

Early Warning of Intraoperative Adverse Events via Transformer-Driven Multi-Label Learning

Xueyao Wang^{1,2}, Xiuding Cai^{1,2}, Honglin Shang^{1,2}, Yaoyao Zhu³, Yu Yao^{1,2*}

¹Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610213, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³China Zhenhua Research Institute Co., Ltd., Guiyang 550014, China

{wangxueyao221, caixiuding20, shanghonglin24, zhuyaoyao19}@mails.ucas.ac.cn, casitmed2022@163.com

Abstract

Early warning of intraoperative adverse events plays a vital role in reducing surgical risk and improving patient safety. While deep learning has shown promise in predicting the single adverse event, several key challenges remain: overlooking adverse event dependencies, underutilizing heterogeneous clinical data, and suffering from the class imbalance inherent in medical datasets. To address these issues, we construct the first Multi-label Adverse Events dataset (MuAE) for intraoperative adverse events prediction, covering six critical events. Next, we propose a novel Transformer-based multi-label learning framework (IAENet) that combines an improved Time-Aware Feature-wise Linear Modulation (TAFiLM) module for static covariates and dynamic variables robust fusion and complex temporal dependencies modeling. Furthermore, we introduce a Label-Constrained Reweighting Loss (LCRLoss) with co-occurrence regularization to effectively mitigate intra-event imbalance and enforce structured consistency among frequently co-occurring events. Extensive experiments demonstrate that IAENet consistently outperforms strong baselines on 5, 10, and 15-minute early warning tasks, achieving improvements of +5.05%, +2.82%, and +7.57% on average F1 score. These results highlight the potential of IAENet for supporting intelligent intraoperative decision-making in clinical practice.

Introduction

Surgery is a cornerstone of modern healthcare, with over 300 million procedures performed annually worldwide (Nepogodiev et al. 2019). However, it carries a disproportionately high risk of harm: 46–65% of all medical adverse events are surgery-related, and 3–22% of surgical patients experience complications (Meara et al. 2015). Intraoperative adverse events are common, when prolonged, and are associated with severe complications such as pulmonary dysfunction, acute kidney injury, and cardiovascular events. These outcomes contribute to higher mortality and worse long-term prognosis (Varghese et al. 2024). Importantly, most are preventable with timely intervention, underscoring the urgent need for accurate risk early warning systems.

Advanced predictive models leveraging Artificial Intelligence (AI) could improve perioperative care by identify-

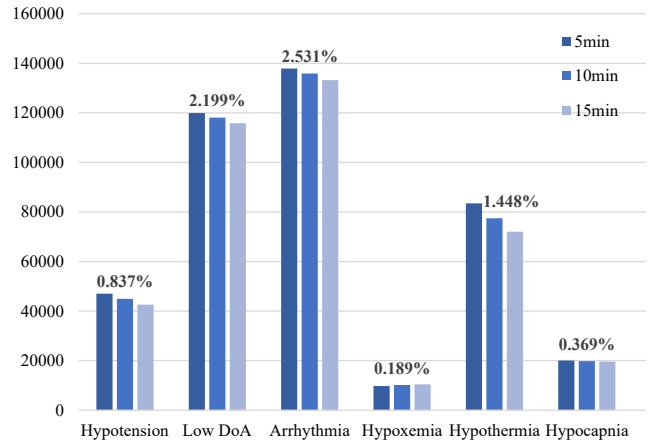


Figure 1: Positive sample distribution by adverse event for 5, 10, and 15 minutes prediction windows in MuAE. Percentages indicate the event occurrence rate among total samples.

ing high-risk patients, enabling early interventions, and ultimately reducing preventable harm (Cai et al. 2024, 2025). Currently, the mainstream approach to perioperative risk modeling focuses on single adverse event prediction using machine learning and deep learning, including tasks like intraoperative Hypotension (IOH) (Chen et al. 2021; Lu et al. 2023; Moon et al. 2024), Hypoxemia (Lundberg et al. 2018; Park et al. 2023), inadequate and excessive Depth of Anesthesia (DoA) prediction (Lee et al. 2018, 2019). Lee et al. used a 1D Convolutional Neural Network for regression and classification tasks to predict intraoperative hypotension events using invasive and non-invasive arterial pressure, thereby reducing postoperative organ dysfunction risks. Chen et al. merely extended this paradigm to multiple events by training separate models for each outcome, albeit with shared self-supervised features. Prevailing single-event approaches overlook the intrinsic dependencies between adverse events. While modeling multiple events jointly can improve individual predictions by leveraging their mutual relationships, existing methods fail to explicitly model the dependencies within the time-series modality. Moreover, the paradigm shift still faces the key challenge: the intra-event imbalance from just 0.189% to 2.531% of the total sam-

*Corresponding author

ples (see **Fig. 1**), severely hinders model generalization. Despite its clinical relevance, multi-adverse-event prediction remains largely unexplored in perioperative risk assessment.

Besides, recent studies have increasingly focused on integrating static covariates and dynamic variables (Schnider et al. 2021; Bahador et al. 2021; Lu et al. 2023; Moon et al. 2024) through various fusion strategies. For instance, Moon et al. integrated dynamic features extracted from both time and frequency domains with static variables via simple concatenation for downstream classification. Meanwhile, Lu et al. fused dynamic variables with static counterparts into four distinct modalities, employing attention mechanisms for long-range interaction in hypotension prediction. However, such direct fusion strategies often introduce feature redundancy and noise, thereby increasing the complexity of representation learning and hindering model performance.

