# When Attention Fails: Pitfalls of Attention-based Model Interpretability for High-dimensional Clinical Time-Series

**Shashank Yadav** SHASHANK@ARIZONA.EDU
*University of Arizona, USA*

**Vignesh Subbian** VSUBBIAN@ARIZONA.EDU
*University of Arizona, USA*

## Abstract

Attention-based deep learning models are widely used for clinical time-series analysis, largely due to their perceived ability to enhance model interpretability. However, the reliability and consistency of attention mechanisms as an interpretability tool in high-dimensional clinical time series data require further investigation. We conducted a comprehensive evaluation of consistency of attention mechanisms in deep learning models applied to high-dimensional clinical time-series data. Specifically, we trained 1000 different variants[1] of an attention-based LSTM model architecture with random initializations to analyze the consistency of attention scores across mortality prediction and patient severity group classification. Our findings revealed significant inconsistencies in attention scores for individual samples across the thousand model variants. Visual inspection of attention weight distributions indicated that the attention mechanism did not consistently focus on the same feature-time pairs, challenging the assumption of reliability in model interpretability. The observed inconsistencies in per-sample attention weights suggest that attention mechanisms are unreliable as an interpretability tool for clinical decision-making tasks involving high-dimensional time-series data. While attention mechanisms may enhance model performance metrics, they often fail to produce clinically meaningful and consistent interpretations, limiting their utility in healthcare settings where transparency is critical for informed decision-making.

1. We define **variant** as the same model architecture being trained to differ only in the random seed used for parameter initialization.

**Data and Code Availability** This paper uses the HiRID dataset (Yèche et al., 2021), which is available on the PhysioNet repository (Goldberger et al., 2000). The code is available at https://github.com/xinformatics/inconsistentattention

**Institutional Review Board (IRB)** This research did not require IRB approval.

## 1. Introduction

Deep learning (DL) models have become integral to healthcare analytics, where they are used to predict patient outcomes, detect abnormalities, and assist in clinical decision-making (Begoli et al., 2019). However, given the high stakes involved, it is insufficient for these models to merely deliver accurate predictions. Although interpretability is frequently portrayed as a means to foster trust in DL models, the assumption that it automatically leads to trust remains to be proven (Weber et al., 2024). At the same time, many interpretability methods lack standardized evaluation frameworks, raising concerns about whether they accurately capture a model's behavior. In the absence of rigorous computational validation or relevant ground-truth benchmarks, such methods may inadvertently create additional risks, such as misleading clinical judgments, overconfidence in unverified outputs, and the potential for wasted resources spent on repeatedly verifying uncertain results in clinical settings. As Weber et al. (2024) emphasize, rigorous evaluation is needed before end users can reliably depend on model interpretability for critical decision-making.

Interpretability methods are essential to bridge the gap between model performance and their clinical utility. These methods can be broadly cat-

egorized into two types: intrinsic interpretability, where the model's architecture is inherently understandable, and post-hoc interpretability, where interpretations are generated after the model has made its predictions (Jacovi and Goldberg, 2020; Lipton, 2018). One promising approach within intrinsic interpretability is incorporating the attention mechanism, which calculates a weighted sum of the vector representations from a layer in a neural network model. By assigning weights to different input features, attention mechanisms are thought to highlight the most important factors influencing a model's prediction (Bahdanau, 2014). Though initially developed for natural language processing (NLP) tasks, attention-based DL models have been applied to clinical multivariate time-series tasks such as retrospective prediction of adverse events in intensive care units (Gandin et al., 2021). Such attention-based models have become popular mainly because of their perceived interpretability and excellent performance (Chen et al., 2020). However, several studies, including those on NLP models, have critiqued the reliability of attention mechanisms as a tool for model interpretability (Jain and Wallace, 2019; Serrano and Smith, 2019). While some research suggests that attention weights can provide insights into model behavior, other studies reveal inconsistencies, showing that attention weights do not always correlate with feature importance and produce inconsistent changes in predictions when modified (Vashishth et al., 2019; Clark, 2019; Vig and Belinkov, 2019).

To our knowledge, the attention mechanism's reliability as an inherent interpretability technique has not been rigorously tested in the context of models for high-dimensional clinical time-series data. In this work, we empirically investigate the relationship between attention weights, inputs, and outputs in the context of clinical multivariate time-series data. We investigate the claim that attention mechanisms automatically help interpret high-dimensional clinical time series data. In particular, we check if attention weights actually indicate why a model makes its predictions by: I) Evaluating the consistency of these attention patterns across different model variants. ii) Examining whether the attention weights are temporally coherent. Our hypothesis is that attention mechanisms cannot always help achieve this alignment between true model behavior and the interpretations provided by the model itself. The attention weights can sometimes highlight features that do not genuinely influence the model's predictions, leading to potential misinterpretations. Moreover, causality, transferability, fairness, and informativeness are potential failure points for attention-based interpretability (Lipton, 2018). We investigate the validity of these assumptions across different clinically relevant tasks by addressing the following questions:

1. How consistent are the per-sample attention weights when aggregated over multiple model variants?
2. How sparse and visually interpretable are the per-sample attention weight distributions?

## 2. Related Work

The application of the attention mechanism within Long-Short Term Memory (LSTM) networks for clinical prediction tasks has gained traction due to the perceived interpretability of the attention mechanism, such as highlighting critical feature-time pairs in clinical data. Kaji et al. (2019) employed an attention-based LSTM model for predicting daily sepsis, myocardial infarction, and vancomycin antibiotic administration using the MIMIC-III dataset. Their results indicated that the attention mechanism could improve prediction performance while also identifying the time periods during which input features had the greatest impact on predictions, offering clinicians a level of interpretability regarding the model's decision-making process. Similarly, Gandin et al. (2021) used a cohort of 10,616 cardiovascular patients from the MIMIC-III dataset to predict mortality within seven days using an LSTM model equipped with an attention layer, which highlighted features— norepinephrine, phenylephrine, creatinine, gender, blood urea nitrogen, etc. as strong predictors of patient mortality. However, while these studies emphasize the potential utility of the attention mechanism for interpretability, they often implicitly assume that the weights assigned by the attention layer provide meaningful interpretations without thoroughly validating the consistency of these interpretations across multiple model variants.

