuous PICU monitoring data from this setting, nor formally assessed the calibration of international severity scores against local outcomes.

2 Methods

2.1 Data

Data were collected from pediatric patients treated at AKUH from September 2018 to September 2021, comprising continuously recorded vital signs (one-minute resolution) from bedside monitors and laboratory test results. The study was approved by the AKUH Ethical Review Committee (2022-7200-20695). After merging and cleaning, the final cohort comprised 1,586 unique patients (1,851 visits): 1,394 who survived and 191 who died. The median PICU length of stay was 2.0 days (IQR: 1.0–4.2). Cohort characteristics by age group are presented in Table 1.

The bedside monitoring system initially recorded 131 vital-sign fields; PICU specialist physicians narrowed these to 54 clinically meaningful variables, producing a

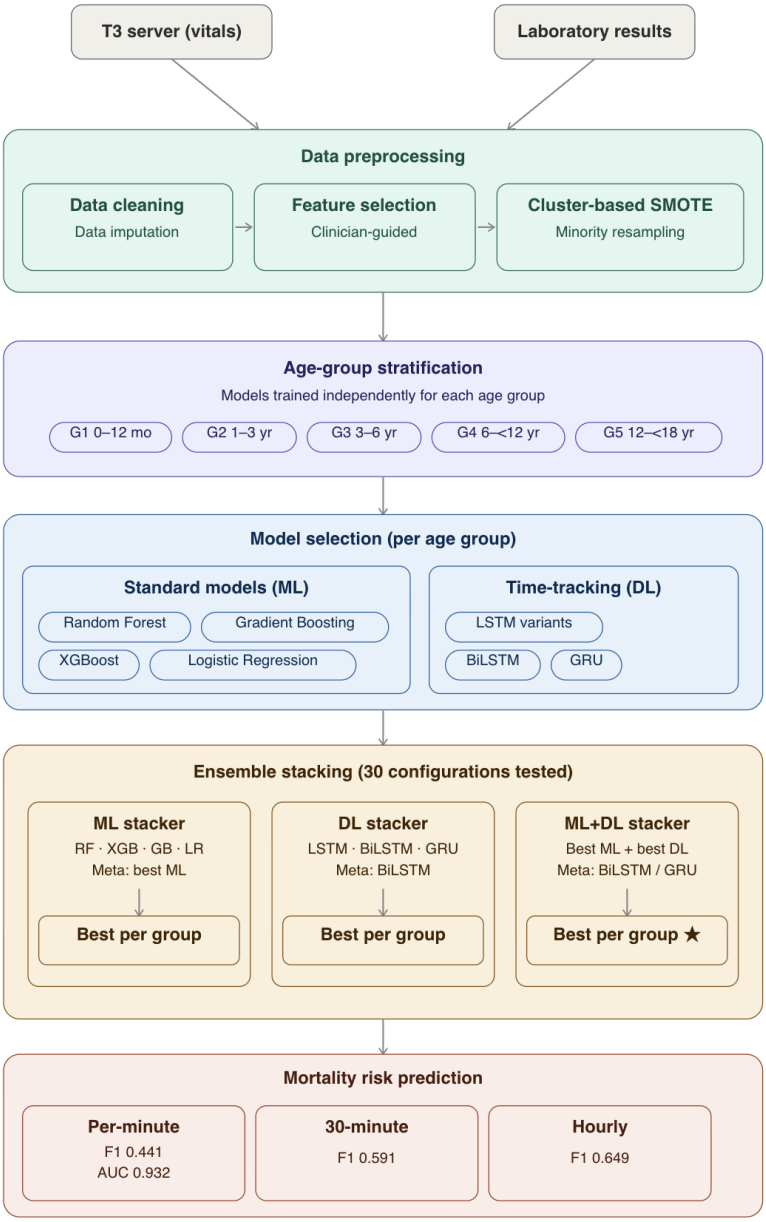


Fig. 1: PICTOR framework overview: from raw PICU monitoring data through data cleaning, clinician-guided variable selection, cluster-based resampling, model training, and ensemble stacking. All data-dependent steps were performed within the training partition only.

dataset of over 8.6 million readings. Laboratory data were similarly narrowed from 261 to 80 tests through expert review. Combining both sources yielded 134 candidate variables. The overall framework is summarised in Figure 1.

Table 1: Baseline characteristics of the study cohort by age group. Continuous variables are presented as median (IQR). Sex was not recorded in structured form for a subset of visits in G1 (47.6%) and G2 (43.0%) owing to missing data in the monitoring dataset.

Characteristic	G1 (0–12 mo)	G2 (1–3 yr)	G3 (3–6 yr)	G4 (6–<12 yr)	G5 (12–<18 yr)	Overall (N = 1,586)
n (visits)	544	372	262	339	334	1,851
Sex, n (%)						
Male	188 (34.6%)	128 (34.4%)	220 (84.0%)	266 (78.5%)	264 (79.0%)	1,066 (57.6%)
Female	97 (17.8%)	84 (22.6%)	42 (16.0%)	73 (21.5%)	70 (21.0%)	366 (19.8%)
Not recorded	259 (47.6%)	160 (43.0%)	0	0	0	419 (22.6%)
Age (months), median (IQR)	5.0 (2.6–8.0)	20.7 (16.0–26.6)	49.1 (40.9–60.0)	108.0 (87.2–125.4)	175.9 (156.1–192.0)	36.0 (9.1–119.4)
PICU LOS (days), median (IQR)[a]	1.9 (1.0–4.0)	2.1 (1.0–4.4)	2.4 (1.4–4.2)	2.0 (0.9–4.6)	1.9 (1.0–4.0)	2.0 (1.0–4.2)
Outcome, n (%)						
Survived	473 (86.9%)	338 (90.9%)	241 (92.0%)	312 (92.0%)	296 (88.6%)	1,660 (89.7%)
Died	71 (13.1%)	34 (9.1%)	21 (8.0%)	27 (8.0%)	38 (11.4%)	191 (10.3%)

[a] Estimated from monitoring timestamp range; may underestimate true PICU stay. LOS, length of stay; IQR, interquartile range.

2.2 Proposed framework

Figure 1 summarises the complete pipeline. The dataset contained substantial missing values, typically from monitoring gaps when patients were off the bedside monitor or undergoing procedures. Missing values were filled per patient using forward and backward imputation within clinically plausible ranges; backward imputation was applied exclusively within each patient's individual admission window to prevent information leakage across admissions. Related readings from different devices were merged into

eleven composite variables. Variables missing more than 15% of values were excluded; remaining gaps were filled with age-group means.

Variable selection combined a physician-assigned clinical relevance score with a predictive importance score derived from Random Forest feature importance, weighted equally to produce a combined ranking. Clinical relevance scores were assigned independently by a senior PICU physician (N.U.R.S.) with over a decade of paediatric critical care experience, using a five-point scale (1 = low clinical relevance, 5 = high), and scoring was performed blind to the predictive importance scores to prevent circularity. The top 15 variables were selected, spanning vital signs (heart rate, respiratory rate, SpO_2, blood pressure), patient demographics (age, height, weight, sex), and laboratory values (whole blood lactate, PaO_2, temperature). Figure 2 shows the combined importance scores for all 15 selected variables. The combined score was calculated as:

$$\begin{aligned}\text{Combined Score} = 0.5 \left(\frac{S_{\text{stat}}}{\max(S_{\text{stat}})} \times 100 \right) \\ + 0.5 \left(\frac{S_{\text{clin}}}{\max(S_{\text{clin}})} \times 100 \right),\end{aligned} \tag{1}$$

where S_{stat} is the predictive importance score and S_{clin} is the clinical relevance score.