To bridge existing gaps, we construct the first intraoperative **Multiple Adverse Event** (MuAE) dataset based on the public VitalDB dataset to address the research gap in early warning of multi-adverse events. Then, we propose an **Intraoperative Adverse Events Network** (IAENet) for multi-label time-series classification. It deploys a **Time-Aware Feature-wise Linear Modulation** (TAFiLM) module on the transformer encoder to dynamically modulate static covariates and dynamic variables for better feature fusion to reduce redundant noise. The transformer encoder is used to extract the temporal feature for the classification task, where the inverted embedding of variables as input tokens inherently captures multivariate correlations. Additionally, we also design a **Label-Constrained Reweighting Loss** (LCRLoss), which dynamically reweights prediction outputs based on batch-wise label frequency and incorporates a co-occurrence regularization term to model structured label dependencies, mitigating intra-event imbalance and improving model robustness and generalization. To summarize, our main contributions are the following:

- We construct the first multi-label dataset for early warning of intraoperative adverse events, the MuAE dataset, comprising six adverse event types derived from the VitalDB dataset.
- We propose a transformer-based model for intraoperative multi-adverse events classification, named IAENet. It employs a Time-Aware Feature-wise Linear Modulation (TAFiLM) module for modulation fusion of static covariates and dynamic vital signs to enhance feature quality and reduce redundant noise.
- We design a Label Constraint Reweighting Loss function (LCRLoss) to effectively mitigate the intra-event imbalance problem and capture structured label dependencies.
- Extensive experiments and ablation studies demonstrate the effectiveness of the proposed framework.

Related Works

Intraoperative Single-event Early Warning. AI-driven early warning for intraoperative adverse events enables earlier clinical intervention, enhancing patient safety and improving surgical outcomes (Cai et al. 2024). Existing research (Lee et al. 2018; Lundberg et al. 2018; Hwang et al.

2023; Park et al. 2023; Lu et al. 2023; Moon et al. 2024) predominantly focuses on single-event prediction through time-series forecasting or classification. For instance, Hwang et al. predicted the event solely from arterial blood pressure waveforms, employing discrete wavelet transforms and CNN-based feature extraction. Lundberg et al. used ensemble models with SpO₂-derived features for hypoxemia prediction, incorporating risk factor analysis for interpretability. Such isolated event prediction fails to capture perioperative complexity comprehensively and neglects clinically significant interdependencies between co-occurring adverse events. To address the current gap in multi-adverse-event prediction research, we propose the first Transformer-based multi-label learning model that explicitly captures inter-event dependencies to enhance predictive performance.

Class Imbalance in Multi-label Learning. Medical datasets typically suffer from class imbalance, where sustained normal states outweigh brief adverse events. In multi-label settings, the imbalance is more complex due to interdependent label distributions. For example, bradycardia often co-occurs with hypotension (Cheung et al. 2015), highlighting intra-event interaction in multi-label intraoperative prediction. To mitigate this, prior work explores cost-sensitive learning (Lin et al. 2017; Cao et al. 2019) and label-aware resampling (Cui et al. 2019). However, synthetic signals like SMOTE (Chawla et al. 2002) may distort physiological patterns, while static loss weights like Asymmetric loss (Ridnik et al. 2021) fail to adapt to temporal label dynamics or capture co-occurrence. We address these issues with a label-constrained reweighting loss that models inter-label dependencies and adapts to distributional shifts during training.

Medical Time Series Modeling. Medical time-series data from heterogeneous monitoring devices often exhibit misalignment, noise susceptibility, and multiple variables. Deep learning models have been employed to capture the temporal dependencies and inter-variable correlations in time series analysis (Wang et al. 2024). With the inclusion of different covariates and vital signs data, multivariate data present challenges in effective integration. Recent models like iTransformer (Liu et al. 2024) leverage attention mechanisms to model temporal and variable-level dependencies jointly. However, many existing methods concatenate static variables (expanded over time steps) with dynamic inputs directly, which introduces redundancy and hampers model efficiency. So we improve the FiLM module (Perez et al. 2018) based on the transformer to perform dynamic modulation of vital signs series and static variables for better early feature fusion.

Dataset

Due to the lack of publicly available datasets on intraoperative multiple adverse events, we develop the MuAE dataset by processing the VitalDB dataset (Lee et al. 2022) through rigorous data cleaning and preprocessing. VitalDB contains anesthesia records from 6,388 non-cardiac surgeries between August 2016 and June 2017 at Seoul National University Hospital, encompassing comprehensive perioperative physiological signals and anesthetic parameters.

Dataset Cleaning

Case selection was performed first. Inclusion and exclusion criteria were used to ensure data quality through selection criteria guided by clinical experts. Cases were selected for surgical durations of more than 2 hours to ensure adequate monitoring data under general anesthesia. Paediatric cases were excluded if they were aged less than 18 years and weighed less than 35 kg. Patients with an ASA classification higher than grade 6 were also excluded.

Then, vital signs and static features were selected from the intraoperative monitoring data. 15 key physiological dynamic variables are filtered from raw monitoring characteristics, such as drug infusion parameters (PPF20_VOL, RFTN20_VOL, PPF20_CE, and RFTN20_CE), neuromonitoring (BIS), blood pressure (ART_DBP, ART_MBP, ART_SBP), temperature (BT), heart rate (HR), oxygen saturation (PLETH_SPO2) and ECG (ECG_II) can be filtered. Patient's age, weight, height, gender, and ASA classification were selected as 5 static covariates.

Finally, the 15 vital signs variables were resampled at 2-second intervals. Negative or blank values were set to zero and treated as missing data. Missing values were filled using forward and backward interpolation to minimize information loss. For anesthetic drug infusion parameters, we converted the drug unit volume to a rate for indirect use. The final selection of 873 patients was made after data quality control. We spliced 15 dynamic variables \mathbf{x}_d and 5 static covariates \mathbf{x}_s as network inputs \mathbf{x} :

$$\mathbf{x} = \left(\underbrace{\mathbf{x}_{d_0}, \dots, \mathbf{x}_{d_{14}}}_{\text{Dynamic variables}}, \underbrace{\mathbf{x}_{s_0}, \dots, \mathbf{x}_{s_4}}_{\text{Static covariates}} \right) \in \mathbb{R}^{W \times (D+S)}, \quad (1)$$

where W represents the input window size and D, S represent the number of dynamic variables and static covariates.