While the growing use of attention mechanisms in DL models is evident, their reliability as an interpretability tool has been scrutinized, especially in the NLP domain. Jain and Wallace (2019) were among the first to question the assumption that attention weights inherently correlate with feature importance. Through experiments across several NLP tasks, they demonstrated that attention weights did not neces-

sarily align with standard feature importance measures, and attention-based models could achieve similar performance even when attention weights were randomly reassigned. Similarly, Serrano and Smith (2019) conducted perturbation-based experiments to investigate the reliability of attention mechanisms. They found that modifying attention weights often had little impact on the model's output, suggesting that the features highlighted by attention may not be as crucial as previously thought. Furthermore, Pruthi et al. (2019) introduced adversarial manipulation to attention weights. They demonstrated that attention weights could be manipulated without significantly affecting model performance, indicating that the highlighted features might not actually drive the model's decisions. Specifically, they perturbed the attention distributions while keeping the output unchanged, revealing that the weights assigned by the attention mechanism could be decoupled from the model's internal decision process. These findings collectively suggest that attention mechanisms do not always yield consistent or causally linked interpretations to model predictions.

Motivated by the critique of attention mechanisms in NLP, we extend the analysis of attention-based interpretability to high-dimensional clinical time-series data. While attention-based models have been adapted to provide interpretability for mortality and sepsis prediction, there remains a significant gap in validating whether these models consistently highlight the most important clinical feature-time pairs responsible for a model's decision. Our work aims to fill this gap by systematically evaluating the attention mechanism's reliability and consistency within attention-based LSTM models for high-dimensional clinical time-series data. By focusing on high-dimensional clinical time series, we contribute an essential extension to the existing model interpretability literature, examining whether the perceived interpretability of attention-based models holds up under empirical scrutiny for high-stakes intensive care applications.

## 3. Methods

### 3.1. Dataset and prediction task description.

We utilized two prediction tasks from the HiRID-ICU-Benchmark study: (a) prediction of mortality in the intensive care unit after 24 hours of stay and (b) prediction of admission group (phenotyping based on APACHE score) after 24 hours of stay (Yèche et al., 2021). Detailed information on the dataset distribution, including the number of ICU stay records in the training, validation, and test sets, and the number of prediction samples, are provided in Table Table 1. The multivariate time series dataset spans 288 timesteps at 5-minute intervals, totaling 24 hours, and includes 231 clinical features such as vital signs, hemodynamic data, treatments, lab values, and ventilation parameters for critical care management. The mortality task involves binary classification, with a single prediction for each ICU stay. The phenotyping task is a multi-class classification with a total of 15 classes with one prediction per stay. In this task, patients are classified based on their admission diagnosis after 24 hours in the ICU, using APACHE II and IV labels (Zimmerman et al., 2006).

### 3.2. Model Architectures and Training

We used an attention-based Long Short-Term Memory (LSTM) network, identical to the models employed in the previous studies (Gandin et al., 2021; Kaji et al., 2019), which claim that attention mechanisms play a crucial role in enhancing model interpretability by learning a set of weights corresponding to the input features. In attention-based LSTM models, the attention mechanism helps to highlight important feature-time pairs. It operates at the level of input features and involves a dense layer with a softmax activation function as implemented by Kaji et al. (2019). They provide an implementation that is available on Zenodo (Kaji, 2018) and re-utilized by Gandin et al. (2021).

Let $\boldsymbol{X}$ be the input data matrix with dimensions, $T \times D$, where $T$ represents the number of time steps, and $D$ represents the number of features. The dense layer produces attention scores, denoted as $\boldsymbol{S}$, which also have dimensions $T \times D$. This can be expressed as:

$$\boldsymbol{S} = \boldsymbol{W} \cdot \boldsymbol{X} \tag{1}$$

$\boldsymbol{W}$ is the attention weight matrix of the dense layer with dimensions $D \times D$. The result, $\boldsymbol{S}$, contains the attention scores for each feature-time pair for a sample. The attention scores $\boldsymbol{S}$ are then normalized using the softmax function to obtain attention weights $\boldsymbol{A}$. The softmax function is applied along the feature dimension, ensuring that the weights sum to 1 for each time step. This is expressed as,

Table 1: Summary of tasks, prediction types, and dataset splits.

| Task | # ICU Stays | | | Prediction Type |
|---|---|---|---|---|
| | Training | Validation | Test | |
| Mortality | 10,525 | 2,206 | 2,231 | Binary |
| Phenotyping | 10,470 | 2,194 | 2,217 | Multiclass |

$$\boldsymbol{A_{t,j}} = \frac{\exp(\boldsymbol{S_{t,j}})}{\sum_{k=1}^{D} \exp(\boldsymbol{S_{t,k}})} \tag{2}$$

where $\boldsymbol{A_{t,j}}$ represents the attention weight for feature $\boldsymbol{j}$ at time step $\boldsymbol{t}$. The attention weights $\boldsymbol{A}$ are applied to the input data matrix $\boldsymbol{X}$ through element-wise multiplication, as follows:

$$\boldsymbol{X}_{\text{new}} = \boldsymbol{X} \odot \boldsymbol{A} \tag{3}$$

This procedure produces a final output matrix $\boldsymbol{X}_{\text{new}}$, a weighted representation of the input data that guides the model in focusing on the most relevant feature-time pairs across both time and feature dimensions in the subsequent layers. To account for variability in per-sample attention weights and quantify their consistency across model variants, we trained an attention-LSTM model architecture with 1000 different random initializations. Although the number of model variants was limited by training time, this approach was substantially more extensive than previous studies, which typically trained only 10 variants and calculated the average attention weights corresponding to each sample. We investigated the extent of variability in attention weights and gained a more comprehensive understanding of their consistency across different model variants by training 100 times more model variants than prior studies.

### 3.3. Cumulative attention drop analysis

Understanding and visualizing interpretability aspects is key to ensuring reliable DL models, especially in clinical settings. Visualization techniques offer a window into how models reach their decisions. Clinicians can better assess the relevance and accuracy of the model's outputs by visually representing the feature relationships and patterns identified by the model. This transparency is essential in clinical settings, where decisions directly impact patient outcomes (Vellido, 2020; Holzinger et al., 2017). We performed an attention drop analysis, a visualization-based method that evaluates the stability of attention weights across multiple model variants with varying random initialization. For each model variant $\boldsymbol{m}$, we extracted the top K attention weights, where K was set to 10,000, 1,000, and 100. Let $\boldsymbol{P_{m,k}}$ denote the set of top K attention weights for the model $\boldsymbol{m}$. We then calculated how many of these top K weights remained unchanged across all 1,000 model variants. Specifically, we determined the intersection of the top K attention weights from each model variant to compute $C_K$, defined as:

$$\boldsymbol{C_K} = \left| \bigcap_{\boldsymbol{m=1}}^{\boldsymbol{1000}} \boldsymbol{P_{m,K}} \right| \tag{4}$$

where $C_K$ represents the cumulative count of attention weights that persisted across all 1,000 model variants. This approach allowed for visualizing the stability and reliability of per-sample attention weights by evaluating their consistency across different random initializations. Although our analysis was limited to 1,000 model variants, we hypothesize that as more model variants are trained, the common attention weights will continue to decrease and may eventually reach zero.

