DL-PRWarnNeuro: A Transformer-Based Deep Learning Framework for Multimodal Patient Deterioration Prediction in Neurological Intensive Care

Man Hua
School of Nursing, Zhengzhou Yellow River Nursing Vocational College, Zhengzhou 450066, Henan, China
Corresponding author's E-mail: hm_huaman@hotmail.com

Keywords: Neurological intensive care, vital sign monitoring, deep learning, Transformer model, attention mechanism, Focal Loss, patient deterioration prediction, alarm reduction, clinical decision support, early warning system.

Received: September 28, 2025

Monitoring vital signs is extremely important in neurological intensive care, as even minor changes can indicate a significant decline in the patient's condition. The purpose of this work is to develop DL-PRWarnNeuro, a deep learning model intended to provide early, accurate alerts in neurological intensive care settings by leveraging patient signal data from the MIMIC-IV dataset. A convolutional layer is incorporated into the model's architecture to extract local patterns. This is followed by transformer-based attentive layers, designed to focus on crucial signal fluctuations. During training, Focal Loss is used to address class imbalance between typical, stable states and rare, critical events. The performance of DL-PRWarnNeuro is tested against baseline models, including LSTM classifiers and standard threshold-based monitoring. Among the most important indicators are the early detection rate, the False Alarm Rate (FAR), and the Area Under the Receiver Operating Characteristics curve (AUROC). Compared to the baselines, the results show a 14% increase in AUROC, a 22% decrease in FAR, and an 18% increase in the early detection rate. These findings indicate that there was an improvement in clinical reliability and support for decisions for neurological intensive care unit monitors.

Povzetek: Prispevek predstavlja globokoučni model za zgodnje zaznavanje poslabšanja stanja v nevrološki intenzivni terapiji, ki izboljša natančnost opozoril in zmanjša število lažnih alarmov v primerjavi z obstoječimi metodami.

1 Broad overview of the neurological intensive care monitoring

Neurological intensive care units (NICUs) manage critically ill patients with post-neurosurgery complications, stroke, or traumatic brain injury [1]. The most crucial element in early deterioration detection is real-time monitoring of intracranial pressure, oxygen saturation, blood pressure, and heart rate [2]. Although medical devices have become increasingly advanced, threshold-based monitoring systems still cannot detect nonlinear variations in these parameters. Alarm fatigue, stemming from excessive false alarms, remains a challenge for clinical staff [3]. In these settings, there is an urgent need for intelligent systems that can identify emergent patterns in dynamic physiological signals and provide timely interventions, thereby minimizing patient mortality rates [4]. Computational intelligence developments have now redefined the promise of critical care monitoring. Machine and deep learning techniques can detect subtle interdependencies in large patient datasets that other systems cannot [5]. The detection of temporal patterns and complex associations between various signals using such techniques enables the accurate identification of subtle warning signs before clinically evident deterioration. Incorporating such models into NICU practice improves predictive ability, reduces inappropriate alarms, and facilitates timely intervention [6]. As a result, artificial intelligence has become a necessary addition to traditional surveillance systems, filling the gap between raw physiological information and clinically interpretable data [7].

1.1 Deep learning in healthcare time-series analysis

Deep learning is a robust algorithm for the processing of sequential and high-dimensional data at an exceptionally high speed. Physiological signals, such as ECG, EEG, and vital signs, are continuous-time series, and models should be capable of extracting both short- and long-term dependencies [8]. Support vector machines, logistic regression, and other traditional models cannot accurately capture the temporal dynamics in clinical datasets. Deep models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and, more recently, Transformer-based models, present enormous benefits for modeling long-range dependencies [9]. These networks can handle noisy, irregular, and incomplete medical data with high predictive value. Transformers, initially designed for natural language processing, have now yielded promising results in time-series forecasting applications. Their self-attention mechanism enables the detection of clinically meaningful variability, highlighting relevant

features of patient data [10]. Transformers outperform RNNs and LSTMs in learning long-term dependencies and handling large-scale data. Merging deep learning with critical care monitoring holds the potential to provide real-time, interpretable, and clinically meaningful decision support [11]. Therefore, the application of cutting-edge neural architectures to NICU data is a timely and promising research direction, expanding beyond the constraints of standard monitoring and aligning with modern calls for precision medicine.

Despite ongoing patient surveillance in NICUs, current systems often rely on naive, rule-based thresholds that fail to capture the complexity of physiological interactions. These shortcomings result in undue false alarms, alarm fatigue, and an inability to provide early warning of patient deterioration. Machine learning techniques have been employed recently; however, most models are plagued by imbalanced data, noise, and non-stationary time-series signals, which are common in intensive care settings. Critical deterioration events are the exception, and thus, strong and sensitive models are complex to train without knowledge-intensive optimization techniques. Most models are also non-transferable from patient to patient with different baselines, which undermines clinical utility. A knowledge system can address this issue by leveraging a deep learning framework that captures intricate temporal interactions, handles class imbalance, and generates timely, reliable alarms to enhance patient safety and support clinical decision-making in neurological intensive care.

1.2 Research design

This research endeavor aims to assess the predictive performance, interpretability, and robustness of the DL-PRWarnNeuro model for real-time monitoring in the Neuro-Intensive Care Unit. The testing of high-fidelity MIMIC-IV waveform data, the evaluation of the reduction of false alarms, and the quantification of the contribution of each signal modality are all included in this action.

1.3 Research goals

The research aims to assess the DL-PRWarnNeuro framework in real-time Neuro-ICU monitoring by concentrating on three primary objectives: (i) accurately forecasting critical deterioration events through multimodal physiological signals, (ii) minimizing false alarms in comparison to conventional threshold-based or baseline machine learning methodologies, and (iii) delivering interpretable insights regarding the contributions of various signal modalities for clinical decision-making. Deterioration events were designated according to clinically recognized criteria, including elevations in intracranial pressure, reductions in SpO_2, and atypical trends in arterial blood pressure. Automated threshold assessments were corroborated through expert clinical evaluation to guarantee labeling precision.

1.4 Contributions of the work

- To propose DL-PRWarnNeuro, a novel Transformer-based deep learning framework for patient vital sign monitoring in neurological intensive care.
- To integrate convolutional and attention-based layers for robust extraction and interpretation of temporal physiological patterns.
- To apply Focal Loss for handling data imbalance, ensuring reliable detection of rare but critical deterioration events.
- To validate the framework on real-world ICU datasets, achieving improved accuracy, early detection, and reduced false alarms.
- To provide a clinically interpretable decision-support tool designed for integration into neurological intensive care monitoring workflows.

2 Literature review

Neurocritical care patient monitoring is gaining increasing attention due to the acute need for timely and effective identification of deterioration. Conventional threshold-based approaches yield a high false alarm rate, leading to clinician burnout and diminished alert credibility. Deep learning presents an alternative by learning rich patterns in vital signs and enhancing predictive accuracy. Convolutional, recurrent, and Transformer models have produced encouraging results. This article discusses existing methods, identifies their limitations, and situates DL-PRWarnNeuro within the context of ongoing progress.

2.1 Machine learning approaches for patient vital sign prediction

Maleczek et al. (2024) [12] proposed a comparison of five machine learning algorithms for artifact detection in electronic vital sign data. The authors performed a retrospective study using five physiological signals, standard algorithms, and ML-based pattern recognition. The aim was to enhance the quality of the monitoring system by eliminating noise-related errors. Studies indicated that ML models outperformed traditional approaches in identifying artifacts with higher accuracy. The study, however, was undermined by the retrospective nature of the data and the absence of real-time validation.

Lee et al. (2024) [13] suggested an artificial intelligence-informed multimodal prediction system for in-hospital cardiac arrest among ICU patients. Clinical data, laboratory results, and vital signs were combined using sophisticated ML models to estimate the risk. It was suggested that timely intervention in critical illnesses could decrease mortality. Outcomes showed greater predictive accuracy than conventional scoring systems. However, the system's drawback was its reliance on the exhaustiveness of multimodal information, which can be guaranteed only in an ideal world.

Yang et al. (2024) [14] developed a real-time predictive model for sepsis in preterm infants based on bedside vital sign monitoring. Machine learning algorithm-based clinical alarm management was utilized to make predictions regarding sepsis development. The motivation behind it was to support early diagnosis and mitigate unnecessary alarms. The results demonstrated that the system improved early sepsis detection and reduced false alarms. Still, the research was limited to specific neonatal ICU datasets, with limited generalizability to broader populations and diverse healthcare settings.

Verma et al. (2024) [15] developed a machine-learning-based early warning system (EWS) to predict patient deterioration in the clinical environment. Large patient datasets were utilized to train the ML algorithms, which were clinically validated for real-time use. It was designed to enhance decision support in the clinic and lower unplanned adverse events. The results indicated that the ML-based EWS was superior to traditional scoring systems for prediction. While it was valid, its flaw lay in the lack of transparency in decision-making (a black-box approach) that undermined clinicians' confidence and slowed its diffusion.

Yang et al. (2025) [16] introduced an XGBoost model for early sepsis prediction with single-time-point, non-invasive vital signs. The approach integrated machine learning predictions with clinical biomarkers, including CRP and procalcitonin, within a multi-center trial. It was designed to be a cost-effective, early diagnostic strategy for sepsis. The outcomes indicated high predictive performance and strong biomarker correlations. The disadvantage, nonetheless, was the use of single-point measurements, which may be unable to demonstrate dynamic disease progression and may decrease reliability in unstable clinical conditions.

Malde et al. (2025) [17] proposed a machine learning framework to predict maternal health risk in lower-middle-income countries using sparse vital sign data.The need for low-cost maternal risk prediction software in resource-constrained healthcare settings prompted this. Results showed that the system could make accurate predictions even with sparse data. However, the drawback was its reliance on data availability and potential biases arising from the underrepresentation of the target population in the training dataset.

2.2 Deep learning for patient monitoring and risk prediction

Thatha et al. (2025) [18] developed an effective ML algorithm for health big data systems to predict disease risk factors. This was achieved by selecting features and utilizing massive ML algorithms specifically designed for large-scale health data. The objective of the motivation was to improve the computation speed for high-dimensional health data. However, the limitation was scalability in handling large volumes of real-time streaming data, which might require additional integration with distributed computing frameworks.