Event Label Definition

Six major intraoperative adverse events were defined based on clinical guidelines and prior studies (Chen et al. 2021; Lu et al. 2023; Moon et al. 2024): hypotension, low depth of anesthesia, arrhythmia, hypoxemia, hypothermia, and hypocapnia. Event occurrence $\mathbf{y}_{t+\Delta:t+\Delta+T}^i \in \{0, 1\}^6$ was determined when monitored parameters exceeded predefined thresholds:

- Hypotension: MAP < 65 mmHg for at least 1 minute.

$$\max(\mathbf{x}_{\text{MAP}}[t + \Delta : t + \Delta + 30]) < 65. \quad (2)$$

- Low depth of anesthesia: BIS < 40 for at least 1 minute.

$$\max(\mathbf{x}_{\text{BIS}}[t + \Delta : t + \Delta + 30]) < 40. \quad (3)$$

- Arrhythmia: HR < 60 bpm or HR > 100 bpm for at least 1 minute.

$$\begin{aligned} & (\max(\mathbf{x}_{\text{HR}}[t + \Delta : t + \Delta + 30]) < 60) \\ & \vee (\min(\mathbf{x}_{\text{HR}}[t + \Delta : t + \Delta + 30]) > 100). \end{aligned} \quad (4)$$

- Hypoxemia: SpO₂ < 90% for at least 1 minute.

$$\max(\mathbf{x}_{\text{SpO}_2}[t + \Delta : t + \Delta + 30]) < 90. \quad (5)$$

- Hypothermia: BT < 35°C for at least 1 minute.

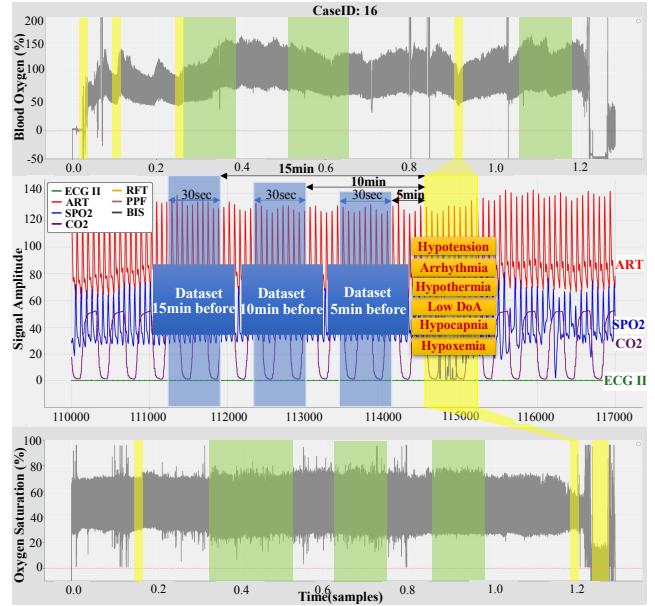


Figure 2: The pipeline of Multi-adverse Events Early Warning

$$\max(\mathbf{x}_{\text{BT}}[t + \Delta : t + \Delta + 30]) < 35. \quad (6)$$

- Hypocapnia: EtCO₂ < 30 mmHg for at least 1 minute.

$$\max(\mathbf{x}_{\text{EtCO}_2}[t + \Delta : t + \Delta + 30]) < 30. \quad (7)$$

Given a continuous Δ -minute sequence of physiological signals preceding time point t , we train a multi-label classification model f_θ to predict whether any of six representative intraoperative adverse events will occur during the subsequent T -minute interval. We define the predicted outcome $\hat{\mathbf{y}}_i = \hat{\mathbf{y}}_{t+\Delta:t+\Delta+T}^i$ as the output of the network function:

$$\hat{\mathbf{y}}_{t+\Delta:t+\Delta+T}^i = f_\theta(\mathbf{x}[t - W : t]), \quad (8)$$

where Δ represents the advance prediction time, T represents the predicted event window length, i represents the adverse events, and t represents the current timespoint.

Methods

Data Preprocessing

The multivariate physiological data were normalized to eliminate scale differences caused by inconsistent feature units. The data were then segmented into overlapping time windows of 30 seconds in length, using a 2-second sliding step. Approximately 4.88 million, 5.66 million, and 5.55 million samples were generated under the 5, 10, and 15 minutes ahead prediction settings, respectively.

The **Fig. 2** shows the pipeline of multi-adverse events early warning. Physiological data segments are labeled as normal (green) or abnormal (yellow). Input data are extracted from 30-second intervals preceding the occurrence of abnormal events (blue) within 5/10/15-minute windows in advance. The output data is presented using binary labels. The top and bottom panels demonstrate label generation for