### 3.4. Cumulative attention rank analysis

We extended our evaluation beyond specific cutoff thresholds and conducted a rank consistency analysis for all 66528(231×288) feature-time pairs for each sample. For each feature-time pair in $\boldsymbol{A}$, denoted by $f_{i,j}$, where $i \in \{1, \ldots, 231\}$ and $j \in \{1, \ldots, 288\}$, we computed the running mean rank of all feature-time pair for each sample for all model variants. The cumulative rank of the feature-time pair $f_{i,j}$ is defined as:

$$\overline{\boldsymbol{R}}(\boldsymbol{f_{i,j}}) = \frac{\boldsymbol{1}}{\boldsymbol{M}} \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{R_m}(\boldsymbol{f_{i,j}}) \tag{5}$$

where,

- $M$ is the total number of model variants (e.g., $M = 1000$),
- $R_m(f_{i,j})$ is the rank assigned to feature-time pair $f_{i,j}$ in the $m$-th model variant.

### 3.5. Cluster Analysis of Attention Weights Across Model Variants

We performed Ward linkage-based clustering on the $\boldsymbol{W}$ matrix in the attention layer for the three model variants chosen in the cumulative attention drop analysis **(Section 3.3), based on predictive performance**. Ward linkage clustering grouped the attention parameters of $\boldsymbol{W}$ to demonstrate how clinical features were weighted across different model variants and which clinical features tend to be clustered together, revealing structured global attention patterns that may align with clinically relevant information.

## 4. Results

### 4.1. Attention Drop For the Mortality Task

The attention drop analysis for the mortality prediction task revealed that the consistency of common attention weights across cumulative model variants diminishes exponentially. Specifically, when we began with the top 10,000 attention weights for each sample, the number of common weights consistently decreased as more trained model variants were considered. For the last (1,000th) variant, the number of common attention weights per sample had dropped to 100 on average, as shown in Figure 1. It illustrates that with increasing model variants considered, the variability in attention weights could become so pronounced that no consistent feature-time pairs remain across the model variants. This further challenges the reliance on attention mechanisms for interpretability, as identifying key feature-time pairs may become entirely model-dependent, leading to potential discrepancies in clinical decision-making. The pattern of exponential decline was also observed when we initiated the analysis with fewer feature-time pairs—starting with the top 1,000 and 100 attention weights per sample. The number of common attention weights per sample had dropped to almost 0 on average upon reaching the 1,000th model variant, as illustrated in Figure 1. The trend remained consistent in each case, highlighting the variability and instability of attention weights as more model variants were incorporated into the analysis. Moreover, we calculated the model metrics for all 1000 model variants on the test set and found the results to be closely aligned, as evidenced by the low standard deviation in both AUC–ROC and AUC–PR values (Figure 9 - Appendix 5). The actual and cumulative mean plots further illustrate that the individual metric values for each model variant and their cumulative means remain largely consistent, as indicated by the flatness of the curves. This consistency in performance metrics implies that the observed variability in attention weights is not driven by differences in model performance, but rather points to an issue inherent in the attention mechanism itself.

### 4.2. Attention Drop For Phenotyping Task

For the phenotyping task, we focused on the cardiovascular class (355 patient samples), as shown in Figure 2 (with additional results for the other three major classes—Gastrointestinal, Respiratory, and Neurological—provided in Figure 5 in Appendix A.1). In this case, we observed that the common attention weights decreased even more rapidly than in the ICU Mortality Task. The samples in the other three major classes—Gastrointestinal (130 patient samples), Respiratory (260 patient samples), and Neurological (503 patient samples), also show a rapid decline in stability of attention weights across cumulative model variants. Specifically, when we began with the top 10,000 attention weights for each sample in the cardiovascular domain, the number of common attention weights dropped to zero even before 100 cumulative model variants were incorporated into the analysis. This rapid decline indicates an even higher level of variability in the per-sample attention weights, further highlighting the instability and potential reliability issues of attention mechanisms as a model interpretability method. The early drop to zero common attention weights emphasizes the limitation of using attention mechanisms to derive consistent, interpretable insights in high-dimensional clinical time-series data, as the results are susceptible to the randomness inherent in model training.

### 4.3. Cumulative Attention Ranks across the Mortality Task

We analyzed the cumulative mean ranks of the 66,528 feature-time pairs as more model variants were aggregated for a patient sample with an adverse event. The histograms in Figure 3 illustrate how these ranks

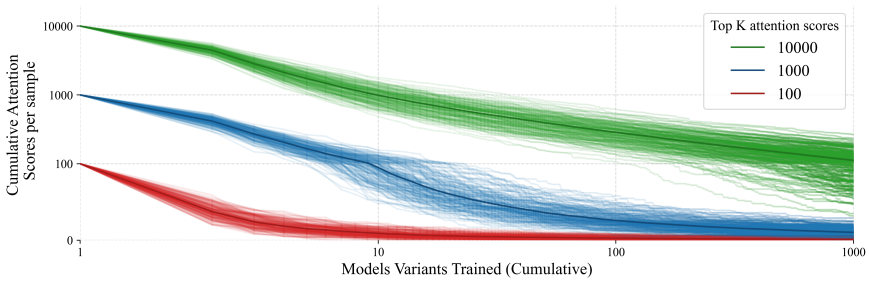



Figure 1: Demonstration of drop in consistent attention weights per sample for patients with adverse events in the test set (N=186). The bolder lines show the average attention drop across the model variants for the same cohort.

