공학박사 학위논문

Development of a Sparsity-Aware Deep Learning Framework for Early Pediatric ICU Readmission Prediction

소아중환자실 조기 재입실 예측을 위한 희소성 인지 딥러닝 프레임워크 개발

2026년 02월

서울대학교 융합과학기술대학원

헬스케어융합학과

지 현 민

Development of a Sparsity-Aware Deep Learning Framework for Early Pediatric ICU Readmission Prediction

지도 교수 김 경 훈

이 논문을 공학박사 학위논문으로 제출함
2026년 02월

서울대학교 융합과학기술대학원
헬스케어융합학과
지 현 민

지현민의 공학박사 학위논문을 인준함
2026년 02월

위 원 장 김 헌 민 (인)

부위원장 유 수 영 (인)

위 원 김 경 훈 (인)

위 원 권 가 진 (인)

위 원 이 승 원 (인)

Abstract

Development of a Sparsity-Aware Deep Learning Framework for Early Pediatric ICU Readmission Prediction

Hyunmin Ji
Department of Health Science and Technology
Graduate School of Convergence Science and Technology
Seoul National University

Background

Early prediction of pediatric ICU readmission remains challenging due to extreme class imbalance, sparse observations, and irregular time-series patterns in real-world clinical data. While early risk stratification may support ICU resource management, existing models often demonstrate limited generalizability and unreliable probability estimates under such conditions. This study aimed to develop a sparsity-aware deep learning framework for early pediatric ICU readmission prediction and to evaluate its performance and reliability using internal and external validation cohorts.

Methods

Early readmission was conceptualized as unplanned ICU readmission occurring shortly after discharge and was operationally defined as readmission within 72 hours for primary model development, with additional sensitivity analyses conducted using a clinically relevant 48-hour threshold. Model development used a single-center pediatric ICU cohort from PhysioNet (2010-2018; n = 9,529), and external validation was performed on a pediatric subset of MIMIC-III (2001-2012; n = 8,200). Dynamic clinical variables from the 24 hours preceding discharge were encoded as a three-channel tensor comprising exponentially decayed values, observation masks, and time-decay features, while static covariates included demographic information. The proposed BAHA-Net integrates a convolutional time-series encoder and a multilayer perceptron with a gate-based feature selection mechanism to explicitly address data sparsity. Model training employed focal loss and exponential moving average optimization. For external validation, model parameters were fixed, normalization parameters were re-estimated, and predicted probabilities were recalibrated using Platt scaling and isotonic regression (evaluated as alternative calibration methods). Model performance was evaluated using discrimination, calibration, and alert budget-based prioritization metrics.

Results

In internal validation, the proposed framework achieved an AUROC

of 0.752 and an AUPRC of 0.185. In the external MIMIC-III pediatric cohort, discriminative performance was maintained with an AUROC of 0.745, while AUPRC was 0.039. Under a 10% alert-budget constraint, the framework identified 41.4% and 25.5% of early readmissions in internal and external evaluations, respectively.

Conclusion

This study presents a sparsity-aware deep learning framework for early pediatric ICU readmission prediction that demonstrates stable discriminative performance across internal and external validation cohorts. These findings suggest that the proposed approach may serve as a reliable decision-support framework for early risk stratification in pediatric intensive care settings, particularly under constrained monitoring resources.

Keywords: Pediatric Intensive Care Unit (PICU), Early Readmission, Deep Learning, Time-Series Modeling

Student Number: 2023-30649

Table of Contents

LIST OF TABLES

[APPENDIX TABLES]

LIST OF FIGURES

Chapter 1. Introduction

1.1. Background

Unplanned readmission to the Pediatric Intensive Care Unit (PICU) presents a critical challenge in pediatric healthcare. Such events not only reflect inefficient allocation of limited medical resources but also correlate strongly with increased patient morbidity and mortality rates [1−3]. The disruption of continuity of care and escalation in healthcare expenditures caused by PICU readmissions highlight their clinical and operational importance [1,4].

Evidence indicates that the mortality rate among readmitted PICU patients is two to three times higher than that of their non−readmitted counterparts [2]. As a result, PICU readmission rates have been recognized as a key quality indicator in intensive care settings. Despite their relatively low frequency–only 2−3% of all discharges–these events are classified as low−frequency, high−impact outcomes that are associated with longer hospital stays, heightened risk of complications, and substantial cost burdens [5]. This paradoxical nature renders the prediction of PICU readmissions both clinically essential and methodologically challenging [1,5]. Accordingly, the development of robust predictive models that can effectively identify high−risk patients in advance, while overcoming issues of class imbalance and data complexity, is of paramount importance.

1.2. Gaps in Existing Research

To address the problem of hospital readmissions, a variety of predictive models have been developed over the years [4,6,7]. Particularly in adult Intensive Care Units (ICUs), these models range from traditional statistical approaches such as logistic regression [3,8] to more advanced machine learning methods like random forests and boosting algorithms [4,5]. Recently, the advent of deep learning techniques, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures, has enabled researchers to leverage complex temporal patterns in electronic health records (EHRs) and improve predictive performance beyond the capabilities of conventional models [9-11]. In particular, recent work has extended these approaches to pediatric intensive care settings, where machine-learning-based models have been applied to predict length of stay and clinical outcomes, demonstrating the growing feasibility of data-driven prediction in pediatric critical care [12,28]. Despite these advances, several key limitations remain that restrict the applicability of current models in the PICU setting. First, there is a notable lack of studies that focus specifically on pediatric populations. Most existing models are trained on adult ICU data [5,13], which may not generalize well due to physiological, developmental, and epidemiological differences between adults and children [14]. Pediatric patients often exhibit age-specific trajectories and heterogeneous disease profiles, necessitating dedicated models that

capture these nuances [13].

Second, EHR data from critical care settings are inherently sparse, irregularly sampled, and noisy [1,9]. Traditional models typically fail to handle such complexities effectively, especially when clinical variables are measured at inconsistent intervals. In some cases, low-frequency features are treated with equal importance as highly reliable signals, potentially degrading overall model performance [4]. Third, severe class imbalance remains a persistent challenge. Given that PICU readmissions are rare events (2-3% of cases) [5], most learning algorithms tend to favor the majority class (non-readmission) and perform poorly in detecting the clinically critical minority class (readmission) [1]. This often results in high overall accuracy but low recall or sensitivity for high-risk patients, limiting the clinical utility of the model.

Fourth, a lack of rigorous external validation undermines the generalizability of many proposed models. The vast majority of existing studies rely on single-center datasets [1,15], leading to models that may be overfitted to specific institutional practices or patient populations. Without testing on independent external datasets, it is difficult to ensure consistent performance across diverse healthcare settings [6,15].

1.3. Objectives and Contributions

To overcome these limitations, this study proposes a deep learning-based prediction framework for identifying patients at risk of early unplanned PICU readmission. The main objectives of this study are to develop a prediction model tailored to the clinical characteristics of pediatric patients, to address data sparsity and class imbalance through methodological innovation, and to evaluate the model's generalizability using independent external datasets. To achieve these goals, this study introduces a sparsity-aware learning mechanism that incorporates data reliability directly into the model architecture, thereby enhancing robustness in the presence of missing or irregularly sampled data. In addition, comprehensive internal and external validation was conducted to assess the stability of the proposed framework across independent pediatric ICU cohorts.

These processes help ensure that the predictions remain applicable across independent clinical settings. Ultimately, this study aims to support evidence-based discharge decision-making and promote more efficient resource utilization in pediatric critical care.

Chapter 2. Theoretical Background and Literature Review

2.1. EHR Data and Clinical Prediction

Electronic Health Records (EHRs) have fundamentally transformed modern medical research, enabling large-scale and data-driven approaches to clinical prediction. These digital records encompass a wide spectrum of patient information—including demographics, diagnoses, medication use, vital signs, and laboratory measurements—collected longitudinally and stored as time-stamped sequences [17–19]. While EHRs offer an exceptionally rich foundation for predictive modeling, their intrinsic complexity introduces multiple methodological challenges. Clinical variables are often measured at irregular intervals that depend on patient condition, care protocols, and clinical judgment. Many observations are missing not at random, producing sparse and incomplete datasets that can bias model performance if handled improperly. Moreover, the coexistence of continuous, categorical, and textual data adds substantial heterogeneity to model development, complicating the assumptions of traditional statistical frameworks [18,19]. Ongoing harmonization efforts have therefore prioritized the use of standardized terminologies—such as ICD, LOINC, and RxNorm—and common data models including the OMOP CDM to enhance interoperability across institutions and patient cohorts [21]. In pediatric populations, however, these challenges are amplified by

developmental variability, distinct physiological norms, and age-dependent patterns of disease manifestation. Previous studies have reported that the determinants of pediatric ICU readmission and length of stay differ fundamentally from those observed in adults, revealing unique temporal and clinical dynamics that must be modeled separately [2,10]. The temporal irregularity of EHR data further complicates predictive tasks. Measurements are taken at discrete yet uneven time points, and the intervals between observations carry clinically meaningful information regarding monitoring intensity and disease progression. Recurrent neural architectures—such as GRU-D and Phased-LSTM—introduce explicit mechanisms to encode time gaps and decay functions, allowing models to learn both the trajectory of values and the temporal density of observations [18,32]. Collectively, these approaches represent a paradigm shift from static to temporally aware representations, highlighting that the timing of data acquisition is as informative as the data itself.

In summary, effective EHR-based temporal modeling requires representational strategies capable of managing irregular sampling and missingness, capturing complex longitudinal dependencies, and preserving interpretability under high-dimensional conditions.

2.2. Deep Learning for EHR Time-Series Modeling

Deep learning has fundamentally transformed the field of

clinical prediction by enabling models to automatically extract high-level representations from complex, heterogeneous medical data. Early applications in healthcare primarily focused on sequential modeling, leveraging architectures such as recurrent neural networks (RNNs) to capture temporal dependencies in electronic health records (EHRs). Among these, long short-term memory (LSTM) and gated recurrent unit (GRU) networks emerged as foundational approaches for representing patient trajectories over time [16,18,43]. These models successfully addressed the vanishing gradient problem that constrained traditional RNNs, allowing them to capture long-range dependencies in physiological time series such as heart rate, blood pressure, or oxygen saturation. By learning the temporal evolution of these variables, RNN-based architectures demonstrated improved performance in predicting clinical deterioration and readmission compared with static regression models [16,18,26]. However, despite their success, conventional RNNs have notable limitations. They assume uniformly sampled sequences and often require explicit imputation of missing values, which may distort the underlying clinical signal. To address this, several variants were proposed. GRU-D introduced a decay-based gating mechanism that incorporates both observed values and the time elapsed since the last observation, allowing the network to explicitly represent missingness as an informative feature [18]. Similarly, Phased-LSTM extended this framework by introducing time gates that operate on asynchronous or event-based sampling, making it more efficient for irregularly observed clinical data [32].

These extensions mark a major conceptual advance—treating missingness and observation time as intrinsic components of the data, rather than as noise to be imputed. Parallel to RNN developments, convolutional neural networks (CNNs) were adapted for modeling temporal patterns in EHR data by reinterpreting multivariate time series as structured "images." Each patient record can be represented as a two-dimensional matrix of time steps and variables, or even as a three-channel tensor encoding clinical values, observation masks, and temporal decay factors [19]. Through localized convolutional filters, CNNs capture short-term and inter-variable dependencies, revealing joint patterns across physiological measurements. In practice, this enables recognition of temporal clusters or bursts of instability in vital signs that may signal early clinical deterioration. Moreover, CNNs can efficiently handle fixed-length segments, making them computationally tractable for large-scale ICU datasets. Building upon these sequential and spatial modeling paradigms, hybrid neural architectures were introduced to integrate multimodal inputs—combining dynamic time-series features with static patient covariates such as age, sex, and comorbidities. For instance, CNN or RNN modules can process temporal data while a multilayer perceptron (MLP) concurrently encodes static features. The resulting embeddings are fused to form a comprehensive representation of both temporal evolution and baseline risk factors. Such hybrid networks have been particularly effective in ICU outcome prediction tasks, where static and dynamic signals jointly

determine patient prognosis [30]. More recently, attention-based architectures have emerged as the next evolutionary step in EHR modeling. Attention-based models, such as the Clinical Transformer, leverage self-attention to model temporal dependencies without sequential recurrence, offering greater parallelism and modeling flexibility in clinical time-series data [43]. The Clinical Transformer (2023), for example, demonstrated that attention weights can provide interpretable mappings between temporal patterns and clinical outcomes, enhancing transparency while maintaining predictive performance [43]. Similarly, TabNet introduced sparse attentive feature selection tailored to structured data, balancing interpretability and efficiency by learning which variables are most relevant to prediction at each decision step [45]. These methods have been instrumental in promoting explainability in deep learning; however, their computational demands and extensive data requirements can limit their applicability in pediatric and rare-event settings, where datasets are typically smaller, more heterogeneous, and temporally irregular. Recent studies have therefore emphasized the importance of integrating complementary architectural principles to balance accuracy, interpretability, and practicality. Three lines of development have proven particularly influential. First, time-aware encoding strategies explicitly represent observation frequency and time gaps, preserving the temporal granularity of clinical data. Second, sparsity-inducing mechanisms—through regularized gating or feature selection—enhance interpretability and reduce overfitting in high-dimensional

EHR settings [34,36]. Third, resource-aware modeling approaches seek to optimize performance under operational constraints, such as limited alert capacity in intensive care units [35,42].

Table 2-1. Representative Deep Learning Architectures for EHR

Model	Core Mechanism	Missingness	Interpretability	Clinical Use
LSTM / GRU	Sequential	Partial	Low	Moderate
GRU-D [18]	Decay gate	Yes	Moderate	High
Phased-LSTM [32]	Time gate	Yes	Low	Moderate
Transformer [43]	Self-attention	Indirect	High	High (costly)
TabNet [31]	Sparse attention	Partial	High	High
Hybrid attention-sparse frameworks	Temporal + feature attention	Explicit	High	Triage-fit

This table provides a qualitative summary based on representative studies describing each model's mechanism, handling of missingness, and interpretability characteristics.

The ratings ("High," "Moderate," "Low") denote relative conceptual evaluations derived from prior literature, not quantitative performance metrics.

2.3. Challenges in PICU Readmission Prediction and Strategies

Predicting unplanned readmission in the pediatric intensive care unit (PICU) remains one of the most complex and clinically significant problems in pediatric critical care. Empirical studies have shown that only two to six percent of pediatric ICU discharges result in readmission, yet these rare cases are associated with markedly higher mortality and morbidity rates compared with non-readmitted patients [33,38]. This extreme rarity poses an inherent challenge of class imbalance, whereby predictive models tend to overfit to the majority class and fail to identify the small subset of high-risk patients. Such imbalance not only degrades sensitivity but also masks clinically meaningful patterns that could inform early intervention strategies. In addition to imbalance, data sparsity and temporal irregularity are major impediments to reliable modeling. Pediatric ICU data are often incomplete due to the irregular timing of vital sign measurements and laboratory tests, which vary substantially depending on patient condition, disease severity, and hospital workflow. The heterogeneity of pediatric physiology across developmental stages further complicates model training by introducing nonlinear and age-dependent variability in clinical parameters [18]. Consequently, the construction of robust temporal models requires architectures capable of representing missingness as an informative feature, rather than treating it as a nuisance to be imputed. Generalizability represents a further methodological

challenge. Pediatric cohorts differ significantly across institutions due to variations in data acquisition systems, monitoring frequency, and case mix. As a result, models trained in one institution often exhibit marked performance degradation when applied to another, a phenomenon known as domain shift [40,42]. Despite advances in deep learning, few studies have achieved rigorous cross-institutional validation in pediatric critical care, limiting the practical adoption of predictive systems in real-world settings. Recent methodological progress has partially addressed these issues through loss function engineering, temporal encoding, and validation frameworks. The focal loss function has been introduced to counteract class imbalance by emphasizing difficult minority-class examples during training [20], while time-decay-based architectures such as GRU-D explicitly encode the elapsed time since the last observation, capturing informative missingness [18]. External validation on independent pediatric cohorts has been recognized as an essential step to ensure generalizability and reproducibility [40,42]. However, these techniques are typically applied in isolation. A comprehensive framework that integrates data sparsity management, calibration reliability, and interpretability within a unified learning architecture remains largely unexplored in pediatric settings. This methodological gap underscores the need for models that are not only accurate but also resilient to irregularity, imbalance, and institutional variability.

2.4. Model Validation and Clinical Reliability

Rigorous validation is a cornerstone of clinically reliable machine learning, particularly in high-stakes environments such as intensive care. The TRIPOD and CONSORT-AI guidelines emphasize transparent methodology, reproducibility, and external validation on independent datasets as prerequisites for establishing generalizability [29]. While internal validation can indicate model potential, external validation is essential to confirm that performance is not driven by local institutional characteristics. Model performance is commonly evaluated using discrimination metrics such as the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), which quantify the model's ability to rank patients according to risk. However, discrimination alone does not fully capture the clinical reliability of a prediction model, particularly when predicted probabilities are used to support risk-based decision-making. The correspondence between predicted risks and observed outcome frequencies is commonly assessed through calibration analysis. Poorly calibrated models may produce overconfident or misleading probability estimates, even when discrimination performance is acceptable, potentially limiting their clinical utility. Quantitative measures such as the Brier Score and the Expected Calibration Error (ECE) have therefore been used to summarize calibration performance in prior studies [24,25,37]. In

practice, post-hoc calibration techniques—such as temperature scaling, Platt scaling, and isotonic regression—are often applied to improve probability alignment without altering the underlying discriminative ranking [24,25]. In the context of ICU care, where predicted risks may inform monitoring priorities under limited resources, appropriate validation of both discrimination and probability reliability is necessary to support safe and interpretable clinical use.

