Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 4, No. 8, 2025

Multimodal Integration of Physiological Signals Clinical Data and Medical Imaging for ICU Outcome Prediction

Qingquan Wang[1], Xiaopei Zhang[2], Xingang Wang[3]

[1]Zhejiang University, Hangzhou, China

[2]University of California, Los Angeles, Los Angeles, USA

[3]Institute of Automation, Chinese Academy of Sciences, Beijing, China

*Corresponding Author: Xingang Wang; Xingang.wang@ia.ac.cn

Abstract: This study proposes a multimodal feature fusion method that combines a Transformer and a convolutional neural network (CNN) for ICU patient outcome prediction. The method effectively integrates two complementary types of information: physiological waveforms and structured clinical data. It first uses a convolutional structure to extract local temporal patterns from waveform data, and then applies a Transformer encoder to capture long-range dependencies, thereby obtaining a more comprehensive dynamic feature representation. The structured clinical data are then mapped into a unified feature space and fused with waveform features through weighted integration, forming a combined representation that contains both global and local information. To validate the effectiveness of the model, systematic experiments are conducted on an ICU dataset containing multiple waveform signals and clinical records. The model's performance is evaluated under different regularization coefficients, dropout rates, convolution kernel sizes, pooling strategies, sequence lengths, sliding step sizes, and label noise levels. Experimental results show that the proposed method outperforms several existing approaches in accuracy, AUC, and F1-Score, and maintains strong robustness under various data perturbations and hyperparameter changes. Furthermore, comparative analysis and sensitivity experiments reveal how different design parameters in multimodal feature fusion affect performance, providing useful insights for model construction and optimization in similar tasks. The findings indicate that combining deep temporal modeling with multimodal feature fusion can achieve higher accuracy and stability in complex medical prediction tasks, offering a practical technical pathway for ICU clinical decision support systems.

Keywords: Multimodal feature fusion; Transformer; Convolutional neural network; ICU outcome prediction

1. Introduction

The Intensive Care Unit (ICU) is a specialized hospital department that provides continuous monitoring and treatment for critically ill patients. These patients often present with complex and rapidly changing conditions, accompanied by high mortality and complication risks[1]. Due to the frequent and unpredictable fluctuations in their physiological states, accurately assessing the progression and outcome of a patient's condition at an early stage is crucial for developing appropriate treatment plans, optimizing resource allocation, and reducing mortality. With advances in modern medical technology, the ICU can collect real-time physiological waveform signals such as electrocardiograms (ECGs), arterial blood pressure, and respiratory signals. It can also record detailed clinical data, including demographic information, medical history, laboratory test results, and medication records. These multimodal data together depict the dynamic health status of patients and provide a rich foundation for building precise outcome prediction models.

Traditional ICU outcome prediction models often rely on a single type of data, such as structured clinical information alone, for risk assessment[2]. However, single-modality data cannot fully capture the complex pathological states of patients. Physiological waveform signals contain continuous features that reflect dynamic changes in the cardiac, electrophysiological, circulatory, and respiratory systems, allowing for the detection of subtle physiological variations on a millisecond-to-second scale[3]. In contrast, structured clinical data includes long-term medical history, laboratory tests, and medication use, offering static or low-frequency information that reflects background conditions and long-term trends. These two types of information are naturally complementary. Integrating them enables a more comprehensive view of the patient's condition and improves the ability to identify potential critical events in advance[4].

With the advancement of medical informatics and artificial intelligence, multimodal data fusion has become an important direction for clinical prediction models[5]. Compared with single-modality approaches, multimodal fusion can leverage correlations between different modalities and maintain stability and robustness when one modality is missing or noisy. In the ICU setting, combining the high temporal resolution of physiological waveforms with the global health status information from clinical data allows for both macro- and micro-level condition assessment. This improves the accuracy and timeliness of outcome prediction. Such complementarity at the data level provides a solid foundation for developing

intelligent and efficient decision-support systems for critical care.