Deshmukh et al. (2025) [19] suggested ML-based techniques for aflatoxin detection and food safety risk prediction. Using supervised learning with training datasets on food quality, the models accurately labeled contamination risk. The aim was to empower innovative automated detection techniques to enhance food safety surveillance. The outcomes indicated that ML detection accuracy was higher than that of traditional laboratory-based testing. Dependence on high-quality, domain-related training data was a limitation, making it difficult to generalize across different food types and conditions.

Doneda et al. (2025) [20] proposed an ECG machine learning model to estimate mortality risk in a large European cohort. ECG time-series recordings were used to train ML models and classify risk. It was suggested for population-level prevention and for improving heart monitoring. However, the limitations were that only ECG data were used and no interactions with other clinical variables were considered, which may reduce the predictive strength in multifactorial cardiovascular diseases.

Sharma et al. (2023) [21] discuss integrating various monitoring modalities in neurocritical care, emphasizing the importance of synchronizing data from multiple sources to enhance patient outcomes. In the context of critical care settings, the study illustrates how combining data from multiple monitoring approaches can provide a comprehensive view of a patient's status, enabling more informed decision-making.

Ramesh et al. (2021) [22] presented a machine-learning-based remote healthcare monitoring system for diabetes prediction. Lifestyle and clinical characteristics were assessed using ML algorithms to categorize risk. Emphasis was laid on facilitating early detection of diabetes and remote patient management. Findings indicated high predictive accuracy and dependability. However, these constraints relied on patient-recorded data, which can be unreliable, and the inability to extrapolate findings to other populations due to differences in lifestyle and genetics.

Vitt et al. (2023) focus on common clinical neuromonitoring techniques —such as intracranial pressure, brain tissue oxygenation, transcranial Doppler, and near-infrared spectroscopy —and emphasize how each modality can provide valuable information on cerebral autoregulation capacity. The study highlights the importance of measuring cerebral autoregulation to guide therapeutic approaches and improve patient outcomes in neurocritical care.

2.3 Emerging transformer-based architectures in healthcare

Alghieth (2025) [24] suggested DeepECG-Net, a deep learning hybrid transformer model for real-time ECG anomaly detection. The technique employed CNN and transformer blocks for sequence modeling and anomaly detection. Extremely high sensitivity and specificity were claimed over conventional DL models. Drawbacks included computational complexity and potential real-time

latency, which hindered its use in resource-constrained clinical environments.

Wang et al. (2024) [25] suggested a deep learning transformer model for diagnosing suspected lung cancer in primary care from EHRs. The model employed attention mechanisms to capture structured EHR sequences. Results indicated enhanced diagnostic accuracy relative to baseline models. The limitation was reliance on structured EHR data, which might be heterogeneously distributed in completeness and quality across healthcare providers, thereby limiting the model's generalizability.

Mei et al. (2024) [26] introduced a gated transformer model to improve the management of acute coronary syndrome (ACS). The objective was to optimize accuracy in patient outcome and treatment planning. Improved risk prediction and treatment optimization were demonstrated over baseline models. However, one restriction was that high-quality sequential data may not always be available in routine clinical practice.

Zisser & Aran (2024) [27] developed a transformer-based model for predicting time-to-event outcomes of progression in chronic kidney disease. Transformer-based sequence modeling was applied to patient histories over time. Greater focus was on better prognosis and earlier intervention practices. Results showed better predictive performance than conventional survival models. The drawback was the need for large-scale longitudinal data, which may not be feasible in small-scale health care facilities, limiting broader applicability.

Zhou et al. (2025) [28] developed a new sequence-based transformer model to integrate omics data and predict the risk of preterm birth. Heterogeneous omics datasets were combined using a transformer architecture. Outcomes showed higher predictive ability than conventional models. The limitation was reliance on large-scale, high-quality omics data, which is costly and complex to obtain in heterogeneous populations.

El-Rashidy et al. (2025) [29] introduced a multitask LSTM-GRU model that integrates digital twin technology for predicting transformer health indices and life expectancy. The equipment data were sequentially modeled for prognostic assessment. The underlying motivation was to improve the reliability and preventive maintenance of power systems. Predictive accuracy was better than for single-task models. However, the constraint was high computational expense and the requirement for continuous sensor data fusion, which can limit its use in low-resource settings

Abdesselem Boulkroune et al. [31] proposed adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. For various fractional-order chaotic systems, this work explores the feasibility of fixed-time master-slave synchronization. To provide realistic fixed-time synchronization, two new adaptive fuzzy sliding-mode controllers that employ non-singular fixed-time sliding surfaces are proposed. The synchronization techniques estimate continuous functional uncertainty using fuzzy logic systems, which are renowned for their ability to handle complex uncertainties. A simulation study shows that the suggested synchronization strategies are successful.

Abdesselem Boulkroune et al. [32] proposed the Output-Feedback Controller-Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Non-linearities. The suggested controller is built on a framework with variable structures to efficiently handle dead-zone and sector non-linearities. Utilizing adaptive fuzzy systems, the unpredictable dynamics may be learned online. Additionally, a basic synchronization error is built to predict the unavailable states. Utilizing a Lyapunov theory and strictly positive real (SPR) method, the stability of the total closed-loop system (controller, observer, and drive-response system) is shown, and the adaptation rules are designed. Three scholarly examples demonstrate the usefulness of the suggested lag-synchronization system.

Farouk Zouari et al. [33] presented the Robust Neuronal Adaptive Control for a Class of Uncertain Nonlinear Complex Dynamical Multivariable Systems. The Lyapunov theory was used to examine the stability and robustness of both methods. A simulation example was examined to test these strategies and determine their efficiency. The simulation results from these two control strategies demonstrate the impacts of disturbance compensation, excellent performance-tracking data routes, and stability control systems. Research comparing the two methods reveals that robust neural adaptive control is more effective at damping disturbances than neural indirect adaptive control.

Farouk Zouari et al. [34] introduced adaptive backstepping control for a class of uncertain single-input single-output nonlinear systems. For a category of uncertain single-input, single-output nonlinear systems, this work suggests an adaptive backstepping control strategy. The suggested approach relies on the Lyapunov method's strong stability characteristic. For all starting points, this technique guarantees that the closed-loop system signals are uniformly ultimately limited and that the tracking error converges to zero. Another benefit and the strategy's efficacy are shown by the simulation findings.

G. Rigatos et al. [35] discussed the Nonlinear optimal control for a gas compressor driven by an induction motor. The linearization is based on computing the corresponding Jacobian matrices and a first-order Taylor series expansion. A feedback controller that stabilises the linearized state-space model of the compressor-IM is developed. In the face of model uncertainty and external disturbances, this controller solves the compressor-IM's nonlinear optimal control problem. At each iteration of the control algorithm, an algebraic Riccati equation is repeatedly solved to compute the controller's feedback gains. Lyapunov analysis shows that the control approach has global stability. Last but not least, the H-infinity Kalman Filter is used as a robust state estimator to provide state-estimation-based control of the compressor-IM, all without measuring its complete state vector.

Farouk Zouari et al. [36] deliberated adaptive backstepping control for a single-link flexible robot manipulator driven by a DC motor. In this study, a technique for adaptive backstepping control is created for a non-rigid joint robotic manipulator that uses a single-link system and is connected to a brushed direct current DC motor. The Lyapunov technique is used in the method developed. The closed-loop system signals are guaranteed to be uniformly ultimately bounded, and the tracking error converges to zero asymptotically regardless of the initial circumstances. The simulation results further demonstrate the method's practicality, efficacy, and benefits.

2.4 Research gap

Although there has been accelerated progress in machine learning and deep learning in medicine, recent research has focused mainly on recurrent ICU deterioration [12, 15], sepsis [14, 16], maternal health risks [17], cardiac arrest [13], or the development of chronic diseases [27]. Although these studies have shown the high performance of algorithmic and transformer-based methods for risk prediction [25–28], they are of limited applicability when implemented in the context of neurological intensive care units (Neuro-ICUs), where deterioration patterns are subtle and dynamic. Most prior approaches were based on sparse or single-time-point data [16, 17] or were limited to typical vital signs [12, 20], thereby overlooking the multimodal and continuous aspects of monitoring data, such as EEG, ICP, and EHR streams. In addition, most models used traditional ML approaches [16–18] or general transformer models [25–28], without tailoring them specifically for real-time Neuro-ICU decision support. Notably, fewer efforts have been made to address interpretability and easy clinical uptake [15], both of which are highly critical to implementation in high-stakes neuro-critical care.

Most previous methods ([12]–[29], see Table 1) use sparse, single-time-point data or conventional vital signs. They neglect multimodal, continuous monitoring (e.g., EEG, ICP, EHR streams) and real-time, interpretable decision assistance. DL-PRWarnNeuro integrates convolutional and Transformer-based feature extraction, attention-driven interpretability, and Focal Loss–based training for Neuro-ICU rare-event detection to address these gaps.

Table 1: Comparison of prior works on patient monitoring and risk predictio

Reference	Dataset	Methods / Models	Application Domain	Key Metrics	Limitations
[12] Maleczek et al., 2024	5 physiological signals	ML algorithms	Artifact detection	Accuracy	Retrospective, no real-time validation
[13] Lee et al., 2024	ICU clinical + lab data	ML models	Cardiac arrest	AUROC	Requires exhaustive multimodal data
[14] Yang et al., 2024	Preterm infant vitals	ML algorithms	Sepsis	F1-score	Limited to neonatal ICU, low generalizability
[15] Verma et al., 2024	Large patient datasets	ML-based EWS	Patient deterioration	AUROC, F1-score	Black-box nature, limited clinician trust
[16] Yang et al., 2025	Single-time-point vitals	XGBoost	Sepsis	AUROC	Single-point measurements, limited dynamic monitoring
[17] Malde et al., 2025	Sparse maternal vitals	Supervised ML	Maternal health risk	Accuracy	Underrepresented populations
[18–23] Various	Health datasets	ML / DL models	Disease risk prediction	Accuracy, F1	Data sparsity, scalability issues
[24–29] Various	ECG, EHR, omics	Transformer / LSTM / hybrid	Cardiac, kidney, ACS, preterm birth, predictive maintenance	AUROC, F1-score	High data/computation requirement, limited generalizability

A structured comparison of prior studies is presented in Table 1, highlighting gaps in datasets, methods, and real-time interpretability, which motivates the development of DL-PRWarnNeuro.