2.5. Budget-Aware and Clinical Applicability

While predictive performance metrics such as AUROC and accuracy are useful for algorithmic benchmarking, they do not fully reflect the operational realities of intensive care. In practice, clinicians operate under strict constraints of time, attention, and available resources. An effective predictive model must therefore balance accuracy with practicality by identifying the subset of patients most likely to benefit from additional surveillance within a limited alert capacity. This paradigm, commonly referred to as budget-aware evaluation, assesses model performance under a predefined "alert budget," typically defined as the top α% of predicted risk scores [35,50]. Within this framework, metrics such as PPV@α (precision among the top α%) and Recall@α (coverage within the alert budget) quantify how efficiently a model prioritizes high-risk patients for clinical review [42]. These measures align

predictive performance with real-world decision-making processes, where only a limited number of alerts can be acted upon. In pediatric ICU settings, excessive false alarms may contribute to alarm fatigue and inefficient use of clinical attention. Budget-aware evaluation therefore provides a pragmatic lens through which model utility can be assessed under realistic constraints. When combined with appropriate validation of discriminative performance and probability reliability, this approach supports the development of clinically applicable prediction models that are both operationally feasible and interpretable. Such frameworks represent an important step toward the real-world deployment of resource-aware artificial intelligence in pediatric critical care [14,35,40].

2.6. Integrative Summary

This chapter reviewed the conceptual and methodological foundations underlying the development of predictive models for pediatric intensive care. Section 2.1 outlined the clinical context of unplanned PICU readmission, emphasizing its significance as both a quality indicator and a proxy for post-discharge stability. Section 2.2 surveyed the evolution of deep learning architectures for electronic health record (EHR) modeling, highlighting their growing capacity to handle sequential data, missingness, and heterogeneous clinical information. Subsequent sections examined the practical and methodological challenges unique to pediatric ICU prediction. Section 2.3 identified data sparsity, class imbalance, and inter-

institutional variability as major barriers to model generalizability. Section 2.4 discussed the role of rigorous validation in assessing clinical reliability, noting the importance of evaluating both discriminative performance and the consistency of predicted risk estimates. Section 2.5 introduced the paradigm of budget-aware evaluation, which extends model assessment beyond statistical performance to operational feasibility, ensuring that predictions remain clinically actionable within constrained healthcare resources. Collectively, these discussions emphasize that predictive modeling in pediatric critical care requires a balanced integration of methodological rigor and clinical practicality. The reviewed concepts—including temporal modeling, validation strategies, interpretability considerations, and resource-aware triage—provide the theoretical basis for the framework proposed in the following chapter. Chapter 3 builds upon these foundations to present the design, development, and validation of the proposed deep learning model, structured to address the aforementioned challenges while supporting clinical applicability in real-world pediatric ICU settings.

Chapter 3. Methods

3.1. Study Design

This study is a retrospective, electronic health record (EHR)–based prediction model development study designed to forecast early unplanned readmission to the Pediatric Intensive Care Unit (PICU) after discharge. The study was conducted in three sequential stages.

First, an internal development cohort was constructed using publicly available EHR data from pediatric ICU admissions. Clinical measurements collected during the final 24 hours prior to discharge were used as model inputs.

Second, a deep learning–based predictive model was developed and internally validated using a temporal hold-out split, reflecting a realistic prospective prediction setting.

Third, an independent external dataset, derived from a pediatric subset of the MIMIC-III database, was used to assess the generalizability of the trained model.

The overarching objective of this study was to develop a deep learning framework designed to estimate early readmission risk at the time of PICU discharge, while maintaining robustness under sparse and irregular observational conditions and supporting clinically feasible triage prioritization in resource-constrained settings.

3.2. Dataset Description

3.2.1 Internal (Development) Cohort

The internal development cohort was derived from the PhysioNet Pediatric Intensive Care Unit Database (PICU v1.1.0), which contains single-center PICU records collected from Zhejiang University School of Medicine Children's Hospital in China between 2010 and 2018. Among 13,941 ICU admissions, cases with a length of stay (LOS) shorter than 24 hours were excluded to ensure a sufficient observation window for temporal feature construction and to avoid incomplete clinical trajectories. After applying the inclusion criteria, tables containing admission information (ICUSTAYS), demographics (PATIENTS), vital signs (CHARTEVENTS), and laboratory results (LABEVENTS) were integrated for model development and internal validation.

3.2.2 External (Independent Validation) Cohort

The external validation cohort was extracted from the publicly available MIMIC-III (Medical Information Mart for Intensive Care III) database, which contains de-identified ICU data collected at Beth Israel Deaconess Medical Center (Boston, MA, USA) between 2001 and 2012. Only patients aged under 18 years were included to form a pediatric subset for external validation. To assess the generalizability of the trained model, inference was performed

across the entire pediatric cohort without additional sampling or subgroup restriction. The core schema of MIMIC-III—including demographics, admission and discharge timestamps, vital signs, and laboratory tests—was aligned with that of the PhysioNet PICU dataset, allowing identical preprocessing and model input structures. To account for institutional distributional differences, variable-wise normalization parameters (means, standard deviations, and winsorization thresholds) were re-estimated using the external cohort, while all learned model parameters were fixed, enabling a consistent evaluation of model generalizability across datasets.

3.2.3 Data Governance and Ethical Considerations

Both datasets used in this study were publicly available and fully de-identified electronic health record (EHR) resources.
The author completed the Collaborative Institutional Training Initiative (CITI) program and obtained data access approval from PhysioNet.
Because no identifiable patient information was accessed or analyzed, institutional ethical review (Institutional Review Board; IRB) and informed consent were not required.

3.3. Cohort Selection and Outcome Definition

Eligible cases included Pediatric Intensive Care Unit (PICU) admissions for patients younger than 18 years with valid discharge timestamps (OUTTIME), allowing temporal linkage of consecutive admissions within the same individual (SUBJECT_ID). Admissions with a length of stay (LOS) shorter than 24 hours were excluded to ensure adequate temporal information for outcome definition and feature construction. The primary outcome variable, READMIT_72H, was defined as a binary indicator of early unplanned PICU readmission. For each patient, PICU admissions were chronologically ordered, and a readmission event was identified when a subsequent admission occurred within 72 hours after discharge from the index admission. Specifically, admissions satisfying the condition

$$0<(\text{NEXT\_INTIME}-\text{OUTTIME})\leq 72 \text{ hours}$$

were labeled as READMIT_72H = 1, while all other admissions were labeled as READMIT_72H = 0. Planned or elective admissions were excluded from outcome labeling. This binary variable served as the target outcome for all prediction tasks in the study.

3.4. Feature Definitions

3.4.1 Dynamic Variables

Time-series variables were extracted from the 24-hour lookback window prior to ICU discharge. Initially, 90 candidate features were selected from the *CHARTEVENTS* and *LABEVENTS* tables of the PhysioNet PICU dataset.

Feature selection was informed by literature review and clinical consultation to ensure inclusion of physiologically relevant variables in the pediatric critical care setting.

3.4.2 Static Variables

Static (non-temporal) variables included patient-level characteristics such as age, sex, and length of stay (LOS).

These were processed through a dedicated feed-forward branch (MLP) in BAHA-Net and concatenated with time-series embeddings in the final output layer.

3.5. Model Architecture and Training Procedure

The BAHA-Net framework was designed as a dual-branch deep learning architecture that integrates both temporal and static information to predict early readmission within 72 hours after PICU discharge. Prior to model training, all time-series data were standardized and formatted into three complementary channels per variable—representing observed values adjusted by a decay factor (val_cfZ × decay_factor), observation masks (mask), and the

elapsed time since the last measurement (Δt). Each patient's final 24-hour segment was reconstructed into hourly bins, and irregular sampling was handled using exponential temporal decay-based carry-forward imputation. This encoding strategy transformed each variable into a structured tensor of shape ($24 \times F \times 3$), capturing both measurement intensity and temporal dynamics. Static covariates, including age, sex, and length of stay, were processed separately and later fused with temporal embeddings. The temporal branch employed two convolutional layers (kernel size 3×3, ReLU activation) to learn interactions along both temporal and variable dimensions. Through adaptive average pooling, these features were condensed into a 64-dimensional embedding vector that effectively summarized the 24-hour physiological trajectory preceding discharge. In parallel, static variables were passed through a two-layer multilayer perceptron (MLP) with a hidden dimension of 64 and dropout rate of 0.3, producing an equally sized embedding vector that captured patient-specific characteristics. To improve interpretability and reduce overfitting under sparse observational conditions, a gate-based feature selection module was incorporated. Each clinical variable was assigned a learnable gate parameter that controlled its contribution to the model output through a sigmoid activation. An L1 regularization term weighted by the observation frequency of each feature was added to the total loss, encouraging the model to focus on consistently observed and clinically relevant inputs while suppressing noise from infrequently measured variables. The outputs from the temporal and static

branches were concatenated and passed through two fully connected layers (128→1), followed by a sigmoid activation to yield the probability of early readmission.

For optimization, all admissions were chronologically ordered by discharge time. The earliest 80% were allocated to the training set and the most recent 20% to the internal validation set, ensuring temporal independence between training and evaluation data. The training data were further divided into a core subset for model learning and a calibration subset for post-hoc probability alignment. To further mitigate the impact of extreme class imbalance, an event-balanced sampling strategy was employed during the training process [27]. Model training was conducted using the AdamW optimizer with a learning rate of 2×10^{-3} and a weight decay coefficient of 1×10^{-4}. A cosine annealing scheduler and Exponential Moving Average (EMA, decay=0.999) were applied to stabilize convergence. Gradient clipping (threshold=5.0) and early stopping (patience=6 epochs) prevented overfitting and training instability.

The overall loss function combined focal binary cross-entropy with a sparsity-inducing regularization term, defined as:

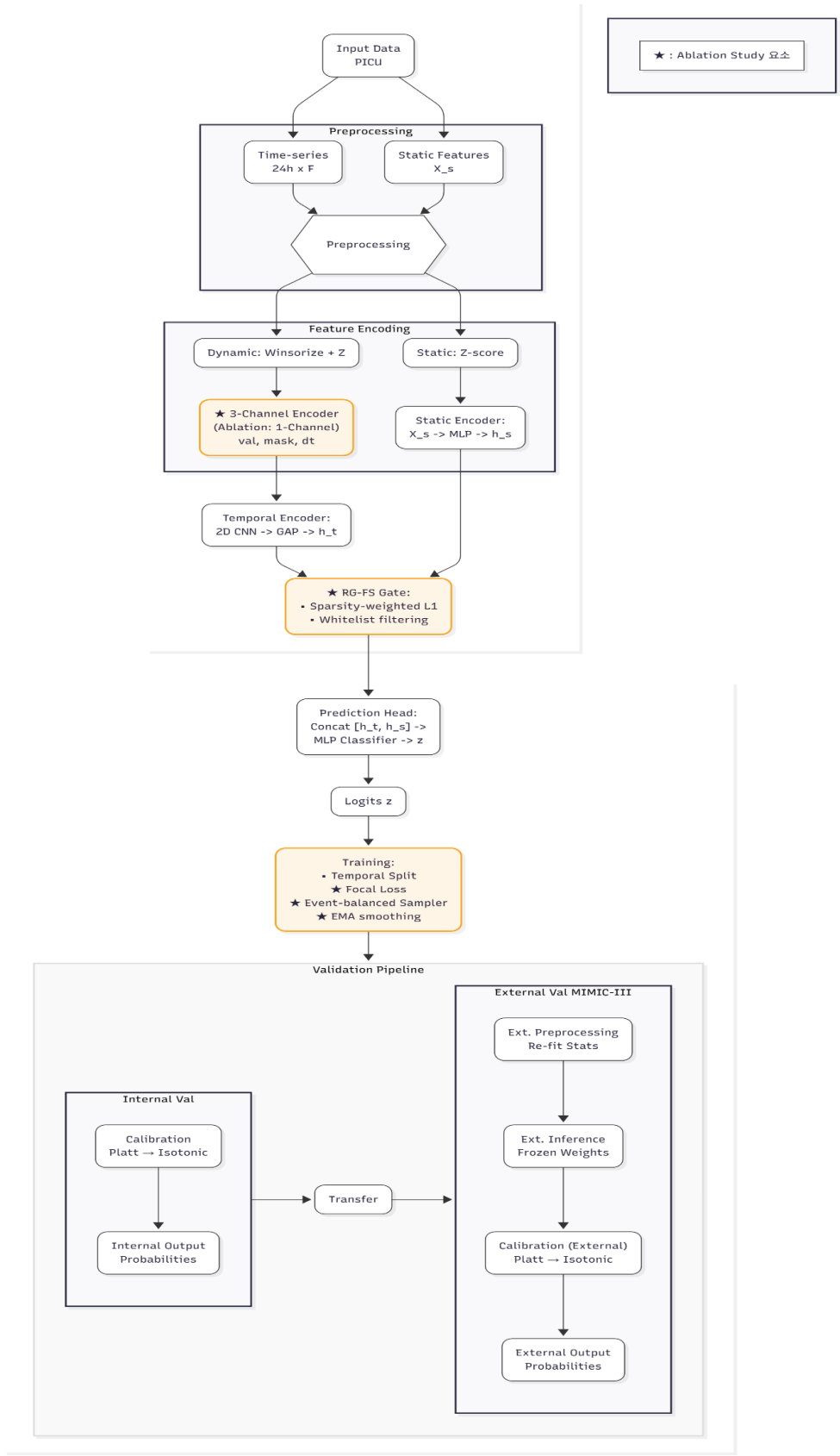
$$L_{total} = L_{focal} + \lambda \sum_i w_i |\sigma(g_i)|$$

where the focal loss emphasized rare positive outcomes and the gate regularization promoted sparsity in feature selection. The

focusing parameter γ was set to 1.5, $\lambda=1\times10^{-4}$, and deterministic random seeds ensured reproducibility. After training, the model weights were frozen, and isotonic calibration was performed using the reserved calibration subset to align predicted probabilities with observed outcomes.

Through this integrated architecture and training pipeline, BAHA-Net effectively captured both the temporal evolution and static context of pediatric ICU patients, while maintaining robustness under sparse, irregular, and imbalanced data conditions.

Figure 3-1. Overall Architecture of BAHA-Net (Budget-Aware Hybrid-Attention Network)


★ : Ablation Study 요소
Input Data
PICU
Preprocessing
Time-series
24h x F
Static Features
X_s
Preprocessing
Feature Encoding
Dynamic: Winsorize + Z
Static: Z-score
★ 3-Channel Encoder
(Ablation: 1-Channel)
val, mask, dt
Static Encoder:
X_s -> MLP -> h_s
Temporal Encoder:
2D CNN -> GAP -> h_t
★ RG-FS Gate:
▪ Sparsity-weighted L1
▪ Whitelist filtering
Prediction Head:
Concat [h_t, h_s] ->
MLP Classifier -> z
Logits z
Training:
▪ Temporal Split
★ Focal Loss
★ Event-balanced Sampler
★ EMA smoothing
Validation Pipeline
Internal Val
Calibration
Platt → Isotonic
Internal Output
Probabilities
Transfer
External Val MIMIC-III
Ext. Preprocessing
Re-fit Stats
Ext. Inference
Frozen Weights
Calibration (External)
Platt → Isotonic
External Output
Probabilities

3.6. Sparsity-Aware Feature Regularization and Budget-Aware Analysis

To address the challenges of high-dimensional and sparsely observed EHR data, the proposed framework incorporates a regularized gate-based feature selection (RG-FS) mechanism. Each input variable i is associated with a learnable gate parameter gi, which is transformed via a sigmoid activation function σ(gi) to modulate its contribution to the model output. The overall training objective combines focal binary cross-entropy loss with a sparsity-inducing regularization term:

$$L = L_{focal} + \lambda \sum_i w_i | \sigma(g_i) |$$

where Lfocal emphasizes rare readmission events, and the regularization term encourages selective use of informative features. The regularization strength λ was set to 1×10^{-4} . Feature-specific weights wi were determined based on observation frequency, with higher penalties assigned to infrequently measured variables to mitigate overfitting under sparse observational conditions. After model training, gate activation values σ(gi) were used to derive a ranked list of features reflecting their relative contribution to prediction. To examine the robustness of the model under varying levels of feature availability, a series of post-hoc analyses were conducted in which prediction performance was evaluated using

subsets of the top-K ranked features. Discriminative performance (AUROC) was assessed across different values of K using the temporal validation split, and performance plateaus were used to identify a parsimonious feature subset. This analysis provides an interpretable and resource-aware perspective on model behavior, demonstrating that predictive performance can be preserved under constrained feature availability through sparsity-aware regularization, without enforcing hard feature budget constraints during training.

