Predicting Circulatory Failure in ICU Patients with Deep Sequential Models

Gabriel Rolla Ferreira[1]
Lucas Pimenta Braga[1]
Samuel L. V. Miranda[1]
Wagner Meira Jr.[1]
José Gabriel V. de Souza[1]
Luísa Barros Ribeiro Andrade[1]
Anisio Mendes Lacerda[1]
Alexandre Guimarães de Almeida Barros[2]
Laura Martins Froede[1]
Matheus Cândido Teixeira[1]
Gisele L. Pappa[1]

[1]Computer Science Department, Federal University of Minas Gerais (UFMG).
[2]Internal Medicine Department, INCT-NeuroTec-R, Federal University of Minas Gerais (UFMG).
{gabrielrolla, jose.souza, laurafroede, lucaspimenta, luisabarros, matheus.candido, samuel.miranda, anisio, glpappa, meira}@dcc.ufmg.br, xandebarros@gmail.com

***Abstract*—Healthcare professionals working in intensive care units (ICUs) receive a large volume of measurements from numerous monitoring systems. However, due to human cognitive limitations, it becomes difficult to process all this information and promptly recognize early signs of patient deterioration. Critically ill patients are typically admitted to ICUs, where alarm systems are implemented for the early detection of circulatory failure — a condition associated with high mortality, especially among severely ill individuals. Current alarm systems, which rely on fixed thresholds, often generate a high number of false positives, contributing to alarm fatigue and potentially delaying appropriate clinical interventions. This study proposes the development of an early warning system based on machine learning (ML) techniques to integrate multivariate data and predict circulatory failure up to 8 hours in advance. The model was trained using the MIMIC-IV database. The dataset includes over 70,000 ICU admissions, and the exploratory analysis was conducted with a temporal resolution of 5 minutes, enabling fine-grained modeling of patient trajectories.**

***Index Terms*—Intensive Care Unit, Circulatory Failure, Transformer, Deep Learning, Real-Time Prediction.**

I. Introduction

Circulatory failure, like many other circulatory diseases, represents a critical clinical condition characterized by inadequate tissue perfusion and insufficient oxygen delivery to vital organs, often associated with high mortality and serious consequences unless promptly treated. Treatment in intensive care units (ICUs) and early identification of this condition is important to improve outcomes, especially among hemodynamically unstable patients [1] [2]. However, anticipating circulatory deterioration remains challenging due to the complexity and volume of continuously generated high-frequency clinical data.

Conventional ICU alarm systems typically rely on static threshold-based rules—such as triggering alerts when mean arterial pressure (MAP) drops below 65 mmHg. While simple, these systems often produce excessive false positives and fail to account for temporal trends or variable interactions, contributing to alarm fatigue and delayed clinical response [3]. Machine learning (ML) approaches offer a promising alternative by leveraging high-frequency, multivariate time series to detect early signs of deterioration, as seen in other studies [4]. However, classical models such as recurrent neural networks (RNNs) may struggle with long-range dependencies and real-time scalability.

Hyland et al. [5] proposed the circEWS system – an early warning model trained on HiRID and MIMIC-III datasets using gradient-boosted decision trees applied to manually engineered temporal features. Despite its strengths, circEWS relies on handcrafted features and flattened time series, which may limit its generalizability and scalability across diverse clinical settings. In this study, we build upon the circEWS framework and apply it to the updated and more comprehensive MIMIC-IV dataset [6], aiming to enhance the model's applicability in current ICU scenarios. We propose a deep learning pipeline centered on the Transformer architecture [7], which enables the modeling of long-range dependencies in high-resolution clinical sequences. Our approach replaces manual static feature extraction with dynamically calculated temporal features integrated into the data sequence. To contextualize its performance, we benchmark the Transformer against Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models.

Our pipeline includes the structured extraction of ICU data—such as vital signs, laboratory tests, and vasopressor administration—along with imputation at regular 5-minute intervals and the derivation of temporal statistics like instability and accumulation. These data are organized into windows of up to 7 days. They are used to perform timestep-level binary classification. We evaluate the performance of the Transformer model in comparison with established recurrent models, using metrics such as the Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision-Recall Curve (AUPRC), and F1-score across multiple random seeds.

The main contribution of this study is to demonstrate the

feasibility of an end-to-end deep learning pipeline for predicting circulatory failure, with a focus on accuracy, real-time applicability, and robustness across different clinical scenarios. Our results highlight the potential of such approaches to support clinical decision-making in intensive care units.

II. Related Work

Early warning systems based on machine learning (ML) have shown great potential in ICU settings by predicting adverse events using high-frequency physiological data. These models go beyond threshold-based alarms by capturing complex temporal patterns and variable interactions, thus improving sensitivity and reducing false positives [8] [9].

Among the early efforts in this area, the circEWS model proposed by Hyland et al. [5] stands out for demonstrating promising results in predicting circulatory failure. The approach employed gradient-boosted decision trees trained on thousands of manually engineered features derived from multivariate time series, including descriptors such as measurement intensity, instability history, multi-resolution temporal summaries, and shapelet patterns. The outcome definition combined hypotension, vasopressor or inotrope administration, and elevated lactate levels. Despite its strong performance, the model heavily relied on handcrafted features and a tabular data representation, which may limit its scalability and its ability to capture complex temporal dependencies in real time.