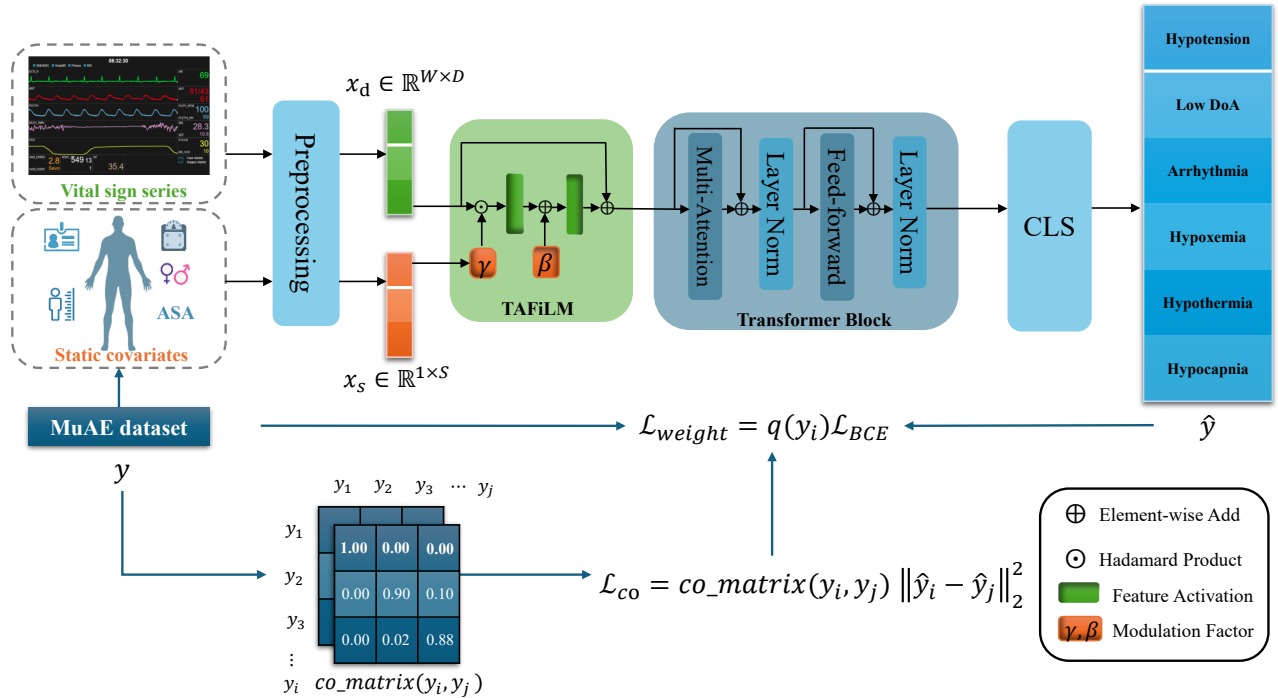


Figure 3: An overview of the IAENet framework for time-series in multi-label classification. Given an sample about vital sign series and static covariate in MuAE dataset $\mathbf{x} = \{\mathbf{x}_{d_0}, \dots, \mathbf{x}_{d_{14}}, \mathbf{x}_{s_0}, \dots, \mathbf{x}_{s_4}\}$, our goal is to predict whether the values in the following Δ time steps will be normal or abnormal $\mathbf{y} = \{y_1, y_2, \dots, y_6\}$. Firstly, the preprocessed dynamic variables and static covariates are respectively fused through the TAFiLM module for feature early fusion. Then, a Transformer encoder is employed to capture temporal correlations among multivariate variables. Finally, the model is trained using the proposed LCRLoss, which combines a batch-wise label frequency-weighted BCE loss with a co-occurrence constraint term based on label dependencies.

hypotension (blood pressure) and hypoxemia (oxygen saturation) events, highlighting distinct patterns between hemodynamic and respiratory instability.

Time Series Classification Model

The overview of the proposed IAENet is shown in **Fig. 3**. It integrates a Transformer encoder for multivariate temporal modeling and a TAFiLM module to fuse static covariates with dynamic vital signs through conditional modulation. The input data \mathbf{x} is derived from the MuAE dataset. After preprocessing, the features are first passed through the TAFiLM module for early fusion. The fused representations are then fed into the Transformer encoder to capture multivariate temporal dependencies. During training, the model adopts a label-constrained reweighting loss as the core optimization strategy.

TAFiLM. Inspired by the Feature-wise Linear Modulation (FiLM) module (Perez et al. 2018), we propose the Time-Aware FiLM (TAFiLM) module for time series data. The FiLM module enables conditional feature modulation through learnable affine transformations, allowing dynamic adjustment of feature representations based on specific contextual conditions. While FiLM has demonstrated strong performance in multimodal tasks, its application to time se-

ries data remains underexplored.

To achieve effective feature fusion, the static conditional features $\mathbf{x}_s \in \mathbb{R}^{B \times S}$ are processed by a conditional network to generate time-varying scaling and shifting factors $\gamma_s, \beta_s \in \mathbb{R}^{B \times W \times D}$. The parameters dynamically modulate the dynamic features through affine transformation within the TAFiLM module, formulated as:

$$\text{TAFiLM}(\mathbf{x}_d) = ((\gamma_s + 1) \odot \mathbf{x}_d + \beta_s) \in \mathbb{R}^{B \times W \times D}, \quad (9)$$

where \odot represents the hadamard product, $\mathbf{x}_d \in \mathbb{R}^{B \times W \times D}$ represents the dynamic feature, and the conditional network chooses the MLP layer.

Transformer Block. Following in time-series tasks (Liu et al. 2024), we reverse the input sequence before token embedding, addressing the timestamp misalignment issues in standard Transformer architectures. After embedding, the tokens are applied via a multi-head self-attention mechanism in the encoder-only transformer, capturing the dynamic interactions of the variables. To ensure variable independence, each variable is individually normalized by the LayerNorm. Finally, they are processed individually using a feed-forward network to create a sequential representation for the downstream classification task.

LCRLoss Definition. To mitigate label imbalance, we propose a batch-wise reweighting strategy that scales loss weights inversely to label frequencies. For batch b with N samples, the weighted BCE loss is:

$$\mathcal{L}_{weight} = \frac{1}{C} \sum_{c=1}^C \left(q_{pos}^b \sum_{i \in \mathcal{P}_c} \mathcal{L}_{BCE} + q_{neg}^b \sum_{i \in \mathcal{N}_c} \mathcal{L}_{BCE} \right), \quad (10)$$

where \mathcal{P}_c and \mathcal{N}_c denote positive and negative samples for class C , and weights q_{pos}^b, q_{neg}^b are updated per batch. Standard BCE loss is used for validation and testing.

Among the evaluated frequency-based reweighting strategies, the square root inverse scheme is adopted due to its overall stability, as demonstrated in the ablation study.

$$q(\mathbf{y}_i) = \begin{cases} \frac{1}{\sqrt{\mathcal{P}_c/N}} & \text{if } \mathbf{y}_i = 1, \\ \frac{1}{\sqrt{\mathcal{N}_c/N}} & \text{otherwise.} \end{cases} \quad (11)$$

To overcome the independence assumption in per-class reweighting, we introduce a co-occurrence loss \mathcal{L}_{co} that encourages consistency among labels that frequently co-occur.

Firstly, we compute a label co-occurrence matrix \mathbf{M} from the entire training set, capturing the correlations among adverse events. Assume the training set contains N samples, which are divided into B batches. The b -th batch contains n_b samples. The computation is formally expressed as:

$$\mathbf{M} = \sum_{b=1}^B (\mathbf{y}^{(b)})^\top \mathbf{y}^{(b)}, \mathbf{M} \in \mathbb{R}^{C \times C}. \quad (12)$$

Secondly, the aggregated co-occurrence matrix *co_matrix* is further normalized by its maximum value to ensure numerical stability and mitigate scale discrepancies:

$$co_matrix = \frac{\mathbf{M}}{\max_{i,j} \mathbf{M}_{ij} + \epsilon}, \quad (13)$$

where ϵ is a small constant added to prevent division by zero. The co-occurrence matrix \mathbf{M} is symmetric, i.e., $\mathbf{M}_{ij} = \mathbf{M}_{ji}$.