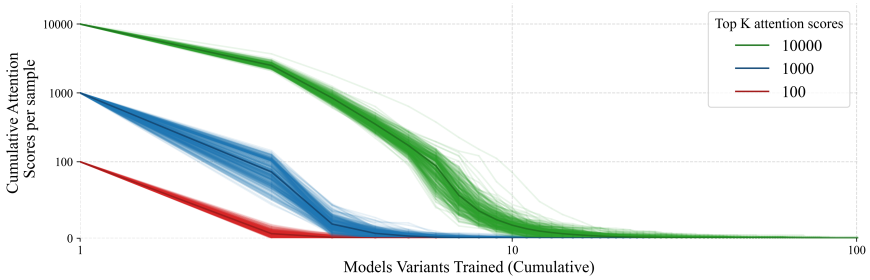



Figure 2: Demonstration of attention weights per sample for the cardiovascular domain in the test set (N=355). The bolder lines show the average attention drop across the model variants for the same cohort.

evolve cumulatively. For a single model variant (Cumulative Model 1), the ranks appeared to be uniformly distributed across the entire range, which reflected that no particular concentration of importance among feature-time pairs. This uniformity is expected due to the randomness of rankings in a single model variant. However, starting from Cumulative Model 2 (when Model Variant #1 and Model Variant #2 both are considered together), the cumulative ranks begin to average out, and the distribution starts shifting towards the theoretical mean rank of 33,265. This shift demonstrates the influence of aggregating feature ranks over multiple model variants, where individual model-specific variability is reduced. As more model variants are considered in the cumulative analysis (e.g., from 4 to 512), the rank distributions further stabilize and shift toward the right. This rightward shift highlights that the ranks converge toward the theoretical mean, with most feature-time pairs averaging around 33,265. Since the plot is on a log scale, the long tail of the mean rank distribution, however, reflects the sparsity of consistently high-ranking feature-time pairs, which remain in the lower-rank region. This observation reinforces that as more model variants are considered, the variability in attention weights can become so significant that no consistent feature-time pairs persist across the different model variants.

Even after considering cumulative ranks across 1,000 model variants, only a single-digit number of feature-time pairs achieve a mean rank below 100. This suggests that despite the presence of 231 clinical variables and 288 timesteps, only a handful of variables consistently exhibit low ranks, and even those have mean ranks close to 100. Such a pattern indicates that attention-based interpretability fails to highlight meaningful clinical insights, as the rankings do not sufficiently differentiate between important and unimportant feature-time pairs. This lack of clarity renders attention-based model architectures ineffective as a reliable interpretability tool. In addition to the mean rank analysis, we analyzed the standard deviation of these ranks for the mortality task as illustrated in Figure 6 in Appendix A2. We observed that

as more model variants are considered, the standard deviation of feature ranks starts to stabilize. However, even the minimum standard deviation remains on the order of 100. This reinforces the observation that the rankings do not sufficiently differentiate between important and unimportant feature-time pairs. Furthermore, for the patient phenotyping task, the attention drop becomes even more pronounced. Despite evaluating cumulative ranks across only 4 model variants for patients in Circulatory, Gastrointestinal, Respiratory and Neurological, none of the clinical feature-time pairs achieve a mean rank below 100 as shown in Figure 7 (Appendix A.3).

### 4.4. Sparsity of per-sample attention weights

Another critical aspect of our analysis focussed on visualizing attention weights across three randomly selected model variants that all yielded high prediction probabilities for a particular patient with an adverse outcome in the mortality task. Although these models predict the event with high probability (prediction probability greater than 0.9), a closer examination of the attention weights reveals striking inconsistencies. Figure 4 visualizes attention weights for the same patient time-series data across the three model variants. We selected this representative patient (index 1394 of 2231, see Table 1.) from the adverse outcome class consisting of 186 ICU stays. For this patient, we examined predictions from all 1000 model variants, ultimately selecting the three variants (ids: 968, 356, 350) that produced the highest mortality prediction probabilities (0.9816, 0.9578, and 0.9362, respectively). Ideally, if the attention mechanism were reliable and consistent, we would expect similar patterns of attention weight distribution across these model variants, mainly because they all reach a similar prediction outcome, close to the ground truth label. However, the visualizations show that the attention weights vary significantly between the model variants, even though the prediction probabilities are similar and higher. The dashed band in Figure 4 at the start of the observation period is a particularly highlighted feature in the visualization. This band represents one of the several areas where the attention mechanism assigns weights inconsistently across the three chosen models. This early period may receive high attention in one model (bottom), suggesting that the model considers the initial conditions critical for the prediction. In contrast, other models (top and middle) have assigned little to no attention to the same feature-time pairs, focusing instead on a different portion of the initial observation period.

In contrast, when applying ExtremalMask, a state-of-the-art time-series model interpretability method (Enguehard, 2023) to the same three model variants, it consistently identified the final time steps and the same feature indexes as critical (Figure 10). Moreover, the Extremal Mask results consistently indicate that maximum attribution is assigned to the end-time steps validated by several studies (Deasy et al., 2020; Shickel et al., 2019; Johnson and Mark, 2018). This comparison highlights that attention fails to consistently capture the temporal coherence and post-hoc mask-based methods such as ExtremalMask not only align with clinical expectations but also provide a more robust and reliable measure of feature-time importance.

### 4.5. Visualizing feature-feature structures

Our results indicated a significant variability in the clustering patterns across model variants, even when the attention weight matrices appear to be similar. Despite using the same architecture, different model variants produced distinctly different attention cluster structures, suggesting that the attention mechanism does not consistently focus on the same feature-time pairs across models. Figure 8 (Appendix A.3) illustrates the clustering results for attention weight matrices from the three chosen model variants based on prediction performance. Prominent clusters in one model variant are entirely absent in others, highlighting the lack of consistency in the attention mechanism. Clustering was expected to reveal consistent and meaningful patterns within the attention mechanism. However, our results highlighted a major challenge in using attention mechanisms for consistent model interpretation, as the clustering patterns, which should reflect the model's focus, are not stable across different model variants.

## 5. Discussion

Our study revealed several critical limitations of attention mechanisms in providing interpretable insights for high-dimensional clinical time series. Despite their widespread use, attention mechanisms exhibit significant shortcomings in clinical settings, affecting their utility for clinical decision-making tasks. We identified five primary reasons why attention mechanisms fail to deliver consistent interpretations.