While effective, circEWS requires extensive manual feature engineering and collapses time series into summary statistics, potentially losing temporal dynamics and limiting adaptability across ICU settings. Additionally, the reliance on tabular models constrains their ability to model long-term dependencies inherent in patient trajectories.

To overcome these limitations, deep learning models—particularly recurrent neural networks (RNNs) such as LSTMs and GRUs—have been explored. Shashikumar et al. [8] demonstrated the feasibility of using LSTMs for early sepsis detection based on multivariate physiological signals. However, RNNs are often limited by sequential processing bottlenecks and sensitivity to input window design.

Transformer-based architectures, initially proposed for natural language processing [7], have emerged as powerful tools for sequence modeling without recurrence. In healthcare, Rasmy et al. [10] developed Med-BERT, a Transformer trained on structured electronic health records (EHRs) for disease prediction, and Li et al. [11] introduced BEHRT for longitudinal patient data. Despite their success, these models focus on episodic EHR data and are not optimized for high-frequency ICU time series.

Our work builds on these developments by applying Transformers to continuous multivariate ICU data sampled at 5-minute intervals. Unlike prior approaches, we avoid manual feature extraction by incorporating temporal statistics (instability, intensity, accumulation) directly into the model input. This enables real-time risk assessment across the patient's ICU stay. We benchmark the Transformer against LSTM and GRU baselines using robust evaluation metrics across multiple seeds to assess performance and reproducibility. This approach aims to advance real-time, generalizable clinical decision support systems for circulatory failure prediction.

III. Data Characterisation

A. Dataset Overview

We used data from the publicly available MIMIC-IV database, version 3.1 [6], which contains de-identified electronic health records of patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019. The dataset is organized into modular tables and includes core hospital records, ICU-specific data, and laboratory test results. For this study, we focused exclusively on patients with at least one ICU admission, resulting in a target cohort of 50,916 patients and 73,181 ICU stays (identified by `stay_id`).

Rather than excluding admissions with missing values, we retained them and addressed data gaps through forward-fill imputation across a fixed 5-minute grid. For each variable, the most recent observation was propagated forward until a new value became available. This ensured temporal continuity, required for sequence modeling, without compromising dataset size.

B. Patient and ICU Stay Statistics

After applying the 1-hour minimum duration filter, the final dataset included 50,496 patients and 72,580 ICU stays—a reduction of approximately 0.15%. This cohort presented a mean age of 63.19 years. A higher prevalence of male admissions was observed (55.84%; n=40,468), compared to female admissions (44.16%; n=32,009). Finally, the average length of an ICU stay for this population was 3.48 days.

C. Variable Types and Distributions

The dataset was segmented into 87 standardized batches, each comprising approximately 2.4 million rows, resulting in over 200 million time-stamped observations across 715 columns. This structure was adopted to ensure computational scalability, particularly in the face of RAM constraints during preprocessing and feature extraction. All time series were resampled onto a fixed 5-minute grid to ensure temporal consistency across ICU stays.

Of the 715 columns, 71 represent continuous laboratory variables and 8 correspond to continuous vital signs. Additionally, the dataset includes 1 binary target variable (circulatory failure), 1 binary indicator for vasopressor administration, and 2 identifiers (`stay_id` and `charttime`). The remaining 632 columns are derived features obtained via temporal summarization of the original signals, including basic statistics and higher-level descriptors across time.

For each continuous variable, we derived a set of temporal descriptors, including minimum, maximum, mean, count of measurements, short-term variability (instability), occurrence of extreme values (intensity), and long-term trends (accumulation). These features were designed to enrich the temporal resolution of the clinical data while preserving interpretability.

Table I summarizes key descriptive statistics for a representative subset of physiological variables, illustrating typical value ranges and population-level variability.

TABLE I
SUMMARY STATISTICS OF SELECTED PHYSIOLOGICAL VARIABLES

Variable	Mean	Std. Dev.	Min	Median	Max
Heart rate (bpm)	82.5	17.5	2	78	198
Mean arterial pressure (mmHg)	82.0	14.2	1	83	299
Respiratory rate (rpm)	18.7	6.1	1	18	68
Temperature (°C)	36.9	0.45	31.0	37.0	40.2
Peripheral O_2 saturation (%)	96.8	2.7	10	98	100
Glucose (mg/dL)	112.5	77.0	5	115	1727
Lactate (mmol/L)	3.41	0.82	0.7	3.4	6.4

Additionally, 79 binary `_imputed` columns indicate whether values were observed (0) or imputed (1), based on a forward-fill method applied over the 5-minute grid. This structured representation enables granular, sequential modeling of physiological dynamics throughout ICU stays and supports robust training of temporal prediction models.