3.7. Model Calibration and Evaluation

3.7.1 Internal Validation

The model’s discriminative performance was evaluated on the internal validation set using standard metrics such as the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). To further assess the model's utility under practical triage scenarios, Positive Predictive Value (PPV) and Recall were calculated at a fixed alerting budget of $\alpha = 10\%$ (i.e., PPV@α and Recall@α). All performance metrics were accompanied by 95% confidence intervals derived from 2,000 bootstrap resamples, ensuring statistical robustness and interpretability.

3.7.2 Probability Calibration

To ensure that the model’s predicted probabilities aligned with observed event frequencies, post-hoc probability calibration was performed using a reserved calibration subset from the internal development cohort. Two calibration strategies were considered: Platt scaling and isotonic regression. Platt scaling was implemented by fitting a logistic regression model to map predicted logits to calibrated probabilities, while isotonic regression was applied as a non-parametric approach without assuming a predefined functional form. Each calibration method was trained independently using the same calibration subset, and their calibrated outputs were evaluated separately alongside uncalibrated (raw) predictions. Calibration performance was assessed using the Expected Calibration Error (ECE) with 10 bins and the Brier Score, complemented by visual inspection using reliability diagrams. This comparative calibration framework was used to determine the calibration strategy used for probability estimation, which was then applied consistently in all subsequent performance evaluations, including external validation and benchmark comparisons.

3.7.3 Benchmark Comparison

To assess the relative performance of BAHA-Net, comparative analyses were conducted against a diverse set of widely used and state-of-the-art baseline models, including Logistic Regression,

Random Forest, Gradient Boosting methods (HistGradientBoosting, LightGBM, XGBoost), and deep learning baselines such as Tiny-CNN and Tiny-LSTM. In addition, the recently proposed iREAD ensemble model was reimplemented for comparison.

For fairness and comparability, all benchmark models were trained using the same input features and identical temporal data splits as BAHA-Net. Hyperparameters for each baseline model were optimized using cross-validation within the training set, without access to the validation or test data. Furthermore, probability calibration using Platt scaling and isotonic regression was applied consistently across all models, and calibrated outputs were evaluated separately alongside uncalibrated predictions. In particular, the iREAD model was adapted to the same input representation and evaluation protocol used in this study to ensure a fair and reproducible comparison. Performance was assessed using discrimination, calibration, and budget-aware prioritization metrics, with all results reported alongside 95% confidence intervals derived from 2,000 bootstrap resamples.

Chapter 4. Results

4.1. Characteristics of the Study Cohort

4.1.1 Internal Dataset (PICU)

The internal cohort used for model development and validation was derived from the publicly available PhysioNet Pediatric Intensive Care Unit (PICU) Database (v1.1.0).

This dataset comprises 13,941 PICU admission cases collected from single pediatric center.

To ensure sufficient temporal context for model input, all admission cases with a length of stay shorter than 24 hours were excluded, as these lacked adequate preceding observations for constructing 24-hour time-series data. After applying this exclusion criterion, a total of 9,529 admission cases remained, corresponding to 8,894 unique patients. Among these, 178 early readmissions (within 72 hours after discharge) were identified, representing 1.86% of all admission cases. This rarity underscores the intrinsic challenge of early PICU readmission prediction as a highly imbalanced classification problem in pediatric critical care. To simulate a realistic clinical deployment scenario, a temporal split was applied based on discharge time (OUTTIME).

The earliest 80% of admission cases were allocated to the development (training) set (N = 7,623 cases; readmissions = 149;

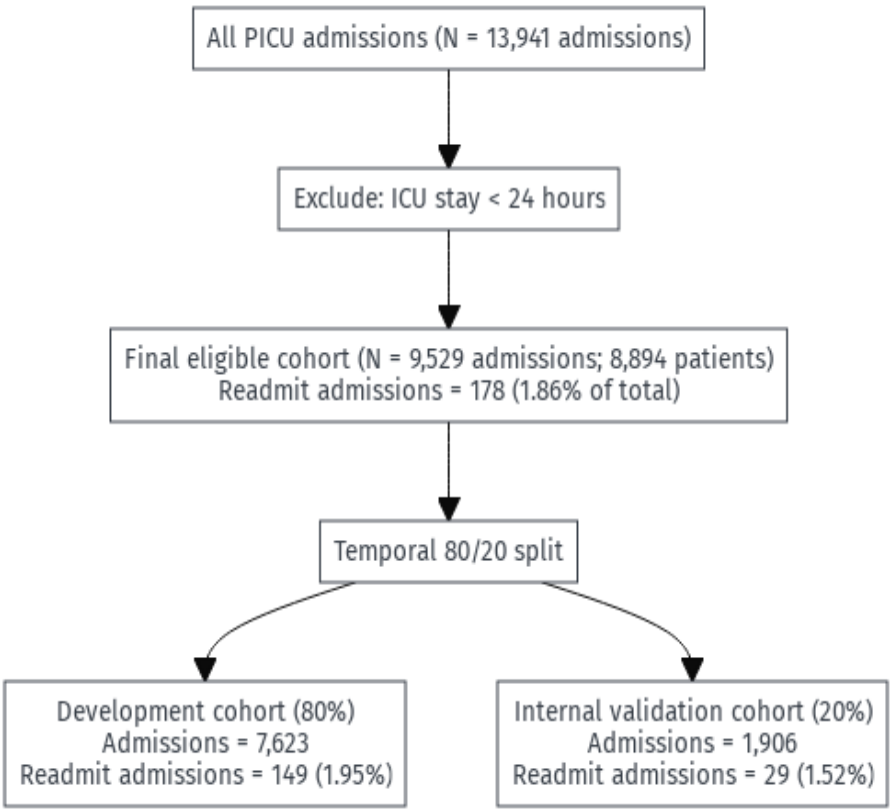
1.95%), and the most recent 20% were designated as the internal validation set (n = 1,906 cases; readmissions = 29; 1.52%).

This chronological partitioning ensured temporal independence between training and validation data, preventing information leakage while reflecting real-world deployment—where models trained on past admissions are applied to predict outcomes for future patients.

Demographic and clinical characteristics of the two subsets were compared using chi-square and t-tests. As presented in Table 4-1, no statistically significant differences were observed between the development and validation cohorts ($p > 0.05$ for all variables), confirming that the temporal division did not introduce systematic bias in key covariates such as sex, age, or diagnostic category.

The entire cohort construction process is summarized in Figure 4-1, which provides a stepwise overview of inclusion and exclusion criteria and illustrates the derivation of the final analytic dataset.

[Figure 4-1] Internal cohort selection flow diagram



A comparison of demographic and clinical characteristics between the development and validation sets revealed no statistically significant differences (Table 4−1; $p > 0.05$ for all variables).

[Table 4−1] Baseline Characteristics and Overall Diagnosis Distribution

Cohort	Variable	Value	P value (α=0.05)
Development	Admissions (N)	7623	
Development	Unique patients (N)	7109	

Development	Readmit+ admissions (n, %)	149 (1.95%)	0.256
Development	Patients with ≥1 readmit (n, %)	135 (1.90%)	0.322
Development	Sex: Male (n, %)	4442 (58.27%)	0.392
Development	Sex: Female (n, %)	3181 (41.73%)	0.392
Development	Age <1 (n, %)	4657 (61.09%)	0.351
Development	Age 1-4 (n, %)	1711 (22.45%)	0.351
Development	Age 5-9 (n, %)	762 (10.00%)	0.351
Development	Age 10-14 (n, %)	463 (6.07%)	0.351
Development	Age 15-17 (n, %)	30 (0.39%)	0.351
Internal-Validation	Admissions (N)	1906	
Internal-Validation	Unique patients (N)	1786	
Internal-Validation	Readmit+ admissions (n, %)	29 (1.52%)	0.256
Internal-Validation	Patients with ≥1 readmit (n, %)	27 (1.51%)	0.322
Internal-Validation	Sex: Male (n, %)	1090 (57.19%)	0.392
Internal-Validation	Sex: Female (n, %)	816 (42.81%)	0.392
Internal-Validation	Age <1 (n, %)	1139 (59.76%)	0.351
Internal-Validation	Age 1-4 (n, %)	461 (24.19%)	0.351
Internal-Validation	Age 5-9 (n, %)	177 (9.29%)	0.351
Internal-Validation	Age 10-14 (n, %)	124 (6.51%)	0.351
Internal-Validation	Age 15-17 (n, %)	5 (0.26%)	0.351
Overall	Diagnosis distribution (subject-level), overall (Chi-square)	—	0.213

4.1.2 Feature Selection Results

To identify the most informative clinical variables for predicting early PICU readmission, Regularized Gate-based Feature Selection (RG-FS) was applied to 89 candidate time-series variables derived from laboratory measurements and vital signs (Table 4-2). Integrated within the BAHA-Net framework, RG-FS assigns a gate activation score ($\sigma(g_i)$) to each input feature,
reflecting its relative contribution through sparsity-regularized optimization.
After training on the internal development (training) cohort, features were ranked according to their activation strength,
and incremental performance analysis was conducted by varying the number of selected features (K).
Figure 4-2 and Table 4-3 illustrate the relationship between K and model discrimination performance (AUROC),
evaluated using both temporal hold-out validation and repeated time-preserved cross-validation (CV).
Temporal validation performance increased as K rose from 20 to approximately 30 features,
reaching its highest AUROC at K = 30. Beyond this point, discrimination performance remained relatively stable,
forming a plateau around $K \approx 30$ without consistent improvement for larger feature sets. In contrast, cross-validation AUROC estimates were more conservative and exhibited greater variability across K values, without demonstrating a clear monotonic trend. This pattern

reflects the impact of severe class imbalance and highlights the importance of temporal validation as the primary criterion for feature selection in this setting.

Based on these observations, K ≈ 30 was identified as a plateau region that consistently achieved the highest discrimination performance across repeated temporal evaluations.

However, when aligning the selected features with the external validation cohort, several variables included in the K ≈ 30 configuration could not be consistently harmonized due to unit discrepancies and excessive missingness in the MIMIC-III pediatric dataset.

Accordingly, cross-dataset harmonization was performed, and a final set of 22 time-series variables that were consistently defined across both datasets was retained and fixed for all subsequent model development and validation. Notably, although the K ≈ 30 configuration yielded the highest point estimates in the internal dataset, comparative analyses showed that the reduced K = 22 configuration preserved discrimination and budget-aware performance, with overlapping confidence intervals and no marked degradation across major evaluation metrics, as confirmed in subsequent internal and external validation analyses. The detailed internal performance of the K ≈ 30 configuration is provided in Appendices A and B for reference, while the internal validation results of the final K = 22 model are presented in Section 4.2.

This refined feature subset captures key physiological domains—including vital signs, gas exchange and acid–base balance,

electrolyte and metabolic regulation, and hepatic and coagulation function—thereby balancing physiological interpretability, model parsimony, and cross-cohort generalizability (Table 4-3).

[Table 4-2] Temporal Clinical Features List (24 Hours Before PICU Discharge)

No.	ITEMID	LABEL
1	5002	Eosinophils
2	5021	ALB/GLB
3	5022	adenosine deaminase
4	5024	Albumin
5	5025	Alkaline Phosphatase
6	5026	Alanine Aminotransferase (ALT)
7	5027	Amylase
8	5031	Aspartate Aminotransferase (AST)
9	5033	Urea
10	5215	Calcium, Total
11	5037	Cholesterol, Total
12	5038	Creatine Kinase (CK)
13	5039	Creatine Kinase, MB Isoenzyme

14	5041	Creatinine
15	5042	Bilirubin, Direct
16	5045	Gamma Glutamyltransferase
17	5046	Globulin
18	5055	Bilirubin, Indirect
19	5057	Lactate Dehydrogenase (LD)
20	5059	Magnesium
21	5072	Phosphate
22	5074	Total Bile Acid
23	5075	Bilirubin, Total
24	5077	Triglycerides
25	5078	Protein, Total
26	5083	Uric Acid, Urine
27	5094	Neutrophils
28	5095	Neutrophils %
29	5225	Hematocrit
30	5257	Hemoglobin
31	5110	Absolute Lymphocyte Count
32	5111	Lymphocytes, Percent
33	5113	MCH
34	5114	MCHC
35	5115	MCV

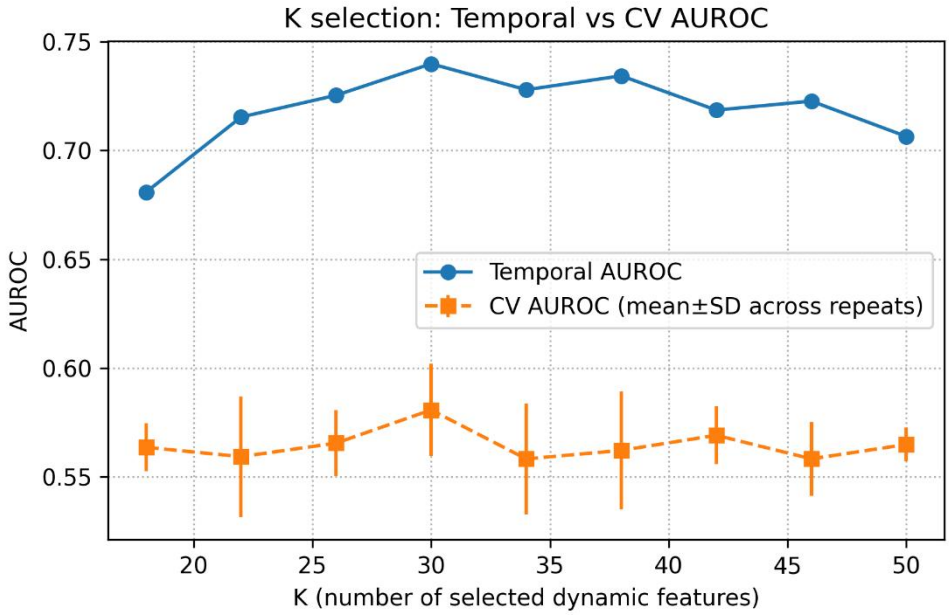
36	5120	Mean Platelet Volume (MPV)
37	5127	PCT
38	5128	Platelet distribution width
39	5129	Platelet Count
40	5132	Red Blood Cells
41	5136	RDW
42	5141	WBC Count
43	5161	PTT
44	5163	D-Dimer
45	5164	Fibrinogen, Functional
46	5174	INR(PT)
47	5175	Reference: APTT
48	5178	Reference:PT
49	5179	Reference:TT
50	5186	PT
51	5192	thrombin time
52	5249	Base Excess
53	5212	Anion Gap
54	5218	Chloride, Whole Blood
55	5219	Carboxyhemoglobin
56	5223	Glucose
57	5224	Calculated Bicarbonate, Whole Blood

58	5226	Potassium
59	5227	Lactate
60	5228	Methemoglobin
61	5230	Sodium, Whole Blood
62	5235	pCO_2
63	5237	pH
64	5239	pO_2
65	5248	Bicarbonate
66	5252	Oxygen Saturation
67	5306	Prealbumin
68	5384	Specific Gravity
69	5492	Basophils
70	5493	Basophils Count
71	5626	C-Reactive Protein
72	6236	Cholinesterase
73	6261	Cystatin C
74	6316	Monocyte Count
75	6317	Monocytes
76	6318	Eosinophil Count
77	6471	Serum hemolytic index
78	6472	Serum icteric index
79	6473	Lipase
80	6857	C1q
81	1001	Temperature

82	1003	Heart Rate
83	1004	Respiratory Rate
84	1007	Excrement
85	1008	Urine Output
86	1009	Input
87	1015	Diastolic Pressure
88	1016	Systolic Pressure
89	1006	Oxygen saturation

These variables were selected to represent key physiological systems, including cardiopulmonary and gas exchange, metabolic and acid–base balance, electrolyte and renal function, hepatic and coagulation function, and overall systemic status.

[Figure 4−2] Relationship between number of selected features and model discrimination (AUROC)



[Table 4-3] AUROC across different K

K	Temporal AUROC	CV mean (±SD)	Remarks
18	0.681	0.563 ± 0.011	Lower performance
22	0.715	0.559 ± 0.028	Reduced feature set
26	0.725	0.565 ± 0.015	Moderate improvement
30	0.740	0.581 ± 0.021	Highest temporal AUROC (plateau)
34	0.728	0.558 ± 0.025	No further gain
38	0.734	0.562 ± 0.027	Similar performance
42	0.719	0.569 ± 0.013	Slight decline
46	0.723	0.558 ± 0.017	Comparable
50	0.706	0.565 ± 0.008	Performance drop

[Table 4-4] Final Temporal Clinical Features (n = 22)

Category	Variable Name
1. Vital Signs	Heart Rate
	Respiratory Rate
	Temperature
	Oxygen Saturation
2. Gas Exchange & Acid-Base Balance	pH
	pCO_2
	pO_2
	Base Excess
	Bicarbonate
	Calculated Bicarbonate (Whole Blood)
	Carboxyhemoglobin
	Methemoglobin
	Chloride (Whole Blood)
3. Electrolyte & Metabolic Status	Calcium (Total)
	Magnesium
	Phosphate
	Potassium
	Lactate
	Specific Gravity
4. Hepatic & Coagulation Function	Aspartate Aminotransferase (AST)
	D-Dimer
	PTT

The final 22 time-dependent clinical features were categorized according to their physiological significance and used as input

variables for the BAHA-Net model.