To enforce consistency among frequently co-occurring labels, we then compute the squared L2 distance between all label prediction pairs $(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j)$, weighted by the corresponding normalized co-occurrence value *co_matrix*($\mathbf{y}_i, \mathbf{y}_j$). The co-occurrence regularization loss, \mathcal{L}_{co} is defined as:

$$\mathcal{L}_{co} = \sum_{i,j}^C co_matrix(\mathbf{y}_i, \mathbf{y}_j) \|\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j\|_2^2, \quad (14)$$

where C is the total number of labels, $\hat{\mathbf{y}}_i$ and $\hat{\mathbf{y}}_j$ are the prediction of the model (e.g., logits or probabilities) for label \mathbf{y}_i and \mathbf{y}_j , respectively.

Finally, the overall loss \mathcal{L}_{LCR} is composed of two components, balanced by a weighting coefficient λ :

$$\mathcal{L}_{LCR} = \mathcal{L}_{weight} + \lambda \mathcal{L}_{co}. \quad (15)$$

Experiments

Baseline Methods. To assess the classification capabilities of the compared methods, we evaluate the IAENet with MLP-based model DLinear (Zeng et al. 2023); RNN-based model SegRNN (Lin et al. 2023); Transformer-based models iTransformer (Liu et al. 2024), PatchTST (Nie et al. 2023), FEDformer (Zhou et al. 2022), Autoformer (Wu et al. 2021), Informer (Zhou et al. 2021), Temporal Fusion Transformer (TFT) (Yèche et al. 2023), Crossformer (Zhang and Yan 2023), and Non-stationary Transformer (Liu et al. 2022).

Implementation Details. IAENet was developed primarily using the Time-Series-Library framework (Wu et al. 2023). The samples were divided according to patients into training (70%), validation (10%), and test (20%) sets for model validation. We used an RAdam optimizer with an initial learning rate of $1e^{-3}$. The optimal value of λ determined by grid search was set to 0.02, where the search range is [0.001, 0.01, 0.02, 0.05, 0.1, 0.2]. The batch size was set to 64 for 10 epochs with an early stopping patience of 3 epochs. The IAENet utilized the proposed LCRLoss, while all other models used the standard BCE Loss.

Evaluation Metrics. To evaluate the performance of the IAENet in the multi-label classification task, we employed six commonly used metrics: Micro F1 (F1), Macro AUC (AUC), Micro Precision (PRE), Micro Recall (REC), Micro Accuracy (ACC), and Hamming Loss (HM). Among them, F1 and AUC serve as the primary evaluation indicators, as they jointly reflect both prediction accuracy and completeness, and maintain robustness under class imbalance.

Experimental Results

Performance Comparison. To evaluate the effectiveness of IAENet, we designed experiments inspired by previous studies (Lee et al. 2021; Yang et al. 2024; Moon et al. 2024), conducting multi-adverse event prediction tasks at 5, 10, and 15 minutes in advance. Among six adverse event classifications, the IAENet consistently outperforms all baseline models in F1 and AUC, as shown in **Table 1**. Compared to the competitive Crossformer and iTransformer models, IAENet achieves average F1 score improvements of 5.05%, 2.82%, and 7.57% across the three classification tasks, respectively. Performance declines as input time intervals increase, indicating that the windows near event onset contain more discriminative features for accurate classification. Additionally, certain adverse events show consistently poor performance across all baselines, likely due to severe label imbalance.

Comparison with Different Loss. To evaluate the effectiveness of the LCRLoss, we compared it against several representative loss functions designed for imbalanced data, including weighted binary cross-entropy (WBCE), binary cross-entropy (BCE), Focal Loss (FL), Polynomial Loss (PolyLoss), Asymmetric Loss (ASL), and SoftMarginLoss (SMLoss). The comparative results are presented in **Table 2**. The results show that LCRLoss achieves the best performance among all compared loss functions. Compared to the most competitive alternative ASL, it improves average F1