IV. METHODOLOGY

A. Data Preprocessing and Imputation

The initial stage of our methodology focused on a rigorous data preprocessing pipeline to construct a reliable dataset for modeling. A primary filtering step was conducted, retaining only ICU admissions with a duration longer than one hour to ensure that each stay contained sufficient data for meaningful sequential analysis. All valid admission IDs were systematically stored to guarantee traceability throughout the entire data processing and modeling workflow. Following this, variables of interest were carefully organized into clinically relevant categories: vital signs, laboratory tests, and vasoactive agents, taking into account their respective collection frequencies and established clinical importance.

A significant challenge in clinical datasets like MIMIC-IV is the prevalence of missing data. This phenomenon is not random; it often results from practical constraints, such as staff or equipment availability, or stems from deliberate clinical decisions where tests are ordered only when deemed necessary for a patient's diagnosis or treatment. Therefore, the very pattern of missingness can affect the dynamic reality of intensive care and patient condition [12]

To address this issue while preserving the temporal integrity of the data, we adopted a forward-fill imputation strategy. This method was specifically chosen over simpler statistical approaches, such as mean or median imputation, because it maintains the temporal causality of the data. By propagating the last known state of a patient, forward-fill provides a more clinically intuitive representation of the patient's trajectory, avoiding the signal distortion that can be introduced by using a single average value across different and dynamic time points. In cases where no prior observation existed at the start of an ICU stay, a predefined global reference value was used. This approach of using clinically plausible references, inspired by the physiological ranges in Hyland et al. [5], offers a more robust starting point than using statistical artifacts like zero or a global mean, which could create artificial drops or spikes in the initial data sequence.

Furthermore, to ensure the model could differentiate between measured and imputed data, a binary mask was generated for each variable. This mask assigns a value of `0` for an observed measurement and `1` for an imputed value.

Table 2 illustrates these two imputation mechanisms. First, at the beginning of the admission (08:00), when no prior heart rate exists, a reference value (70.0 bpm) is imputed, and its corresponding mask is set to 1 (`HR_imputed = 1`). Second, the forward-fill strategy is shown at 08:15, where the last observed HR value (82.0 bpm) is carried over to fill the gap, also with its mask set to 1. In contrast, all directly measured values, like the HR at 08:05, are paired with a mask of 0, signaling their origin as real observations.

TABLE II
EXAMPLE OF IMPUTATIONS FOR HEART RATE (HR) AND TEMPERATURE (TEMP)

Time	HR (bpm)	HR_imputed	Temp (°C)	Temp_imputed
08:00	70.0	1	36.7	0
08:05	80.0	0	36.9	0
08:10	82.0	0	37.0	1
08:15	82.0	1	37.1	0
08:20	85.0	0	37.1	1

In summary, the goal of preprocessing was to select relevant admissions, organize clinical variables, impute missing values using clinically plausible references, and ensure data completeness for modeling. The dataset was also divided into batches to facilitate large-scale processing, and intermediate files were stored to avoid repeated expensive queries to the original database.

B. Outcome Definition

The definition of circulatory failure adopted in this study was based on the criteria proposed by Hyland et al. [5], in which a patient is considered to be in circulatory failure when arterial lactate exceeds 2 mmol/L in conjunction with either hypotension (MAP $\leq$ 65 mmHg) or the use of vasopressor/inotropic agents. This formulation captures tissue hypoperfusion accompanied by hemodynamic compromise and is widely accepted as indicative of incipient circulatory shock.

The agents considered include dopamine, epinephrine, norepinephrine, phenylephrine, vasopressin, dobutamine, and milrinone—drugs commonly used in ICUs. Information on their administration was extracted from the structured temporal records available in the MIMIC-IV database. The binary variable `failure` was assigned a value of 1 when elevated lactate co-occurred with at least one of the additional conditions (reduced MAP or vasopressor use); otherwise, it was set to 0. A summary of the preprocessing steps, including variable

selection, imputation, and outcome labeling, is depicted in Figure 1.

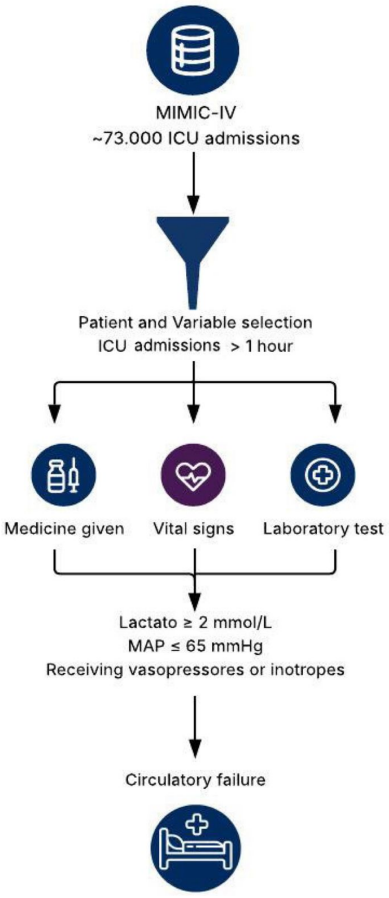

Fig. 1. Simplified pipeline for clinical preprocessing.