4.2. Internal Validation Performance

This section presents a detailed evaluation of the predictive performance of the proposed BAHA-Net model using the internal validation dataset derived from the PhysioNet Pediatric Intensive Care Unit (PICU) cohort (n = 1,906; readmission rate ≈ 1.5%). All performance metrics included 95% confidence intervals, which were estimated from 2,000 stratified bootstrap replications to ensure robustness and statistical reliability. The evaluation focuses on three complementary aspects of model performance—discrimination, calibration, and clinical utility—which collectively determine the model's suitability for real-world deployment in critical care settings.

4.2.1 Discrimination Performance

To assess the model's ability to distinguish between high-risk and low-risk patients, two complementary discrimination metrics were employed: the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision–Recall Curve (AUPRC). On the internal validation dataset, the Platt-calibrated BAHA-Net, used as the primary reference model, demonstrated robust discrimination performance. Specifically, the Platt-calibrated model, which was selected as the primary calibration strategy for

this study, achieved an AUROC of 0.752 (95% CI: 0.643–0.850) and an AUPRC of 0.185 (95% CI: 0.061–0.341). For completeness, isotonic calibration was additionally evaluated as a non-parametric reference method. While the isotonic-calibrated model yielded a slightly higher AUROC (0.768; 95% CI: 0.665–0.854), its AUPRC (0.157; 95% CI: 0.040–0.300) was lower than that of the Platt-calibrated model. Given the severe class imbalance of early readmission events, AUPRC was considered a more clinically relevant discrimination metric. Accordingly, the Platt-calibrated BAHA-Net was used as our model for all subsequent analyses. The full discrimination results across calibration states (RAW, PLATT, and ISOTONIC) with 95% confidence intervals are summarized in Table 4-5. All calibrated models substantially exceeded the random-chance baseline (AUROC = 0.5) and the expected AUPRC based on the baseline readmission prevalence (≈ 0.015), demonstrating that BAHA-Net possesses meaningful discriminatory power even under extreme class imbalance.

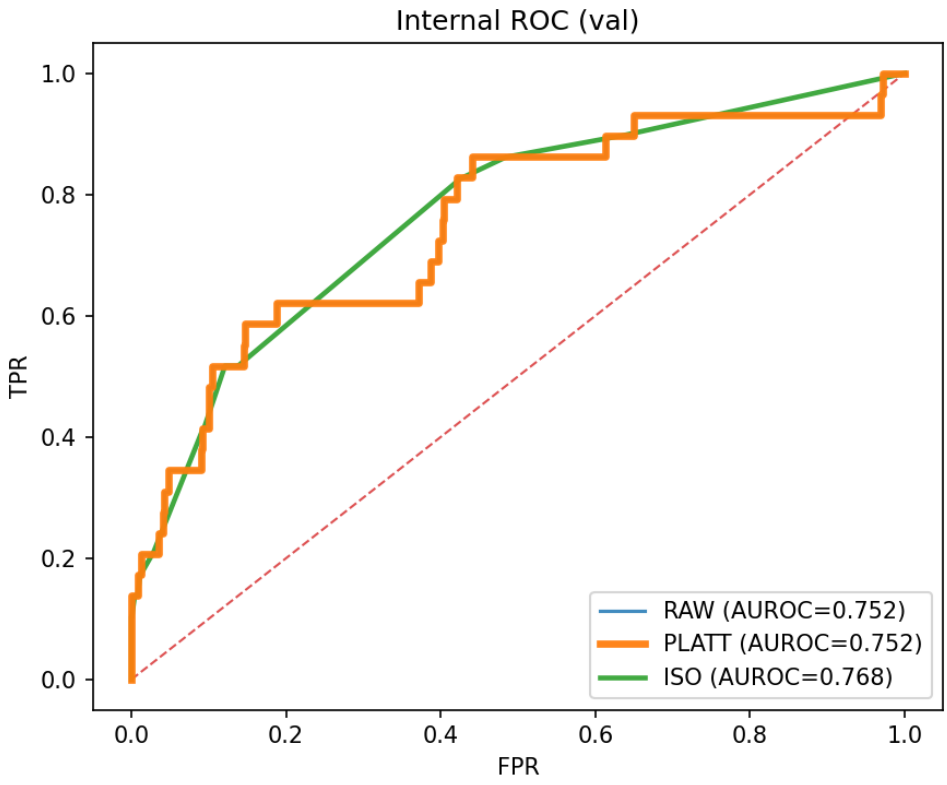
Because early PICU readmission is a rare outcome, AUPRC provides a more sensitive and clinically informative assessment of predictive performance than AUROC alone. A higher AUPRC indicates that the model not only captures a large proportion of true readmissions but also maintains precision when prioritizing high-risk patients. In this respect, BAHA-Net's discrimination performance underscores its ability to identify rare readmission events while limiting false-positive alerts, which is essential for real-world clinical deployment.

The Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves shown in Figure 4-3 and Figure 4-4 further illustrate the discriminative behavior of BAHA-Net across varying decision thresholds. The ROC curves demonstrate consistent separation from the random baseline, indicating stable discrimination between readmitted and non-readmitted patients. Although the PR curves exhibit the expected decline in precision as recall increases due to the rarity of readmission events, the Platt-calibrated BAHA-Net maintains a favorable precision–recall trade-off across clinically relevant operating regions, supporting its robustness under severe class imbalance.

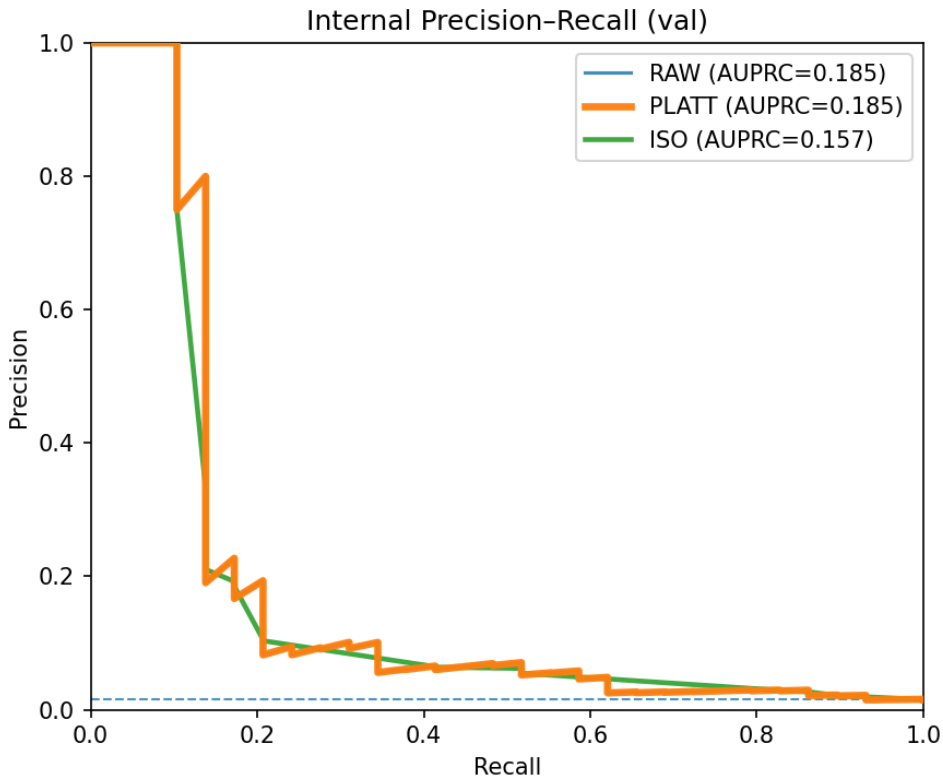
[Table 4-5] Discrimination Performance with 95% Confidence Intervals (Internal Validation)

Calibration	AUROC (95% CI)	AUPRC (95% CI)
RAW	0.752 (0.643–0.850)	0.185 (0.061–0.341)
PLATT	0.752 (0.643–0.850)	0.185 (0.061–0.341)
ISOTONIC	0.768 (0.665–0.854)	0.157 (0.040–0.300)

[Figure 4-3] Internal Receiver Operating Characteristic (ROC) curves



[Figure 4-4] Internal Precision-Recall (PR) curves



4.2.2 Calibration Performance

To evaluate the alignment between predicted probabilities and actual outcomes, calibration performance was assessed using two complementary metrics: the Expected Calibration Error (ECE) and the Brier Score. Calibration was evaluated across three stages—RAW (uncalibrated), PLATT (Platt scaling), and ISOTONIC (isotonic regression)—using the internal validation dataset.

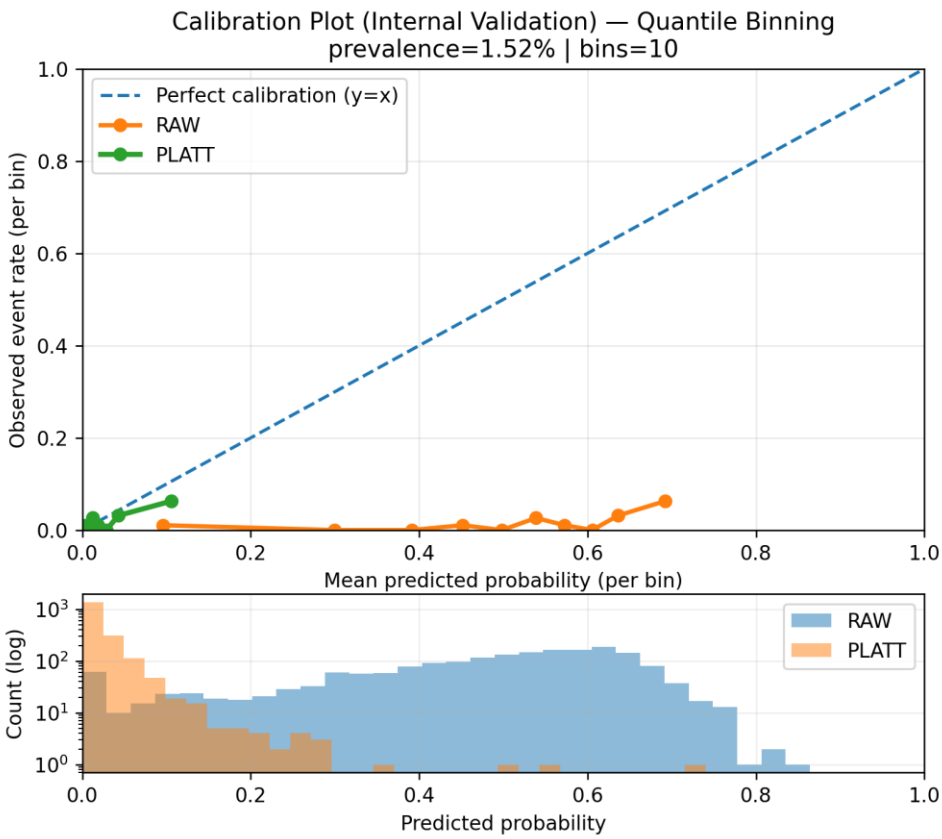
As summarized in Table 4-6, the uncalibrated BAHA-Net (RAW) demonstrated poor calibration, with a high ECE of 0.463 (95% CI: 0.455–0.472) and a Brier Score of 0.255 (95% CI: 0.249–0.261), indicating substantial misalignment between predicted risks and

observed readmission outcomes. After applying Platt scaling, calibration performance improved markedly. The PLATT-calibrated model achieved a substantially reduced ECE of 0.008 (95% CI: 0.003–0.013) and a Brier Score of 0.014 (95% CI: 0.009–0.019), reflecting a close correspondence between predicted probabilities and empirical readmission frequencies. Importantly, this improvement was achieved without altering discrimination performance (AUROC and AUPRC), indicating that Platt scaling effectively enhanced probability calibration while preserving ranking consistency. The isotonic regression model (ISOTONIC) showed a similarly low ECE of 0.008 (95% CI: 0.004–0.014) and a Brier Score of 0.015 (95% CI: 0.011–0.020). However, given its slightly altered discrimination metrics and the risk of overfitting in low-prevalence settings, isotonic calibration was retained for supplementary comparison only. Overall, these results show that post-hoc calibration led to improved probability alignment in the internal validation cohort. Based on the robust improvement in calibration performance and prediction stability, and considering the severe class imbalance of the outcome, the Platt-calibrated BAHA-Net was used as our model for subsequent analyses. The corresponding calibration plots for the internal validation cohort are presented in Figure 4-5. Platt scaling improved the alignment between predicted probabilities and observed event rates across the low-probability range, whereas the uncalibrated model exhibited substantial deviation.

[Table 4-6] Internal Calibration Metrics with 95% Confidence Intervals

Calibration	ECE (95% CI)	Brier Score (95% CI)
RAW	0.463 (0.455–0.472)	0.255 (0.249–0.261)
PLATT	0.008 (0.003–0.013)	0.014 (0.009–0.019)
ISOTONIC	0.008 (0.004–0.014)	0.015 (0.011–0.020)

[Figure 4-5] Calibration Plots of BAHA-Net in Internal Validation Cohort



4.2.3 Clinical Utility (Top 10% Alert Budget)

To evaluate the practical utility of BAHA-Net under limited clinical

monitoring resources, model performance was assessed using an alert–budget framework, in which only a fixed proportion of patients are flagged for potential early readmission. Thresholds corresponding to the top 5%, 10%, and 15% of predicted risk were examined to reflect realistic pediatric ICU discharge triage scenarios, with particular emphasis on identifying an operationally meaningful alert threshold. As summarized in Table 4–7 and illustrated in Figure 4–6, the Platt–calibrated BAHA–Net, used as the model for subsequent analyses, demonstrated a clear concentration of true readmission cases within higher–risk strata, accompanied by a predictable trade–off between precision and coverage as the alert budget increased.

At the operationally relevant alert budget of α = 10%, BAHA–Net achieved PPV = 0.063 (95% CI: 0.032–0.100) and Recall = 0.414 (95% CI: 0.227–0.609), corresponding to a 4.15–fold enrichment over the baseline readmission prevalence (≈ 1.5%). This indicates that more than 40% of early readmission cases can be identified while restricting alerts to one in ten discharged patients, representing a clinically feasible balance between detection sensitivity and alert burden under extreme class imbalance.

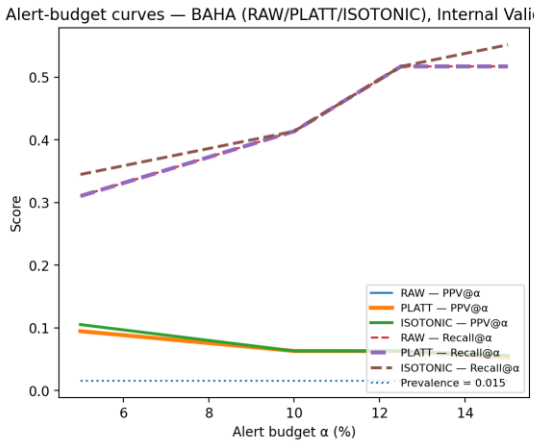
At a more conservative threshold (α = 5%), precision and lift were higher (PPV = 0.095; lift = 6.23), but recall was limited to 31.0%, indicating that a substantial proportion of true readmission cases would remain unidentified. Conversely, increasing the alert budget to α = 15% further improved recall (0.517, 95% CI: 0.355–0.735), but this gain was accompanied by reduced precision (PPV = 0.053)

and lower lift (3.46), reflecting diminishing returns and an increased false-positive burden. Across all evaluated alert budgets, trends in PPV, recall, F1-score, and lift showed consistent and monotonic behavior, with α = 10% emerging as the most clinically balanced operating point. Collectively, these findings demonstrate that BAHA-Net can effectively support resource-aware early-warning and decision-support strategies in pediatric ICU discharge management by prioritizing high-risk patients within realistic clinical capacity constraints, while maintaining an interpretable and actionable alert burden.

[Table 4-7] Budget-Aware Triage Performance under Different α Thresholds

α (%)	PPV (95% CI)	Recall (95% CI)	F1 (95% CI)	Lift (95% CI)
5	0.095 (0.042-0.158)	0.310 (0.138-0.481)	0.145 (0.063-0.223)	6.226 (2.767-9.660)
10	0.063 (0.032-0.100)	0.414 (0.227-0.609)	0.110 (0.055-0.169)	4.151 (2.280-6.107)
15	0.053 (0.032-0.084)	0.517 (0.355-0.735)	0.096 (0.058-0.149)	3.459 (2.373-4.918)

[Figure 4-6] Alert-Budget Curves and Operating Points of BAHA-Net



BAHA-Net (PLATT): Summary at α = 5/10/15%

alpha%	PPV	Recall	F1	Lift
5.0	0.0947	0.3103	0.1452	6.2265
10.0	0.0632	0.4138	0.1096	4.151
15.0	0.0526	0.5172	0.0955	3.4592

Precision@α (solid lines) and Recall@α (dashed lines) across alert budgets (α = 5%, 10%, 15%) for BAHA-Net, with the Platt-calibrated model used for primary interpretation and other calibration states shown for reference.