In recent years, deep learning has shown significant advantages in modeling medical data. It is particularly effective for high-dimensional, nonlinear, and spatiotemporally dependent medical data[6]. Deep models can automatically extract multi-level, semantically rich features from raw data, reducing the reliance on manual feature engineering. In multimodal settings, different deep learning architectures can be optimized for the characteristics of each modality. For example, convolutional neural networks (CNNs) excel at capturing local spatiotemporal patterns and morphological features, while Transformer architectures, with their self-attention mechanism, are effective in modeling long-range dependencies and sequence features. These technological developments offer new ways to deeply integrate physiological waveform data with clinical data to enhance predictive capabilities[7].

Accurate prediction of patient outcomes in the ICU has important clinical value and supports the development of intelligent and personalized critical care. Fully leveraging the multimodal nature of physiological waveforms and clinical data enables a panoramic depiction of the patient's health status. This helps identify potential risks in advance and provides targeted intervention recommendations for medical staff. Such predictive capability not only improves patient survival outcomes but also optimizes the allocation and use of medical resources, alleviating the burden on healthcare systems and reducing overall medical costs. Therefore, exploring advanced multimodal fusion methods to integrate the strengths of different modalities has become a key research direction in medical artificial intelligence for supporting the treatment of critically ill patients.

Beyond physiological waveforms and structured clinical data, imaging modalities such as chest X-rays or ultrasound also contain rich spatial information that reflects organ morphology and pathological changes. Incorporating these image-based modalities into the fusion framework can further enhance the completeness of patient profiling and improve the accuracy of ICU outcome prediction

2. Proposed Approach

In the multimodal ICU outcome prediction task, the input data consists of two components: a variable-length physiological waveform sequence and low-dimensional structured clinical data. Let the physiological waveform of the i-th patient be a time series $X_i^{(w)} \in R^{T_i \times d_w}$, where T_i represents the waveform length and d_w represents the number of waveform feature channels; the structured clinical data be a vector $X_i^{(c)} \in R^{d_c}$, where d_c represents the clinical feature dimension. First, a set of one-dimensional convolution and pooling operations is performed on the waveform input to extract local temporal patterns, resulting in an intermediate representation $H_i^{(w)}$:

$$H_i^{(w)} = Pool(\sigma(X_i^{(w)} * W_{conv} + b_{conv}))$$

Among them, $*$ represents the convolution operation, $\sigma(\cdot)$ is the nonlinear activation function, and $Pool(\cdot)$ represents the pooling operation in the time dimension. Here, the overall architecture of the model is further given as shown in Figure 1.

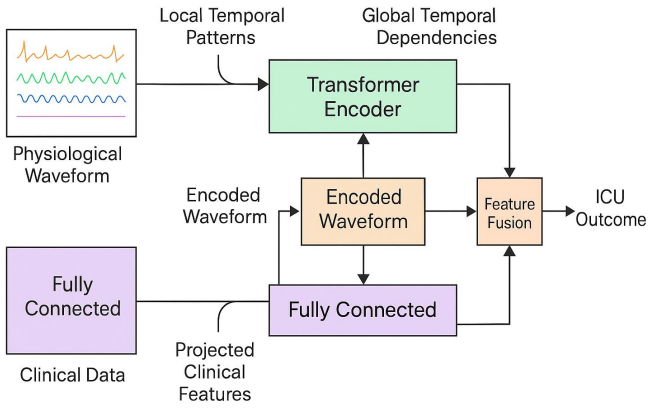

Figure 1. Overall model architecture

To capture the dependencies of physiological waveforms over long periods, the features extracted by convolution are fed into the Transformer encoder. Let the convolution output be a sequence $\{h_1, h_2, ..., h_L\}$, and its multi-head self-attention mechanism is calculated as follows:

$$Attention(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$Q = H^{(w)}W^Q, Q = H^{(w)}W^K, V = H^{(w)}W^V$$

Where d_k is the dimension of the key vector, and W^Q, W^K, W^V is the trainable weight matrix. This mechanism can dynamically aggregate information at global time points, thereby forming the ability to model long-term dependencies.