C. Feature Engineering

To enrich the model's input and capture the temporal dynamics of the patient's condition, a comprehensive feature engineering process was implemented. The use of temporal features demonstrates that the extraction of sequential features is critical for the performance of recurrent neural networks, as they translate physiological evolution into mathematically tractable representations [13]. For each physiological and laboratory variable, the following derived attributes were computed at each timestep:

- **Cumulative statistics:** cumulative count, mean, minimum, and maximum values observed up to that point;
- **Instability:** rolling standard deviation over a one-hour window (12 timesteps of 5 minutes each), reflecting recent variability;
- **Intensity:** absolute difference from the previous value, capturing abrupt changes;
- **Cumulative sum:** total accumulated value across the ICU stay.

After generating the extended feature set, a filtering step was applied: any variable (original or derived) with more than 50% missing values across the dataset was removed to ensure sufficient data coverage and avoid sparse training signals.

D. Data Consolidation and Normalization

Following feature filtering, all eligible ICU admissions were consolidated into a single HDF5 file, organized into training, validation, and test sets. The HDF5 format was selected for its efficiency in storing large datasets and enabling fast sequential access during deep learning workflows. Continuous variables were standardized using Z-score normalization. To prevent data leakage, the mean and standard deviation for each variable were calculated exclusively from the training set and then applied to transform all three subsets. Any missing values remaining after this transformation were imputed with zero.

Each ICU admission was uniquely identified by a `stay_id`. For each data split, a dedicated array of corresponding stay IDs was created and stored in the HDF5 file, e.g., at `/patient_windows/train_stay_ids`. This structure enabled consistent tracking of admissions and facilitated efficient window indexing during data loading. These operations were implemented using PyTables and Pandas, integrating TSV-based split definitions and HDF5 manipulation.

E. Temporal Window Construction

To support sequential modeling and early detection of circulatory failure, the data were segmented into fixed-size observation windows of 5 minutes, aligned with the temporal resolution adopted in prior literature. Each ICU admission was transformed into a multivariate time series containing vital signs, laboratory results, and treatment indicators. The maximum sequence length was set to 2016 timesteps (equivalent to seven days in the ICU), balancing event coverage and computational feasibility. Stays shorter than this threshold were zero-padded and masked during training and inference, while longer stays were truncated at the limit.

At each timestep, the model receives the full temporal history up to that point and predicts the presence of circulatory failure at that moment. No future information is included in the input sequence, ensuring a causal and clinically realistic setup. The data were split by `stay_id`, ensuring that admissions from the same patient were not distributed across different sets. Stratified sampling was used to maintain the proportion of positive events in each subset: 70% training, 15% validation, and 15% test.

F. Modeling Pipeline

To predict circulatory failure, we selected three distinct deep learning models: LSTMNet, GRUNet, and SimpleTransformer, all implemented using PyTorch. Input data was structured into tensors with the shape (`batch_size`, `seq_len`, `n_features`). In this configuration, `batch_size` denotes the number of sequences processed in parallel, `seq_len` represents the maximum length of the time sequences (after padding), and `n_features` refers to the number of variables or features at each time step. This tensor structure is a common convention for time-series analysis with neural networks.

1) Input Data Structure: The forward-fill imputation strategy, combined with the creation of binary imputation masks, follows practices established in previous studies dealing with sparse ICU data [14]. As previously described, the ICU data were preprocessed into fixed-length windows of 5-minute intervals, with a maximum sequence length of 2016 timesteps (equivalent to 7 days). Each timestep contains 620 variables extracted from vital signs, laboratory results, and medication records. The resulting dataset was stored in HDF5 format and partitioned into training, validation, and test sets.

Although the original batch files included 715 columns, the number of variables per timestep was reduced to 620 through a two-step refinement process. First, variables with more than 50% missing values were excluded to ensure adequate data coverage and avoid sparse training signals. Second, auxiliary columns with the `_imputed` suffix—which indicated whether a value was observed or imputed—were deliberately removed. Additionally, the binary variable indicating active vasopressor use was excluded to prevent data leakage, since its presence directly contributes to the definition of the outcome (circulatory failure). Including this variable as an input feature could allow the model to infer the target from an explicit component of its definition, undermining the predictive validity. After these filtering steps, the final dataset consisted of 620 temporally aligned and clinically relevant features per timestep.

2) Model Architecture: We selected LSTM, GRU, and Transformer-based models based on their proven effectiveness in capturing temporal dependencies in clinical time series, as demonstrated in studies such as the HiRID-ICU-Benchmark. These architectures have consistently shown strong performance in ICU prediction tasks involving irregular, multivariate data. However, while such studies validate the general suitability of these models, they often do not prescribe specific architectural configurations such as the number of layers or neurons. Consequently, our model design choices were guided by a combination of insights from related literature, practical computational constraints, and empirical tuning on the validation set. [15] [16].