Data were partitioned using four-fold cross-validation at the visit level: three folds were used for training and one held-out fold for testing. All data-dependent steps—feature selection, oversampling, and hyperparameter optimisation—were performed within the training partition only. The held-out test fold was not accessed until final model evaluation. The cohort comprised 1,586 unique patients across 1,851 visits; for the 265 patients with more than one visit (14.3% of patients), visits could be distributed across training and test partitions, which is acknowledged as a limitation (Section 4.1).

2.2.1 Age-group formation

Because pediatric physiology and risk profile change substantially from birth to age 18, the cohort was divided into five age groups predefined on the basis of established paediatric developmental staging [23]: **G1** neonates and infants (0–12 months); **G2** toddlers (1–3 years); **G3** preschool-age children (3–6 years); **G4** middle childhood (6 to under 12 years); and **G5** adolescents (12 to under 18 years). Principal component analysis confirmed that the five groups formed physiologically distinct clusters (Figure 3), validating the predefined grouping strategy rather than deriving it.

2.2.2 Correcting for class imbalance

With only 12.1% mortality, standard models risk predicting survival for all patients. We generated synthetic minority-class examples using cluster-based SMOTE, applied exclusively within the training partition. Expired-patient records were first clustered by age group to account for their heterogeneous clinical presentations, and synthetic samples were generated separately within each cluster (target ratio 0.3), preventing unrealistic interpolations between dissimilar cases. The held-out test set contained

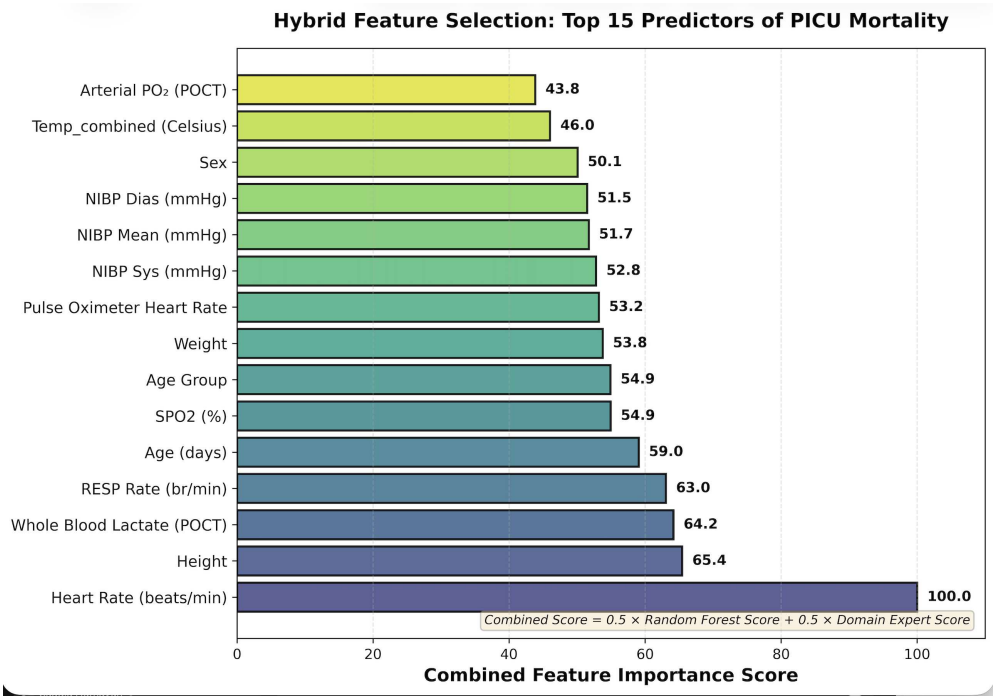
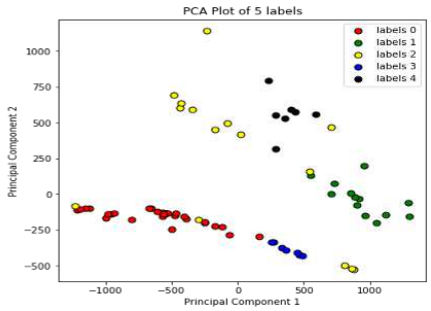
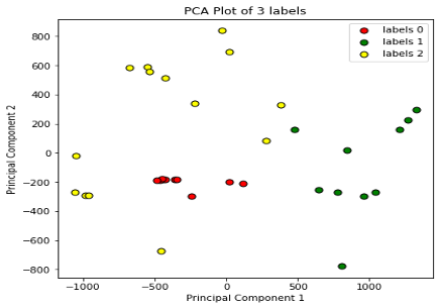


Fig. 2: Hybrid feature selection scores for the top 15 predictors of PICU mortality. The combined score is the equally weighted average of a normalised Random Forest predictive importance score (S_{stat}) and a normalised clinical relevance score (S_{clin}) assigned blind by a senior PICU physician (N.U.R.S.) on a 1–5 scale. Heart rate was the highest-ranked predictor (score 100.0), followed by height (65.4), whole blood lactate (64.2), and respiratory rate (63.0). These scores illustrate feature *selection* importance (which variables entered the model) rather than post-hoc attribution from the trained model.

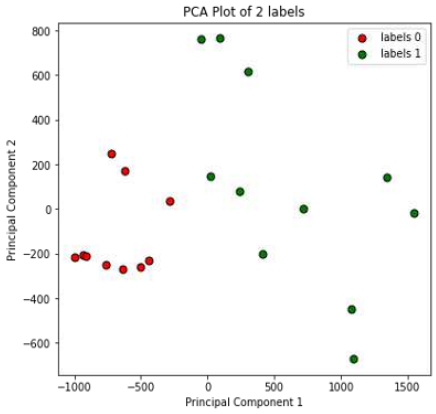
only real patient records and was not accessed until final model evaluation, ensuring synthetic samples did not influence performance estimates. Algorithm 1 summarises the procedure.



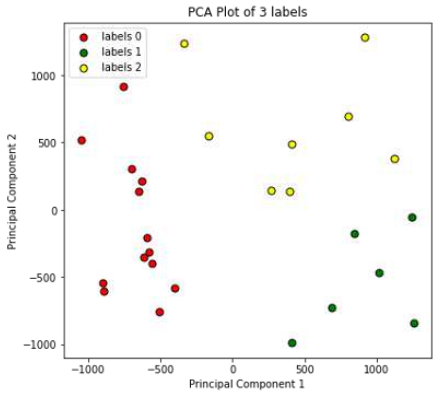
(a) G1: DBSCAN, Silhouette = 0.284



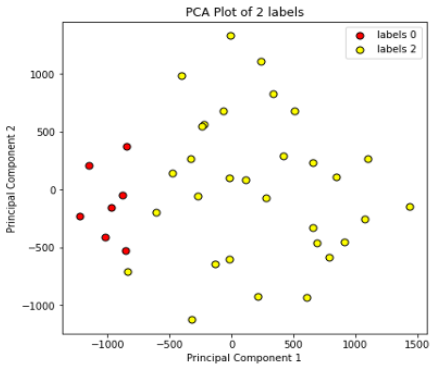
(b) G2: OPTICS, Silhouette = 0.249



(c) G3: DBSCAN, Silhouette = 0.387



(d) G4: DBSCAN, Silhouette = 0.383



(e) G5: OPTICS, Silhouette = 0.225

Fig. 3: PCA visualisations of the five clinically predefined paediatric age groups. Cluster separation, quantified by the Silhouette Score, confirms the physiological distinctness of each group, validating the predefined grouping strategy. The optimal clustering algorithm per group is noted.