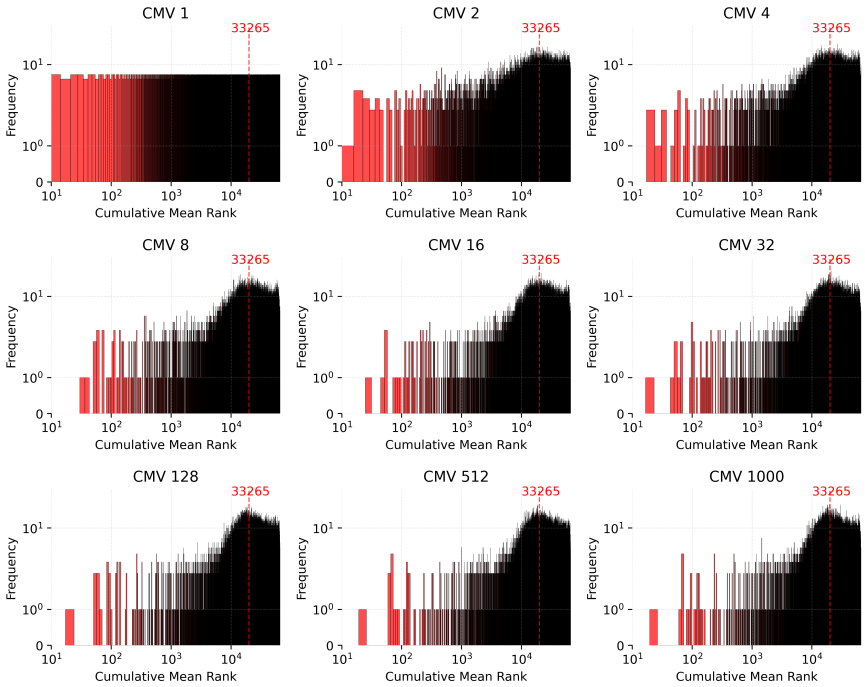



Figure 3: Visualization of cumulative mean ranks for feature-time pairs based on their attention weights for a patient with an adverse outcome as more model variants are taken into consideration. For each subplot, the y-axis denotes the frequency of feature-time pairs, and the x-axis represents their cumulative mean rank. CMV: Cumulative Model Variants. The dashed red line shows the theoretical mean (∼33,265).

**(a) Inconsistency across Model Variants:** A significant issue with attention mechanisms is their inconsistency, as confirmed by the results of the attention drop analysis. Different model variants produce substantially different attention weight distributions for the same input time series. For example, one model might emphasize blood pressure readings. At the same time, another model might focus on cholesterol levels when predicting the risk of a cardiovascular event for the same patient, potentially leading to predictions that diverge from the actual clinical reality.

**(b) Sparsity of Attention Weights:** Attention mechanisms exhibited sparsity in their weight distributions, as illustrated in Figure 4. This sparsity poses significant challenges when applied to high-resolution clinical time-series data, such as those involving five-minute intervals. For acute conditions, a discrete period of five minutes, as highlighted by the attention mechanism, may still be insufficient to capture significant changes. In acute organ failure, for instance, it is essential to have a continuous and contextually relevant distribution of attention weights. This is because the onset of organ failure is typically not abrupt but instead results from a prolonged deterioration of the patient's condition. Our findings indicate that attention weights are frequently sparse and do not adequately cover the entire duration. This sparsity challenges the ability of the attention mechanism to provide a continuous and meaningful representation of the patient's condition over time, leading to a potential misinterpretation of the significance of features within their broader temporal context.

**(c) Lack of Integration with Domain Knowledge:** We believe that the attention mechanism is designed to better optimize model dynamics, leading to performance improvements rather than establishing clinical knowledge. It highlights patterns the model finds important but does not align with clinical relevance. For example, attention might highlight a decrease in blood pressure, but this could be mislead-

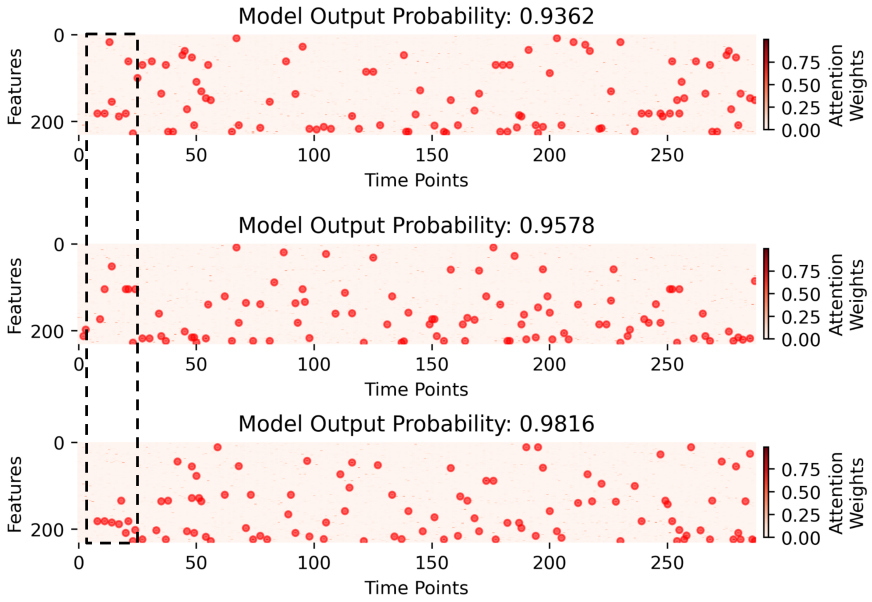



Figure 4: Visualization of attention weights for three selected model variants for a sample patient with an adverse outcome with high prediction probability in the ICU Mortality Task. The red dots represent each model variant's top 100 attention weights. The dashed rectangle represents one of the several regions where attention weights have inconsistent weighting across model variants.

ing without considering other crucial factors, such as cholesterol levels or heart function.

**(d) Neglect of Latent Structures and Interactions:** Clinical time-series data often involve complex and subtle interactions between physiological signals and patient history. For instance, the progression of chronic cardiovascular conditions, as an example from the phenotyping task, involves complex interactions between blood pressure, cholesterol levels, and medication adherence. A simplistic attention mechanism fails to capture these interactions by focusing on individual feature-time pairs in isolation rather than considering how these factors interact and influence each other over time. Recent methods, such as Extremal-Mask, were able to identify the top feature-time pairs in conjunction. (Refer: Appendix: A.6)

**(e) Model Performance vs Informativeness Tradeoff:** Attention mechanisms are primarily designed to improve model performance rather than provide actionable insights for clinical decision-making. While they may enhance prediction metrics, they often fail to give clinicians valuable interpretations. For example, attention might emphasize certain cardiovascular metrics, such as specific changes in blood pressure, cardiac output, or cholesterol levels, as crucial for predicting the risk of a heart attack. However, the resulting interpretations may not provide meaningful insights for effective decision-making if these metrics are not considered within a broader clinical context—such as the patient's overall health status, medical history, and other relevant factors. A possible reason for this issue is due to the use of combinatorial shortcuts by attention, as demonstrated by (Bai et al., 2021). The model embeds extra predictive clues in the pattern of its attention weights instead of highlighting the feature-time pairs with high informativeness about model behavior.