In our implementation, all models use a single hidden layer to reduce complexity and avoid overfitting. The Transformer encoder employs a 231-dimensional input projection and a feedforward layer with 462 neurons. Similarly, LSTM and GRU models use a single unidirectional recurrent layer with 231 hidden units. These values were selected based on grid search over the validation set, considering model convergence and training stability. Although the referenced studies do not report exact hyperparameter settings, our final configuration aligns with simplified architectures commonly used in ICU early warning systems.

SimpleTransformer starts by projecting the data into 231 dimensions, preparing it for the encoder. In this architecture, the encoder consists of a single layer equipped with an attention head; this attention mechanism allows the model to directly assess and weigh the importance of all other time steps in the input sequence when processing each time step, capturing complex dependencies regardless of their distance within the sequence. This design is much lighter than traditional Transformers, which often employ multiple such layers. The encoder is followed by a feedforward layer of 462 neurons (double the input) and includes 10% dropout to avoid overfitting. The final classifier gradually reduces dimensionality, first to 32 neurons with ReLU, then to a single output with sigmoid, which gives the probability of positive class.

The LSTMNet and GRUNet models, on the other hand, are classic recurrent models for processing sequences, offering a more straightforward approach than Transformer, while still being effective. Input data is initially projected into a 231-dimensional embedding, adjusting it to a format compatible with the recurrent layers. Next, a single LSTM or GRU layer processes the sequence unidirectionally. These layers iterate through the sequence step-by-step, maintaining an internal hidden state (or memory cells in LSTMs) that acts as a compressed representation of relevant historical information from previous time steps. This state is updated at each step and used to inform the processing of the current input, allowing the model to capture temporal patterns. This unidirectional, single-layer approach was chosen to keep the models simple while still being efficient. The final classifier is identical to the SimpleTransformer. Comparing the models, we can conclude that Transformers are models that prioritize precision, but require more data and resources. While LSTM/GRU models are lighter and faster, requiring less data, they are less flexible for complex contexts and are less accurate.

3) Training Procedure: The models were trained using the BCELoss (Binary Cross Entropy) loss function, adapted to ignore masked values, ensuring that only valid time steps contributed to the error calculation. The optimizer used was Adam, with a learning rate of 3×10^{-4} and weight decay of 1×10^{-6} for regularization. To avoid overfitting, early stopping was implemented with epoch patience, interrupting training if the validation metric did not improve within this interval. Training was limited to a maximum of 1000 epochs, although early stopping was frequent due to the stopping criterion. The batch size was set at 16 sequences, balancing computational efficiency and gradient stability.

G. Reproducibility and Availability

Due to ethical and legal constraints, the dataset used in this study (MIMIC-IV) is not publicly available. Access is limited to credentialed researchers who complete human subjects training (e.g., the NIH course) and accept PhysioNet's data use terms.

To promote transparency and reproducibility, the full source code is publicly available[1].

The repository includes the full implementation: preprocessing pipeline, time grid construction, imputation strategies, and model configurations. All input variables—vital signs, lab tests, and vasopressors—are mapped to their clinical names

[1]https://github.com/lucaspimentab/CircularyFailure

and `itemid` in MIMIC-IV, ensuring traceability and external validation. The code is extensively documented and organized to facilitate reproduction of results.

V. Results

We evaluated three neural network architectures—Transformer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)—for early detection of clinical deterioration over an 8-hour window. Results are presented in four dimensions: alarm timing, recall over time, classification performance, and discrimination via Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves.

1. Alarm Timing Relative to Deterioration

The temporal distribution of alarms provides valuable information about the model's behavior (Fig. 2). Among patients who experienced clinical deterioration, the model issued alerts an average of 2.34 ± 0.49 hours before the event, offering a clinically meaningful window for early intervention. Such lead time is critical in ICU settings, where timely recognition of physiological instability can significantly impact patient outcomes [2]. For comparison, in the context of myocardial infarction, De Luca et al. [17] demonstrated that each 30-minute delay in angioplasty increases one-year mortality by 7.5%, underscoring the importance of anticipation in acute care.

Most alarms in deteriorating patients occurred within the last two hours before the event, suggesting sensitivity to acute physiological shifts. In contrast, alarms for stable patients were more evenly distributed, indicating persistent monitoring even in non-critical cases. This contrast reflects the model's ability to distinguish physiological trajectories—denser alarms near deterioration, and broader dispersion in stable cases.

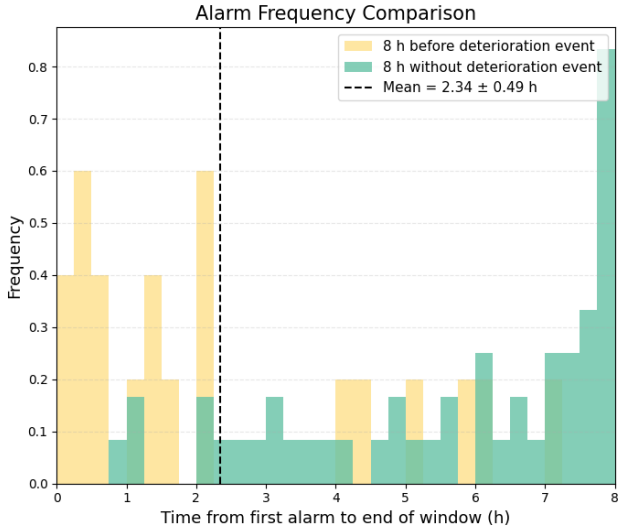

Fig. 2. **Alarm timing distribution relative to deterioration.** Histogram showing the interval between first alarm and the event (or window end), comparing deteriorating and non-deteriorating patients. The dashed line marks the mean alarm time for deteriorating cases.