Algorithm 1 Cluster-Based SMOTE Oversampling

Input: Training dataset D, target minority ratio $r = 0.3$
Output: Balanced dataset D_{final}

Phase 1 — Cluster minority samples
Apply K-means, DBSCAN, OPTICS to D_{minority}; select $C^* = \arg\max(\text{Silhouette Score})$

Phase 2 — Generate synthetic samples
Partition D_{majority} into k subsets $P_1, \ldots, P_k$
for each $C_i \in C^*$: form $D_i = P_i \cup C_i$; apply SMOTE(D_i, $r = 0.3$); remove duplicates

Phase 3 — Merge
$D_{\text{final}} = \bigcup_{i=1}^{k} D_i^{\text{clean}}$

2.2.3 Model development and ensemble stacking

Three standard prediction models were evaluated after systematic hyperparameter optimisation using GridSearchCV: Random Forest (RF), XGBoost (XGB), and Gradient Boosting (GB), with Logistic Regression (LR) as a linear baseline [24]. Three time-tracking (recurrent) architectures were evaluated with hyperparameters tuned using Keras Tuner: a multi-layer LSTM, a Bidirectional LSTM (BiLSTM), and a GRU. A simple single-layer LSTM was also included for comparison. All hyperparameter optimisation was performed using cross-validation within the training set only; the held-out test set was not accessed at any stage of model development or tuning. All models were trained separately within each of the five age groups.

Following individual model evaluation, ensemble stacking combined the best-performing standard and time-tracking models for each age group. Configurations of two, three, and four base models were tested in three modes (standard-only, time-tracking-only, and mixed), with all combinations evaluated exhaustively across 30 configurations.

2.3 PRISM-3 benchmark

A partial PRISM-3 score was computed retrospectively for each age group using 11 of the 17 standard component variables available: systolic and diastolic blood pressure, heart rate, temperature, PCO_2, PaO_2, bicarbonate, glucose, potassium, creatinine, BUN, and platelet count. Four variables were unavailable (GCS, pupillary response, WBC, PT/PTT) and treated as normal. The PRISM-3 logistic equation [25] was applied ($\text{logit}\, p = -6.207 + 0.197 \times \text{score}$) with the classification threshold optimised for F1-score per age group.

2.4 Statistical analysis

Continuous variables are reported as median (IQR) and compared between survivors and non-survivors using the Mann-Whitney U test; $p < 0.05$ was considered significant. Univariate logistic regression estimated odds ratios (OR) with 95% confidence intervals

per one standard deviation increase in each variable. PRISM-3 calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test with patients divided into deciles of predicted mortality probability; $p < 0.05$ indicates poor calibration, meaning predicted probabilities do not accurately reflect observed mortality rates in this cohort.

Ninety-five per cent confidence intervals for reported model performance metrics (F1-score and AUCROC) were computed using the normal approximation method: $\text{CI} = z \times \sqrt{p(1-p)/n}$, where $z = 1.96$ and n is the effective sample size. Because PICTOR generates predictions at one-minute resolution, each cross-validation test fold comprised between 603,105 and 823,007 individual time-point observations (mean 686,709). The mean fold size was used as n, yielding the narrow intervals reported (for example, AUCROC 0.871, 95% CI: 0.870–0.872). These intervals reflect statistical stability at the time-series observation level; they do not account for within-patient clustering of observations and therefore underestimate uncertainty at the patient level. This is noted as a limitation in Section 4.1.

All analyses were performed in Python 3 (SciPy; statsmodels). This study adheres to the TRIPOD+AI reporting guidelines for predictive model development and validation [26].

3 Results

3.1 Clinical characteristics by outcome

Of 1,586 patients (1,851 visits), 191 (12.1%) died. Table 2 presents clinical characteristics stratified by outcome. Fourteen of 18 variables differed significantly between survivors and non-survivors ($p < 0.05$, Mann-Whitney U test), illustrated in Figure 4. Non-survivors showed persistent haemodynamic instability (minimum systolic BP 73.0 vs. 82.0 mmHg; minimum heart rate 97.0 vs. 85.0 bpm, both $p < 0.001$), respiratory failure (peak PCO_2 45.8 vs. 38.0 mmHg; nadir PaO_2 56.4 vs. 67.1 mmHg, both $p \leq 0.004$), hypothermia (minimum temperature 32.9 vs. 34.1°C, $p < 0.001$), stress hyperglycaemia (peak glucose 188.0 vs. 137.3 mg/dL, $p < 0.001$), renal impairment (creatinine and BUN both $p < 0.001$), and severe thrombocytopenia (platelet nadir 102.0 vs. 215.3 $\times 10^3/\mu$L, $p < 0.001$). The partial PRISM-3 score was significantly higher in non-survivors (median 17 vs. 13, $p < 0.001$), confirming its directional validity despite the calibration limitations discussed in Section 3.5.

Univariate logistic regression identified 12 variables significantly associated with in-hospital mortality (Table 3; Figure 5). The strongest predictors of increased mortality risk were the partial PRISM-3 score (OR 1.97 per SD, 95% CI: 1.69–2.29, $p < 0.001$) and peak PCO_2 (OR 1.87, 95% CI: 1.64–2.14, $p < 0.001$). Stress hyperglycaemia (peak glucose OR 1.57, $p < 0.001$), renal impairment (BUN OR 1.26, creatinine OR 1.12, both $p < 0.05$), and thrombocytopenia (platelet count OR 0.56, $p < 0.001$, the strongest inverse predictor) also contributed independently. Temperature and PaO_2 were significantly different between groups in univariate comparison but were not independently predictive in logistic regression ($p = 0.715$ and $p = 0.272$, respectively), suggesting they function as markers of overall severity rather than independent drivers of death.

Table 2: Clinical characteristics by outcome. Values represent the most abnormal measurement recorded in the first 12 hours of admission. Bold: $p < 0.05$, Mann-Whitney U test.