4.2.4 Decision-Curve Analysis

To evaluate the clinical usefulness of the BAHA-Net model beyond conventional discrimination metrics, a decision-curve analysis (DCA) was performed using the internal validation dataset.

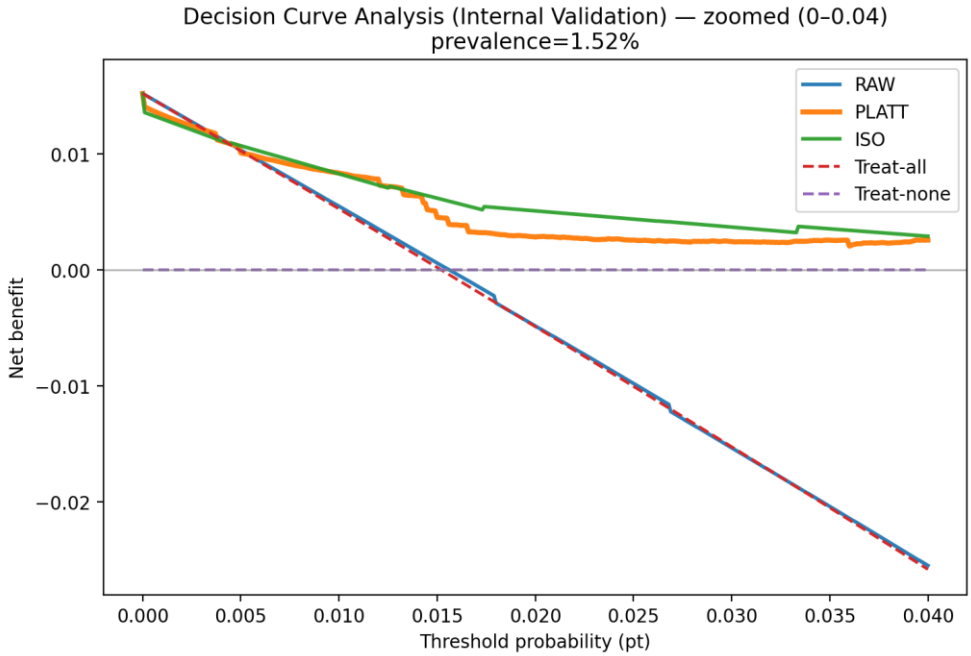
The DCA quantifies the net clinical benefit across a range of threshold probabilities, comparing the model's utility with the default treat-all and treat-none strategies. Here, the net benefit represents the trade-off between correctly identifying early-readmission cases and minimizing unnecessary alerts for low-risk patients, defined as:

Net Benefit = $(TP / N) - (FP / N) \times (p_t / (1 - p_t))$

where pt denotes the decision threshold (i.e., the probability cut-off for triggering an alert).

As illustrated in Figure 4-7, which presents a zoomed view of low threshold probabilities (pt = 0–0.04) corresponding to the observed early readmission prevalence (≈ 1.52%), the Platt-calibrated BAHA-Net demonstrated a stable and interpretable net benefit profile compared with uncalibrated predictions. In contrast, the uncalibrated (RAW) model exhibited a rapidly declining net benefit as the threshold probability increased, closely overlapping with the treat-all strategy, indicating substantial penalties from false-positive alerts under insufficient calibration. Within clinically relevant operating thresholds for pediatric ICU discharge decision-making (approximately pt = 0.01–0.03), the Platt-calibrated BAHA-Net consistently maintained net benefit values comparable to or exceeding the treat-none reference, while outperforming the treat-all strategy. Collectively, these findings indicate that appropriate post-hoc calibration supports decision-level interpretability and avoids net harm under realistic, low-prevalence triage settings.

[Figure 4-7] Decision Curve Analysis of BAHA-Net under Different Calibration Methods



4.2.5 Summary of Internal Validation Performance

In summary, BAHA−Net demonstrated robust and consistent performance across three core evaluation domains—discrimination, calibration, and clinical utility—within the internal validation cohort derived from the PhysioNet Pediatric ICU dataset (Table 4−8).

From a discrimination perspective, the Platt−calibrated BAHA−Net, used as the final model for internal validation, achieved an AUROC of 0.752 (95% CI: 0.643–0.850) and an AUPRC of 0.185 (95% CI: 0.061–0.341), substantially exceeding the expected random baseline performance (AUPRC ≈ 0.015). These results indicate that BAHA−Net effectively differentiates between patients at high and low risk for early PICU readmission, despite the presence of severe

class imbalance. In terms of calibration, the selected post-hoc calibration strategy (Platt scaling) markedly improved the reliability of predicted risk estimates. Compared with the uncalibrated model (ECE = 0.463), the Platt-calibrated BAHA-Net achieved a low Expected Calibration Error (ECE = 0.008, 95% CI: 0.003–0.013) and a favorable Brier Score (0.014, 95% CI: 0.009–0.019), indicating close agreement between predicted probabilities and observed readmission outcomes. Regarding clinical utility, BAHA-Net maintained effective prioritization capacity under constrained alert budgets. At the operationally relevant 10% alert threshold, the model identified approximately 41% of all true readmission events (Recall@10% = 0.414, 95% CI: 0.227–0.609), while achieving a four-fold improvement in precision relative to random selection (PPV@10% = 0.063; Lift ≈ 4.15). These findings highlight the model's practical potential for resource-aware triage and targeted post-discharge monitoring in pediatric intensive care settings. Collectively, these results demonstrate that BAHA-Net combines strong discriminative ability, well-calibrated probabilistic outputs, and clinically meaningful decision performance under realistic operational constraints. The internal validation findings support the robustness of the proposed framework and motivate its evaluation in an external validation cohort, as presented in the following section (Section 4.5).

[Table 4-8] Internal Validation Performance of BAHA-Net

	AUROC	AUPRC	PPV@10%	Recall@10%	Brier	ECE
BAHA-Net	0.752	0.185	0.063	0.414	0.014	0.008

4.3. Comparison with Tuned Baseline Models

4.3.1 Objectives and Methodology of Comparison

To contextualize the performance of the proposed BAHA-Net, we compared its internal validation results with a diverse set of baseline models encompassing conventional statistical methods, ensemble-based machine learning models, and lightweight neural architectures. Specifically, the comparison included Logistic Regression, L1-regularized Logistic Regression, Random Forest, Histogram-based Gradient Boosting (HistGB), LightGBM, XGBoost, as well as Tiny-CNN and Tiny-LSTM. In addition, the recently proposed iREAD ensemble model was reproduced and evaluated under the same experimental setting. All baseline models were trained and evaluated using the same 80:20 temporal split as the proposed BAHA-Net to prevent information leakage across time. To ensure fair comparison, basic class imbalance-aware training strategies and hyperparameter tuning were consistently applied across all baseline models, and post-hoc probability calibration

using isotonic regression or Platt scaling was subsequently performed, depending on model characteristics. Model performance was assessed using AUROC, AUPRC, PPV@10%, and Recall@10%, with 95% confidence intervals estimated via 2,000 stratified bootstrap replicates. This comparative framework enables a comprehensive evaluation not only of discrimination performance but also of calibration quality and clinical utility under resource-constrained triage settings, which is critical for real-world deployment in pediatric intensive care environments.

4.3.2 Discrimination Performance

Figure 4-8(a) and Table 4-9 summarize the discrimination performance across the compared models. The proposed BAHA-Net achieved an AUROC of 0.752 (95% CI: 0.643–0.850), outperforming conventional statistical models such as Logistic Regression (0.666) and ensemble-based methods including Random Forest (0.609). In contrast, more substantial differences were observed in AUPRC, a metric particularly informative for rare-event prediction. BAHA-Net achieved an AUPRC of 0.185 (95% CI: 0.061–0.341), which was markedly higher than that of Logistic Regression (0.061), Random Forest (0.021), HistGB (0.031), and LightGBM (0.030). Even compared with tuned gradient-boosting models and the iREAD ensemble (AUPRC 0.038), BAHA-Net demonstrated superior precision–recall performance. Given the low prevalence of early readmission events in this cohort

(≈1.86%), these AUPRC gains indicate improved robustness under extreme class imbalance, where false positives may impose substantial operational burdens in real-world PICU triage systems.

[Table 4-9] Comparison of Model Discrimination Performance

Model	AUROC (95% CI)	AUPRC (95% CI)
BAHA-Net (ours, PLATT)	0.752 (0.643-0.850)	0.185 (0.061-0.341)
Tiny-CNN (PLATT)	0.712 (0.587-0.829)	0.174 (0.058-0.337)
Logistic L1 (ISO)	0.687 (0.596-0.783)	0.065 (0.026-0.157)
Logistic (ISO)	0.666 (0.537-0.783)	0.061 (0.025-0.146)
XGBoost (ISO)	0.694 (0.585-0.792)	0.049 (0.022-0.087)
Tiny-LSTM (PLATT)	0.667 (0.558-0.765)	0.034 (0.019-0.068)
HistGB (ISO)	0.668 (0.578-0.750)	0.031 (0.016-0.056)
LightGBM (ISO)	0.662 (0.592-0.731)	0.030 (0.015-0.057)
Random Forest (ISO)	0.609 (0.543-0.667)	0.021 (0.013-0.043)
iREAD-Ensemble	0.689 (0.590-0.788)	0.038 (0.020-0.093)

4.3.3 Clinical Utility under a 10% Alert Budget

Under a constrained resource setting simulated by a 10% alert budget—where only the top 10% of patients based on predicted risk are prioritized for intervention—BAHA-Net demonstrated competitive and well-balanced clinical utility when compared with baseline models incorporating hyperparameter tuning and probability calibration Figure 4-8(b) and Table 4-10. Notably, BAHA-Net maintained stable discrimination and robust calibration

while achieving a clinically meaningful level of recall under strict alert constraints. While certain baseline models, such as Logistic Regression and XGBoost, achieved higher point estimates for PPV@10% or Recall@10%, these gains were accompanied by trade-offs in overall discrimination or calibration performance. In contrast, BAHA-Net provided a balanced prioritization profile by jointly maintaining strong discrimination, robust calibration, and clinically meaningful recall. These results suggest that BAHA-Net can effectively support risk-based triage decisions in pediatric intensive care settings, particularly in scenarios where reliable probability estimates and stable prioritization are essential for real-world deployment.

[Table 4-10] Comparison of Clinical Utility Performance

Model	PPV@10 (95% CI)	Recall@10 (95% CI)
BAHA-Net (ours, PLATT)	0.063 (0.032-0.100)	0.414 (0.227-0.609)
Tiny-CNN (PLATT)	0.068 (0.037-0.111)	0.448 (0.273-0.640)
Logistic L1 (ISO)	0.068 (0.032-0.105)	0.448 (0.250-0.609)
Logistic (ISO)	0.074 (0.037-0.116)	0.483 (0.294-0.679)
XGBoost (ISO)	0.074 (0.042-0.116)	0.483 (0.308-0.667)
Tiny-LSTM (PLATT)	0.047 (0.021-0.079)	0.310 (0.133-0.482)
HistGB (ISO)	0.058 (0.021-0.090)	0.379 (0.172-0.520)
LightGBM (ISO)	0.053 (0.016-0.068)	0.345 (0.115-0.429)
Random Forest (ISO)	0.032 (0.005-0.053)	0.207 (0.040-0.320)
iREAD-Ensemble	0.047 (0.016-0.068)	0.310 (0.115-0.433)

4.3.4 Summary

The comparative analysis demonstrated that BAHA-Net achieved strong and competitive discrimination performance, as measured by AUROC, when compared with fully tuned baseline models, including conventional statistical approaches, ensemble-based machine learning methods, and the recently proposed iREAD ensemble. In addition to its favorable AUROC performance, BAHA-Net exhibited a pronounced advantage in AUPRC, underscoring its robustness in rare-event prediction settings characterized by severe class imbalance. Under clinically relevant alert budget constraints, BAHA-Net consistently provided a balanced prioritization profile by jointly maintaining discrimination performance, stable probability calibration, and clinically meaningful recall. While certain tuned baseline models achieved higher point estimates for specific prioritization metrics, these gains were often accompanied by trade-offs in calibration stability or overall robustness. In contrast, BAHA-Net demonstrated stable and reliable performance across complementary evaluation dimensions. When benchmarked against lightweight neural architectures such as Tiny-CNN and Tiny-LSTM, BAHA-Net achieved comparable or improved performance across discrimination and prioritization metrics. Taken together, these findings suggest that the hybrid design of BAHA-Net—integrating temporal representations with static clinical features and calibrated probability estimation—supports stable and clinically applicable early readmission risk stratification in pediatric intensive

care settings, as summarized in Table 4-11.

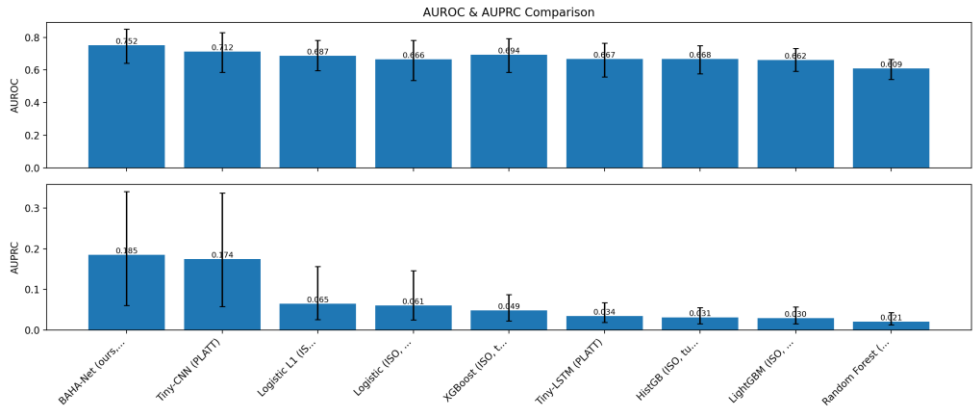
[Table 4-11] Comparison of BAHA-Net and Tuned Baseline Models on Internal Validation

Model	AUROC (95% CI)	AUPRC (95% CI)	PPV@10 (95% CI)	Recall@10 (95% CI)	Brier (95% CI)	ECE (95% CI)
BAHA-Net (ours, PLATT)	0.752 (0.643-0.850)	0.185 (0.061-0.341)	0.063 (0.032-0.100)	0.414 (0.227-0.609)	0.014 (0.009-0.019)	0.008 (0.004-0.013)
Tiny-CNN (PLATT)	0.712 (0.587-0.829)	0.174 (0.058-0.337)	0.068 (0.037-0.111)	0.448 (0.273-0.640)	0.014 (0.010-0.020)	0.010 (0.004-0.015)
Logistic L1 (ISO)	0.687 (0.596-0.783)	0.065 (0.026-0.157)	0.068 (0.032-0.105)	0.448 (0.250-0.609)	0.015 (0.010-0.020)	0.008 (0.003-0.013)
Logistic (ISO)	0.666 (0.537-0.783)	0.061 (0.025-0.146)	0.074 (0.037-0.116)	0.483 (0.294-0.679)	0.015 (0.010-0.020)	0.008 (0.003-0.013)
XGBoost (ISO)	0.694 (0.585-0.792)	0.049 (0.022-0.087)	0.074 (0.042-0.116)	0.483 (0.308-0.667)	0.015 (0.010-0.021)	0.006 (0.001-0.011)
Tiny-LSTM (PLATT)	0.667 (0.558-0.765)	0.034 (0.019-0.068)	0.047 (0.021-0.079)	0.310 (0.133-0.482)	0.015 (0.010-0.021)	0.006 (0.002-0.011)
HistGB (ISO)	0.668 (0.578-0.750)	0.031 (0.016-0.056)	0.058 (0.021-0.090)	0.379 (0.172-0.520)	0.015 (0.010-0.020)	0.006 (0.001-0.011)
LightGBM (ISO)	0.662 (0.592-0.731)	0.030 (0.015-0.057)	0.053 (0.016-0.068)	0.345 (0.115-0.429)	0.015 (0.010-0.020)	0.006 (0.001-0.011)
Random Forest (ISO)	0.609 (0.543-0.667)	0.021 (0.013-0.043)	0.032 (0.005-0.053)	0.207 (0.040-0.320)	0.015 (0.010-0.021)	0.006 (0.001-0.011)

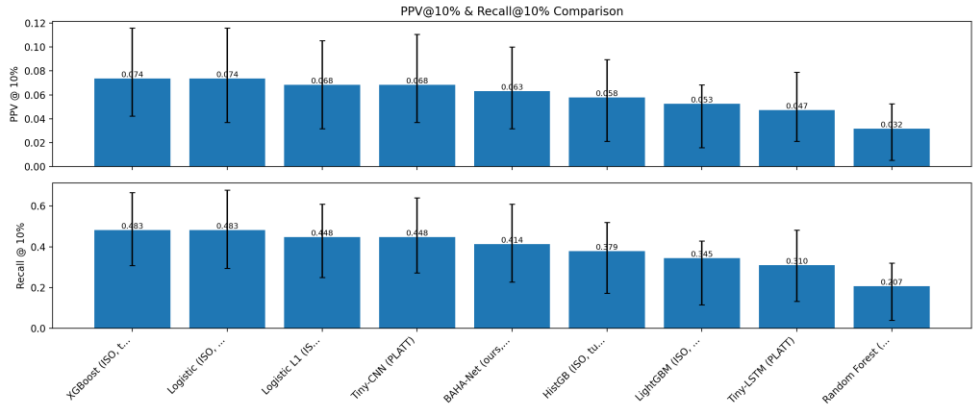
iREAD-Ensemble	0.689 (0.590-0.788)	0.038 (0.020-0.093)	0.047 (0.016-0.068)	0.310 (0.115-0.433)	0.015	0.005

[Figure 4-8] Performance Comparison of BAHA-Net and Baseline Models

(a)



(b)



4.4. Ablation Study

To evaluate the contribution of each architectural and training component of BAHA-Net, an ablation study was conducted using the internal validation cohort. Starting from the full configuration—which integrates multi-channel temporal encoding, gate-based

feature selection with L1 regularization and whitelist constraints, focal loss, event-balanced sampling, and exponential moving average (EMA) smoothing—each key component was individually removed to quantify its impact on performance. All ablation models were evaluated under identical experimental conditions with post-hoc probability calibration to ensure comparability. As shown in Table 4-12, the full BAHA-Net achieved strong overall performance (AUROC = 0.752, AUPRC = 0.185, PPV@10% = 0.063, Recall@10% = 0.414), serving as the reference for subsequent ablations. Removing focal loss or the event-balanced sampling strategy led to consistent declines in discrimination and recall, highlighting the importance of imbalance-aware training. Ablations targeting the feature selection mechanism, including removal of gate-based L1 regularization or the whitelist constraint, resulted in reduced prioritization performance, suggesting that sparsity-aware gating stabilizes learning under irregular and incomplete PICU time-series data. The most pronounced performance degradation was observed when the temporal encoder was simplified from a multi-channel to a single-channel representation, underscoring the critical role of multi-channel temporal encoding in capturing temporal dynamics and missingness patterns. Removing EMA smoothing primarily affected discrimination stability, while recall remained largely unchanged. Overall, these results demonstrate that BAHA-Net's architecture and training pipeline are synergistic, with each component contributing meaningfully to robust early prediction of rare 72-hour

PICU readmission events. In particular, the combination of multi-channel temporal encoding and sparsity-aware feature selection suggests that it is important for maintaining predictive reliability in highly heterogeneous and incomplete clinical data.