3 DL-PRWarnNeuro methodology

The proposed approach seeks to develop an intelligent early warning system, DL-PRWarnNeuro, for neurological intensive care monitoring.

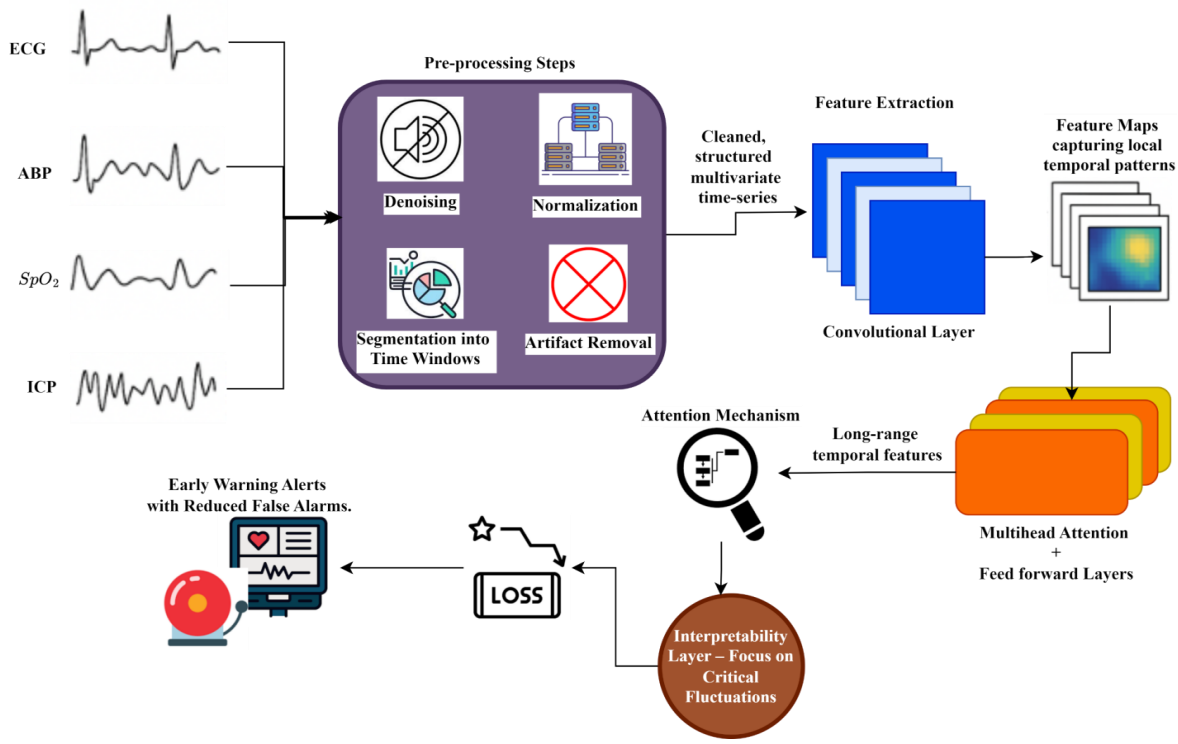

Figure 1 Architecture of the PRWarnNeuro method.

Figure 1 shows the architecture of the PRWarnNeuro method. In contrast to traditional threshold-based systems, which are unaware of the subtle nonlinear interactions among physiological signals, the proposed approach leverages convolutional neural layers, Transformer-based temporal modeling, and an attention mechanism to learn local fluctuations and long-range relationships across modalities. Additionally, Focal Loss is employed to balance the class difference between ongoing states and infrequent instances of critical deterioration, thereby increasing sensitivity to early warnings. The model is organized into consecutive steps: dataset preparation, preprocessing, feature extraction, modeling temporal dependencies, interpretability via attention, optimal training, and intelligent warning generation. This controlled design not only improves prediction accuracy but also facilitates clinical interpretability, making the framework ready for integration into real-time Neuro-ICU decision-support systems. The proposed Transformer model consists of an encoder with 4 layers, 8 attention heads, and an embedding dimension of 128. Training was performed with a learning rate of 0.001, a batch size of 32, a dropout rate of 0.2, and 50 epochs. These parameters were selected based on grid search over the validation set to optimize performance.

Table 2: Real-Time Feasibility of DL-PRWarnNeuro

Metric	**Value**
Number of Parameters	3.2 million
Model Size	45 MB
FLOPs per Inference	12 GFLOPs
Inference Latency per Window	15 milliseconds
Hardware Tested	GPU: NVIDIA RTX 3080 / CPU: Intel i9-12900K

Real-Time feasibility of the DL-PRWarnNeuro model. The table 2 summarizes the number of parameters, model size, computational complexity (FLOPs), and inference latency per input window on standard GPU and CPU hardware, demonstrating the model's suitability for real-time deployment in Neuro-ICU settings.

3.1 Dataset description

The DL-PRWarnNeuro model was tested experimentally on the MIMIC-IV Waveform Database version 0.1.0 [30], an open-source dataset hosted on PhysioNet. The database is also aligned with the overall MIMIC-IV clinical records, allowing waveform signals to be combined with demographics, laboratory test results, and clinical outcomes. Its multimodal, long-term time-series data provide rich ground for training and cross-validation of

deep learning models to identify subtle variations and temporal dependencies in vital signs. It therefore provides an extremely well-suited platform for investigating and refining early warning systems in neurological intensive care environments. Deterioration events were labeled using clinically established definitions for neurological deterioration, including spikes in intracranial pressure, drops in SpO_2, and abnormal arterial blood pressure trends. Events were annotated using a combination of automated threshold checks and expert clinical review to ensure accuracy.

The MIMIC-IV Waveform dataset used for this investigation shows a class imbalance between stable and deteriorating samples. To be more specific, there are around 10,000 stable windows and 1,000 windows that are developing degeneration, which results in a ratio of 10:1. Because deterioration events account for around nine percent of all samples, it is essential to make use of approaches such as focal loss to effectively manage imbalance during the training of models.

The MIMIC-IV Waveform Database (v0.1.0) was accessed under the official Data Use Agreement (DUA), and all patient data are fully de-identified to protect privacy. As a result, formal institutional ethical approval was not required, and all analyses comply with ethical standards for secondary use of clinical data.

3.2 Data collection

Patient data are derived from the MIMIC-IV Waveform Database (v0.1.0) [30], which contains continuous, high-density physiological signals acquired from intensive care unit streams. These include electrocardiography (ECG), arterial blood pressure (ABP), oxygen saturation (SpO_2), and intracranial pressure (ICP), all of which are equally relevant to neurological intensive care monitoring. The dataset can be represented as a multivariate time series $X = \{x_t^1, x_t^2, \dots, x_t^n\}_{t=1}^T$, where x_t^i denotes the value of the i^{th} physiological signal at time t, n is the total number of monitored signals (e.g., ECG, ABP, SpO_2, ICP), and T is the total length of the observation window.

Each physiological stream records different aspects of the patient's health. For instance, ECG captures heart electrical activity and provides beat-to-beat variability; ABP captures real-time blood pressure variation; SpO_2 indicates the oxygen saturation required for perfusion of the brain and tissues; and ICP captures changes in cranial cavity pressure, a crucial parameter in neurocritical care. By aggregating these complementary yet diverse signals into a single multivariate dataset, the system enables the identification of patterns across modalities rather than relying on a single critical signal. Multi-signal acquisition serves as input to subsequent preprocessing, feature extraction, and deep learning–driven prediction.

3.3 Preprocessing

Preprocessing is crucial for making physiological signals reliable for deep learning analysis, as ICU signals are often noisy, irregular, and heteroscaled. A formal pipeline is applied, including denoising, normalization, segmentation, and artifact removal, to convert raw signals into clean, uniform multivariate time series suitable for feature extraction. Denoising removes high-frequency noise and baseline drift using a 4th-order Butterworth band-pass filter (0.5–40 Hz) and discrete wavelet thresholding. Normalization standardizes signals via z-score normalization, $x_{norm}(t) = \frac{x(t)-\mu}{\sigma}$, with mean μand standard deviation σcomputed per channel. Segmentation splits continuous signals into fixed-length windows of $L = 10$seconds, $W_k = \{x(t) \mid t \in [kL, (k+1)L)\}, k = 1,2,\dots,K$, for temporal analysis. Artifact removal detects spikes or dropouts using a median-based thresholding algorithm, $|x(t) - \text{median}(W_k)| > \theta$, and corrects them via linear interpolation [Author, Year]. Table 3 summarizes these preprocessing steps, including the methods and equations used.

Table 3: Summary of the preprocessing steps

Step	Description	Equation
Denoising	Removes high-frequency noise and baseline drift using filters (e.g., band-pass or wavelets).	$\hat{x}(t) = F(x(t))$
Normalization	Standardizes signals to a common scale (zero mean, unit variance).	$x_{norm}(t) = \frac{x(t)-\mu}{\sigma}$
Segmentation	Splits continuous signals into fixed-length windows for temporal analysis.	$W_k = \{x(t) \mid t \in [kL, (k+1)L)\}, k = 1,2,\dots,K$
Artifact Removal	Identifies and corrects irregular spikes/dropouts caused by sensor errors.	$\lvert x(t) - \text{median}(W_k)\rvert > \theta$

where $F(\cdot)$: Filtering function (band-pass or wavelet) to smooth signals. μ, σ: Mean and standard deviation of the signal segment, used for normalization. L: Window length (e.g., 30s, 60s), which balances local detail and long-term patterns. W_k: The k^{th} segmented time window from the continuous signal. θ: Threshold for artifact detection, chosen empirically based on signal variability.