Variable	Survived (n = 1,660)	Died (n = 191)	p-value
Haemodynamic parameters			
Systolic BP — lowest (mmHg)	82.0 (67.0–94.0)	**73.0 (57.0–88.5)**	**<0.001**
Systolic BP — highest (mmHg)	124.0 (112.0–140.0)	122.0 (107.0–142.0)	0.304
Diastolic BP — highest (mmHg)	91.0 (80.0–104.0)	**87.0 (70.5–103.5)**	**0.009**
Heart rate — lowest (bpm)	85.0 (68.0–103.5)	**97.0 (70.0–116.0)**	**<0.001**
Heart rate — highest (bpm)	165.0 (144.0–184.0)	168.0 (148.5–184.0)	0.523
Temperature			
Temperature — lowest (°C)	34.1 (32.3–35.1)	**32.9 (32.0–34.5)**	**<0.001**
Temperature — highest (°C)	36.0 (34.8–37.0)	**35.7 (33.8–36.8)**	**0.027**
Respiratory and blood gas			
PCO_2 — highest (mmHg)	38.0 (31.5–47.3)	**45.8 (34.8–59.7)**	**<0.001**
PaO_2 — lowest (mmHg)	67.1 (46.8–102.9)	**56.4 (40.6–93.8)**	**0.004**
Bicarbonate — lowest (mEq/L)	21.3 (16.9–24.6)	20.6 (15.7–24.9)	0.641
Bicarbonate — highest (mEq/L)	22.0 (18.9–25.7)	**23.0 (19.1–26.1)**	**0.013**
Metabolic and laboratory			
Glucose — lowest (mg/dL)	100.0 (83.0–120.0)	**108.0 (71.5–151.5)**	**0.008**
Glucose — highest (mg/dL)	137.3 (117.0–180.2)	**188.0 (137.3–282.5)**	**<0.001**
Potassium — highest (mEq/L)	3.6 (3.1–3.9)	3.4 (3.1–4.0)	0.844
Creatinine — highest (mg/dL)	0.5 (0.3–0.6)	**0.5 (0.4–0.8)**	**<0.001**
Blood urea nitrogen — highest (mg/dL)	17.0 (9.0–20.4)	**18.0 (11.5–21.0)**	**<0.001**
Platelet count — lowest ($\times 10^3/\mu$L)	215.3 (70.0–251.8)	**102.0 (36.5–220.1)**	**<0.001**
Clinical severity			
PRISM-3 score (partial)[a]	13.0 (9.0–18.0)	**17.0 (13.0–24.0)**	**<0.001**

[a] Computed on 11 of 17 standard variables; see Section 2.3. Bold = $p < 0.05$. BP, blood pressure; bpm, beats per minute.

3.2 Age-group stratification

Table 4 shows that models trained on the undivided cohort performed consistently worse than those trained separately within each age group and averaged, confirming that mortality risk profiles differ meaningfully by age. Note: overall accuracy is not reported as a primary metric because it is inflated by class imbalance—with 87.9%

Table 3: Univariate logistic regression: association of clinical variables with in-hospital mortality. OR per one standard deviation (SD) increase; ordered by p-value. Bold: $p < 0.05$.

Variable	OR (95% CI)	p-value	Direction
Significant predictors ($p < 0.05$)			
PRISM-3 score (partial)	**1.97 (1.69–2.29)**	**<0.001**	Increased risk
PCO_2 — highest (mmHg)	**1.87 (1.64–2.14)**	**<0.001**	Increased risk
Glucose — highest (mg/dL)	**1.57 (1.39–1.76)**	**<0.001**	Increased risk
Platelet count — lowest	**0.56 (0.46–0.67)**	**<0.001**	Decreased risk
Glucose — lowest (mg/dL)	**1.33 (1.18–1.50)**	**<0.001**	Increased risk
Blood urea nitrogen — highest (mg/dL)	**1.26 (1.13–1.41)**	**<0.001**	Increased risk
Systolic BP — lowest (mmHg)	**0.73 (0.63–0.85)**	**<0.001**	Decreased risk
Bicarbonate — highest (mEq/L)	**1.29 (1.10–1.51)**	**0.002**	Increased risk
Diastolic BP — highest (mmHg)	**0.81 (0.69–0.95)**	**0.009**	Decreased risk
Potassium — highest (mEq/L)	**1.20 (1.03–1.39)**	**0.019**	Increased risk
Heart rate — lowest (bpm)	**1.20 (1.03–1.39)**	**0.019**	Increased risk
Creatinine — highest (mg/dL)	**1.12 (1.01–1.25)**	**0.040**	Increased risk
Non-significant ($p \geq 0.05$)			
PaO_2 — lowest (mmHg)	0.91 (0.78–1.07)	0.272	—
Heart rate — highest (bpm)	1.05 (0.90–1.22)	0.517	—
Bicarbonate — lowest (mEq/L)	1.04 (0.89–1.21)	0.612	—
Systolic BP — highest (mmHg)	0.97 (0.83–1.12)	0.661	—
Temperature — highest (°C)	1.03 (0.88–1.20)	0.715	—
Temperature — lowest (°C)	0.98 (0.85–1.13)	0.789	—

OR per 1 SD increase; CI, confidence interval; BP, blood pressure.

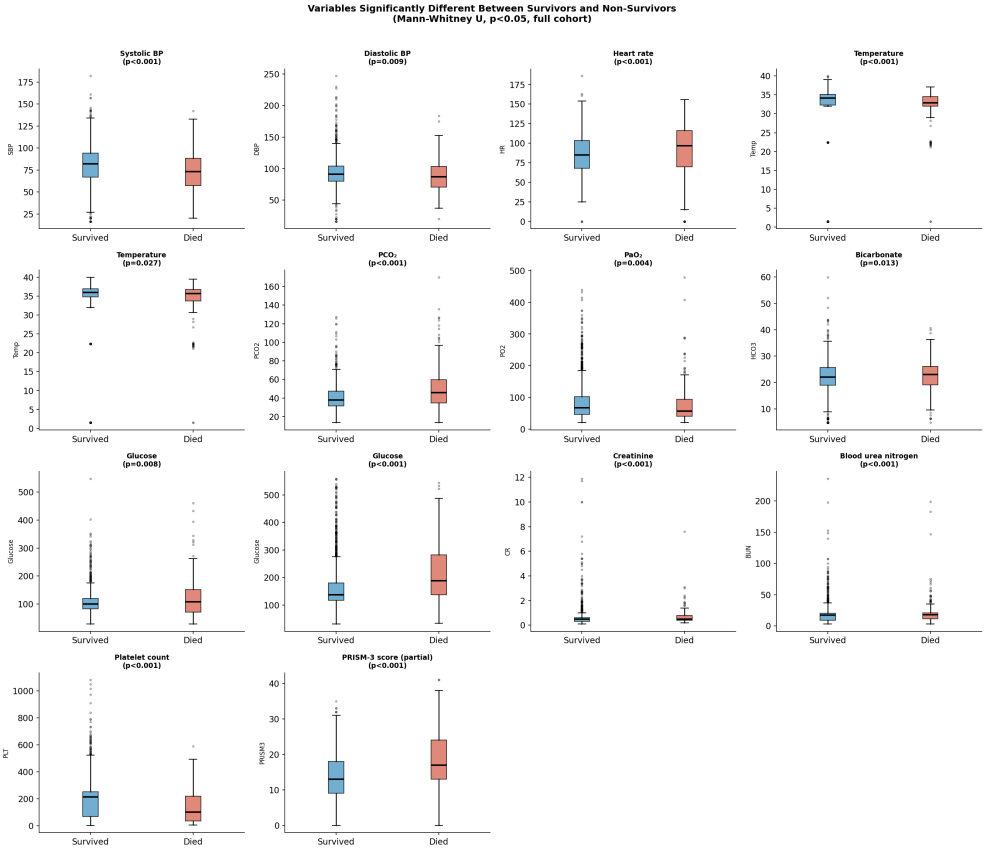


Fig. 4: Distribution of significantly different clinical variables between survivors (blue) and non-survivors (red), full cohort ($p < 0.05$, Mann-Whitney U test). Outliers shown as individual points.

of visits resulting in survival, a model that correctly classifies all surviving patients will achieve accuracy close to 1.000 regardless of its ability to identify deaths. F1-score and AUPRC are the primary evaluation metrics throughout; AUC is retained for completeness.