[Table 4-12] Ablation Study Discrimination and Triage Performance

Setting	AUROC	AUPRC	PPV@10	Recall@10
FULL	0.752	0.185	0.063	0.414
Without FOCAL	0.715	0.139	0.058	0.379
Without SAMPLER	0.687	0.112	0.053	0.345
Without EMA	0.666	0.146	0.063	0.414
Without GATE_L1	0.739	0.148	0.058	0.379
Without WHITELIST	0.676	0.148	0.058	0.379
Without 3CHANNEL	0.630	0.096	0.053	0.345

[Table 4-13] Performance Changes Relative to the Full BAHA-Net

Setting	⊿AUROC	⊿AUPRC	⊿PPV@10%	⊿Recall@10%
Full	0.000	0.000	0.000	0.000
-Focal	-0.037	-0.046	-0.005	-0.035

−Sampler	−0.065	−0.073	−0.010	−0.069
−EMA	−0.086	−0.039	0.000	0.000
−GateL1	−0.013	−0.037	−0.005	−0.035
−Whitelist	−0.076	−0.037	−0.005	−0.035
−3ch	−0.122	−0.089	−0.010	−0.069

4.5. External Validation Performance

4.5.1 Cohort Characteristics

For external validation, a pediatric intensive−care cohort was extracted from the MIMIC−III database, which contains a total of 61,532 ICU stays. Admissions involving patients younger than 18 years at both admission and discharge were selected, yielding a cohort of 8,200 ICU admissions from 7,967 unique pediatric patients. Among these, 125 admissions (1.52 %) satisfied the criterion for unplanned readmission within 72 hours after discharge. This readmission rate is comparable to that of the internal PhysioNet development dataset (combined training and validation sets, 1.89 %), confirming that early PICU readmission remains a rare and highly imbalanced event.

It also indicates that the external validation dataset preserves a class−imbalance structure consistent with the internal development cohort, ensuring comparability and reliability of model inference across institutions. The demographic composition of the MIMIC−III pediatric cohort is summarized in Table 4−14.

[Table 4-14] Baseline characteristics of the independent validation cohort (MIMIC-III Pediatric)

Cohort	Variable	Value
External Validation Cohort (MIMIC-III Pediatric)	Admissions (N)	8200
	Unique patients (N)	7967
	Readmit+ admissions (n, %)	125 (1.52%)
	Patients with ≥1 readmit (n, %)	125 (1.57%)
	Sex: Male (n, %)	4430 (54.02%)
	Sex: Female (n, %)	3770 (45.98%)
	Age <1 (n, %)	8100 (98.78%)
	Age 1-4 (n, %)	0 (0%)
	Age 5-9 (n, %)	0 (0%)
	Age 10-14 (n, %)	1 (0.01%)
	Age 15-17 (n, %)	99 (1.21%)

4.5.2 External Validation Results

The BAHA-Net model, trained exclusively on the internal PhysioNet Pediatric ICU dataset, was directly applied to an independent external validation cohort derived from the MIMIC-III Pediatric database. The external cohort preserved an identical feature structure, consisting of four static variables and twenty-two temporal clinical variables, ensuring full architectural compatibility with the trained model. To address potential distributional discrepancies between the internal and external datasets, feature harmonization procedures were applied prior to

inference, including winsorization based on internal quantile thresholds, z-score restandardization using internal scaling parameters, and the incorporation of observation-mask channels to explicitly represent missingness patterns. Model inference was conducted using the Exponential Moving Average (EMA) weights obtained during internal training, without any fine-tuning on the external data. Post-hoc probability calibration was subsequently applied following the same strategy selected during internal validation. Performance metrics and 95% confidence intervals were estimated using 2,000 stratified bootstrap resamples. As summarized in Table 4-15, BAHA-Net demonstrated stable discrimination performance in the external cohort. The externally validated AUROC was 0.745 (95% CI: 0.712–0.778), indicating preserved risk-ranking capability under institutional domain shift. Given the low prevalence of early readmission events (≈1.5%), the AUPRC was 0.039 (95% CI: 0.027–0.054), exceeding the expected random baseline.

In terms of clinical utility, BAHA-Net maintained meaningful triage performance under constrained alert budgets. At a 10% alert threshold, the model identified approximately 25.5% of true early readmission cases while maintaining precision above random selection (Table 4-15).

[Table 4-15] External Validation Performance with 95% Confidence Intervals

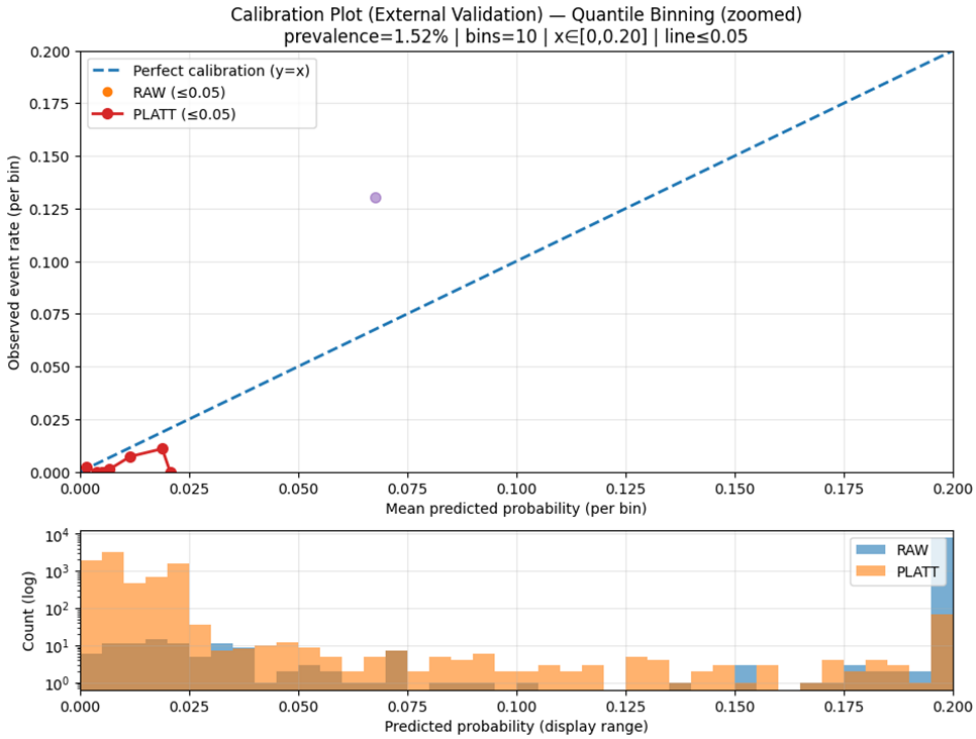
Calibration	AUROC (95% CI)	AUPRC (95% CI)	PPV@10 (95% CI)	Recall@10 (95% CI)	Brier (95% CI)	ECE (95% CI)
BAHA-Net	0.745 (0.712-0.778)	0.039 (0.027-0.054)	0.039 (0.026-0.052)	0.255 (0.181-0.333)	0.015 (0.013-0.018)	0.001 (0.000-0.004)

4.5.3 Calibration Plots of BAHA-Net in External Validation Cohorts

Calibration plots were used to visually examine the probabilistic outputs of BAHA-Net in the external validation cohort, as shown in Figure 4-9. Owing to the extremely low prevalence of early readmission events (≈1.5%) and the infant-dominant composition of the MIMIC-III Pediatric cohort, predicted probabilities were largely concentrated within the low-probability range.

Under these conditions, calibration plots based on quantile binning exhibited limited resolution in higher probability regions due to the small number of observed events. Nevertheless, comparison of the uncalibrated and Platt-calibrated predictions indicated differences in the alignment between predicted probabilities and observed event rates within the low-probability range, which constitutes the majority of predictions in the external cohort.

[Figure 4-9] Calibration Plots of BAHA-Net in External Validation Cohort



4.5.4 Comparison with Tuned Baseline Models (External Validation)

To contextualize the external validation performance of BAHA-Net, its results were compared with a set of tuned baseline models evaluated on the same MIMIC-III Pediatric cohort. As summarized in Table 4-16, the comparison models included L1-regularized logistic regression, random forest, gradient boosting–based methods (XGBoost, LightGBM, HistGB), and lightweight neural network architectures (Tiny-CNN and Tiny-LSTM). All baseline models were refit on the external dataset using the same feature set and evaluated under identical evaluation conditions, including the top 10% alert-budget setting. Overall, baseline models demonstrated attenuated discrimination performance in the external cohort, with

AUROC values ranging from 0.53 to 0.66 and AUPRC values close to the prevalence-based random baseline. In contrast, BAHA-Net exhibited comparatively higher discrimination and a more balanced prioritization profile under the same evaluation setting.

[Table 4-16] Comparison of Model Discrimination Performance

Model	AUROC	AUPRC	PPV@10	Recall@10	Brier	ECE
Logistic (L1)	0.656	0.02297	0.018	0.12	0.015	0.0006
Logistic	0.612	0.01994	0.024	0.16	0.015	0.0004
XGBoost	0.582	0.01986	0.018	0.12	0.015	0.0014
LightGBM	0.589	0.01943	0.018	0.12	0.015	0.0005
Random Forest	0.587	0.01870	0.024	0.16	0.015	0.0005
HistGB	0.530	0.01760	0.024	0.16	0.015	0.0037
Tiny-CNN	0.574	0.02951	0.018	0.12	0.015	0.0019
Tiny-LSTM	0.658	0.02776	0.036	0.24	0.015	0.0031

4.5.5 Interpretation of External Validation Results

Despite being trained exclusively on the internal PhysioNet Pediatric ICU dataset, BAHA-Net demonstrated stable discrimination performance when applied to the independent MIMIC-III Pediatric cohort. The externally validated AUROC remained comparable to that observed during internal validation,

suggesting that the model's ability to rank patients by early readmission risk was largely preserved under institutional domain shift. It should be noted that the external validation setting was characterized by a markedly low event prevalence (≈1.5%), an infant-dominant age distribution, and substantially increased sparsity in longitudinal observations. To mitigate these structural differences, minimal preprocessing steps—such as carry-forward–based handling of sparse temporal measurements and simple imputation for remaining missing values—were applied consistently prior to inference, without any refitting of model parameters. Nevertheless, these conditions are expected to disproportionately affect precision–recall–based metrics. Accordingly, the observed AUPRC in the external cohort was lower than that observed internally, yet remained clearly above the prevalence-based random baseline, indicating preserved precision–recall discrimination under challenging conditions. When contextualized against tuned baseline models evaluated on the same external cohort (Table 4-16), BAHA-Net exhibited comparatively more balanced discrimination and triage performance. While most baseline models showed attenuated AUROC and AUPRC values close to the random baseline, BAHA-Net consistently achieved higher PPV@10 and Recall@10, reflecting a more favorable trade-off between sensitivity and alert burden under a constrained alert budget. These differences were particularly evident when comparing BAHA-Net with simpler linear and tree-based models, which demonstrated limited recall in the low-prevalence setting.

From a clinical perspective, BAHA-Net maintained meaningful triage behavior under realistic operational constraints. At a 10% alert threshold, the model identified a non-trivial proportion of true early readmission cases while maintaining precision above random selection, supporting its potential utility for risk prioritization rather than absolute event prediction. Taken together, the external validation results suggest that BAHA-Net exhibits clinically interpretable risk-ranking behavior in an independent pediatric ICU cohort, even in the presence of severe class imbalance and domain shift. Although these findings do not imply immediate generalizability without site-specific adaptation, they provide supportive evidence that the proposed framework demonstrates relative robustness compared with commonly used baseline models and warrants further investigation in prospective and institution-tailored deployment settings.

Chapter 5. Discussion

5.1. Performance Comparison and Clinical Relevance

The proposed BAHA-Net model achieved an AUROC of 0.752 (95% CI: 0.643-0.850), an AUPRC of 0.185, a PPV@10% of 0.063, and a Recall@10% of 0.414 for predicting early readmission within 72 hours in the Pediatric Intensive Care Unit (PICU). These results indicate that BAHA-Net effectively distinguishes between high- and low-risk patients under extreme class imbalance while maintaining clinically meaningful precision at operationally realistic alert budgets.

5.1.1. Comparative Performance in the Context of Prior Studies

This level of performance is highly competitive compared with previously reported PICU readmission prediction studies. Studies [38,39] reported AUROC values ranging from 0.613 to 0.811 for early PICU readmission prediction, employing logistic regression or nomogram-based approaches that relied heavily on manually selected clinical variables. Unlike those traditional frameworks, BAHA-Net leverages high-frequency time-series data directly extracted from electronic medical records (EMRs), coupled with sparsity-aware feature gating [34,35] and hybrid attention mechanisms [43,44]. This end-to-end design enables the model to

learn complex temporal interactions and missingness patterns that conventional regression-based models cannot capture.

5.1.2. Rare-Event Robustness and Budget-Aware Evaluation

Because early PICU readmission is a rare outcome (approximately 2%), conventional discrimination metrics such as AUROC may overestimate real-world clinical utility [15,50]. Accordingly, this study emphasized budget-aware evaluation, focusing on PPV@10% and Recall@10%, which more closely reflect realistic alert capacities in intensive care environments. Under a fixed 10% alert budget, BAHA-Net identified approximately 40% of true early readmissions while achieving precision several-fold higher than the baseline event rate. This prioritization behavior supports the potential utility of the model for selective monitoring under constrained clinical resources and is consistent with prior studies highlighting the importance of evaluating predictive models under fixed alert budgets in highly imbalanced clinical settings [40,42,50-52].

5.1.3. Comparison with Adult ICU Models

Adult ICU readmission prediction models typically report higher discrimination performance (AUROC ≈ 0.75-0.80), reflecting differences in cohort size, event prevalence, and clinical homogeneity rather than intrinsic model superiority [40-42]. Adult

ICU datasets such as MIMIC-IV and eICU generally contain larger sample sizes, higher readmission rates, and more standardized discharge practices, all of which facilitate stronger feature stability and calibration.

In this study, BAHA-Net was additionally compared with a representative adult ICU readmission framework (iREAD), which was originally developed and validated using adult ICU cohorts. For comparison, the iREAD framework was re-implemented and evaluated on the pediatric ICU dataset under matched experimental conditions. In comparison with the reproduced iREAD model, BAHA-Net demonstrated stable calibration performance and clinically interpretable probability estimates in the pediatric intensive care unit setting. These findings underscore the importance of domain-specific model design for pediatric critical care, where heterogeneous disease profiles, age-dependent physiological variability, and lower event prevalence pose distinct modeling challenges compared with adult ICUs [33]. Within this more demanding context, BAHA-Net achieves clinically meaningful risk stratification, supporting its suitability for real-world deployment in pediatric ICU discharge management.

5.1.4. Sensitivity Analysis Using a 48-Hour Readmission Definition

To assess the robustness of the proposed model to outcome definition, we conducted a sensitivity analysis using a stricter definition of early readmission within 48 hours after PICU discharge.