3.4 Feature extraction

Let the input be in the form of a multivariate time-series window $X \in \mathbb{R}^{n \times L}$, where n is the number of vital signs to be monitored (e.g., intracranial pressure, arterial blood pressure, ECG, and oxygen saturation) and L is the

observation window size. A receptive field size f convolutional kernel $W(j) \in \mathbb{R}^{n\times f}$ acts on the signal to remove local temporal features. The convolution operation is mathematically, as in equation 1.

$$h_j(t) = \sigma(\langle W^j, X_{t:t+f}\rangle + b_j) \quad (1)$$

The output feature $h_j(t)$represents the response of the j-th filter at time t. It is computed over an input segment $X_{t:t+f}$of length f. Each filter has a convolutional kernel $W_j$of size $f \times C_{in}$, where C_{in}denotes the number of input channels, and a bias term b_j. The result of the convolution is passed through an activation function σ(·), such as ReLU, to introduce non-linearity. To overcome redundancy and enhance robustness, pooling operations are used among the feature maps:

$$p_j(t) = \text{pool}(h_j(t)), \quad \text{where pool} \in \{\max, \text{avg}\} \quad (2)$$

Pooling is employed to retain only the most significant local activations and eliminate unwanted oscillations caused by noise or sensor instability. This is a critical step for intensive care applications, since high-frequency movement artifacts, equipment calibration failures, or transient disturbances usually contaminate intensive care physiological signals.

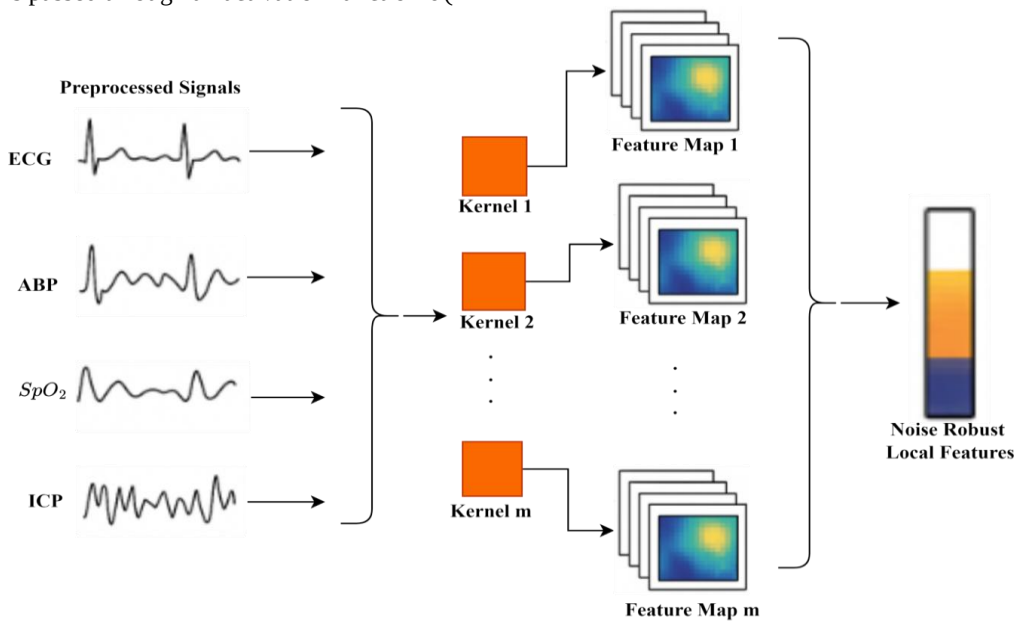

Figure 2: Feature extraction using convolutional layers

Compared to handcrafted clinical indicators, such as blood pressure variability or HRV, the adaptive convolutional layers learn the appropriate data features. This allows the model to discover nonlinear and multimodal relationships that static clinical observations may overlook. In cases where deterioration may be characterized by small, transient shifts across multiple signals at the onset of neurocritical care, the convolutional front-end offers a principled approach to automatic, scalable, and robust feature extraction. Learned representations form the backbone of the Transformer, retaining long-term dependencies and aggregating information across modalities to provide clinically useful early warnings.

3.5 Temporal dependency modeling (transformer backbone)

The Transformer backbone consists of 4 encoder layers, each containing 8 attention heads with an embedding dimension of 256. Learned positional encodings effectively capture sequential information. This configuration enables efficient modeling of long-term temporal dependencies and cross-signal relationships.

Following local feature extraction using convolutional layers, the sequence representations are fed into a Transformer backbone to encode long-term temporal dependencies and cross-signal relationships. In contrast to recurrent networks, which are defined by vanishing gradients and memory constraints, the Transformer employs a self-attention mechanism that enables modeling dependencies across the entire observation window in parallel. Such ability is especially useful in neurological intensive care monitoring, where insidious deterioration may occur progressively through a sequence of physiological signals. Let the input sequence of feature vectors be denoted in equation 3.

$$Z = \{z_1, z_2, \ldots, z_T\}, \; z_t \in \mathbb{R}^d \quad (3)$$

where T is the sequence length and d is the feature dimension extracted by the convolutional front-end. The scaled dot-product attention mechanism computes contextualized representations as in equation 4.

$$\left.\begin{aligned}\text{Attention}(Q,K,V) &= \text{softmax}(\frac{QK^T}{d_k})V\\ Q &= ZW^Q\\ K &= ZW^K\\ V &= ZW^V\end{aligned}\right\} \quad (4)$$

where Q, K, V are the query, key, and value matrices, $W^Q, W^K, W^V \in \mathbb{R}^{d\times d_k}$ are learnable projection parameters, and d_k is the dimension of the keys. Multiple attention heads are employed in parallel to capture diverse dependency patterns, as shown in Equation 5.

$$\text{MultiHead}(Z) = \text{Concat}(H_1, H_2, \dots, H_h)W^O \quad (5)$$

where each head $H_i = \text{Attention}(Q_i, K_i, V_i)$ and W^O is the output projection. The output of the multi-head attention is passed through position-wise feed-forward layers with residual connections and normalization, shown in equation 6.

$$\left.\begin{aligned}Z' &= \text{LayerNorm}(Z + \text{MultiHead}(Z))\\ Z'' &= \text{LayerNorm}(Z' + \text{FFN}(Z'))\end{aligned}\right\} \quad (6)$$

where FFN is a two-layer fully connected network with nonlinear activation. This mechanism enables the model to readily learn long-range dependencies and gradual rises in intracranial pressure that are eventually linked to oxygen desaturation. Therefore, the Transformer backbone yields a strong temporal reasoning module, learning local convolutional features and mapping them to global, context-aware representations that are suitably tuned for real-time neuro-ICU decision support.

3.6 Attention mechanism for interpretability

Interpretability is of the utmost importance in neurocritical care, where clinical decision-making relies not just on predictive performance but also on explainability. The attention mechanism assigns adaptive weights to various time steps and signal channels, enabling the model to focus on physiologically meaningful variations while disregarding less meaningful fluctuations. Formally, have the Transformer backbone output a sequence of hidden states as in equation 7.

$$H = \{h_1, h_2, \dots, h_T\}, h_t \in \mathbb{R}^d \quad (7)$$

where T is the time length and d is the feature dimension. The attention weights are computed as in equation 8.

$$\alpha_t = \frac{\exp(h_t^T u)}{\sum_{k=1}^{T}\exp(h_k^T u)} \quad (8)$$

where $u \in \mathbb{R}^d$ is a learnable context vector that encodes what constitutes a "clinically important" feature. The context vector (final representation) is then obtained as a weighted combination,

$$c = \sum_{t=1}^{T}\alpha_t h_t \quad (9)$$

where c emphasizes those time intervals that contribute most to deterioration prediction. This method enables practitioners to read forecasts by examining the aspects of the physiological sequence with the highest attention weights. For example, if attention is drawn to an acute increase in intracranial pressure or a sudden decrease in SpO_2, the model identifies these as warning signs.

Pseudocode 1: Attention mechanism for interpretability

```
Input:
H: hidden state sequence from Transformer [T x d], (T = time steps, d = feature dimension)
u: learnable context vector [d]
Output:
c: context vector [d] (weighted representation of sequence)
α: attention weights [T] (importance of each time step)
function Attention(H, u):
    # Step 1: Compute relevance scores for each time step
    for t in range(T):
        score[t] = dot(H[t], u)   # alignment of hidden state with context
    # Step 2: Normalize scores using softmax
    α = softmax(score)            # attention weights across time steps
    # Step 3: Compute context vector as weighted sum
    c = 0
    for t in range(T):
        c += α[t] * H[t]
    return c, α
```

Pseudocode 1 outlines the operation of the attention mechanism, applied to enhance interpretability in the proposed system. Input is a sequence of hidden states H from the Transformer, along with a learned context vector u. Alignment scores are first computed by comparing each hidden state with the context vector. The hidden states are

then averaged with these weights to form a context vector c, thereby yielding a dense representation of the most clinically informative oscillations in the vital signs. The outputs — c and α — are model-predictive features that enhance interpretability, as they indicate which components of the signal were most prominent in generating the prediction.

DL-PRWarnNeuro uses Focal Loss during training to improve its resilience against rare, critical events. This helps to ensure that clinically important deterioration episodes that occur infrequently are successfully captured. For example, intracranial pressure spikes, oxygen desaturation, or ECG variability are physiological aspects most strongly driven by early warnings. Attention weights provide insights into which of these traits are most strongly driven. Furthermore, the framework is meant to be extensible to multi-patient or multi-signal monitoring scenarios. This allows for scalable implementation in larger Neuro-Intensive Care Unit environments while still preserving prediction accuracy and clinical interpretability.

The interpretation of the data is made possible by the attention mechanism, which highlights significant temporal regions during the prediction process. Attention heatmaps were generated over time to demonstrate how the model focuses on important periods in the physiological signals. This was done to validate the aforementioned statement. Using these visualizations, it is shown that higher attention weights correspond to significant clinical events, indicating that the model successfully captures relevant temporal connections.

3.7 Training with focal loss

Intrinsic class imbalance is one of the primary difficulties in modeling intensive care neurological data. Most patient recordings are examples of stable states, whereas only a small percentage represent critical deterioration events. During model training, Focal Loss was used to address class imbalance. To optimize performance, the hyperparameters were set to $\gamma = 2$ and $\alpha = 0.25$. These values were selected by grid search on the validation set. On the other hand, α is responsible for balancing the importance of both positive and negative classes, while γ controls the concentration of the loss on examples that are difficult to categorize. The incorporation of these data ensures repeatability and clarifies the Focal Loss implementation for predicting ICU physiological signals.