3.3 Variable selection and oversampling

The clinician-guided 15-variable set outperformed the full 36-variable set on F1-score (0.727 vs. 0.375), precision (+0.394), and AUPRC (0.806 vs. 0.222) for a representative stacker model (Table 5, panel a). Among three oversampling methods, SMOTE outperformed ADASYN by 6% on F1-score (Table 5, panel b). Adding the cluster-based refinement improved F1-score by a further 30% in G3 (Table 5, panel c), and this approach was adopted for all subsequent experiments.

Table 4: Standard prediction models with and without age-group stratification (expired class). Age-stratified models consistently outperform those trained on the undivided cohort.

Model	Precision	Recall	F1	AUC
RF (without age groups)	0.004	0.535	0.008	0.858
RF (age-stratified avg.)	0.671	0.390	0.465	0.813
GB (without age groups)	0.002	0.549	0.003	0.910
GB (age-stratified avg.)	0.218	0.381	0.197	0.719
LR (without age groups)	0.000	0.000	0.000	0.500
LR (age-stratified avg.)	0.211	0.054	0.051	0.675
XGB (without age groups)	0.017	0.465	0.033	0.908
XGB (age-stratified avg.)	0.778	0.306	0.420	0.892

Table 5: Optimisation results: (a) clinician-guided vs. all-variable feature set; (b) oversampling methods; (c) cluster-based vs. standard SMOTE (age group G3). All results for the expired class. All oversampling applied within the training partition only.

(a) Feature Selection Impact

Feature Set	Model	F1	Precision	Recall	AUC	AUPRC
All 36 features	RF_XGB_f_RF	0.375	0.273	0.600	0.810	0.222
15 selected features	RF_XGB_f_RF	0.727	0.667	0.800	0.856	0.806
Improvement		+0.352	+0.394	+0.200	0.046	+0.584

(b) Oversampling Methods

Method	F1	Precision	Recall	AUC	AUPRC
SMOTE	0.429	0.521	0.483	0.917	0.381
ADASYN	0.403	0.477	0.483	0.913	0.388
BoltGAN	0.242	0.510	0.442	0.914	0.159

(c) SMOTE Variants (G3)

SMOTE Type	Model	F1	Precision	Recall	AUC
Cluster-based	XGB	0.272	0.279	0.633	0.879
Standard	XGB	0.208	0.289	0.433	0.889
Difference		+0.064	−0.010	+0.200	−0.010

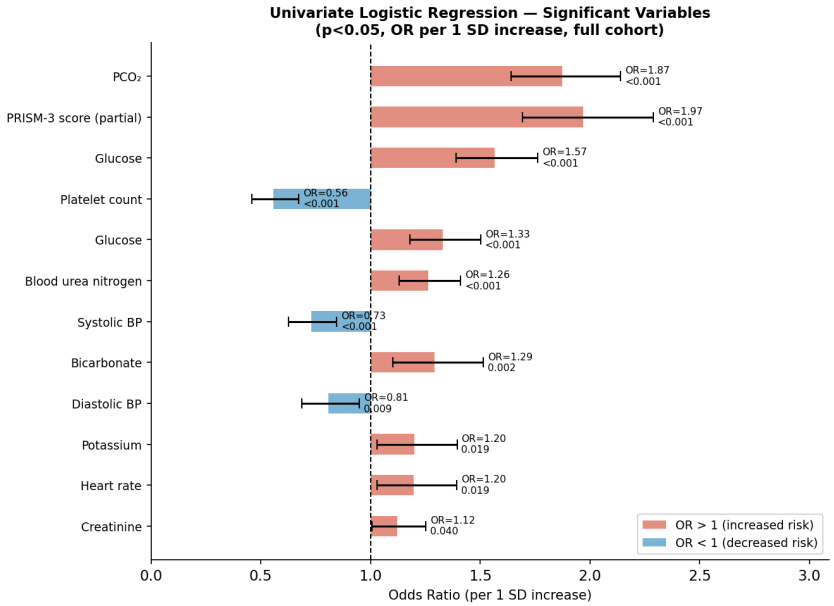


Fig. 5: Forest plot of significant univariate predictors of in-hospital mortality (full cohort, $p < 0.05$). OR per one SD increase with 95% CI. Red: increased risk; blue: decreased risk.

3.4 Model performance

Table 6 presents the best results for each age group across all model families. Among standard models, RF was the top performer in G1, G4, and G5 (F1: 0.192–0.519), and XGB performed best in G2 and G3 (F1: 0.272–0.345). Among time-tracking models, LSTM_complex_1 led in G1, BiLSTM in G2, LSTM_3_LSTM in G3, GRU in G4, and LSTM_complex_2 in G5 (F1: 0.095–0.531). Ensemble stacking consistently outperformed individual models in every age group; the best combined configuration per group is shown in the bottom panel of Table 6, achieving an average F1-score of 0.441 and AUCROC of 0.871 across all age groups. Given the pronounced class imbalance, AUC-ROC should be interpreted with caution and F1-score and AUPRC are the primary metrics.

3.5 Comparison with PRISM-3

Table 7 and Figure 6 summarise the comparison. PICTOR outperformed partial PRISM-3 on both F1-score and AUPRC in every age group. Gains were largest in older children: F1 improvements of 50%, 85%, and 88% in G3, G4, and G5, respectively. In G1 the margin was narrow (F1: 0.367 vs. 0.360), consistent with PRISM-3's stronger historical calibration in neonates and infants. PRISM-3 achieved modestly higher recall in most groups (0.238–0.535 vs. 0.233–0.508) but at the cost of substantially lower precision (0.210–0.271 vs. 0.389–1.000), producing more false-positive alerts per true

Table 6: Best model results by age group: standard models (expired vs. alive classes), time-tracking models (expired class), and best combined ensemble (expired class). F1 and AUPRC are primary metrics. AUC values close to 1.000 reflect class imbalance (12.1% mortality) and should be interpreted with caution.

Standard Models — Best Result by Age Group (Expired Class)				
Group	**Model**	**F1**	**AUC**	**AUPRC**
G1	RF	0.192	0.867	0.129
G2	XGB	0.345	0.876	0.278
G3	XGB	0.272	0.879	0.314
G4	RF	0.434	0.804	0.353
G5	RF	0.519	0.736	0.429
Average		0.352	0.832	0.301

Time-Tracking Models — Best Result by Age Group (Expired Class)				
Group	**Model**	**F1**	**AUC**	**AUPRC**[a]
G1	LSTM_complex_1	0.181	0.910	0.101
G2	BiLSTM	0.095	0.805	0.035
G3	LSTM_3_LSTM	0.349	0.878	0.250
G4	GRU	0.310	0.940	0.225
G5	LSTM_complex_2	0.531	0.910	0.395
Average		0.293	0.889	0.201

[a]AUPRC for G1 (0.101) and G2 (0.035) are near the random-classifier baseline for those groups (0.131 and 0.091 respectively). Individual DL models were superseded by the ML–DL ensemble in these age groups (Table 6, bottom panel).