2. Recall Variation Over Time and Impact of Decision Thresholds

The ability of the models to detect impending clinical deterioration varies across the 8-hour prediction window, depending on the decision threshold applied. Figure 3 illustrates how recall changes over time for three representative probability thresholds: 0.80, 0.90, and 0.95. As anticipated, recall values decrease significantly as the prediction is made further in advance of the deterioration event.

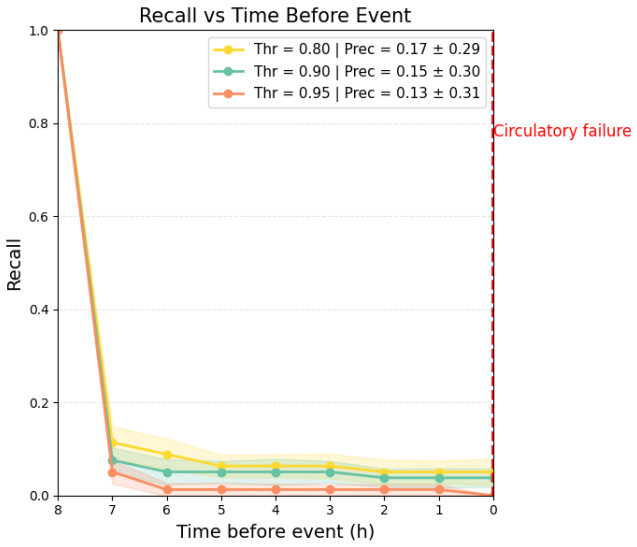

Fig. 3. **Recall at different times before circulatory failure.** Recall values for alarm thresholds (0.80, 0.90, 0.95) across the 8-hour prediction window. A sharp drop in recall is observed as the prediction horizon increases.

As shown in Fig. 3, lower decision thresholds result in higher recall values, indicating a more permissive alarm configuration that captures a greater proportion of true positives. However, this increase in sensitivity comes at the cost of reduced precision due to higher false positive rates. Additionally, the decline in recall as the prediction horizon lengthens reflects the intrinsic challenge of anticipating clinical events well in advance.

3. Classification Performance

To assess the performance of the selected Transformer model (seed = 43), we computed key classification metrics. The model achieved a recall of 0.77, precision of 0.85, and a false positive rate of 0.025, demonstrating strong discriminative ability. It effectively identified most deterioration events while minimizing false alarms—critical in ICU environments, where alarm fatigue can undermine clinical decision-making [3].

These results highlight the importance of using evaluation metrics that consider both false positives and false negatives, as emphasized by recent studies [18]. The low false positive rate, in particular, reinforces clinical trust in automated decision-support systems.

Figure 4 presents the confusion matrix for the best-performing Transformer model, offering an intuitive summary

of its classification outcomes. The model correctly classified 21,317 deterioration cases as positive (77.5% of actual positives) and 151,724 stable cases as negative (97.5% of actual negatives), while misclassifying 3,879 stable cases as false positives (2.5%) and 6,189 deterioration events as false negatives (22.5%). This distribution demonstrates the model's robustness in critical care scenarios. The high number of true positives indicates strong sensitivity, while the low number of false positives is key to mitigating alarm fatigue.

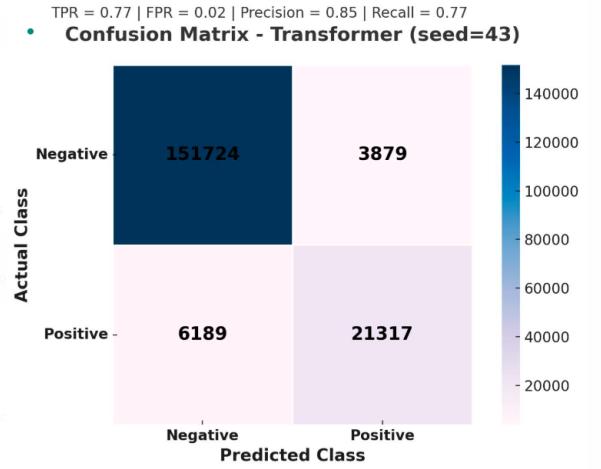

Fig. 4. **Confusion matrix for the best-performing model.**

4. Model Discrimination: PR and ROC Analysis

To evaluate the model's ability to distinguish between classes across various decision thresholds, we analyzed both the Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves. As shown in Table III, the Transformer achieved the highest AUPRC and AUROC values among the evaluated models for the best-performing seed. Although the GRU and LSTM models showed competitive performance—particularly in AUROC—the Transformer consistently outperformed both in this configuration.