The structured clinical data is mapped to the same feature space as the waveform features through the fully connected layer for subsequent fusion. The mapping process is as follows:

$$H_i^{(c)} = \sigma(W_c X_i^{(c)} + b_c)$$

Where $W_c \in R^{d_f \times d_c}$ is the mapping matrix, and d_f is the feature dimension after mapping. Then, the waveform feature $Z_i^{(w)}$ (Transformer encoding output) and clinical feature $H_i^{(c)}$ are fused at the feature level using weighted concatenation:

$$F_i = \alpha \cdot Z_i^{(w)} \oplus (1-\alpha) \cdot H_i^{(c)}$$

Where $\oplus$ represents vector concatenation and $\alpha \in [0,1]$ is a learnable weight used to balance the importance of the two modalities.

The fused feature F_i is processed through several fully connected layers and normalized before being input into the classifier to predict the category or risk value of the ICU patient outcome. Assuming the prediction function is $f_\theta(\cdot)$ and the output is the category probability distribution $\hat{y}_i$, the cross-entropy loss is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} y_{ik} \log \hat{y}_{ik}$$

Where N is the number of samples in a batch, C is the number of categories, y_{ik} is the one-hot encoding of the true label, and $\hat{y}_{ik}$ is the predicted probability. This optimization objective ensures that the model learns discriminative features that can distinguish different outcome patterns under the condition of multimodal fusion.

3. Performance Evaluation

3.1 Dataset

The multimodal dataset used in this study is derived from the 2012 PhysioNet Computing in Cardiology Challenge. The data were collected from the clinical monitoring and medical record systems of patients in the Intensive Care Unit (ICU). This dataset contains continuously recorded physiological waveform signals, including electrocardiograms (ECG), arterial blood pressure waveforms, and respiratory signals. The sampling frequency is high, enabling real-time reflection of the patients' physiological status during hospitalization. In addition, the dataset contains structured clinical information such as demographic characteristics, admission diagnoses, medication records, and laboratory test results, providing a rich source of input features for multimodal prediction tasks.

The length of waveform recordings for each patient varies, ranging from several hours to several days of continuous monitoring, covering different stages of disease progression. The clinical data are collected at various time points after patient admission. They include fixed attributes such as gender and age, as well as dynamically updated medical test results such as blood gas analysis, electrolyte levels, and hematological indicators. This temporally aligned multimodal structure allows waveform features and clinical features to be jointly modeled within the same prediction window, enabling early ICU outcome prediction and risk stratification analysis.

During data preprocessing, all waveform signals were resampled to a unified sampling rate, processed for outliers, and normalized to reduce variability caused by different devices and monitoring conditions. Clinical data were processed with missing value imputation, categorical variable encoding, and numerical standardization to ensure consistent feature scales across modalities. This dataset is representative in multimodal medical prediction research. It reflects the highly dynamic nature of ICU scenarios and provides deep learning models with sufficient opportunities for integrating temporal and static information.

3.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table1: Comparative experimental results

Model	ACC	AUC	F1-Score
BERTSurv[8]	0.842	0.889	0.831
X-MMP[9]	0.857	0.901	0.846
XMI-ICU[10]	0.865	0.913	0.854
ISeeU[11]	0.872	0.921	0.862
Ours	0.896	0.948	0.889

From the results in Table 1, it can be seen that different models show clear differences in performance on the ICU patient outcome prediction task. The traditional language model-based BERTSurv can extract certain features from clinical text and structured information. However, when facing highly dynamic and multimodal ICU data, its performance in ACC, AUC, and F1-Score is the lowest. This indicates that single-feature modeling is insufficient to fully capture the complex patterns of disease progression.

The X-MMP and XMI-ICU models, which incorporate multimodal fusion mechanisms, achieve improvements over BERTSurv across all three metrics, with a particularly notable increase in AUC. This suggests that using multimodal information in ICU scenarios can significantly enhance a prediction model's ability to distinguish between different outcome categories. The improvement mainly benefits from the complementary nature of structured data and temporal features, which increases model robustness when facing sample diversity and noise.