Models	IAENet		iTransformer		Crossformer		PatchTST		FEDformer		Autoformer		Informer		DLinear		Non-stationary		TFT		SegRNN		
	Δ	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Hypotension	5	51.45	74.83	45.64	65.76	<u>48.32</u>	67.29	24.30	57.58	45.90	66.69	3.02	50.69	47.65	67.96	47.65	<u>67.69</u>	48.37	68.66	36.71	62.59	29.23	59.19
	10	36.21	65.85	16.91	54.79	19.14	55.50	10.58	52.80	11.59	53.11	1.07	50.10	<u>25.16</u>	<u>58.35</u>	17.81	55.20	23.10	57.40	19.09	55.54	14.47	54.55
	15	28.54	62.08	9.79	52.57	15.24	54.28	5.45	51.29	10.13	52.67	0.03	49.95	11.17	53.02	8.01	52.03	16.10	54.84	<u>16.81</u>	<u>55.31</u>	10.17	52.68
Low DoA	5	61.45	76.69	54.03	69.49	37.70	66.77	2.28	50.43	51.42	68.08	6.11	50.31	49.53	67.19	49.53	67.19	47.92	66.19	6.18	51.06	<u>54.65</u>	<u>70.35</u>
	10	56.54	73.42	<u>49.80</u>	<u>67.35</u>	43.18	63.80	0.95	50.17	43.85	64.14	11.62	50.57	37.89	61.17	43.77	64.09	31.35	58.56	7.08	51.21	43.59	64.01
	15	50.21	68.16	28.95	57.81	39.50	62.00	0.96	50.16	30.91	58.48	14.71	50.73	27.19	56.53	28.87	57.85	30.18	57.74	8.90	50.72	<u>32.34</u>	<u>59.07</u>
Arrhythmia	5	65.04	77.40	62.76	75.19	<u>63.68</u>	<u>76.05</u>	5.13	51.02	56.29	70.63	3.79	50.00	52.35	68.47	52.35	68.47	59.93	73.43	26.23	56.17	60.82	73.76
	10	59.12	73.77	<u>57.91</u>	<u>72.28</u>	56.22	70.72	5.76	51.06	55.15	70.08	21.97	51.76	51.06	67.86	0.67	50.12	47.84	66.01	15.77	52.60	57.28	72.00
	15	54.79	71.04	48.36	66.28	<u>53.47</u>	<u>69.86</u>	3.01	50.52	48.99	66.58	2.08	50.19	41.08	62.51	0.57	50.11	41.94	62.92	15.71	51.79	51.78	68.37
Hypoxemia	5	57.79	88.25	<u>56.34</u>	75.68	49.65	71.57	55.18	77.22	51.81	73.41	1.98	50.44	55.88	<u>77.95</u>	49.83	69.86	51.90	74.62	56.17	76.80	46.51	69.69
	10	51.56	84.77	<u>45.58</u>	<u>68.70</u>	43.61	66.96	36.95	63.43	34.39	61.67	3.88	50.99	39.25	66.17	36.56	63.02	42.59	69.55	35.98	62.92	33.19	61.28
	15	43.65	81.37	17.07	55.08	30.49	61.14	15.66	54.55	24.60	58.15	0.00	50.00	24.01	59.03	17.19	55.08	19.66	56.45	<u>28.79</u>	<u>63.36</u>	21.02	56.75
Hypothermia	5	84.64	92.77	72.40	82.69	<u>71.22</u>	<u>84.67</u>	34.97	61.02	75.19	85.12	3.06	50.59	70.62	81.79	70.62	81.80	67.19	81.84	42.41	64.63	77.19	86.17
	10	75.59	86.38	<u>73.64</u>	<u>86.19</u>	60.08	73.05	31.20	59.31	59.67	74.01	6.64	51.00	54.17	72.79	47.64	66.55	51.95	70.07	35.97	61.37	65.28	79.05
	15	59.93	77.00	37.96	62.37	45.13	65.48	22.95	56.27	42.61	64.52	1.50	50.25	41.47	65.72	33.03	60.13	36.28	62.43	30.87	59.37	<u>51.36</u>	<u>71.18</u>
Hypocapnia	5	42.29	75.31	36.49	62.21	<u>37.95</u>	<u>63.04</u>	19.26	55.83	28.80	59.03	1.96	50.39	37.98	65.84	37.98	65.84	35.87	64.31	2.34	50.59	30.51	60.96
	10	28.17	62.14	7.80	52.04	5.22	51.32	2.33	50.57	1.43	50.34	0.56	50.09	6.76	51.74	10.25	52.74	<u>13.52</u>	<u>54.13</u>	7.47	51.95	4.42	51.11
	15	25.89	63.04	1.29	50.29	2.02	50.47	0.36	50.08	6.19	51.58	0.10	50.01	8.13	52.16	2.48	50.58	11.31	53.21	<u>11.27</u>	<u>53.17</u>	3.21	50.79
Mean	5	65.36	80.86	60.06	<u>71.84</u>	<u>60.31</u>	71.56	16.24	58.85	57.51	70.49	4.24	50.41	54.76	71.54	41.03	64.57	56.21	71.51	26.30	60.31	59.07	70.02
	10	58.51	74.39	<u>55.69</u>	<u>66.89</u>	48.94	63.58	12.40	54.56	48.23	62.22	13.66	50.75	44.60	63.02	30.41	58.62	41.01	62.62	19.04	55.93	51.00	63.58
	15	50.52	70.45	35.74	57.40	<u>42.95</u>	<u>60.45</u>	7.67	52.14	37.67	58.66	6.62	50.19	33.64	58.16	19.00	54.30	33.64	57.93	17.72	55.62	41.82	59.81

Table 1. Classification Accuracy results across all tasks and methods. Δ denotes the forecasting horizon time steps. All metrics are in %. The best results are in **bold** font and the second best are underlined.

Loss	ACC	PRE	REC	F1	AUC	HM
LCRLoss	91.72	61.77	69.39	65.36	80.88	0.0828
WBCE	92.17	68.22	56.99	62.10	71.12	0.0783
BCE	92.69	71.70	57.83	64.03	74.21	0.0731
FL	92.73	73.76	54.91	62.96	73.75	0.0727
ASL	91.22	58.98	72.08	<u>64.87</u>	<u>79.86</u>	0.0878
PolyLoss	92.66	70.90	58.91	<u>64.35</u>	<u>75.07</u>	0.0734
SMLoss	92.59	71.96	55.93	62.94	73.65	0.0741

Table 2. Performance of IAENet under different losses.

score by 0.49% and AUC by 1.02%, demonstrating its effectiveness in handling intra-label imbalance.

Ablation Analysis. To evaluate the effectiveness of TAFiLM and LCRLoss, we integrated them into several baseline models. In the baselines, static covariates and dynamic variables are directly concatenated as input and trained using the BCE loss. As shown in **Table 3**, adding only FiLM, TAFiLM, and LCRLoss improves the average F1 and AUC across baselines, indicating that the components act as effective plug-and-play modules. Notably, TAFiLM outperforms FiLM, indicating its superior ability to modulate dynamic features through static inputs while miti-

gating noise from direct concatenation. TAFiLM slightly reduces precision but significantly improves recall, which is important in medical diagnosis to minimize underdiagnosis.

We further conducted a comprehensive ablation study to assess the impact of different reweighting strategies, batch size, and the coefficient λ on model performance, as measured by F1 and AUC (see **Table 4**). We first evaluated several reweighting schemes in LCRLoss, including inverse, log inverse, square root inverse, and cube root inverse on global frequency and batch-wise local frequency. The batch-wise frequency-based *sqrt.inverse* strategy achieves the highest F1, indicating a strong balance between precision and recall. The inverse and log inverse strategies achieve high AUC but suffer from low F1, suggesting overfitting or instability. Overall, batch-wise label frequency-based reweighting consistently outperforms global static reweighting, indicating the importance of incorporating local data distributions under imbalanced label distributions. Based on the trade-off, the *sqrt.inverse* is selected as the default.