TABLE III
AUPRC AND AUROC SCORES FOR EACH MODEL)

Model	AUPRC	AUROC
Transformer	0.88 ± 0.01	0.98 ± 0.002
GRU	0.86 ± 0.00	0.97 ± 0.002
LSTM	0.84 ± 0.01	0.97 ± 0.002

Figures 5 and 6 provide visual confirmation of these results, illustrating the discriminative performance of all models across decision thresholds. The ROC curves, in particular, show that the Transformer maintains higher sensitivity and specificity throughout the operating range.

Although the ROC curves in Figure 6 suggest that all models exhibit high discriminative power, the differences among them are less visually distinguishable compared to the PR curves. This is expected, as AUROC can sometimes overestimate model performance under class imbalance. In contrast, the

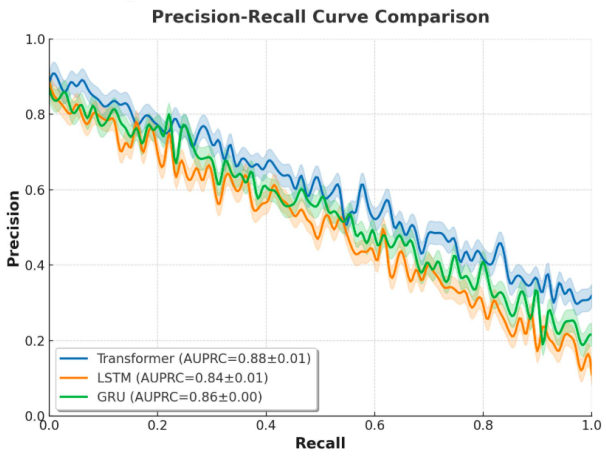

Fig. 5. **Precision-recall curves for all models.** The Transformer consistently achieves higher AUPRC values across thresholds.

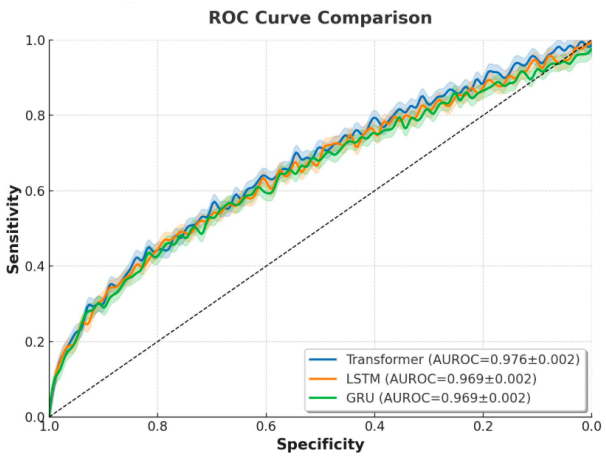

Fig. 6. **ROC curves for all models.** The Transformer model maintains superior discrimination across a wide range of operating points.

precision-recall curves in Figure 5 more clearly separate the models, reinforcing the superiority of the Transformer in identifying positive cases with higher precision across a wide range of recall values. This distinction is particularly relevant for clinical applications, where the prevalence of the positive class tends to be low and high precision is essential to avoid alarm fatigue.

TABLE IV
PERFORMANCE METRICS FOR EACH MODEL USING SEED 43

Model	AUROC	AUPRC	F1	Precision	Recall	Loss
Transformer	0.976	0.884	0.804	0.844	0.768	0.177
LSTM	0.969	0.858	0.745	0.839	0.671	0.207
GRU	0.969	0.839	0.782	0.812	0.755	0.192

To provide a more detailed comparison, Table IV presents the classification metrics for all models using seed 43, which was selected as the best-performing run. These results confirm the consistent advantage of the Transformer, particularly in AUROC and AUPRC, while also revealing that the GRU

achieves a higher F1-score than LSTM, suggesting better overall balance between precision and recall in this configuration.

5. Summary and Discussion

The results obtained demonstrate that the models developed were effective in the task of early prediction of circulatory failure in intensive care unit settings. The strong performance observed across key evaluation metrics — including AUROC, AUPRC, F1-score, and the timing of predictive alarms — validates the accuracy and reliability of the proposed approach. These outcomes confirm that the central goals of the study were successfully achieved. Additionally, the consistency of results observed across multiple training runs and distinct neural network architectures reinforces the robustness and generalizability of the methodology employed. Together, these findings highlight the potential of the proposed models to support real-time clinical decision-making in critical care environments, as has been shown in other studies. [19]

VI. Conclusions and Future Work

This study demonstrates the feasibility and robustness of deep learning models—specifically Transformer, LSTM, and GRU architectures—for the early prediction of circulatory failure in intensive care unit (ICU) settings. All three models exhibited strong and consistent performance across multiple evaluation metrics, including AUROC, AUPRC, F1-score, and alarm timing. Although overall results were similar, the Transformer architecture achieved slightly higher scores in classification and discrimination tasks, suggesting marginal benefits in capturing complex temporal patterns.