## 6. Conclusion

Attention mechanisms are valuable in enhancing the performance of deep clinical time-series models. However, their role in interpretability is questionable. They can be inconsistent across model variants, sparsely distributed, and fail to capture complex clinical interactions. Moreover, they are inherently model-centric rather than user-centric, significantly

limiting their utility as clinical decision-making tools. Hence, it is crucial to distinguish between attention as a model mechanism and attention as an interpretability tool in clinical settings.

## Acknowledgments

The authors acknowledge the investigators of the HiRID-ICU-Benchmark Study for providing access to data used in this work. This work was partly supported by the National Science Foundation under grant #1838745.

## References

Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21, page 25–34. ACM, August 2021. doi: 10.1145/3447548.3467307. URL http://dx.doi.org/10.1145/3447548.3467307.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Peipei Chen, Wei Dong, Jinliang Wang, Xudong Lu, Uzay Kaymak, and Zhengxing Huang. Interpretable clinical prediction via attention-based neural network. *BMC Medical Informatics and Decision Making*, 20:1–9, 2020.

Kevin Clark. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Jacob Deasy, Pietro Liò, and Ari Ercole. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. *Scientific Reports*, 10 (1):22129, 2020.

Joseph Enguehard. Learning perturbations to explain time series predictions. In *International Conference on Machine Learning*, pages 9329–9342. PMLR, 2023.

Ilaria Gandin, Arjuna Scagnetto, Simona Romani, and Giulia Barbati. Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit. *Journal of biomedical informatics*, 121:103876, 2021.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Alistair EW Johnson and Roger G Mark. Real-time mortality prediction in the intensive care unit. In *AMIA Annual Symposium Proceedings*, volume 2017, page 994, 2018.

Deepak Kaji. Mimic-lstm: Initial release, October 2018. URL https://doi.org/10.5281/zenodo.1473691. Software available from Zenodo.

Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*, 2019.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Benjamin Shickel, Tyler J Loftus, Lasith Adhikari, Tezcan Ozrazgat-Baslanti, Azra Bihorac, and Parisa Rashidi. Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, 9(1): 1879, 2019.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention interpretability across nlp tasks. *arXiv preprint*

*arXiv:1909.11218*, 2019.

Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083, 2020.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

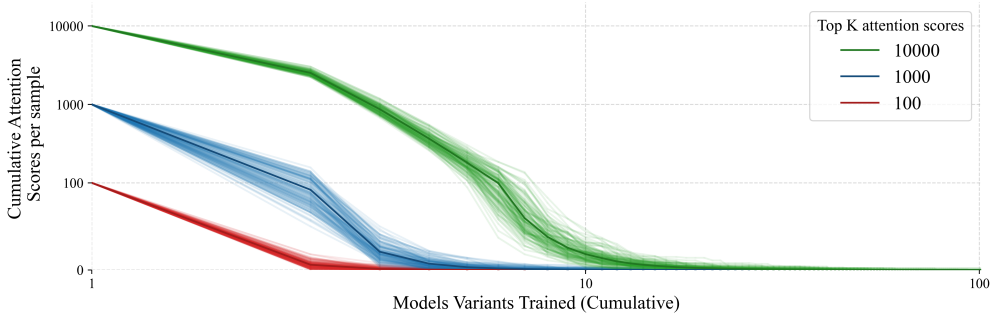
Rosina O Weber, Adam J Johs, Prateek Goel, and João Marques Silva. Xai is in trouble. *AI Magazine*, 45(3):300–316, 2024.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data. *arXiv preprint arXiv:2111.08536*, 2021.

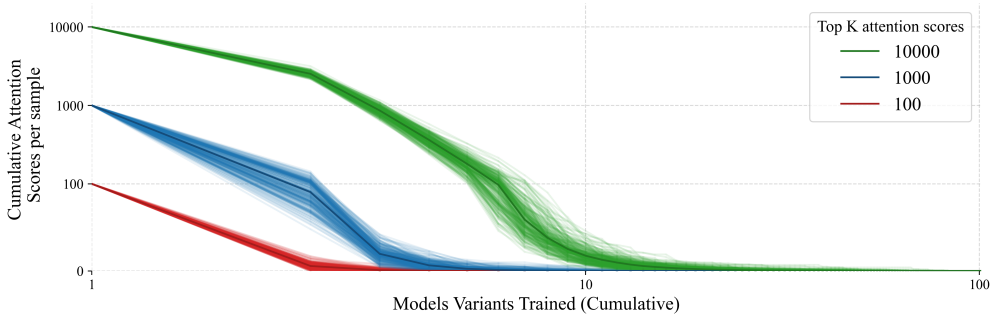
Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.

# Appendix A.

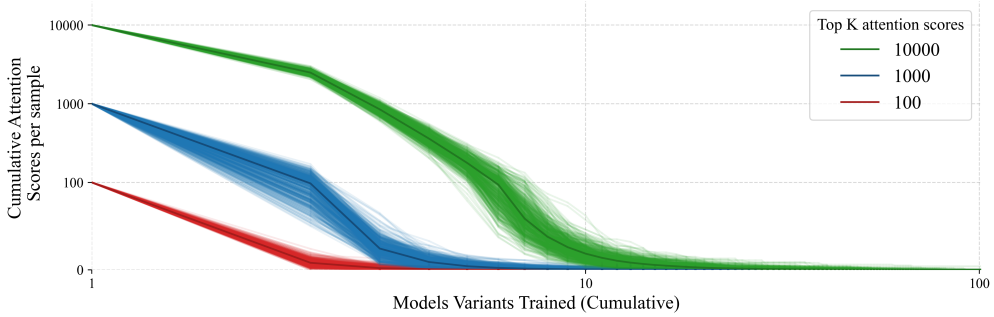
## A.1. Attention Drop Across the Phenotyping class for Gastrointestinal, Respiratory, and Neurological Class




(*a*) Gastrointestinal class (N=130).