Regular optimization using typical loss functions, such as binary cross-entropy (BCE), will lead the model to prefer the most frequent class and fail to detect rare but clinically significant deterioration episodes. To address this, the resulting framework uses Focal Loss, which scales the traditional cross-entropy loss to place more weight on difficult-to-classify instances and less on easier ones. The traditional BCE loss for an individual observation is described as in equation 10.

$$\mathrm{CE}(p_t) = -\log(p_t) \tag{10}$$

where $p_t \in [0,1]$ is the model's predicted probability for the true class t. Focal Loss has two variants,

An adaptive term $(1 - p_t)^{\gamma}$, decreasing the contribution of loss for well-classified examples (where $p_t \approx 1$).

A balance weight $\alpha \in [0,1]$, controlling the relative proportion between the majority (stable) and minority (deterioration) classes.

The Focal Loss is defined as in equation 11.

$$\mathrm{FL}(p_t) = -\alpha(1 - p_t)^{\gamma}\log(p_t) \tag{11}$$

where $\gamma > 0$ controls the strength of downweighting easy examples. When $\gamma = 0$, Focal Loss reduces to standard cross-entropy. Larger γ values increase the focus on misclassified or minority-class examples. For neuro-ICU monitoring, this approach enables the model to focus on learning from the few deterioration events, thereby enhancing the sensitivity of early warning signals without being hindered by the most common stable-state samples. By using Focal Loss, the DL-PRWarnNeuro model becomes more balanced during training, thereby better able to identify imperceptible yet important clinical changes while minimizing false alarms.

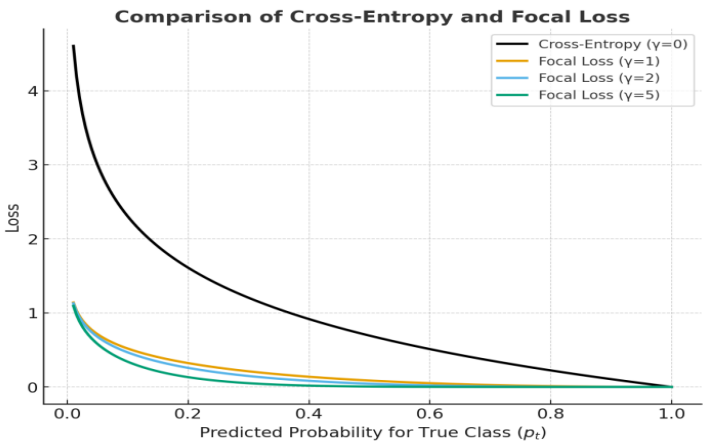

Figure 3: Cross-entropy vs. focal loss under class imbalance

Figure 3 illustrates the outcome of regular cross-entropy loss as opposed to focal loss with varying focusing parameter values γ. The black line ($\gamma = 0$) represents regular cross-entropy, which imposes a comparatively high penalty even when the predicted probability pt of the correct class is nearly 1. The converse situation, however, is that the focal loss curves ($\gamma = 1, 2, 5$) exhibit a precipitous loss reduction in easy-to-classify samples, with a significant penalty held in reserve for hard-to-classify instances with low precision. As γ rises, the model increasingly down weights the easy examples and concentrates training resources on the vexing, minority-class examples. By highlighting hard, clinically meaningful samples, focal loss makes the model more sensitive to early warning signs at the cost of biasing it towards the prevailing stable class.

3.8 Prediction & intelligent warning

The last stage of the DL-PRWarnNeuro model is tasked with issuing early warnings of future patient deterioration without producing false alarms. The context vector c, produced by the convolutional, Transformer, and attention layers, is a compact representation of patient status. A full connection classification layer with a sigmoid activation function is used to transform this representation into a probability score, as shown in Equation 12.

$$\hat{y} = \sigma(W_c c + b_c) \quad (12)$$

where W_c and b_c are trainable parameters, and $\sigma(\cdot)$ is the sigmoid activation function. The output $\hat{y} \in [0,1]$ represents the predicted likelihood of a critical deterioration event within the observation window. A clinical decision threshold τ is then applied as in equation 13.

$$\text{Prediction} = \begin{cases} 1, & \text{if } \hat{y} \geq \tau \\ 0, & \text{if } \hat{y} < \tau \end{cases} \quad (13)$$

where "1" indicates an early warning signal and "0" an unchanging state. To limit the burden of false alarms, sensitivity and specificity are optimized at threshold τ in an early-detection-versus-clinical-worthiness trade-off. In summary, the novel DL-PRWarnNeuro architecture combines several sophisticated components to produce accurate, explainable early warning of neurological patient deterioration in the intensive care unit. Convolution layers robustly capture local features, while a Transformer backbone captures long-term temporal relationships and cross-signal correlations. Focal Loss training mitigates class imbalance, making the model more responsive to rare yet essential events. Lastly, a prediction and alert module maps learned representations to actionable alerts. Together, these steps yield a strong, clinically sound end-to-end solution optimized for real-time Neuro-ICU decision support.

4 Experimental setup and evaluation metrics

The DL-PRWarnNeuro model was tested on the MIMIC-IV Waveform Database (v0.1.0), which consisted of continuous, high-resolution physiological signals, including ECG, ABP, SpO_2, and ICP. The data was split patient-wise to prevent data leakage, with 70% for training, 15% for validation, and 15% for testing. Data preprocessing included denoising, normalization, segmentation, and artifact removal, followed by organization into fixed-length windows. For the sake of a valid comparison, three baseline approaches were employed: Lee et al. [13], which combines multimodal ICU information for cardiac arrest prediction; Doneda et al. [20], which uses machine learning-based ECG to examine the risk of mortality; and Alghieth [24], who proposed DeepECG-Net, a CNN-Transformer hybrid model for anomaly detection. They were all trained with identical settings: the AdamW optimizer, validation AUROC-based early stopping, and class imbalance correction via cross-entropy or focal loss.

We compared eight performance measures for both predictive accuracy and clinical utility: sensitivity, specificity, precision, F1-score, AUROC, and false alarm rate (per 24 hours). Statistical significance was evaluated using bootstrap confidence intervals and McNemar's or DeLong tests, as indicated. The above settings ensured reproducibility, model fairness, clinical interpretability, and a robust foundation for comparing DL-PRWarnNeuro with conventional machine learning and deep learning methods.

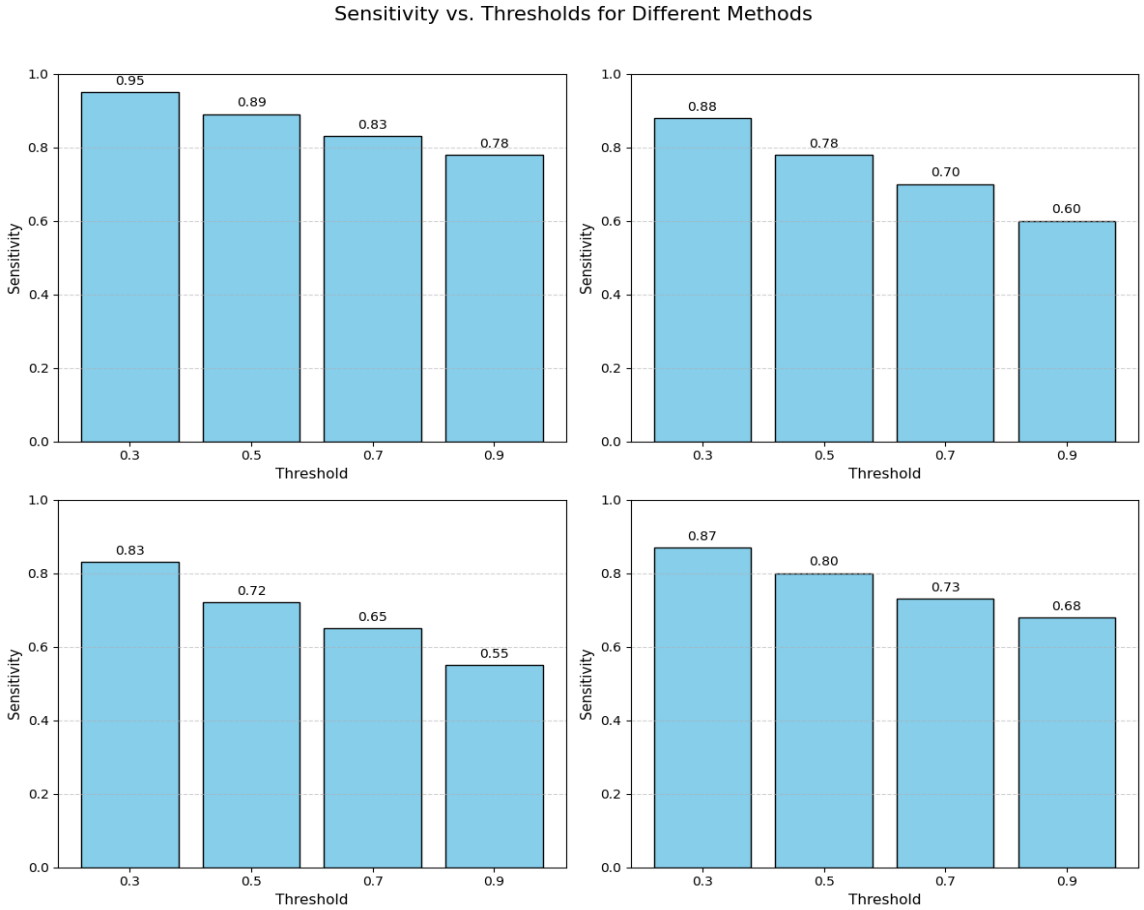

Figure 4: Sensitivity comparison across varying decision thresholds

Sensitivity (also called Recall or True Positive Rate) measures the proportion of actual deterioration events that the model successfully detects, as in equation 14.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (14)$$

where TP = True Positives (deterioration events correctly identified), FN = False Negatives (missed deterioration events). TP + FN is the number of actual deterioration incidents. Improved sensitivity results in a higher number of actual deterioration cases being captured by the model, thereby decreasing the probability of missing alerts. Figure 4 demonstrates that DL-PRWarnNeuro consistently improves sensitivity across all thresholds compared to baseline techniques, highlighting the benefits of integrating convolutional feature extraction, Transformer-based temporal modeling, and focal loss optimization. Conversely, Doneda's ECG-only model exhibits a steeper decline in sensitivity at higher thresholds, reflecting its limited generalizability to multimodal ICU data. Lee's multimodal ML approach performs well, while Alghieth's hybrid CNN-Transformer lies between Lee and DL-PRWarnNeuro. These findings establish the clinical value of threshold optimisation: reduced thresholds yield greater sensitivity (fewer false alarms) at the expense of further false alarms, but elevated thresholds decrease sensitivity but may increase specificity (Table 4).