Despite these promising results, most alarms were concentrated near the time of clinical deterioration, highlighting the challenge of forecasting over longer horizons. Future work will explore advanced temporal modeling strategies—such as multi-scale attention, memory-augmented mechanisms, or hybrid recurrent-transformer architectures—to enhance early warning capabilities over extended prediction windows. Furthermore, future iterations will evaluate the impact of retaining the `_imputed` columns in the feature set, allowing the model to explicitly identify imputed values. This approach may enhance predictive accuracy by incorporating information about data provenance, as suggested by prior studies on missing clinical data [20].

Another important research direction involves evaluating model performance across clinically relevant subgroups. Inspired by approaches like circEWS, we plan to conduct stratified analyses based on diagnostic categories from the Acute Physiology and Chronic Health Evaluation (APACHE), severity scores, age groups, admission types, and ICU length of stay [5]. These evaluations aim to ensure equitable model performance and to guide subgroup-specific calibration where needed.

References

[1] S. H. Mahmoud, S. S. Soussa, A. M. Hassan, H. F. Abdelrazik, S. M. Hashem, and A. M. Mari, "Smart prediction of circulatory failure: Machine learning for early detection of patient deterioration," in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, Giza, Egypt, 2023, pp. 199–203.

[2] G. Duke, J. Green, and J. Briedis, "Survival of critically ill medical patients is time-critical," *Critical Care and Resuscitation*, vol. 6, no. 4, pp. 261–267, 2004, iSSN: 1441-2772.

[3] M. Cvach, "Monitor alarm fatigue: An integrative review," *Biomedical Instrumentation & Technology*, vol. 46, no. 4, pp. 268–277, 2012. [Online]. Available: https://array.aami.org/doi/abs/10.2345/0899-8205-46.4.268

[4] A. D. C. Bezerra, N. S. Maciel, L. S. da Silva Filho, A. S. Mendes, F. N. B. Gois, and L. M. S. da Silva, "Efetividade de algoritmos de inteligência artificial para predição de sepse em adultos de unidades de terapia intensiva: Revisão de escopo," *Rev. Interfaces*, vol. 11, no. 4, pp. 3180–3190, Dec. 2023.

[5] S. L. Hyland, M. Faltys, M. Hüser, J. R. Ledsam, A. Meyer, A. Schwaighofer, M. Moor, C. Bock, M. Horn, B. Rieck *et al.*, "Early prediction of circulatory failure in the intensive care unit using machine learning," *Nature Medicine*, vol. 26, no. 3, pp. 364–373, 2020.

[6] A. E. W. Johnson, T. J. Pollard, L. Shen *et al.*, "Mimic-iv (version 2.2)," https://physionet.org/content/mimiciv/, 2023.

[7] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[8] S. P. Shashikumar, M. D. Stanley, I. Sadiq, Q. Li, A. Holder, G. D. Clifford, and S. Nemati, "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," *Journal of Electrocardiology*, vol. 50, no. 6, pp. 739–743, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022073617302546

[9] S. Yuan, Z. Yang, J. Li, C. Wu, and S. Liu, "Al-Powered early warning systems for clinical deterioration significantly improve patient outcomes: a meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 25, no. 203, 2025.

[10] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[11] Y. Li, Y. Rao, D. Solares *et al.*, "Behrt: Transformer for electronic health records," *Scientific Reports*, vol. 10, no. 1, p. 7155, 2020.

[12] J. Wang, W. Du, Y. Yang, L. Qian, W. Cao, K. Zhang, W. Wang, Y. Liang, and Q. Wen, "Deep learning for multivariate time series imputation: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2402.04059

[13] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1511.03677

[14] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," 2016. [Online]. Available: https://arxiv.org/abs/1606.01865

[15] J. Huang, Y. Cai, X. Wu, X. Huang, J. Liu, and D. Hu, "Prediction of mortality events of patients with acute heart failure in intensive care unit based on deep neural network," *Computer Methods and Programs in Biomedicine*, vol. 256, p. 108403, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260724003961

[16] L. Su, X. Zuo, R. Li, X. Wang, H. Zhao, and B. Huang, "A systematic review for transformer-based long-term series forecasting," *Artificial Intelligence Review*, vol. 58, no. 3, p. 80, 2025.

[17] G. De Luca, H. Suryapranata, J. P. Ottervanger, and E. M. Antman, "Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction," *Circulation*, vol. 109, no. 10, pp. 1223–1225, 2004. [Online]. Available: https://www.ahajournals.org/doi/10.1161/01.CIR.0000121424.76486.20

[18] A. Karrar, "The effect of using data pre-processing by imputations in handling missing values," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 10, pp. 375–384, 06 2022.

[19] M. D. Aldhoayan and Y. Aljubran, "Prediction of icu patients' deterioration using machine learning techniques," *Cureus*, vol. 15, no. 5, p. e38659, may 2023. [Online]. Available: https://doi.org/10.7759/cureus.38659

[20] Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with rnns," 2016. [Online]. Available: https://arxiv.org/abs/1606.04130