(*b*) Respiratory class (N=260).




(*c*) Neurological class (N=503).

Figure 5: Drop in attention weights per sample for the other three classes in the test set. The bolder lines show the average attention drop across the model variants.

### A.2. Cumulative Rank variance analysis

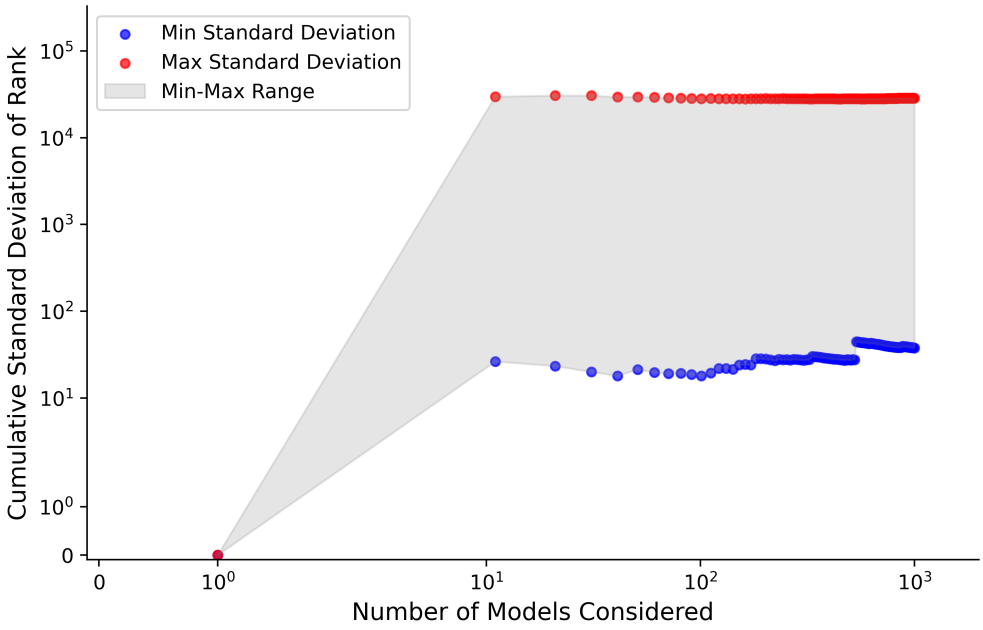



Figure 6: Scatter plot showing how the minimum (blue) and maximum (red) standard deviation of feature ranks change with the number of models considered. The gray area highlights the range between these values.

### A.3. Attention Rank Analysis for the Patient Phenotyping Task

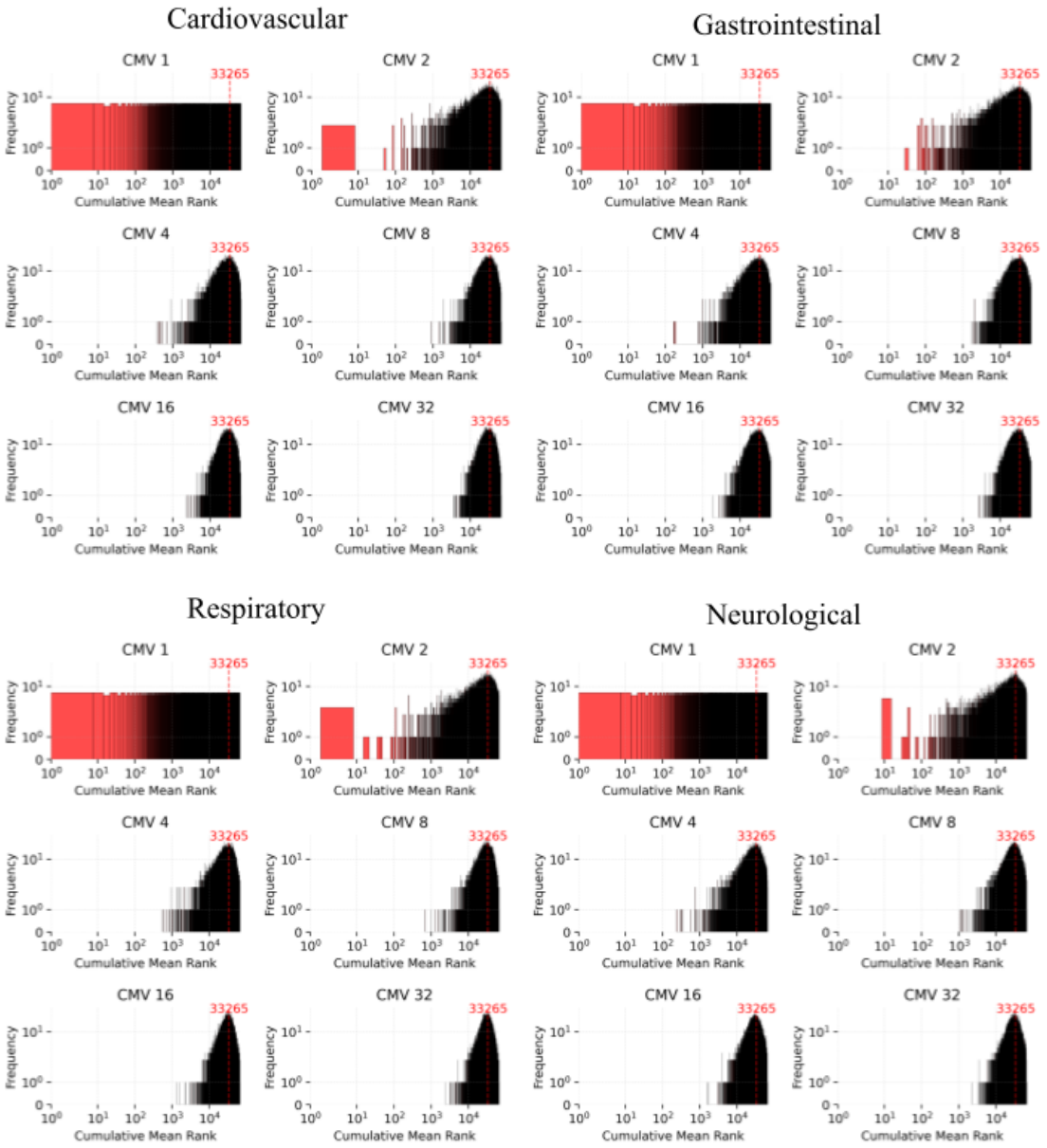



Figure 7: Visualization of cumulative mean ranks for feature-time pairs based on their attention weights for a patient in the **Cardiovascular cohort (top left)**, **Gastrointestinal cohort (top right)**, **Respiratory cohort (bottom left)** and **Neurological cohort (bottom right)** for cumulative model variants. For each plot, the y-axis denotes the frequency, and the x-axis represents their cumulative mean rank. CMV: Cumulative Model Variants. The dashed red line shows the theoretical mean (∼33,265).