Best Combined Ensemble by Age Group (Expired Class)				
Group	**Model**	**F1**	**AUC**	**AUPRC**
G1	XGB_GB_f_GB	0.367	0.909	0.331
G2	GRU_4	0.359	0.821	0.302
G3	BiLSTM_4	0.356	0.886	0.328
G4	RF_XGB_LR_f_R	0.541	0.860	0.425
G5	BiLSTM_3	0.580	0.879	0.496
Average		0.441	0.871	0.376

high-risk patient identified. Hosmer-Lemeshow calibration testing showed severe miscalibration of PRISM-3 across all five age groups (HL = 272.57, $p < 0.001$ for the full cohort; $p < 0.05$ in all individual groups). To address the known sensitivity of the Hosmer-Lemeshow test to bin choice, we additionally applied Spiegelhalter's Z-test, which directly quantifies the direction of miscalibration. The Z-statistic was 3.58 ($p < 0.001$), confirming that PRISM-3 systematically *underestimates* mortality risk

in this cohort. The overall observed-to-expected (O/E) ratio was 1.17 (191 observed deaths vs. 163.5 expected). Miscalibration was most severe in older children: O/E ratios were 2.10 in G4 and 4.38 in G5, meaning PRISM-3 underestimated mortality by a factor of two to four in school-age children and adolescents.

Table 7: PICTOR vs. partial PRISM-3 across five age groups. PRISM-3 computed on 11 of 17 standard variables; threshold optimised for F1. F1 and AUPRC are primary metrics. Bold: better-performing model.

Age Group	Model	F1	Precision	Recall	AUPRC	Died	Total
G1 (0–12 mo)	PICTOR	**0.367**	0.389	0.508	**0.331**	71	544
	PRISM-3	0.360	0.271	0.535	0.300		
G2 (1–3 yr)	PICTOR	**0.359**	1.000	0.233	**0.302**	34	372
	PRISM-3	0.333	0.243	0.529	0.260		
G3 (3–6 yr)	PICTOR	**0.356**	0.751	0.333	**0.328**	21	262
	PRISM-3	0.238	0.238	0.238	0.169		
G4 (6–<12 yr)	PICTOR	**0.541**	0.784	0.506	**0.425**	27	339
	PRISM-3	0.292	0.210	0.481	0.308		
G5 (12–<18 yr)	PICTOR	**0.580**	1.000	0.417	**0.496**	38	334
	PRISM-3	0.308	0.217	0.526	0.327		

PRISM-3 unavailable variables: GCS, pupillary response, WBC, PT/PTT, FiO_2. PICTOR results from Table 6 (best ensemble per group).

4 Discussion

The central finding of this study is that PRISM-3 is severely miscalibrated in a Pakistani tertiary PICU. A Hosmer-Lemeshow statistic of 272.57 ($p < 0.001$) and a Spiegelhalter Z of 3.58 ($p < 0.001$) both confirm severe PRISM-3 miscalibration across all five age groups. Critically, the Spiegelhalter test identifies the *direction*: PRISM-3 systematically underestimates mortality, with an overall O/E ratio of 1.17 and ratios reaching 2.10 and 4.38 in G4 and G5 respectively. This means a child assessed as low-risk by PRISM-3 in an older age group may in fact be at substantially elevated risk—a clinically dangerous failure mode. HL values above 20 are generally regarded as strong evidence of poor fit; HL = 272 represents a profound calibration failure. The most

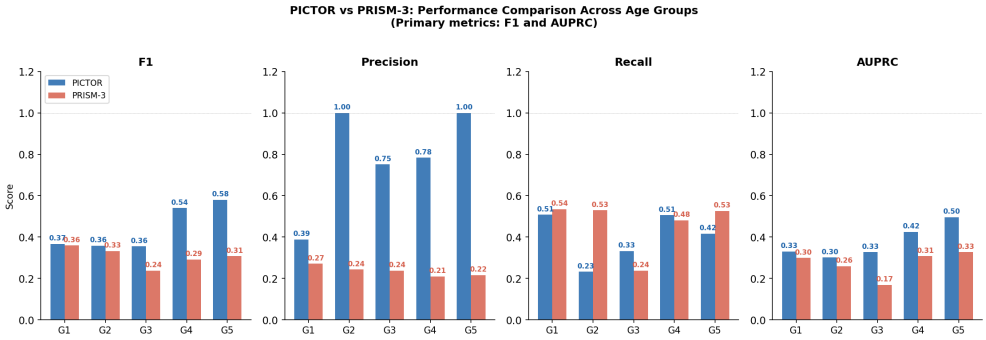


Fig. 6: Performance comparison of PICTOR and partial PRISM-3 across five paediatric age groups on F1-score, precision, recall, and AUPRC. PICTOR (blue) outperforms PRISM-3 (red) on both primary metrics (F1 and AUPRC) in all age groups.

likely explanation is that the PRISM-3 logistic equation—derived from North American PICU data in the 1990s [25]—encodes population-specific baseline mortality rates and case-mix patterns that do not transfer to a contemporary Pakistani tertiary care setting. These findings warrant similar calibration assessments in other LMIC PICUs where PRISM-3 is routinely applied.

The clinical characterisation of the cohort confirms that children who died presented with evidence of multiorgan compromise from the first hours of admission. The strongest univariate predictors were hypercapnia (OR 1.87, $p < 0.001$) and an elevated partial PRISM-3 score (OR 1.97, $p < 0.001$), underscoring that overall physiological severity and respiratory failure are the dominant pathways to death in this population. Thrombocytopenia was the strongest inverse predictor (platelet count OR 0.56, $p < 0.001$), likely reflecting sepsis-related consumption or disseminated intravascular coagulation. Notably, temperature and PaO_2—both significantly different between groups in simple comparison—were not independently predictive in logistic regression, suggesting they function as markers of overall severity rather than independent drivers of death [27].

PICTOR outperformed partial PRISM-3 on both primary metrics (F1 and AUPRC) in every age group, with the largest gains in children aged three years and older (F1 improvements of 50%, 85%, and 88% in G3, G4, and G5, respectively). Unlike PRISM-3, PICTOR was trained directly on local AKUH data, allowing it to learn the population-specific mortality patterns that PRISM-3's globally derived equation cannot capture. The comparison is conservative: four standard PRISM-3 variables were unavailable and treated as normal, likely underestimating what a fully computed PRISM-3 would achieve.

PICTOR's three design choices each contributed independently measurable, quantified performance gains. Age-group stratification raised the best model F1 by 45% and the average F1 across all standard models by 8% [10, 28]. The clinician-guided 15-variable set—selected using a blind scoring procedure that prevented circular reliance on the statistical ranking—improved F1 by almost 90% compared with the full variable

pool [29]. Cluster-based SMOTE oversampling, applied exclusively within the training partition, improved F1 by a further 30% over standard SMOTE [30]. Ensemble stacking delivered a 25% F1 improvement over standard models used alone [31].