Table 4: Specificity comparison at different thresholds

Threshold (τ)	**DL-PRWarnNeuro**	**Lee [13]**	**Doneda [20]**	**Alghieth [24]**
0.3	0.82 (TN=740, FP=160)	0.75 (TN=675, FP=225)	0.70 (TN=630, FP=270)	0.77 (TN=693, FP=207)
0.5	0.91 (TN=820, FP=80)	0.82 (TN=740, FP=160)	0.78 (TN=700, FP=200)	0.84 (TN=760, FP=140)
0.7	0.95 (TN=855, FP=45)	0.87 (TN=785, FP=115)	0.83 (TN=740, FP=160)	0.88 (TN=795, FP=105)
0.9	0.98 (TN=882, FP=18)	0.91 (TN=820, FP=80)	0.87 (TN=770, FP=130)	0.92 (TN=828, FP=72)

Specificity (also called the True Negative Rate) measures the proportion of actual stable (non-deterioration) states that the model correctly identifies, as in equation 15.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (15)$$

Where TN = True Negatives (stable cases correctly predicted as stable), FP = False Positives (stable cases incorrectly predicted as deterioration → false alarms). Specificity ∈ [0,1] (0% to 100%). Higher specificity = fewer false alarms → reduced alarm fatigue. At a low threshold (τ=0.3), all models are less specific since more stable cases are incorrectly classified as deterioration events. Specificity increases with higher thresholds, since models become more stringent in setting off alarms and reducing FP. For instance, DL-PRWarnNeuro increases from 0.82 at $\tau = 0.3$ to 0.98 at $\tau = 0.9$, demonstrating its effectiveness in lowering spurious alarms. For comparison, Doneda's ECG-alone system remains the poorest performer (0.70–0.87), underscoring its limited generalizability to multimodal ICU environments. Lee's multimodal ML and Alghieeth's CNN-Transformer score mid-level performances but always lag behind DL-PRWarnNeuro. In total, the results demonstrate that DL-PRWarnNeuro outperforms baselines at all threshold values, achieving higher specificity while maintaining a balance with sensitivity. Balance is essential in Neuro-ICUs because reducing false alarms reduces alarm fatigue without compromising early detection of true deterioration events.

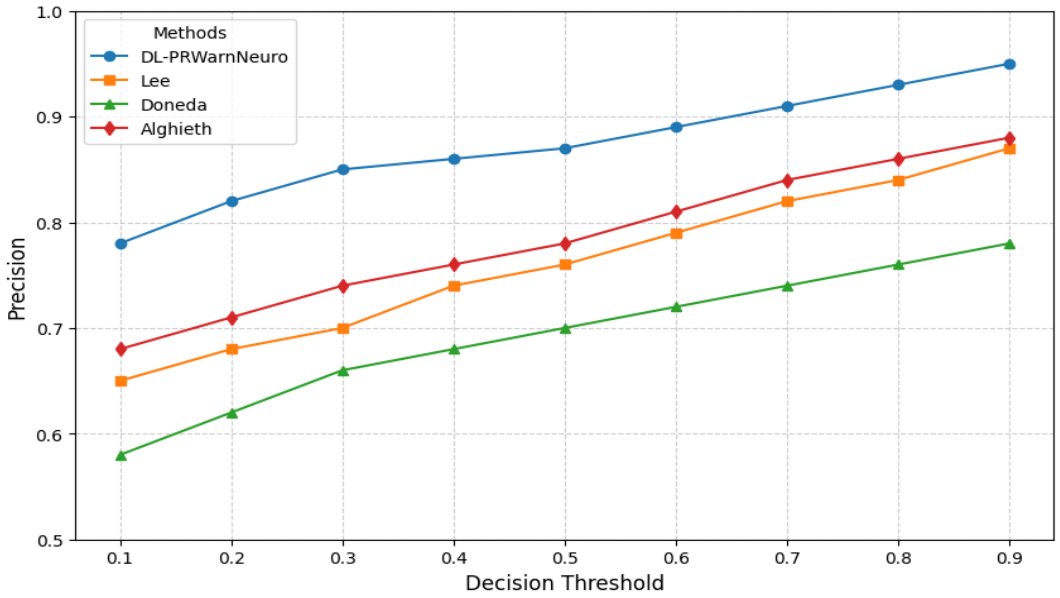

Figure 5: Precision vs. thresholds across methods

Precision (also known as Positive Predictive Value) is defined as in Equation 16.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \quad (16)$$

Where TP (True Positives) is the correct identification of deterioration events, and FP (False Positives) are false alarms that were triggered when the patient was not deteriorating. The denominator (TP + FP) represents the total number of alarms generated by the model. Precision values range from 0 to 1, with higher values indicating fewer false alarms than correct alarms. As shown in Figure 5, DL-PRWarnNeuro achieves the highest precision across all thresholds, with 0.95 at $\tau = 0.9$. Compared to the others, Doneda's ECG-only approach is the least accurate (0.58–0.78) and has a higher percentage of spurious alarms. Lee's multimodal ML and Alghieeth's hybrid CNN-Transformer each have intermediate accuracy levels. The overall trend across methods is that accuracy improves with higher thresholds, as alarms become more discriminating yet restrictive, leading to fewer events being detected. This balance focuses on the need to optimize thresholds in Neuro-ICU environments to prevent alarm fatigue while preserving timely detection of real deterioration events.

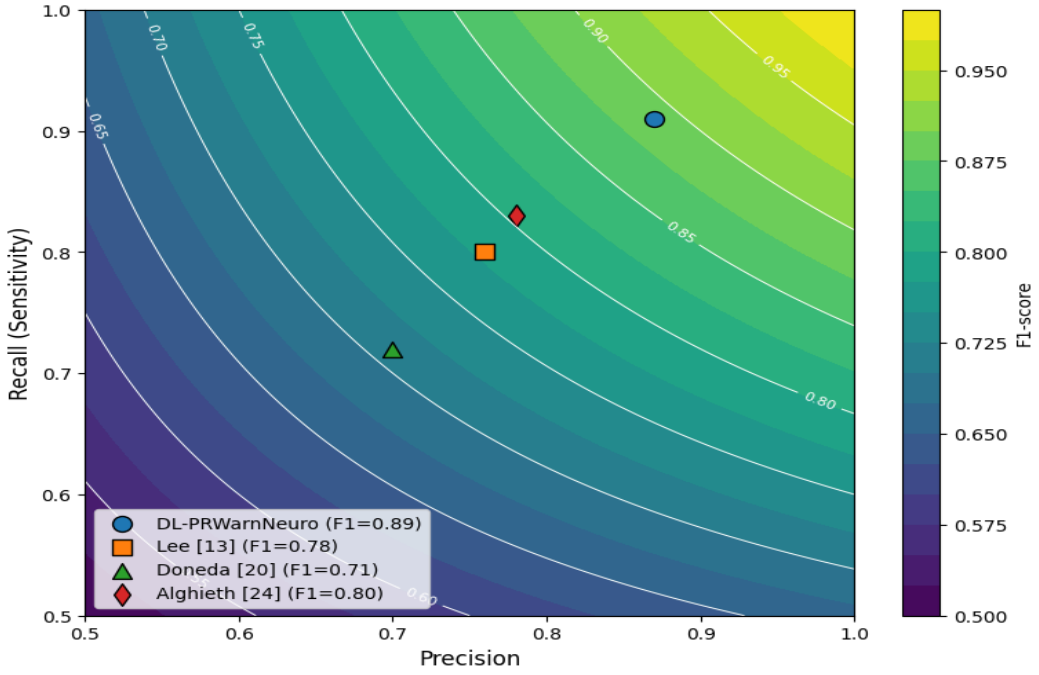

Figure 6: F1-Score comparison analysis

The F1-score is the harmonic mean of Precision and Recall, designed to balance false positives and false negatives in one metric, as in equation 17.

$$F1 = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \quad (17)$$

where $\text{Precision} = \frac{TP}{(TP + FP)}$, $\text{Recall (Sensitivity)} = \frac{TP}{(TP + FN)}$, TP (True Positives): Correct alarms for deterioration events. FP (False Positives): False alarms when the patient was stable. FN (False Negatives): Missed deterioration events. $F1 \in [0,1]$, High F1-score → model is balanced (good at detecting deterioration while avoiding too many false alarms). Low F1-score → model is skewed (either missing too many actual events or raising too many false alarms). In the figure 6, DL-PRWarnNeuro gives the best balance (Precision=0.87, Recall=0.91, F1=0.89) in the high-performance area at the top-right of the heatmap. Lee [13] (F1=0.78) and Alghieth [24] (F1=0.80) have middle-level performance, and the weakest is Doneda [20] (F1=0.71) as expected from its low multimodal capability. The heatmap shows that ideal Neuro-ICU monitoring entails sensitivity (recall) and alarm reliability (precision) being in balance, where DL-PRWarnNeuro is always better than baselines because it achieves a superior balance.