The analysis was performed on the internal cohort using the optimal variable configuration (K = 30) identified in the primary 72-hour analysis. Despite increased class imbalance under the 48-hour definition, BAHA-Net maintained stable discrimination performance (AUROC = 0.772), with precision-recall performance comparable to the primary analysis. This indicates that the model's ability to prioritize high-risk patients is preserved under a more stringent clinical definition of early readmission. Probability calibration remained stable after Platt scaling, with substantial reductions in Brier score and ECE compared with uncalibrated outputs (Appendix C, Table C1). Overall, these findings suggest that BAHA-Net demonstrates robust and consistent performance across clinically relevant definitions of early PICU readmission, supporting its applicability in real-world settings with varying operational criteria.

5.2. Model Architecture and Clinical Interpretability

Building upon the performance findings in Section 5.1, this section elaborates the design and interpretability principles underpinning BAHA-Net, focusing on how architectural choices translate into reliable, clinically actionable behavior.

5.2.1 Architectural overview and design rationale.

BAHA-Net adopts a hybrid dual-branch architecture that

integrates short-term physiological dynamics with static clinical context. A time-series encoder processes 24-hour laboratory and vital-sign trajectories immediately preceding discharge, while a static encoder represents demographic and admission-level characteristics. The temporal representation is extracted using 2D convolutions applied to a 3-channel encoding (value-with-decay, observation mask, and elapsed time since the last measurement), which collectively capture measurement intensity, observation patterns, and data timeliness under irregular sampling. To improve interpretability and promote robust learning in sparse observational conditions, BAHA-Net incorporates Regularized Gate-based Feature Selection (RG-FS). Each clinical variable is modulated through a learnable gate parameter, serving as a feature-level attention mechanism that enforces sparsity. This emphasizes physiologically meaningful signals while suppressing noise from infrequently sampled inputs. This mechanism aligns with feature-sparse neural formulations (e.g., LassoNet) and cost-aware elicitation strategies that discourage reliance on unnecessary variables in resource-limited scenarios [34,35]. Training leverages focal loss to address rare early-readmission outcomes [20], and temporal splitting is applied to ensure practical deployment fidelity. The architecture unifies established practices in deep learning for EHR data — including convolution-based trajectory modeling and missingness-aware temporal encoding [16–19,22] — with pediatric readmission-specific design choices [33], implemented on standardized critical-care data principles [21].

5.2.2 Intrinsic interpretability from feature-level attention

Interpretability in BAHA-Net is achieved directly through its architectural design rather than post-hoc surrogate methods. While convolutional layers capture complex temporal dynamics (e.g., trend shifts or volatility) across the 24-hour pre-discharge window, the RG-FS mechanism provides explicit feature-level attention. This enables the model to assign meaningful importance weights to each variable and yields clinically meaningful and interpretable feature contributions. Prior ICU studies have reported that intrinsic attention mechanisms can improve transparency and reliability in clinical time-series modeling [43,44]. Compared with perturbation-based post-hoc explainers such as SHAP and LIME, this built-in feature selection reduces fidelity loss and computational burden — advantages that are critical under real-time constraints in ICU settings [45,46,56]. In pediatric critical care, where operational trust requires clinically plausible rationale, BAHA-Net's sparsity-driven explanations help clinicians trace physiologic risk signatures — such as hemodynamic instability or inflammatory surges — and align these patterns with discharge-time decision-making [14,43,44].

5.2.3 Probability calibration and reliability

Because deployment decisions are made on probabilities, BAHA-Net applies a two-stage post-hoc calibration (temperature/Platt →

isotonic) to stabilize risk estimates. Modern neural networks are often miscalibrated [24,37]; Platt/temperature adjustments followed by non-parametric isotonic correction provide well-founded improvements over naïve logistic calibration in the imbalanced regime [24,25,37]. In this thesis, calibrated outputs are used directly to set operating points and to report ECE and Brier as reliability indicators (metrics definitions in Chapter 3; numerical results in Chapter 4), without re-stating performance figures here. This aligns with TRIPOD guidance emphasizing transparent reporting of discrimination and calibration for clinical prediction models [29]. In rare-event triage, calibrated probabilities enable budget-aware operation—selecting an alert budget α (e.g., 10%) and optimizing PPV@α vs. Recall@α under resource constraints [50–52]—while keeping thresholds clinically interpretable.

5.2.4 Clinical workflow integration (CDS) and human–AI teaming.

Within an EHR/CDS dashboard, BAHA-Net supports: (i) discharge hold for top-risk α% patients, (ii) targeted post-discharge monitoring/call-back, and (iii) ward-level intensified surveillance. Variable- and time-level attention maps allow clinicians to validate alerts, prioritize resources, and document the rationale behind actions, improving team situational awareness across physicians, nurses, and respiratory therapists [46,48,49]. This human-AI collaboration view—evaluating systems not only by AUROC/AUPRC but by decision utility and workflow fit—reflects current

recommendations for pragmatic, user–centered evaluation of clinical ML systems [54,55]. BAHA–Net's explanations are therefore positioned as actionable context, not merely visual artifacts.

5.2.5 Governance: trust, accountability, and lifecycle management.

Gate/attention pathways make the model auditable: users can trace prediction origins, examine feature/time attributions, and monitor drift across updates, enabling recalibration and bias checks consistent with trustworthy–AI roadmaps in healthcare [46–49,53,57]. Logging explanation artifacts facilitates change management (e.g., post–update parity checks) and supports site–specific policy controls over alert budgets and thresholds to mitigate alarm fatigue. Consistent with the broader literature on trustworthy and collaborative clinical AI, explanations serve as an interactive oversight channel; final decisions remain clinician–led, with the model functioning as a cognitive copilot [53–55].

5.2.6 Summary.

BAHA–Net integrates (i) attention–based temporal reasoning, (ii) sparsity–regularized gating for feature transparency, and (iii) calibrated probability outputs into a single deployable stack. By tying model mechanics to clinically meaningful operations—budget–aware triage, transparent justification of alerts, and auditable

deployment—the framework balances accuracy, interpretability, and reliability for pediatric intensive care [14,16–18,20–22,29,33–35,37,43–49,50–55,56–58].

5.3. Strengths, Limitations, and Future Directions

This study proposed BAHA-Net, a deep learning–based prediction framework explicitly designed to address time-series missingness and the challenges of rare-event modeling in pediatric critical care [18, 19, 33, 38]. A key methodological strength lies in the dual-cohort design, which integrates internal development using a single-center PhysioNet pediatric ICU dataset with external validation on the MIMIC-III pediatric subset. This cross-institutional validation supports generalizability and mitigates spectrum bias [15, 21, 29, 33]. The external cohort included all pediatric ICU stays, regardless of length of stay, thereby encompassing both short-stay and high-acuity cases that are often excluded in prior research. This inclusive design enhances external validity and reflects real-world deployment conditions [15, 33]. Another notable strength is the budget-aware and clinically interpretable evaluation, which emphasizes PPV@10% and Recall@10% to reflect operational priorities in intensive care environments. These metrics simulate realistic alert-budget constraints and demonstrate that the model can selectively identify high-risk cases while maintaining a manageable false-alert rate

[14, 40, 51]. This approach highlights the framework's suitability for resource-constrained triage and its practical utility in guiding precision-based monitoring and decision support. Together, these methodological choices ensure both robust discrimination and operational relevance within real-world PICU workflows [14, 33, 40, 51]. Notwithstanding these advantages, several limitations warrant discussion and motivate future work. First, although the two cohorts differ institutionally, the data are geographically limited to East Asia and North America; multi-regional, multi-ethnic validation—potentially via privacy-preserving or federated approaches—would strengthen fairness, robustness, and reproducibility across diverse health systems [15,33,55]. Second, the external validation cohort exhibited notable demographic and structural differences relative to the internal development data. As shown in Table 4-14, the MIMIC-III pediatric subset was heavily skewed toward infants, with 98.78% of patients being under one year of age, and demonstrated substantially increased sparsity in longitudinal measurements, reflecting differences in monitoring intensity and data availability. These characteristics represent a clinically meaningful domain shift rather than a simple reduction in model performance. Under such conditions, precision–recall–based metrics are highly sensitive to changes in population composition and data density, even when outcome prevalence remains similar [15,50–52]. Accordingly, the lower absolute AUPRC observed in the external cohort should be interpreted in the context of altered feature distributions and measurement patterns, rather than as a

loss of discriminative capability. Consistent with this interpretation, although AUPRC decreased in the external cohort, AUROC remained stable across internal and external validation, indicating that patient-level risk ranking was largely preserved despite demographic and structural domain shifts. When evaluated against tuned baseline models under identical alert-budget constraints, BAHA-Net continued to demonstrate comparatively favorable triage-oriented performance, suggesting that the model retained the capacity to identify patients at elevated risk of early readmission even in the presence of domain shift. Third, the granularity of available signals constrained temporal sensitivity: continuous high-resolution streams (e.g., waveforms, ventilator traces) were unavailable, and some laboratory or intervention variables were sparsely sampled; incorporating high-frequency time series and exploring streaming or online architectures (e.g., attention-based temporal models, time-aware RNN variants, or graph-based temporal learners) may enable earlier and more precise risk estimation [18,32,43,44]. Fourth, the present framework focuses on a short-horizon endpoint; while clinically impactful for discharge safety and bed management, extending to longer-term and multi-facet outcomes—such as 30-day readmission, in-hospital mortality, and nosocomial infection—would broaden clinical and policy relevance. Multi-task or multi-outcome learning can capture dependencies among endpoints and improve sample efficiency [15,22,23,33]. Relatedly, the inclusion of diagnostic information as static inputs provided limited incremental

benefit in supplementary analyses. This suggests that, for short-horizon PICU readmission prediction, dynamic physiologic signals may play a more prominent role than diagnostic categories in capturing clinically relevant risk patterns. From an interpretability standpoint, BAHA-Net's regularized gate-based feature selection (RG-FS) provides intrinsic transparency by learning feature salience under sparsity constraints, reducing reliance on post-hoc surrogate explainers and supporting clinically plausible reasoning [31,34,36]. Future work should progress from internal interpretability to prospective, clinician-centered evaluations—assessing how explanations influence trust, decision confidence, and workflow usability—and incorporate interactive dashboards for feedback-driven threshold tuning, mitigation of alarm fatigue, and alignment with bedside practice [46–49,53–55].

Although neuro-symbolic learning represents a compelling long-term direction for reliable and knowledge-guided clinical AI, applying such approaches directly in the present study was not feasible for several methodological and practical reasons. Early PICU readmission lacks well-defined, universally accepted symbolic clinical rules or guideline-based decision pathways that could be readily formalized into a symbolic reasoning module. Introducing symbolic constraints without stable domain knowledge carries substantial risk of model mis-specification, optimization instability, and overfitting—particularly in rare-event settings such as this one. Moreover, constructing differentiable logical components or medical knowledge graphs requires extensive

engineering effort and iterative clinical validation that extend beyond the scope of this study. For these reasons, establishing a robust neural backbone was a necessary first step before integrating symbolic reasoning mechanisms.

Even so, the architectural design of BAHA-Net—particularly its gate-based feature regularization and structured temporal attention—already embodies early elements of neuro-symbolic integration. By explicitly learning sparse, interpretable representations that align with clinically meaningful variables and temporal logic, BAHA-Net reflects an inherent tendency toward knowledge-guided reasoning and approximates symbolic inference pathways within pediatric intensive-care data. Beyond purely data-driven architectures, future iterations of BAHA-Net could evolve further toward a neuro-symbolic direction that combines deep neural representation learning with structured clinical reasoning. Such integration would allow the model not only to detect statistical patterns but also to leverage curated domain knowledge encoded in medical ontologies (e.g., SNOMED CT, ICD hierarchies) and rule-based reasoning engines. By aligning learned attention patterns with symbolic clinical concepts—such as organ systems, treatment pathways, or diagnostic categories—the system can move from correlation-driven prediction toward knowledge-grounded inference, narrowing the gap between black-box models and human clinical reasoning [59–61]. This direction may further enhance interpretability, generalizability, and ethical transparency in high-stakes pediatric care [53,57].

Finally, advancing transparency through concept-based reasoning, contrastive/counterfactual analysis, and human-aligned attention visualization may deepen clinical insight and strengthen ethical assurances, while ongoing calibration governance helps maintain probability fidelity during deployment drift [24,37,53]. In summary, this work illustrates how rigorous model design, multi-cohort validation, and intrinsic interpretability can establish a scientifically robust and ethically grounded foundation for decision support in pediatric intensive care. By broadening data inclusion, enriching physiologic signals, conducting prospective evaluations, and sustaining calibration governance, future research can evolve BAHA-Net toward a clinically integrated, globally scalable, and trustworthy neuro-symbolic AI system for critical-care medicine [29,45,53,59–61].

Chapter 6. Conclusion

This study developed and validated BAHA-Net, a deep learning-based framework for early unplanned pediatric ICU readmission prediction, with intrinsic interpretability and neuro-symbolic potential. In this study, early readmission was defined as unplanned PICU readmission occurring within 72 hours after discharge, while a 48-hour window was additionally evaluated as a sensitivity analysis to examine the consistency and robustness of the primary findings. The work addressed a central challenge in pediatric critical care: anticipating adverse short-term outcomes using sparse, irregular, and severely imbalanced clinical data while maintaining interpretability and operational feasibility. Through its methodological, empirical, and clinical dimensions, this research contributes both to the scientific understanding of EHR-based modeling and to the practical design of trustworthy AI systems in intensive-care medicine.

BAHA-Net was developed using a large-scale single-center dataset (PhysioNet PICU v1.1.0) and externally validated on an independent multi-institutional cohort (MIMIC-III Pediatric). This dual-cohort design ensured rigorous assessment of generalizability, calibration, and clinical reliability. Dynamic features from the 24 hours preceding discharge were encoded into a three-channel temporal tensor—capturing observed values, missingness masks, and time-decay context—while static variables such as age, sex,

and length of stay were processed in a parallel branch. The proposed Regularized Gate-based Feature Selection (RG-FS) mechanism learned variable-level importance in a sparsity-aware manner, offering intrinsic interpretability and resilience to incomplete observations. By integrating temporal attention, focal-loss optimization, and post-hoc calibration, the model achieved both discriminative strength and probabilistic fidelity.

Across internal and external validations, BAHA-Net demonstrated robust and clinically meaningful performance. The model achieved an AUROC of 0.752 and demonstrated stable calibration performance (Expected Calibration Error $\leq$ 0.01, Brier Score $\approx$ 0.01–0.02). Under a 10% alert-budget constraint, the model consistently identified approximately 40% of true early readmissions (Recall@10% $\approx$ 0.41) while maintaining precision several-fold higher than the baseline event prevalence (PPV@10% $\approx$ 3–6%), corresponding to a three- to four-fold improvement over baseline prevalence. These findings confirm that the model can effectively function as a budget-aware clinical triage system capable of prioritizing high-risk patients within limited monitoring resources.

Furthermore, external validation without model re-training—using only re-fit normalization and isotonic recalibration—demonstrated stable discrimination and calibration, affirming that BAHA-Net is not overfitted to a single institution and retains applicability across distinct healthcare environments. Beyond quantitative metrics, this study contributes conceptually to the evolving paradigm of

interpretable and resource-aware deep learning in healthcare. BAHA-Net embodies early characteristics of a neuro-symbolic architecture, as its gate-based feature regularization and structured attention modules encode knowledge-guided reasoning and variable-level logical relations. This fusion of statistical representation and implicit symbolic structure bridges data-driven learning and clinical reasoning, laying the foundation for future neuro-symbolic AI in critical-care medicine.

BAHA-Net also demonstrates how data sparsity and class imbalance—long considered obstacles—can be structurally encoded into the learning process through gated regularization and temporal decay mechanisms. This approach shifts the modeling philosophy from post-hoc correction toward intrinsic robustness, embedding interpretability, sparsity awareness, and calibration directly into the architecture. In doing so, the framework moves beyond the pursuit of raw predictive accuracy to prioritize trustworthiness, transparency, and operational viability—qualities that are indispensable for clinical adoption.

Clinically, the findings have direct implications for discharge decision-making and post-ICU monitoring. By identifying high-risk patients likely to deteriorate within 72 hours, the model can assist physicians in refining discharge timing, allocating step-down-unit beds, and planning follow-up surveillance. Because the system operates under explicit alert-budget constraints, it aligns with real-world workflow capacity, minimizing alarm fatigue while maintaining safety. In this respect, BAHA-Net illustrates how AI

can augment rather than replace human judgment—serving as a transparent decision-support tool that informs resource allocation and patient prioritization in critical-care environments.

From a methodological standpoint, this research advances the field in several ways. First, it demonstrates that multi-channel temporal encoding can effectively represent irregular EHR time series without heavy imputation, preserving both timing and uncertainty information. Second, it establishes regularized gate-based feature selection as a unified mechanism for interpretability and sparsity control, enabling stable learning even with incomplete pediatric data. Third, by coupling budget-aware evaluation metrics with conventional discrimination and calibration analyses, the study provides a comprehensive framework for assessing clinical utility under operational constraints. Finally, through systematic external validation and post-hoc calibration, it sets a reproducible standard for verifying the real-world reliability of deep learning models in medicine.