Figure 2 provides clinician-interpretable feature importance through the hybrid selection scores. Heart rate ranked highest (combined score 100.0), followed by height (65.4), whole blood lactate (64.2), and respiratory rate (63.0)—a pattern clinically coherent with paediatric critical illness, where tachycardia and poor tissue perfusion are early warning signals. Although these scores reflect selection importance rather than post-hoc attribution, the dominance of physiological stability markers over demographic variables is consistent with findings from SHAP analyses in comparable PICU prediction studies [32]. Post-hoc SHAP analysis with preserved model objects remains a priority for future work.

A distinctive feature of PICTOR not present in any existing paediatric severity score is its multi-horizon prediction capability. Generating continuously updated mortality risk estimates at per-minute (F1 0.441), 30-minute (F1 0.591), and hourly (F1 0.649) windows from routinely collected vital signs and laboratory values allows clinical teams to track deterioration as it unfolds rather than relying on a single admission snapshot—particularly relevant in resource-constrained PICUs where the capacity for frequent manual reassessment is limited.

4.1 Strengths and limitations

Key strengths include minute-by-minute vital-sign data from a real tertiary PICU over three years, with each design choice contributing a verifiable performance gain. The blind clinician scoring procedure ensured the 15-variable set reflects genuine clinical judgment rather than redundant confirmation of the statistical ranking. The PRISM-3 comparison is supported by formal calibration testing, providing a quantitative rationale for locally trained models beyond simple metric comparison.

Limitations include the retrospective, single-centre design, which restricts generalisability; conclusions about performance in other LMIC PICUs await multicentre external validation. Data partitioning was performed at the visit level rather than the patient level. Of 1,586 unique patients, 265 (16.7%) had more than one PICU visit; for these patients, different visits could appear in both training and test partitions, violating strict patient-level independence and potentially introducing a modest degree of information leakage. The effect is bounded by the proportion of multi-visit patients and is unlikely to explain the magnitude of observed performance gains, but reported metrics should be interpreted with this caveat in mind. Reported confidence intervals for performance metrics were computed at the time-point observation level and do not account for within-patient clustering; they may therefore underestimate true uncertainty at the patient level. Patient-level re-partitioning and external validation remain necessary before clinical deployment. Class imbalance makes F1 and AUPRC the appropriate primary metrics; AUC-ROC values should be interpreted cautiously. The partial PRISM-3 implementation and missing FiO_2 data may underestimate PRISM-3's true performance. Decision-curve analysis was not performed, so the net clinical benefit of deployment remains to be established.

5 Conclusions

This study provides the first formal evidence that PRISM-3 is severely miscalibrated in a Pakistani PICU ($HL = 272.57$, $p < 0.001$), systematically underestimating mortality risk across all five paediatric age groups. Clinicians relying on PRISM-3 probability estimates in this setting are receiving systematically misleading prognostic information. This finding warrants similar calibration assessment in other LMIC PICU settings before continued use.

PICTOR, trained on local AKUH data and designed for age-stratified, continuous-monitoring-based prediction, outperformed partial PRISM-3 on both primary metrics in every age group, achieving an AUCROC of 0.871 and an F1-score of 0.441 for minute-by-minute mortality prediction—improving to 0.649 at one hour. Its multi-horizon prediction capability, blind clinician-guided feature selection, and cluster-based resampling applied within the training partition each contributed independently verified performance gains.

External validation across independent LMIC PICU datasets and prospective evaluation within clinical workflows are the necessary next steps before deployment.

Acknowledgements. We express our gratitude to Aga Khan University Hospital, Karachi, for providing the data that made this study possible, and to the PICU clinical team for their invaluable insights and time spent reviewing the study. We also thank the Institute of Business Administration, Karachi, for providing computational resources essential for training and validating the models.

Abbreviations

AUPRC	Area under the precision-recall curve
AUCROC	Area under the receiver-operating-characteristic curve
BiLSTM	Bidirectional long short-term memory
CI	Confidence interval
GRU	Gated recurrent unit
HL	Hosmer-Lemeshow statistic
IQR	Interquartile range
LMIC	Lower-middle-income country
LSTM	Long short-term memory
O/E	Observed-to-expected ratio
OR	Odds ratio
PICTOR	Paediatric Intensive Care Outcome Risk Predictor
PICU	Paediatric intensive care unit
PRISM-3	Paediatric Risk of Mortality score (version 3)
RF	Random Forest
SMOTE	Synthetic Minority Oversampling Technique
XGB	XGBoost

Declarations

Ethics approval and consent to participate. This study was conducted in accordance with the Declaration of Helsinki. It was approved by the Aga Khan University Hospital Ethical Review Committee (approval number: 2022-7200-20695). Patient data were de-identified prior to analysis. Requirement for individual informed consent was waived by the Ethics Committee owing to the retrospective nature of the study.

Consent for publication. Not applicable. No individual patient data or identifying information are included in this manuscript.

Availability of data and materials. The datasets analysed during the current study are not publicly available owing to patient privacy restrictions and institutional data governance policies at Aga Khan University Hospital, Karachi. Anonymised data may be made available from the corresponding author on reasonable request, subject to applicable data governance approvals.

Competing interests. The authors declare that they have no competing interests.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions. M.A.G. designed and implemented the machine learning pipeline, carried out all model training and evaluation experiments, and wrote the original draft. S.H. led the PRISM-3 calibration analysis, contributed to study conceptualisation, and wrote sections of the original draft. T.M. provided computational and methodological supervision throughout the study. N.U.R.S. provided clinical supervision and paediatric critical care domain expertise, and independently scored clinical variable relevance. H.N., H.S., R.O., Q.A., and F.J. contributed to data curation, formal analysis, and critical revision of the manuscript. All authors read and approved the final manuscript.

References

[1] Gemke RJ, Van Vught JA. Scoring systems in pediatric intensive care: PRISM III versus PIM. Intensive Care Medicine. 2002 Feb;28(2):204–207. https://doi.org/10.1007/s00134-001-1185-2.

[2] Slater A, Shann F. The suitability of the Pediatric Index of Mortality (PIM), PIM2, the Pediatric Risk of Mortality (PRISM), and PRISM III for monitoring the quality of pediatric intensive care in Australia and New Zealand:. Pediatric Critical Care Medicine. 2004 Sep;5(5):447–453. https://doi.org/10.1097/01.PCC.0000138557.31831.65.

[3] Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. Critical Care. 2019 Dec;23(1):279. https://doi.org/10.1186/s13054-019-2561-z.

[4] Myers KC, Sauter S, Zhang X, Bleesing JJ, Davies SM, Wells SI, et al. Impaired immune function in children and adults with Fanconi anemia. Pediatric Blood &

Cancer. 2017 Nov;64(11):e26599. https://doi.org/10.1002/pbc.26599.

[5] Akinkuotu AC, Peace A, Maduekwe UN, Hayes AA. Surgical outcomes and survival of adrenocortical carcinoma between children and adults. The American Journal of Surgery. 2022 Oct;224(4):1090–1094. https://doi.org/10.1016/j.amjsurg.2022.05.033.