Considering that the representativeness based on batch statistics may be sensitive to the batch size, we then computed a co-occurrence matrix from training data to obtain stable estimates of adverse event correlations. Results demonstrate that global co-occurrence statistics computed on the full training set more accurately reflect event corre-

Module	ACC	PRE	REC	F1	AUC	HM
Informer	91.28	65.74	46.92	54.76	71.54	0.0872
+ LCRLoss	90.24	55.87	62.89	59.17	79.36	0.0976
+ FiLM	92.05	70.10	51.14	<u>59.13</u>	70.62	0.0795
+ TAFiLM	91.95	67.89	53.96	60.13	<u>71.34</u>	0.0805
Crossformer	92.39	73.03	51.36	<u>60.31</u>	71.56	0.0761
+ LCRLoss	90.80	57.88	66.80	62.02	81.17	0.0920
+ FiLM	92.21	71.11	51.36	59.99	71.46	0.0779
+ TAFiLM	92.49	72.45	53.63	61.63	<u>71.55</u>	0.0751
SegRNN	91.84	67.86	52.29	59.07	70.02	0.0816
+ LCRLoss	90.93	58.11	69.51	63.30	80.50	0.0907
+ FiLM	92.30	72.40	50.99	<u>59.83</u>	<u>70.74</u>	0.0770
+ TAFiLM	92.34	70.84	54.26	61.45	71.40	0.0766
iTransformer	92.19	70.76	52.16	60.06	71.84	0.0781
+ LCRLoss	91.02	59.48	63.31	61.34	77.99	0.0898
+ FiLM	92.46	72.90	52.55	<u>61.07</u>	71.86	0.0754
+ TAFiLM	92.69	71.70	57.83	64.03	74.21	0.0731

Table 3. Ablation studies on the components of IAENet.

Factor	Setting	F1	AUC
Weighting (G / L)	none	63.97	74.30
	inverse	40.55 / 60.82	78.80 / 85.35
	log_inverse	24.40 / 51.02	60.23 / 83.94
	sqrt_inverse	59.51 / 65.36	83.68 / 80.88
	cubic_inverse	63.91 / <u>64.26</u>	82.33 / 79.44
Batch Size	64	<u>65.21</u>	81.35
	128	64.64	<u>81.38</u>
	256	64.78	81.76
	All	65.36	80.88
λ	0.001	62.28	72.58
	0.01	64.89	75.06
	0.02	65.36	80.88
	0.05	63.33	74.41
	0.1	62.43	73.2

Table 4. Ablation studies on reweighting strategies, batch size, and λ of LCRLoss.

lations than small batches of estimates. Furthermore, we examined the effect of the hyperparameter λ , which balances the contributions of the BCE loss and the co-occurrence loss.

Gradient Analysis. To analyze the optimization properties of LCRLoss, we further conducted the loss gradients. The gradient curves of different loss functions are depicted in Fig. 4. While conventional losses (e.g., BCE, FL) exhibit vanishing gradients for high-confidence predictions, LCR-Loss maintains informative gradients, enabling effective optimization even in confident regions. Notably, LCR-Loss promotes similar logits for frequently co-occurring events. For example, arrhythmia shares consistent gradient trends with low DoA and hypothermia. In contrast to, hypoxemia and hypocapnia display more independent patterns. This aligns with the label co-occurrence matrix shown in Fig. 5, highlighting LCR-Loss’s capacity to capture inter-label dependencies and enhance recall in multi-label classification.

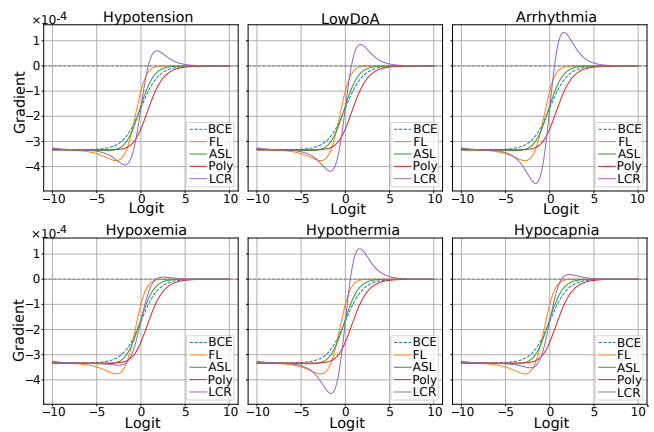


Figure 4: Derivatives of the loss functions. The X-axis denotes the logit of positive labels, and the Y-axis is the corresponding gradients.

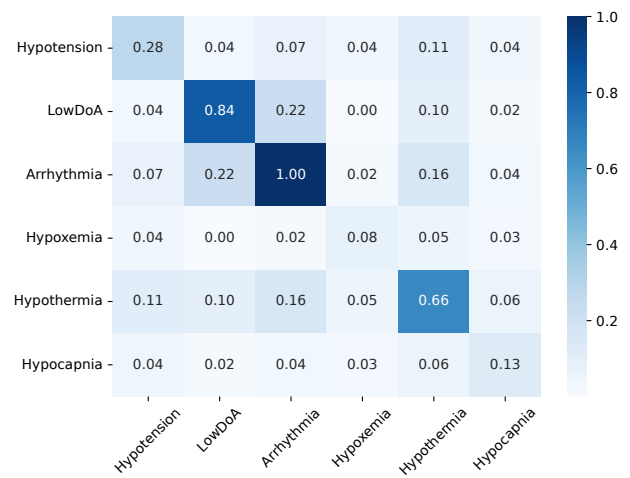


Figure 5: Co-occurrence matrix of adverse events

Conclusion

In this work, we propose a novel multi-label classification framework for early warning of adverse events based on the proposed MuAE dataset. By integrating static covariates and dynamic vital signs via the TAFiLM module and leveraging the Transformer architecture to capture complex temporal dependencies, IAENet effectively models heterogeneous clinical time series. To address label imbalance and inter-label dependencies, we introduce a label-constrained reweighting loss that combines batch-wise dynamic weighting and label co-occurrence constraints.