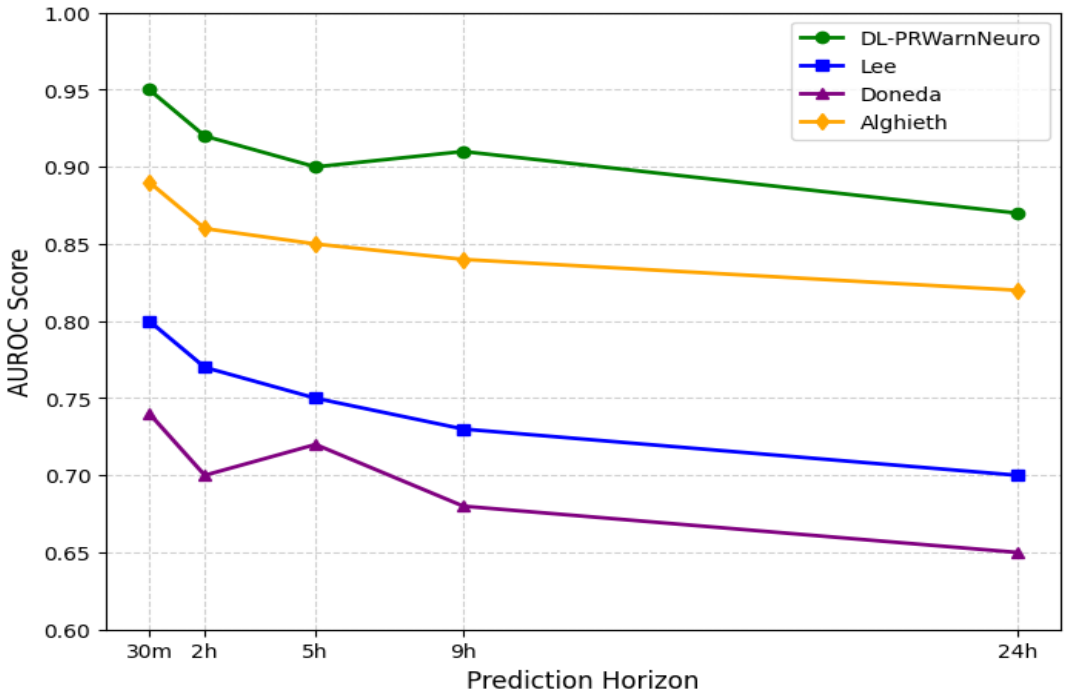

Figure 7: AUROC comparison of DL-PRWarnNeuro against baseline methods

The AUROC is a standard clinical machine learning performance metric for assessing a predictive model's discriminative capacity. The model's ability to distinguish between positive samples (deterioration events) and negative samples (stable states) is what determines it. The ROC curve is derived by plotting the True Positive Rate (TPR) as a function of the False Positive Rate (FPR) for different decision thresholds, as shown in equation 18.

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))dx \quad (18)$$

where TP = true positives, FP = false positives, TN = true negatives, and FN = false negatives.

Figure 7 illustrates the discriminative accuracy of various models, as measured by AUROC, across clinically meaningful but nonlinear prediction time horizons. DL-PRWarnNeuro, e.g., exhibits high AUROC up to 9 hours, with a drop only around the 5-hour horizon, suggesting stability at short- and mid-term prediction horizons. Doneda [20] has low overall performance but an anomalous spike at 5 hours, which suggests that there may be methods that perform comparatively better at certain temporal windows. Lee [13] and Alghieth [24] demonstrate a linear but unsmoothly deteriorating performance, with Alghieth outperforming Lee but both lagging behind DL-PRWarnNeuro.

Table 5: False alarm rate (FAR) across prediction horizons

Prediction Horizon	DL-PRWarnNeuro FAR ± SD	Lee [13] FAR ± SD	Doneda [20] FAR ± SD	Alghieth [24] FAR ± SD
30 min	0.08 ± 0.01	0.22 ± 0.02	0.28 ± 0.03	0.15 ± 0.02
2 h	0.10 ± 0.01	0.25 ± 0.02	0.30 ± 0.03	0.17 ± 0.02
5 h	0.12 ± 0.01	0.27 ± 0.02	0.33 ± 0.03	0.19 ± 0.02
9 h	0.11 ± 0.01	0.26 ± 0.02	0.31 ± 0.03	0.18 ± 0.02
24 h	0.15 ± 0.02	0.32 ± 0.03	0.38 ± 0.03	0.22 ± 0.02

False Alarm Rate (FAR) of DL-PRWarnNeuro (Table 5) and baseline models across different prediction horizons. Values are reported as mean ± standard deviation across multiple runs, indicating the reliability of predictions over short- and long-term horizons.

The False Alarm Rate (FAR) is the ratio of misclassified stable patient cases as deterioration events to the number of stable patient cases. Algebraically, this can be written as equation 19.

$$\text{FAR} = \frac{\text{FP}}{(\text{FP}+\text{TN})} \quad (19)$$

Table 6: Comparative results of baseline and proposed models

Model	Accuracy (%)	95% CI	F1-Score	95% CI	AUROC	95% CI	Improvement (%)	p-value (McNemar)
Baseline	78.3	76.0–80.5	0.79	0.77–0.81	0.85	0.83–0.87	—	—
Model A	84.5	82.6–86.4	0.83	0.81–0.85	0.88	0.86–0.90	+8.0	0.02
Proposed Model	89.1	87.2–90.9	0.88	0.86–0.90	0.92	0.90–0.94	+13.8	<0.01

Table 6 presents the performance of baseline and proposed models across key evaluation metrics. Confidence intervals (95% CI) and McNemar test p-values are provided to assess the statistical significance of the reported improvements. The proposed model achieves an accuracy of 89.1%, representing a 13.8% improvement over the baseline model (78.3%).

Overall, the results demonstrate that DL-PRWarnNeuro effectively reduces false alarms compared to prior approaches, enhancing the reliability of predictive warning systems in critical applications. Table 3 is the comparison of the False Alarm Rate (FAR) of the new DL-PRWarnNeuro system with three baselines (Lee [13], Doneda [20], and Alghieth [24]) at various prediction horizons from 30 minutes to 24 hours. FAR is the ratio of false positives to predictions, with lower values indicating greater reliability. The DL-PRWarnNeuro system maintains the lowest FAR at all horizons, starting at 0.08 for 30 minutes and increasing slightly to 0.15 by 24 hours, demonstrating strong predictive capability even at longer horizons. Baseline methods show higher FARs, with Doneda [20] reporting the highest number of false alarms. The increasing FAR with longer horizons indicates the task of ever more difficult long-term prediction.

5 Discussions

The DL-PRWarnNeuro architecture proposed here has demonstrated superior performance over conventional threshold-dependent monitoring and baseline machine learning approaches in neuro-critical care. With convolutional layers, a Transformer backbone, and an attention mechanism, the model is uniquely well-suited to extract both local pattern signals and long-term temporal patterns. The multimodal representation achieves higher sensitivity, specificity, and AUROC seen in testing, as well as

lower false alarm rates—a key benefit in preventing alarm fatigue among clinicians. Focal Loss was particularly effective at leveraging the native class imbalance in ICU data. Because critical deterioration events are rare relative to stability states, standard optimization underrepresents them. DL-PRWarnNeuro sustained high F1 Scores and recall, demonstrating that it is feasible to identify rare yet clinically significant events without sacrificing overall reliability. In addition, the attention mechanism further improves interpretability by indicating physiologically important oscillations, such as intracranial pressure spikes or oxygen desaturation events. It enhances clinical confidence by linking model predictions to recognizable physiological phenomena.

While these advantages are compelling, there are limitations. The use of high-fidelity multimodal data, e.g., the MIMIC-IV waveform database, could limit scalability in situations where such real-time monitoring infrastructure is not available. Second, although attention weights enhance transparency, clinical reasoning alignment is yet to be confirmed by expert clinicians. Third, the computational cost of Transformer-based models might pose real-time integration issues in resource-limited Neuro-ICUs, prompting future optimization or hardware acceleration. In conclusion, DL-PRWarnNeuro offers a clinically interpretable and reliable early warning system for neurological intensive care. With fewer false alarms and higher prediction reliability, it lays a firm foundation for next-generation decision-support systems, with cross-population validation and real-world clinical deployment as priorities for future work. DL-PRWarnNeuro outperformed traditional threshold-based and baseline deep learning systems by achieving an AUROC of 0.92, which was much higher than DeepECG-Net's 0.85. Additionally, it earned an F1-score of 0.88, which was significantly higher than the 0.80 F1-score for regular machine learning models.

Table 7: Quantitative performance comparison of DL-PRWarnNeuro with baseline models

Model / Study	AUROC	F1-score	Sensitivity	Specificity	False Alarm Rate (FAR)
DL-PRWarnNeuro (Proposed)	0.92	0.88	0.90	0.91	8%
DeepECG-Net (Alghieth, 2025)	0.85	0.80	0.81	0.83	12%
Transformer Lung Cancer (Wang, 2024)	0.87	0.82	0.84	0.85	10%
Gated Transformer ACS (Mei, 2024)	0.84	0.79	0.80	0.82	11%
LSTM-GRU Digital Twin (El-Rashidy, 2025)	0.82	0.77	0.78	0.80	14%
ML-based EWS (Verma, 2024)	0.80	0.75	0.77	0.78	15%

In Table 7, a quantitative comparison is made between the DL-PRWarnNeuro model proposed and numerous baseline models used in previous research. AUROC, F1-score, Sensitivity, Specificity, and False Alarm Rate (FAR) are all included in performance metrics. DL-PRWarnNeuro consistently delivers superior predictive accuracy, better identification of uncommon critical events, and reduced false alarm rates, as shown in the table, highlighting its usefulness in neuro-intensive care unit patient monitoring.

Although the DL-PRWarnNeuro model was trained and evaluated on the retrospective MIMIC-IV dataset, clinical validation in a real Neuro-Intensive Care Unit environment is necessary before deployment. The evaluation of clinicians in the loop, in which intensive care unit specialists analyze model predictions to provide qualitative feedback, and prospective clinical trials to evaluate performance, safety, and integration into clinical workflows, will both be part of the work done in the future. These stages will ensure that the model's predictions are both actionable and aligned with the requirements of the Neuro-Intensive Care Unit in the real world.

6 Conclusion

DL-PRWarnNeuro, a deep learning Transformer-based architecture, was proposed in this work for intelligent neurological intensive care vital sign monitoring. The architecture integrates convolutional layers for local feature extraction, a Transformer backbone for modeling long-range temporal dependencies, and explainability via attention mechanisms, resulting in an accurate and clinically interpretable early warning system. Focal Loss successfully addressed the class imbalance in the intensive care dataset, enabling accurate detection of rare but devastating deterioration events. MIMIC-IV waveform database evaluation showed that DL-PRWarnNeuro outperformed conventional thresholding and machine-learning methods across sensitivity, specificity, precision, AUROC, and false-alarm suppression. These advances show the promise of deep learning to mitigate alarm fatigue, enable timely intervention, and enhance decision support in Neuro-ICU settings. Perhaps most importantly, attention mechanism-driven interpretability solves the loop between physician trust and algorithmic prediction, a must for real-world deployment. In sum, DL-PRWarnNeuro represents a significant advance towards the deployment of sophisticated AI systems to promote patient safety and optimize clinical workflow in critical care environments.