In conclusion, the BAHA-Net framework represents a neuro-symbolically inspired, interpretable, and clinically validated deep learning model for early pediatric ICU readmission prediction. It demonstrates that the rigorous integration of sparsity-aware modeling, calibrated probability estimation, and operational feasibility can yield an AI system that is not only accurate but also trustworthy and clinically actionable. By bridging data-driven learning with structured reasoning and human-centered design, this research lays the groundwork for the next generation of neuro-

symbolic, precision-oriented critical-care systems—models capable of enhancing patient safety, optimizing hospital operations, and advancing the broader goal of ethical, transparent, and global healthcare AI.

Chapter 7. Appendix

7.1. Appendix A — Supplementary Validation Result

Table A1. Internal Validation Performance of the K = 30 Feature Configuration

Calibration	AUROC (95% CI)	AUPRC (95% CI)	PPV@10 (95% CI)	Recall@10 (95% CI)	Brier (95% CI)	ECE (95% CI)
RAW	0.777 (0.676–0.867)	0.193 (0.077–0.362)	0.079 (0.037–0.116)	0.517 (0.300–0.679)	0.238 (0.233–0.243)	0.451 (0.443–0.459)
PLATT	0.777 (0.676–0.867)	0.193 (0.077–0.362)	0.079 (0.037–0.116)	0.517 (0.300–0.679)	0.014 (0.009–0.019)	0.008 (0.003–0.014)
ISOTONIC	0.789 (0.694–0.871)	0.132 (0.046–0.289)	0.079 (0.037–0.116)	0.517 (0.296–0.676)	0.014 (0.010–0.020)	0.007 (0.002–0.012)

7.2. Appendix B — Supplementary Training Result

Table B1. Apparent Training Performance of the K = 30 Feature Configuration (For Reference)

Calibration	AUROC	AUPRC	PPV@10%	Recall@10%	Brier Score	ECE
RAW	0.845	0.153	0.094	0.483	0.240	0.452
PLATT	0.845	0.153	0.094	0.483	0.018	0.005
ISOTONIC	0.844	0.130	0.096	0.356	0.018	0.003

7.3. Appendix C — Supplementary Sensitivity Analysis

Table C1. Sensitivity Analysis for 48–Hour PICU Readmission Using the Internal Cohort (K = 30)

Calibration	AUROC	AUPRC	PPV@10%	Recall@10%	Brier Score	ECE
RAW	0.772 (0.661–0.877)	0.154 (0.025–0.338)	0.037 (0.011–0.068)	0.368 (0.150–0.591)	0.283 (0.279–0.286)	0.518 (0.512–0.523)
PLATT	0.772 (0.661–0.877)	0.154 (0.025–0.338)	0.037 (0.011–0.068)	0.368 (0.150–0.591)	0.009 (0.006–0.014)	0.006 (0.002–0.010)
ISOTONIC	0.760 (0.647–0.866)	0.143 (0.020–0.322)	0.037 (0.011–0.068)	0.368 (0.143–0.588)	0.010 (0.006–0.014)	0.006 (0.002–0.010)

Bibliography

1. Rojas, E., et al. (2023). Explainable machine learning for ICU readmission prediction. arXiv preprint arXiv:2309.13781v4.
2. Kalzén, H., Larsson, B., Eksborg, S., Lindberg, L., Edberg, K. E., & Frostell, C. (2018). Survival after PICU admission: The impact of multiple admissions and complex chronic conditions. PloS One, 13(4), e0193294. https://doi.org/10.1371/journal.pone.0193294
3. Zimmerman, L. P., et al. (2022). Development and validation of an interpretable 3-day intensive care unit readmission prediction model using explainable boosting machines. Frontiers in Medicine, 9, 960296. https://doi.org/10.3389/fmed.2022.960296
4. Desautels, T., et al. (2021). Prediction of 30-day hospital readmission with clinical notes and EHR information. arXiv preprint arXiv:2103.14050v1.
5. Sharp, E. A., Wang, L., Hall, M., Berry, J. G., & Forster, C. S. (2023). Frequency, characteristics, and outcomes of patients requiring early PICU readmission. Hospital Pediatrics, 13(8), 678-688. https://doi.org/10.1542/hpeds.2022-007100
6. Alam, M. R., et al. (2024). Development and evaluation of a machine learning model for predicting 30-day readmission in general internal medicine. Computers, 14(5), 177.

https://doi.org/10.3390/computers14050177

7. Wang, S., & Zhu, X. (2022). Predictive modeling of hospital readmission: Challenges and solutions. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(5), 2975–2995. https://doi.org/10.1109/TCBB.2021.3089682
8. Chowdhury, M. Z. I., et al. (2022). Predictive model for ICU readmission based on discharge summaries using machine learning and natural language processing. Informatics, 9(1), 10. https://doi.org/10.3390/informatics9010010
9. Shin, Y., et al. (2022). Multicenter validation of a deep-learning-based pediatric early-warning system (pDEWS) for prediction of deterioration events. Acute and Critical Care, 37(2), 181–191. https://doi.org/10.4266/acc.2022.00976
10. da Silva, N. C., et al. (2024). Machine learning for hospital readmission prediction in pediatric population. Computer Methods and Programs in Biomedicine, 244, 107980. https://doi.org/10.1016/j.cmpb.2023.107980
11. Chiu, C. C., et al. (2024). Predicting ICU readmission from electronic health records via BERTopic with long short-term memory network approach. Journal of Clinical Medicine, 13(18), 5503. https://doi.org/10.3390/jcm13185503
12. Ganatra, H. A., et al. (2024). Pediatric intensive care unit length of stay prediction by machine learning. Bioengineering,

11(10), 962.
https://doi.org/10.3390/bioengineering11100962
13. Kim, M., et al. (2024). Development of a deep learning model for predicting critical events in a pediatric intensive care unit. Acute and Critical Care, 40(2), 123-134.
https://doi.org/10.4266/acc.2023.01424
14. Loftis, L. L., et al. (2022). The use of machine learning and artificial intelligence within pediatric critical care. Frontiers in Pediatrics, 10, 966002.
https://doi.org/10.3389/fped.2022.966002
15. Van Calster, B., et al. (2021). Published models that predict hospital readmission: A critical appraisal. BMC Medicine, 19(1), 190.
https://doi.org/10.1186/s12916-021-02010-8
16. Choi, E., et al. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. JMLR Workshop and Conference Proceedings, 56, 301-318.
17. Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1, 18.
https://doi.org/10.1038/s41746-018-0029-1
18. Che, Z., et al. (2018). Recurrent neural networks for multivariate time series with missing values. Scientific Reports, 8, 6085.
https://doi.org/10.1038/s41598-018-24271-9
19. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018).

Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063.

20. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

21. Johnson, A. E. W., et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

22. Harutyunyan, H., et al. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6(1), 96. https://doi.org/10.1038/s41597-019-0103-9

23. Steyerberg, E. W. (2009). Clinical prediction models: A practical approach to development, validation, and updating. Springer.

24. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, 70, 1321–1330.

25. Kull, M., de Menezes e Silva Filho, T., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 54:623–631.

26. Futoma, J., et al. (2017). An improved multi-output Gaussian process RNN with real-time validation for early sepsis detection. arXiv preprint arXiv:1708.05894.

27. Lemaitre, G., et al. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1-5.

28. Shah, N., Arshad, A., Mazer, M. B., Carroll, C. L., Shein, S. L., & Remy, K. E. (2023). The use of machine learning and artificial intelligence within pediatric critical care. Pediatric Research, 93(2), 405–412. https://doi.org/10.1038/s41390-022-02380-6

29. Moons, K. G. M., et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Annals of Internal Medicine, 162(1), W1-W73.

30. Suresh, H., et al. (2017). Clinical intervention prediction and understanding using deep networks. In Proceedings of the Machine Learning for Healthcare Conference (MLHC).

31. Shukla, A., & Ghosh, H. (2022). Continuous patient state attention model for processing irregularly sampled clinical time series. BMC Medical Informatics and Decision Making, 22, 141. https://doi.org/10.1186/s12911-022-01868-5

32. Neil, D., Pfeiffer, M., & Liu, S. C. (2017). Phased LSTM:

Accelerating recurrent network training for long or event-based sequences. In Advances in Neural Information Processing Systems 29 (NIPS 2017) (pp. 3882–3890).

33. Fernandez, A., et al. (2021). Risk prediction models for unplanned 72-hour readmission in pediatric critical care: A systematic review. Critical Care Medicine, 49(6), 972-983.

34. Lemhadri, I., et al. (2021). LassoNet: A neural network with feature sparsity. Journal of Machine Learning Research, 22(33), 1-29.

35. Wang, J., et al. (2021). P-CAFE: A personalized, cost-aware feature elicitation framework for medical diagnosis. arXiv preprint arXiv:2106.03847.

36. Ning, X., Zhao, T., Li, W., Lei, P., Wang, Y., & Yang, H. (2020). DSA: More Efficient Budgeted Pruning via Differentiable Sparsity Allocation. https://doi.org/10.48550/arxiv.2004.02164.

37. Minderer, M., et al. (2021). Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems, 34, 15682-15694.

38. Kaur, R., et al. (2020). Predicting early pediatric ICU readmissions using machine learning. Journal of Critical Care, 55, 163-170.

39. Ding, Y., et al. (2021). A nomogram for predicting unplanned pediatric ICU readmission: A single-center experience. Scientific Reports, 11, 8323.

40. Lim, L., et al. (2025). Multicenter validation of a machine

learning model to predict ICU readmission within 48 hours after discharge. eClinicalMedicine, 75, 103264.

41. Rojas, J. C., et al. (2023). Predicting adult ICU readmission using gradient boosting machines. Critical Care Medicine, 51(3), 271–281.

42. Koumantakis, E., et al. (2025). Deep learning models for ICU readmission prediction: A systematic review and meta-analysis. Critical Care, 29, 442.

43. Kaji, D. A., et al. (2019). An attention-based deep learning model of clinical events in the intensive care unit. PLOS One, 14(2), e0211057.

44. Gandin, I., et al. (2021). Interpretability of time-series deep learning models: A study in cardiovascular patients. Journal of Biomedical Informatics, 121, 103883.

45. Arik, S. O., & Pfister, T. (2020). TabNet: Attentive Interpretable Tabular Learning. arXiv.Org.

46. Qaiser, A., et al. (2024). XAI in ICU decision making: A critical review. Journal of Medical Systems, 48(2), 15.

47. Fathy, W., Émériaud, G., & Cheriet, F. (2025). A comprehensive review of ICU readmission prediction models: From statistical methods to deep learning approaches. Artificial Intelligence in Medicine, 103126. https://doi.org/10.1016/j.artmed.2025.103126

48. Agard, G., et al. (2025). Improving sepsis prediction with explainable AI. Journal of Clinical Medicine, 14(3), 548.

49. Okada, Y., et al. (2023). Explainable AI in emergency

medicine. Clinical and Experimental Emergency Medicine, 10(2), 103–115.

50. Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) (pp. 1–8).

51. Zhang, L., et al. (2023). Optimizing alert precision in imbalanced clinical event prediction: A triage-budget perspective. Artificial Intelligence in Medicine, 143, 102631.

52. Gupta, R., et al. (2024). Budget-aware evaluation metrics for clinical decision support under class imbalance. Journal of Biomedical Informatics, 155, 104571.

53. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. Nature Medicine, 25(10), 1629–1639. https://doi.org/10.1038/s41591-019-0621-3.

54. Ribeiro, M. T., et al. (2023). Beyond interpretability: Evaluating human-AI collaboration in medical decision support. Journal of Biomedical Informatics, 142, 104590.

55. Wiens, J., et al. (2023). Human-centered evaluation of clinical machine learning systems: A call for pragmatic trials. npj Digital Medicine, 6, Article 89.

56. Sadeghi, S., et al. (2024). A review of explainable artificial intelligence in healthcare. Computers & Electrical Engineering, 118, 109076.

57. Ahadian, S., et al. (2025). Ethics of trustworthy AI in healthcare: Challenges and frameworks. Neurocomputing, 599, 128–139.
58. Bohlen, L., Rosenberger, J., Zschech, P., & Kraus, M. (2025). Leveraging interpretable machine learning in intensive care. Annals of Operations Research, 347(2), 1093–1132. doi:10.1007/s10479-024-06226-8.
59. DeLong, L. N., Mir, R. F., & Fleuriot, J. D. (2025). Neurosymbolic AI for Reasoning Over Knowledge Graphs: A Survey. IEEE Transaction on Neural Networks and Learning Systems, 36(5), 7822–7842. https://doi.org/10.1109/TNNLS.2024.3420218.
60. Zhang, J., Chen, B., Zhang, L., Ke, X., & Ding, H. (2021). Neural, symbolic and neural-symbolic reasoning on knowledge graphs. AI Open, 2, 14–35. https://doi.org/10.1016/j.aiopen.2021.03.001.
61. Liang, B. (2025). AI reasoning in deep learning era: From symbolic AI to neural-symbolic AI. Mathematics, 13(4), 723. https://doi.org/10.3390/math13040723

Abstract in Korean

배경

소아중환자실(Pediatric Intensive Care Unit, PICU) 환자의 퇴실 후 조기 재입실을 예측하는 것은 병상 운영의 효율화, 고위험 환자에 대한 선별적 모니터링, 그리고 시의적절한 임상 개입을 가능하게 함으로써 환자 안전과 임상 성과 향상에 기여할 수 있습니다. 그러나 실제 임상 환경에서 생성되는 PICU 전자의무기록(EHR) 데이터는 사건 발생률이 매우 낮고, 관측이 희소하며 시계열 구조가 불규칙하다는 특성을 가지므로, 기존 예측 모델들은 일반화 가능성과 확률 신뢰성 측면에서 한계를 보여왔습니다. 이에 본 연구는 이러한 특성을 고려한 희소성 인지(sparsity-aware) 딥러닝 기반 조기 재입실 예측 프레임워크를 개발하고, 내부 및 외부 독립 코호트를 활용하여 모델의 성능과 신뢰성을 체계적으로 평가하였습니다.

방법

조기 재입실은 PICU 퇴실 후 단기간 내 발생하는 계획되지 않은 재입실로 개념화하였으며, 주 분석에서는 퇴실 후 72시간 이내 재입실을 예측 대상으로 정의하였습니다. 추가적으로 임상적으로 의미 있는 더 엄격한 기준인 48시간 재입실을 적용한 민감도 분석을 수행하였습니다. 모델 개발에는 PhysioNet 단일 기관 PICU 코호트(2010–2018; n = 9,529)를 사용하였고, 외부 검증은 MIMIC-III에서 추출한 소아중환자 코호트(2001–2012; n=8,200)를 이용하였습니다.

퇴실 전 24시간 동안의 시계열 임상 변수는 지수 감쇠값, 관측 마스크, 마지막 관측 이후 경과 시간을 포함하는 3채널 텐서로 인코딩하였으며, 연령 및 성별과 같은 정적 인구통계 변수는 별도의 분기로 처리하였습니다. 제안된 BAHA-Net은 합성곱 기반 시계열 인코더와 다층 퍼셉트론

을 통합한 구조로 설계되었으며, 데이터 희소성을 명시적으로 반영하기 위해 게이트 기반 특성 선택 메커니즘을 포함합니다. 학습 과정에서는 희귀 사건을 고려한 Focal Loss와 지수가중이동평균(Exponential Moving Average, EMA)을 적용하였습니다. 외부 검증 시에는 모델 가중치를 고정한 채 정규화 재적합과 사후 확률 보정을 수행하였습니다. 성능 평가는 판별력, 보정 지표, 그리고 경보 예산(alert budget)을 고려한 우선순위화 지표를 중심으로 이루어졌습니다.

결과

내부 검증에서 제안된 모델은 AUROC 0.752, AUPRC 0.185의 성능을 보여, 조기 재입실 예측에서 신뢰할 수 있는 판별력을 나타냈습니다. 외부 MIMIC-III 소아 코호트에서도 데이터 분포의 차이에 따라 AUPRC는 일부 하락하였으나, AUROC 0.745로 판별 성능은 유지되었습니다. 경보 예산을 10%로 제한한 평가에서는 내부 및 외부 검증에서 각각 전체 조기 재입실 사례의 41.4%와 25.5%를 식별하여, 희귀 사건 환경에서도 임상적으로 의미 있는 우선순위화 성능을 보였습니다.

결론

본 연구는 조기 PICU 재입실 예측을 위한 희소성 인지 딥러닝 프레임워크 BAHA-Net을 제안하고, 내부 및 외부 독립 코호트를 통해 일관된 판별 성능과 안정적인 우선순위화 성능을 확인하였습니다. 특히, 시계열 희소성과 불균형이라는 임상 데이터를 구조적으로 반영한 모델 설계와 게이트 기반 특성 선택 메커니즘은 해석 가능성과 실용성을 동시에 확보하였으며, 이는 향후 뉴로심볼릭(neuro-symbolic) 기반의 신뢰 가능한 임상 의사결정 지원 도구로의 확장 가능성을 시사합니다.

주제어: 소아중환자실(PICU), 재입실 예측, 심층학습, 시계열 데이터

학 번: 2023 – 30649