Experimental results show that our approach consistently outperforms strong baselines across multiple evaluation metrics, highlighting its potential for clinical risk prediction. In future work, we plan to incorporate external knowledge and evaluate the framework on broader clinical datasets and event types. This study offers a foundation for developing intelligent early warning systems to support real-time decision-making in perioperative care.

Acknowledgments

This work was supported in part by the Sichuan Provincial Science and Technology Department (No.2022YFQ0108) and in part by the Chengdu Science and Technology Program (No.2022-YF04-00078-JH).

References

- Bahador, N.; Jokelainen, J.; Mustola, S.; and Kortelainen, J. 2021. Multimodal spatio-temporal-spectral fusion for deep learning applications in physiological time series processing: A case study in monitoring the depth of anesthesia. *Information Fusion*, 73: 125–143.
- Cai, J.; Li, P.; Li, W.; and Zhu, T. 2024. Outcomes of clinical decision support systems in real-world perioperative care: a systematic review and meta-analysis. *International Journal of Surgery*, 110(12): 8057–8072.
- Cai, X.; Wang, X.; Zhu, Y.; Yao, Y.; and Chen, J. 2025. Advances in automated anesthesia: a comprehensive review. *Anesthesiology and Perioperative Science*, 3(1): 3.
- Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.
- Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.
- Chen, H.; Lundberg, S. M.; Erion, G.; Kim, J. H.; and Lee, S.-I. 2021. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. *NPJ digital medicine*, 4(1): 167.
- Cheung, C. C.; Martyn, A.; Campbell, N.; Frost, S.; Gilbert, K.; Michota, F.; Seal, D.; Ghali, W.; and Khan, N. A. 2015. Predictors of intraoperative hypotension and bradycardia. *The American journal of medicine*, 128(5): 532–538.
- Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.
- Hwang, E.; Park, Y.-S.; Kim, J.-Y.; Park, S.-H.; Kim, J.; and Kim, S.-H. 2023. Intraoperative hypotension prediction based on features automatically generated within an interpretable deep learning model. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10): 13887–13901.
- Lee, H.-C.; Park, Y.; Yoon, S. B.; Yang, S. M.; Park, D.; and Jung, C.-W. 2022. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1): 279.
- Lee, H.-C.; Ryu, H.-G.; Chung, E.-J.; and Jung, C.-W. 2018. Prediction of bispectral index during target-controlled infusion of propofol and remifentanyl. *Anesthesiology*, 128(3): 492–501.
- Lee, H.-C.; Ryu, H.-G.; Park, Y.; Yoon, S. B.; Yang, S. M.; Oh, H.-W.; and Jung, C.-W. 2019. Data driven investigation of bispectral index algorithm. *Scientific reports*, 9(1): 13769.
- Lee, S.; Lee, H.-C.; Chu, Y. S.; Song, S. W.; Ahn, G. J.; Lee, H.; Yang, S.; and Koh, S. B. 2021. Deep learning models for the prediction of intraoperative hypotension. *British journal of anaesthesia*, 126(4): 808–817.
- Lin, S.; Lin, W.; Wu, W.; Zhao, F.; Mo, R.; and Zhang, H. 2023. Segrnn: Segment recurrent neural network for long-term time series forecasting. *arXiv preprint arXiv:2308.11200*.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.
- Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.
- Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35: 9881–9893.
- Lu, F.; Li, W.; Zhou, Z.; Song, C.; Sun, Y.; Zhang, Y.; Ren, Y.; Liao, X.; Jin, H.; Luo, A.; et al. 2023. A composite multi-attention framework for intraoperative hypotension early warning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14374–14381.
- Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.-W.; Newman, S.-F.; Kim, J.; et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10): 749–760.
- Meara, J. G.; Leather, A. J.; Hagander, L.; Alkire, B. C.; Alonso, N.; Ameh, E. A.; Bickler, S. W.; Conteh, L.; Dare, A. J.; Davies, J.; et al. 2015. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *The lancet*, 386(9993): 569–624.
- Moon, J.-H.; Lee, G.; Lee, S. M.; Ryu, J.; Kim, D.; and Sohn, K.-A. 2024. Frequency domain deep learning with non-invasive features for intraoperative hypotension prediction. *IEEE Journal of Biomedical and Health Informatics*, 28(10): 5718–5728.
- Nepogodiev, D.; Martin, J.; Biccard, B.; Makupe, A.; Bhangu, A.; Nepogodiev, D.; et al. 2019. Global burden of postoperative death. *Lancet*, 393(10170): 401.
- Nie, Y.; H. Nguyen, N.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*.
- Park, J.-B.; Lee, H.-J.; Yang, H.-L.; Kim, E.-H.; Lee, H.-C.; Jung, C.-W.; and Kim, H.-S. 2023. Machine learning-based prediction of intraoperative hypoxemia for pediatric patients. *PLoS One*, 18(3): e0282303.
- Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

- Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91.
- Schnider, T. W.; Minto, C. F.; Egan, T. D.; and Filipovic, M. 2021. Relationship between propofol target concentrations, bispectral index, and patient covariates during anesthesia. *Anesthesia & Analgesia*, 132(3): 735–742.
- Varghese, C.; Harrison, E. M.; O’Grady, G.; and Topol, E. J. 2024. Artificial intelligence in surgery. *Nature medicine*, 30(5): 1257–1268.
- Wang, Y.; Wu, H.; Dong, J.; Liu, Y.; Long, M.; and Wang, J. 2024. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.
- Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.
- Yang, K.; Ren, M.; Xu, J.; and Zeng, X. 2024. Dynamic Prediction of Intraoperative Hypotension Based on Hemodynamic Monitoring Data With a Transformer-Based Deep Learning Model. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 5166–5173. IEEE.
- Yèche, H.; Pace, A.; Ratsch, G.; and Kuznetsova, R. 2023. Temporal label smoothing for early event prediction. In *International Conference on Machine Learning*, 39913–39938. PMLR.
- Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.
- Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.
- Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, 27268–27286. PMLR.