[6] Lee B, Kim K, Hwang H, Kim YS, Chung EH, Yoon JS, et al. Development of a machine learning model for predicting pediatric mortality in the early stages of intensive care unit admission. Scientific Reports. 2021 Jan;11(1):1263. https://doi.org/10.1038/s41598-020-80474-z.

[7] Aczon MD, Ledbetter DR, Laksana E, Ho LV, Wetzel RC. Continuous Prediction of Mortality in the PICU: A Recurrent Neural Network Model in a Single-Center Dataset*. Pediatric Critical Care Medicine. 2021 Jun;22(6):519–529. https://doi.org/10.1097/PCC.0000000000002682.

[8] Pienaar MA, Sempa JB, Luwes N, Solomon LJ. An Artificial Neural Network Model for Pediatric Mortality Prediction in Two Tertiary Pediatric Intensive Care Units in South Africa. A Development Study. Frontiers in Pediatrics. 2022 Feb;10:797080. https://doi.org/10.3389/fped.2022.797080.

[9] Prince RD, Akhondi-Asl A, Mehta NM, Geva A. A Machine Learning Classifier Improves Mortality Prediction Compared With Pediatric Logistic Organ Dysfunction-2 Score: Model Development and Validation. Critical Care Explorations. 2021 May;3(5):e0426. https://doi.org/10.1097/CCE.0000000000000426.

[10] El-Rashidy N, El-Sappagh S, Abuhmed T, Abdelrazek S, El-Bakry HM. Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model. IEEE Access. 2020;8:133541–133564. https://doi.org/10.1109/ACCESS.2020.3010556.

[11] Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. International Journal of Medical Informatics. 2017 Dec;108:185–195. https://doi.org/10.1016/j.ijmedinf.2017.10.002.

[12] Xia J, Pan S, Zhu M, Cai G, Yan M, Su Q, et al. A Long Short-Term Memory Ensemble Approach for Improving the Outcome Prediction in Intensive Care Unit. Computational and Mathematical Methods in Medicine. 2019 Nov;2019:1–10. https://doi.org/10.1155/2019/8152713.

[13] Na Pattalung T, Ingviya T, Chaichulee S. Feature Explanations in Recurrent Neural Networks for Predicting Risk of Mortality in Intensive Care Patients. Journal of Personalized Medicine. 2021 Sep;11(9):934. https://doi.org/10.3390/jpm11090934.

[14] Yu R, Zheng Y, Zhang R, Jiang Y, Poon CCY. Using a Multi-Task Recurrent Neural Network With Attention Mechanisms to Predict Hospital Mortality of Patients. IEEE Journal of Biomedical and Health Informatics. 2020 Feb;24(2):486–492. https://doi.org/10.1109/JBHI.2019.2916667.

[15] Jun E, Mulyadi AW, Choi J, Suk HI. Uncertainty-Gated Stochastic Sequential Model for EHR Mortality Prediction. IEEE Transactions on Neural Networks and Learning Systems. 2021 Sep;32(9):4052–4062. https://doi.org/10.1109/TNNLS.2020.3016670.

[16] Azhir A, et al. BEHRTDAY: Dynamic Mortality Risk Prediction using Time-Variant COVID-19 Patient Specific Trajectories. In: AMIA Joint Summits on Translational Science Proceedings. vol. 2022; 2022. p. 120–129.

[17] Finch A, Crowell A, Chang YC, Parameshwarappa P, Martinez J, Horberg M. A comparison of attentional neural network architectures for modeling with electronic medical records. JAMIA Open. 2021 Jul;4(3):ooab064. https://doi.org/10.1093/jamiaopen/ooab064.

[18] Huang R, Liu J, Wan TK, Siriwanna D, Woo YMP, Vodencarevic A, et al. Stroke mortality prediction based on ensemble learning and the combination of structured and textual data. Computers in Biology and Medicine. 2023 Mar;155:106176. https://doi.org/10.1016/j.compbiomed.2022.106176.

[19] Emdad FB, Tian S, Nandy E, Hanna K, He Z. Towards Interpretable Multimodal Predictive Models for Early Mortality Prediction of Hemorrhagic Stroke Patients. In: AMIA Joint Summits on Translational Science Proceedings. vol. 2023; 2023. p. 128–137.

[20] Huang Y, Wang M, Zheng Z, Ma M, Fei X, Wei L, et al. Representation of time-varying and time-invariant EMR data and its application in modeling outcome prediction for heart failure patients. Journal of Biomedical Informatics. 2023 Jul;143:104427. https://doi.org/10.1016/j.jbi.2023.104427.

[21] Iqbal F, Satti MI, Irshad A, Shah MA. Predictive analytics in smart healthcare for child mortality prediction using a machine learning approach. Open Life Sciences. 2023 Jul;18(1):20220609. https://doi.org/10.1515/biol-2022-0609.

[22] Diallo AH, Sayeem Bin Shahid ASM, Khan AF, Saleem AF, Singa BO, Gnoumou BS, et al. Characterising paediatric mortality during and after acute illness in Sub-Saharan Africa and South Asia: a secondary analysis of the CHAIN cohort using a machine learning approach. eClinicalMedicine. 2023 Mar;57:101838. https://doi.org/10.1016/j.eclinm.2023.101838.

[23] Fedullo AJ, Swinburne AJ. Relationship of patient age to cost and survival in a medical ICU:. Critical Care Medicine. 1983 Mar;11(3):155–159. https://doi.org/10.1097/00003246-198303000-00001.

[24] Hutter F, Lücke J, Schmidt-Thieme L. Beyond Manual Tuning of Hyperparameters. KI - Künstliche Intelligenz. 2015 Nov;29(4):329–337. https://doi.org/10.1007/s13218-015-0381-0.

[25] Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated Pediatric Risk of Mortality score. Critical Care Medicine. 1996;24(5):743–752.

[26] et al CG. TRIPOD+AI. BMJ. 2024;385:e078378.

[27] Wang N, Wang M, Zhou Y, Liu H, Wei L, Fei X, et al. Sequential Data–Based Patient Similarity Framework for Patient Outcome Prediction: Algorithm Development. Journal of Medical Internet Research. 2022 Jan;24(1):e30720. https://doi.org/10.2196/30720.

[28] Kostaki A, Lanke J. Degrouping mortality data for the elderly. Mathematical Population Studies. 2000 Jan;7(4):331–341. https://doi.org/10.1080/08898480009525465.

[29] Sun F, Wu H, Luo Z, Gu W, Yan Y, Du Q. Informative Feature Selection for Domain Adaptation. IEEE Access. 2019;7:142551–142563. https://doi.org/10.1109/ACCESS.2019.2944226.

[30] Mohammed AJ. Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. International Journal of Advanced Trends in Computer Science and Engineering. 2020 Jun;9(3):3161–3172. https://doi.org/10.30534/ijatcse/2020/104932020.

[31] Pavlyshenko B. Using Stacking Approaches for Machine Learning Models. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). Lviv: IEEE; 2018. p. 255–258. Available from: https://ieeexplore.ieee.org/document/8478522/.

[32] Maheshwari S, Agarwal A, Shukla A, Tiwari R. A comprehensive evaluation for the prediction of mortality in intensive care units with LSTM networks: patients with cardiovascular disease. Biomedical Engineering / Biomedizinische Technik. 2020 Aug;65(4):435–446. https://doi.org/10.1515/bmt-2018-0206.