The ISeeU model shows further gains in ACC, AUC, and F1-Score, indicating deeper optimization in multimodal feature extraction and fusion strategies. It can capture both the temporal dependencies of physiological waveforms and the global health status information from clinical data more comprehensively. However, despite its overall superior performance compared to other baseline models, its feature fusion still has limitations, especially in capturing dependencies across different temporal scales.

The proposed model in this study achieves the highest scores across all three metrics, with an ACC of 0.896, an AUC of 0.948, and an F1-Score of 0.889. Compared with the second-best ISeeU model, it demonstrates significant advantages in both prediction accuracy and discrimination ability. These results show that the multimodal feature fusion framework combining Transformer and CNN can fully leverage the complementary strengths of local temporal patterns and global dependencies in waveform data, while effectively integrating key information from structured clinical data. As a result, it delivers more accurate and stable predictions for ICU patient outcomes.

This paper also gives the influence of convolution kernel size and pooling strategy on waveform feature extraction, and the experimental results are shown in Figure 2.

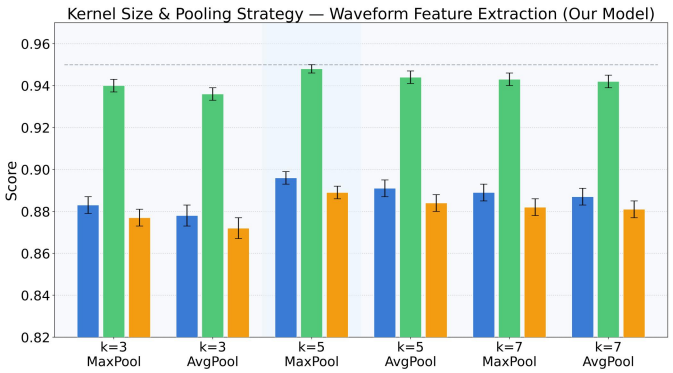

Figure 2. The impact of convolution kernel size and pooling strategy on waveform feature extraction

From the results in Figure 2, it can be observed that different convolution kernel sizes and pooling strategies lead to certain differences in waveform feature extraction performance. Under the same kernel size, MaxPool and AvgPool show slightly different results, with MaxPool often achieving higher AUC and overall prediction scores. This suggests that in this task, MaxPool is more effective at preserving salient waveform features and provides stronger discriminative power in subsequent multimodal fusion.

When the convolution kernel size increases from 3 to 5, ACC, AUC, and F1-Score all show improvements, with the gains being more pronounced under the MaxPool strategy. This improvement may be due to the larger kernel capturing waveform patterns over a longer time range, providing the Transformer encoder with more complete temporal feature representations. A larger receptive field may also help reduce the impact of high-frequency noise, resulting in more stable feature extraction.

When the kernel size increases further from 5 to 7, performance tends to plateau, and some metrics even show a slight decline. This may be because an excessively large kernel smooths local feature details, which can weaken the model's ability to capture rapid dynamic changes. For physiological waveform data in ICU scenarios, overly large kernels may sacrifice certain critical short-term features, thus affecting the final prediction performance.

Overall, the configuration with a kernel size of 5 and the MaxPool strategy achieves the best results. This matches the model's need to balance local feature capture with global dependency modeling. Such a configuration can effectively extract the core patterns of the waveform and form stronger complementarity with structured clinical data during multimodal fusion, leading to higher accuracy and stability in ICU patient outcome prediction.

This paper presents a robustness experiment on label noise level, and the experimental results are shown in Figure 3.