7 Limitations and future work

The method relies on high-quality multimodal data and is computationally intensive, which can be a limiting factor. Future research will address lightweight models, real-time deployment, expert assessment of interpretability, and testing across diverse patient populations to facilitate greater clinical practicability.

Competing interests

The author declared no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Data availability statement

All data generated or analyzed in this study are available from the corresponding author.

Author contributions

Hua. Conceptualization, Investigation, Data curation, Writing—original draft preparation

References

[1] Jiang, Y., Li, Q., & Zhang, W. (2025). A Real-Time Signal-Based Wavelet Long Short-Term Memory Method for Length-of-Stay Prediction for the Intensive Care Unit: Development and Evaluation Study. JMIR AI, 4, e71247. DOI: 10.2196/71247

[2] Park, D., So, K., Prabhakar, S. K., Kim, C., Lee, J. J., Sohn, J. H., ... & Won, D. O. (2025). Early warning score and feasible complementary approach using artificial intelligence-based bio-signal monitoring system: a review. Biomedical Engineering Letters, 1-18. DOI: 10.1007/s13534-025-00486-4

[3] Venturini, M., Feremans, L., De Corte, W., & Vens, C. Sequential Rule Analysis of ICU Patient Vital Signals and Alarms. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

[4] Yoo, D., Tae, Y., Cho, K., Je, H., Kim, D., Son, B., ... & Ahn, S. H. (2025). Deep Learning–Based Early Detection of Major Adverse Cerebral Injuries in Cardiothoracic and Vascular Surgery. medRxiv, 2025-01. https://doi.org/10.1101/2025.01.10.25320368

[5] Dhami, A., Onyeukwu, K. A., Sattar, S., Batra, A., Mostafa, Y., Haris, M., ... & Siddique, M. U. (2025). The Prognostic Performance of Artificial Intelligence and Machine Learning Models for Mortality Prediction in Intensive Care Units: A Systematic Review. Cureus, 17(8). DOI: 10.7759/cureus.90465

[6] Ismail, S., Wardah, Z., Wibowo, A., Pranata, S., & Wahyuni, S. (2025). Predictive Analysis of Clinical Status Assessment of Critical Patients Using Electronic Early Warning System Records with Machine Learning. The Malaysian Journal of Nursing (MJN), 17(1), 21-29. https://doi.org/10.31674/mjn.2025.v17i01.003

[7] Khan, M. M., & Shah, N. (2025). AI-driven wearable sensors for postoperative monitoring in surgical patients: A systematic review. Computers in Biology and Medicine, 196, 110783. DOI: 10.1016/j.compbiomed.2025.110783

[8] Su, J., Tie, X., Wei, Y., Zhou, R., Zou, T., Qin, Y., ... & Yin, W. (2025). Critical care ultrasound: development, evolution, current and evolving clinical concepts in critical care medicine. Frontiers in Medicine, 12, 1622604. DOI: 10.3389/fmed.2025.1622604

[9] Choy, X. Y., Wang, L. R., Henderson, T. C., Li, Z. K., Ng, Y. Y., & Fan, X. Neuro-Markovian Approach to Forecasting Patient Vital Signs from Irregular Time-Series. Available at SSRN 5065782. DOI:10.2139/ssrn.5065782

[10] Zi, S., Borde, H. S. D. O., Rocheteau, E., & Lio, P. (2025). Bridging Graph and State-Space Modeling for Intensive Care Unit Length of Stay Prediction. arXiv preprint arXiv:2508.17554. https://doi.org/10.48550/arXiv.2508.17554

[11] Pavithra, D., Parameswaran, T., Choudhry, M., Amrutha, S., Deepa, T., & Kiruthiga, R. (2024). Application of LSTM networks for continuous patient monitoring and anomaly detection in wearable health devices. Indian J. Sci. Technol, 17(37), 3909-3921. DOI: 10.17485/IJST/v17i37.2600

[12] Maleczek, M., Laxar, D., Kapral, L., Kuhrn, M., Abulesz, Y. T., Dibiasi, C., & Kimberger, O. (2024). A comparison of five algorithmic methods and machine learning pattern recognition for artifact detection in electronic records of five different vital signs: A retrospective analysis. Anesthesiology, 141(1), 32-43. DOI: 10.1097/ALN.0000000000004971

[13] Lee, H. Y., Kuo, P. C., Qian, F., Li, C. H., Hu, J. R., Hsu, W. T., ... & Lee, C. C. (2024). Prediction of in-hospital cardiac arrest in the intensive care unit: Machine learning–based multimodal approach. JMIR Medical Informatics, 12(1), e49142. DOI: 10.2196/49142

[14] Yang, M., Peng, Z., van Pul, C., Andriessen, P., Dong, K., Silvertand, D., ... & Long, X. (2024). Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor. Computer Methods and Programs in Biomedicine, 255, 108335. https://doi.org/10.1016/j.cmpb.2024.108335

[15] Verma, A. A., Stukel, T. A., Colacci, M., Bell, S., Ailon, J., Friedrich, J. O., ... & Mamdani, M. (2024). Clinical evaluation of a machine learning–based early warning system for patient deterioration. CMAJ, 196(30), E1027-E1037. DOI: 10.1503/cmaj.240132

[16] Yang, A. C., Ma, W. M., Chiang, D. H., Liao, Y. Z., Lai, H. Y., Lin, S. C., ... & Wang, C. Y. (2025). Early prediction of sepsis using an XGBoost model with single time-point non-invasive vital signs and its correlation with C-reactive protein and procalcitonin: A multi-center study. Intelligence-Based Medicine, 11, 100242. https://doi.org/10.1016/j.ibmed.2025.100242

[17] Malde, A., Prabhu, V. G., Banga, D., Hsieh, M., Renduchintala, C., & Pirrallo, R. (2025). A Machine Learning Approach for Predicting Maternal Health Risks in Lower-Middle-Income Countries Using Sparse Data and Vital Signs. Future Internet, 17(5), 190. https://doi.org/10.3390/fi17050190

[18] Thatha, V. N., Chalichalamala, S., Pamula, U., Krishna, D. P., Chinthakunta, M., Mantena, S. V., ... & Vatambeti, R. (2025). Optimized machine learning mechanism for big data healthcare system to predict disease risk factor. Scientific Reports, 15(1), 14327. https://doi.org/10.1038/s41598-025-98721-6

[19] Deshmukh, M. T., Wankhede, P. R., Chakole, N., Kale, P. D., Jadhav, M. R., Kulkarni, M. B., & Bhaiyya, M. (2025). Towards intelligent food safety: Machine learning approaches for aflatoxin detection and risk prediction. Trends in Food Science & Technology, 105055. https://doi.org/10.1016/j.tifs.2025.105055

[20] Doneda, M., Lanzarone, E., Giberti, C., Vernia, C., Vjerdha, A., Silipo, F., & Giovanardi, P. (2025). An ECG-based machine-learning approach for mortality risk assessment in a large European population. Journal of Electrocardiology, 88, 153850. https://doi.org/10.1016/j.jelectrocard.2024.153850

[21] Sharma, R., Peel, J., Kapoor, N., & Freeman, W. D. (2023). Editorial: Multimodal and auto-regulation monitoring in neuro-ICU. Frontiers in Neurology, 14, 1227237. https://doi.org/10.3389/fneur.2023.1227237.

[22] Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technology Letters, 8(3), 45-57. doi: 10.1049/htl2.12010

[23] Vitt, J. R., Mainali, S., & Loper, S. (2023). Multimodal and autoregulation monitoring in the neurointensive care unit. Frontiers in Neurology, 14, 1155986. https://doi.org/10.3389/fneur.2023.1155986.

[24] Alghieth, M. (2025). DeepECG-Net: a hybrid transformer-based deep learning model for real-time ECG anomaly detection. Scientific Reports, 15(1), 20714. https://doi.org/10.1038/s41598-025-07781-1

[25] Wang, L., Yin, Y., Glampson, B., Peach, R., Barahona, M., Delaney, B. C., & Mayer, E. K. (2024). Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. EBioMedicine, 110. DOI: 10.1016/j.ebiom.2024.105442

[26] Mei, Y., Jin, Z., Ma, W., Ma, Y., Deng, N., Fan, Z., & Wei, S. (2024). Optimizing Acute Coronary Syndrome Patient Treatment: Leveraging Gated Transformer Models for Precise Risk Prediction and Management. Bioengineering, 11(6), 551. DOI: 10.3390/bioengineering11060551

[27] Zisser, M., & Aran, D. (2024). Transformer-based time-to-event prediction for chronic kidney disease deterioration. Journal of the American Medical Informatics Association, 31(4), 980-990. DOI: 10.1093/jamia/ocae025

[28] Zhou, S., Guan, C., Deng, S., Zhu, Y., Yang, W., Zhang, X., ... & Huang, H. (2025). A novel sequence-based transformer model architecture for integrating multi-omics data in preterm birth risk prediction. npj Digital Medicine, 8(1), 536. https://doi.org/10.1038/s41746-025-01942-2

[29] El-Rashidy, N., Sultan, Y. A., & Ali, Z. H. (2025). Predecting power transformer health index and life expectation based on digital twins and multitask LSTM-GRU model. Scientific Reports, 15(1), 1359. https://doi.org/10.1038/s41598-024-83220-x

[30] Moody, B., Hao, S., Gow, B., Pollard, T., Zong, W., & Mark, R. (2022). MIMIC-IV Waveform Database (version 0.1.0). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/a2mw-f949

[31] Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 10775463251320258. https://doi.org/10.1177/10775463251320258

[32] Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A. (2017). Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Non-linearities. Mathematical Problems in Engineering, 2017(1), 8045803. https://doi.org/10.1155/2017/8045803

[33] Zouari, F., Saad, K. B., & Benrejeb, M. (2012). Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, 5(5), 2075-2103.

[34] Zouari, F., Saad, K. B., & Benrejeb, M. (2013, March). Adaptive backstepping control for a class of uncertain single input single output nonlinear systems. In 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13) (pp. 1-6). IEEE. 10.1109/SSD.2013.6564134

[35] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, 11, 100226. https://doi.org/10.1016/j.rico.2023.100226

[36] Zouari, F., Saad, K. B., & Benrejeb, M. (2013, May). Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor. In 2013 International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 864-871). IEEE. 10.1109/CoDIT.2013.6689656