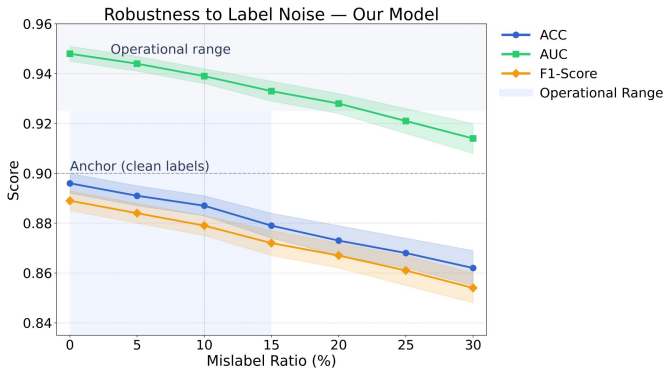

Figure 3. Robustness experiments on label noise levels

From the results in Figure 3, it can be seen that as the proportion of label noise increases, the model's performance in ACC, AUC, and F1-Score decreases. However, the overall decline is relatively gradual, indicating that the proposed multimodal fusion method maintains strong robustness to some extent. Within the “operational range” of 0%–15%, the AUC remains above 0.94, and both ACC and F1 stay at relatively high levels, showing that the model can retain good predictive performance under mild label errors.

In the low-noise range (0%–10%), the performance drop is small, especially for the AUC curve, which decreases most slowly. This suggests that the model's discriminative ability in distinguishing different outcome categories is less affected. This result is related to the multimodal feature fusion strategy, where the complementary information between waveform and clinical data helps offset the interference caused by label errors, thus maintaining stable classification boundaries.

When the label noise proportion exceeds 15%, the decline in performance becomes more pronounced, with ACC and F1 showing larger drops. This indicates that in high-noise environments, the model's ability to discriminate between classes and classify samples accurately is more severely impacted. The main reason is that label errors directly guide the model to learn decision boundaries that deviate from the true distribution in the feature space, reducing the effectiveness of multimodal information fusion.

Overall, the proposed method can maintain stable performance under low to moderate levels of label noise, which is important in ICU scenarios where label errors are difficult to completely avoid in real clinical data. Although performance decreases under high noise, the method still maintains a relatively leading AUC performance, reflecting the advantages of the multimodal fusion architecture in information redundancy and feature complementarity, and providing a solid foundation for further improving robustness.

This paper presents a sensitivity analysis of sequence length and time window sliding step size, and the experimental results are shown in Figure 4.

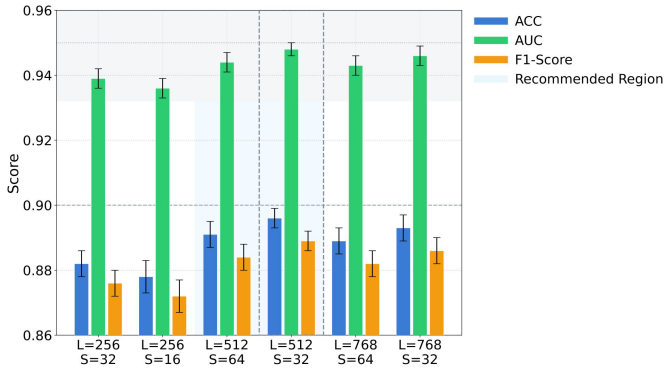

Figure 4. Sensitivity analysis of sequence length and time window sliding step

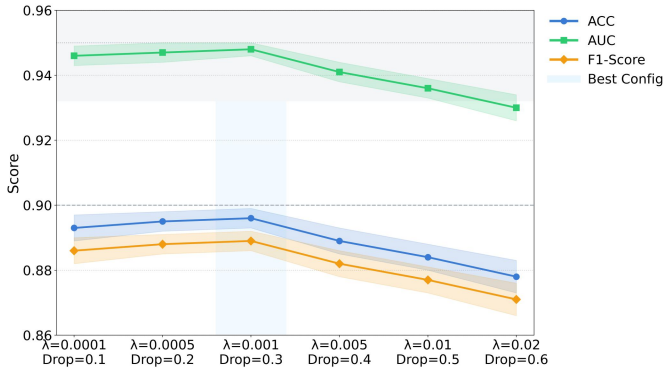

Figure 5. Robustness evaluation of the regularization coefficient and dropout ratio

From the results in Figure 4, it can be seen that different sequence lengths and sliding step sizes of the time window have a clear impact on model performance. In general, moderate sequence lengths and step sizes achieve more balanced results across the three metrics, while overly short or long sequences lead to decreases in ACC, AUC, and F1-Score. This indicates that in multimodal ICU outcome prediction tasks, the time coverage of the input sequence needs to balance capturing sufficient temporal information and avoiding redundant noise.

With a short sequence length (L=256), the model shows relatively low ACC and F1 performance. The main reason is that the time coverage of the sequence is limited, making it difficult to capture long-range dependencies in the progression of the patient's condition, which restricts feature representation capability. Although AUC remains at a relatively high level, the overall discriminative ability is slightly lower than the best configuration. This suggests that in this task, relying solely on short time segments is insufficient to support high-precision multimodal fusion prediction.

When the sequence length increases to a medium level (L=512) combined with a smaller sliding step size (S=32), all three metrics reach their best values, with AUC approaching 0.95. This configuration can fully utilize the temporal details of waveforms while sampling more densely along the time axis through the sliding window, enabling more comprehensive and fine-grained feature capture. It achieves a good balance between information richness and computational cost, which is beneficial for improving the model's generalization ability and stability.

When the sequence length further increases to L=768, AUC remains high, but ACC and F1 show slight declines. This may be due to the introduction of more redundant information and noise, which can dilute the discriminative power of key features. In addition, the larger period increases modeling complexity and computational cost, which may be less efficient for real-time prediction scenarios. Therefore, from the perspective of both predictive performance and computational efficiency, a medium sequence length combined with a moderate sliding step size is the optimal choice for this task.

Finally, this paper also gives a robustness evaluation of the regularization coefficient and dropout ratio, and the experimental results are shown in Figure 5.

From the results in Figure 5, it can be seen that changes in the regularization coefficient and dropout rate have a clear impact on the model's ACC, AUC, and F1-Score. Within a small regularization coefficient range ($\lambda \leq 0.001$) and a moderate dropout rate range (0.2–0.3), the model maintains high performance across all three metrics. In particular, when $\lambda = 0.001$ and dropout = 0.3, the model achieves its best performance. This indicates that moderate regularization and dropout can effectively suppress overfitting while preserving sufficient feature representation ability, thereby improving stability in ICU outcome prediction tasks.

When the regularization coefficient increases further ($\lambda \geq 0.005$) or the dropout rate exceeds 0.4, model performance declines significantly, especially in ACC and F1. This may be because strong regularization reduces the model's ability to learn important features, while an excessively high dropout rate leads to overly sparse feature representations, weakening the modeling of temporal dependencies. In multimodal data fusion scenarios, this loss is particularly pronounced during the joint modeling of waveform and clinical features.

The AUC curve remains generally higher than the other metrics and shows a smaller decline during parameter changes. This suggests that the model can maintain strong class discrimination ability under different regularization and dropout configurations. This stability is closely related to the redundancy of multimodal information and the complementarity between global and local feature capture in the Transformer-CNN architecture, which gives the model a certain level of robustness when adjusting regularization strategies.

Overall, setting the regularization coefficient and dropout rate appropriately is critical to enhancing the model's generalization ability and resistance to overfitting. For this task, a moderate parameter range can balance suppressing noise and preserving key information, ensuring high predictive accuracy while maintaining good stability and transferability for real clinical deployment.

4. Conclusion

This study proposes a multimodal feature fusion method that combines a Transformer and a CNN for ICU patient outcome prediction. The method effectively integrates the

temporal patterns of physiological waveforms with the global information from structured clinical data. Across multiple comparative experiments and sensitivity analyses, the model demonstrates excellent performance in metrics such as ACC, AUC, and F1-Score, and shows strong robustness under different hyperparameters, data distributions, and environmental conditions. These results verify the effectiveness of the proposed architecture in capturing complex multimodal relationships and provide a feasible technical solution for early risk identification and decision support in ICU scenarios.

From a methodological perspective, this study investigates the impact of key factors, including convolution kernel size, pooling strategy, sequence length, sliding step size, regularization coefficient, and dropout rate, on model performance. It also systematically evaluates the model under data perturbations such as label noise and sampling rate changes. The results show that reasonable structural and parameter configurations can improve the model's generalization ability and stability while maintaining high accuracy. This systematic analysis offers optimization insights for multimodal temporal modeling and establishes a reproducible experimental paradigm for similar clinical prediction tasks, promoting refined development in model design and deployment in related fields.

From an application perspective, the proposed method can provide real-time and stable prediction support in high-risk and high-complexity clinical environments such as the ICU. It can assist medical staff in identifying potential critical conditions in advance, thereby optimizing intervention strategies and resource allocation. Integrated with existing medical information systems, the model can be seamlessly connected to patient monitoring and electronic medical record systems to achieve a closed-loop process from data collection to risk warning. This not only improves the scientific basis and timeliness of clinical decisions but also helps reduce patient mortality and complication rates, offering broad prospects for promotion in public health management and smart healthcare development.

Future research can be expanded in several directions. One direction is to incorporate more types of modalities, such as imaging data, temporal laboratory test results, or genomic data, to further enrich the feature space and improve prediction accuracy and interpretability. Another direction is to explore more efficient model compression and inference acceleration techniques to meet the needs of real-time applications in edge computing or resource-constrained environments. In addition, the proposed method can be applied to other time-critical and multimodal-dependent domains, such as emergency triage, remote medical monitoring, and industrial process safety monitoring. This would extend the technical achievements of this study to broader application scenarios and further unlock its potential in critical task prediction and intelligent decision support.

In addition to waveform and structured data, future extensions can consider integrating medical images such as radiographs or CT scans. These imaging modalities provide complementary spatial and morphological cues that are often critical for ICU decision-making, and their fusion with temporal and structured features may further improve robustness and interpretability of outcome prediction models.

References

[1] Zhang J, Contreras M, Bandyopadhyay S, et al. Mango: Multimodal acuity transformer for intelligent icu outcomes[J]. arXiv preprint arXiv:2412.17832, 2024.

[2] Yang Z, Mitra A, Liu W, et al. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records[J]. Nature communications, 2023, 14(1): 7857.

[3] Xu Y, Xu S, Ramprassad M, et al. TransEHR: self-supervised transformer for clinical time series data[C]//Machine Learning for Health (ML4H). PMLR, 2023: 623-635.

[4] Wang C, Yang X, Sun M, et al. Multimodal fusion network for ICU patient outcome prediction[J]. Neural Networks, 2024, 180: 106672.

[5] Jeanselme V, Agarwal N, Wang C. Review of language models for survival analysis[C]//AAAI 2024 Spring Symposium on Clinical Foundation Models. 2024.

[6] Lyu W, Dong X, Wong R, et al. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction[C]//AMIA Annual Symposium Proceedings. 2023, 2022: 719.

[7] Karami H, Atienza D, Ionescu A. Tee4ehr: Transformer event encoder for better representation learning in electronic health records[J]. Artificial Intelligence in Medicine, 2024, 154: 102903.

[8] Zhao Y, Hong Q, Zhang X, et al. Bertsurv: Bert-based survival models for predicting outcomes of trauma patients[J]. arXiv preprint arXiv:2103.10928, 2021.

[9] Li X, Gu J, Wang Z, et al. XAI for In-hospital Mortality Prediction via Multimodal ICU Data[J]. arXiv preprint arXiv:2312.17624, 2023.

[10] Mesinovic M, Watkinson P, Zhu T. XMI-ICU: Explainable Machine Learning Model for Pseudo-Dynamic Prediction of Mortality in the ICU for Heart Attack Patients[J]. arXiv preprint arXiv:2305.06109, 2023.

[11] Caicedo-Torres W, Gutierrez J. ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU[J]. Journal of biomedical informatics, 2019, 98: 103269.