### A.4. Attention Mechanism Weight Matrices and Clustered Attention Patterns Across Model Variants

The three models in Figure 8 correspond to model variants of the attention-LSTM model architecture with different parameter initializations selected during the sparsity analysis (described in Section 4.4). Each model variant captures variations in how the model variant globally attends to clinical features. This enables a comparison of attention weight patterns across different variants of the model architecture, illustrating how feature importance and associated clustering patterns vary with parameter initializations. These clusters demonstrate how attention is distributed across clinical features, highlighting which features tend to be grouped together by each model variant. This clustering approach provides a form of global interpretability, revealing variant-wide patterns in feature importance.

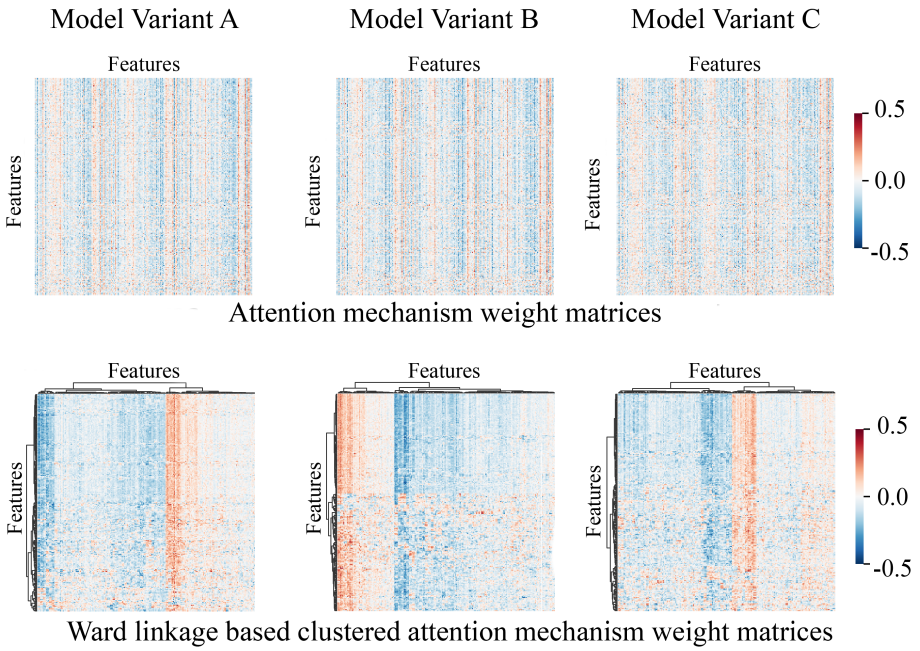



Figure 8: Comparison of attention weight matrices $\boldsymbol{W}$ across three model variants and their clustered counterparts. Here, we illustrate visually similar attention weight matrices from three different model variants (top), showing completely different patterns when clustered to find global attention structures (bottom).

### A.5. Performance comparison of model variants.

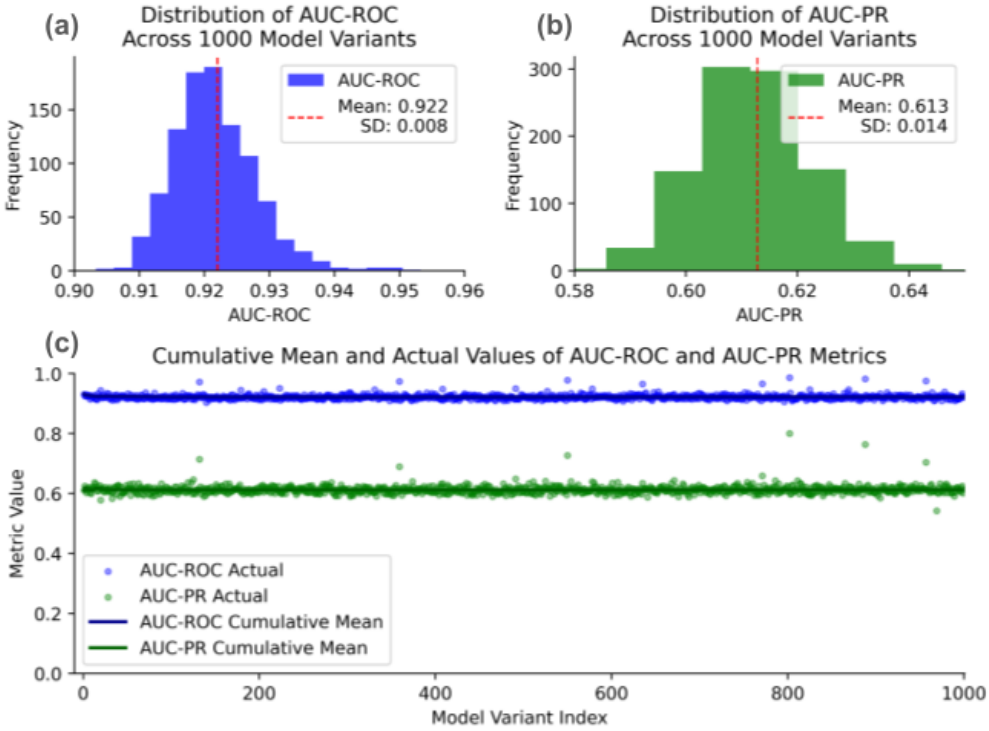



Figure 9: (a) Distribution of AUC–ROC: A histogram of AUC-ROC scores for 1000 model variants. (b) Distribution of AUC–PR: A histogram of AUC-PR scores for the same 1000 model variants. (c) Cumulative Mean and Actual Values: Scatter plots show the individual metric values for each model variant and their cumulative means. The flatness of the curves suggests that the AUC–ROC and AUC–PR metrics are consistent across model variants.

### A.6. Mask-based model interpretability using ExtremalMask

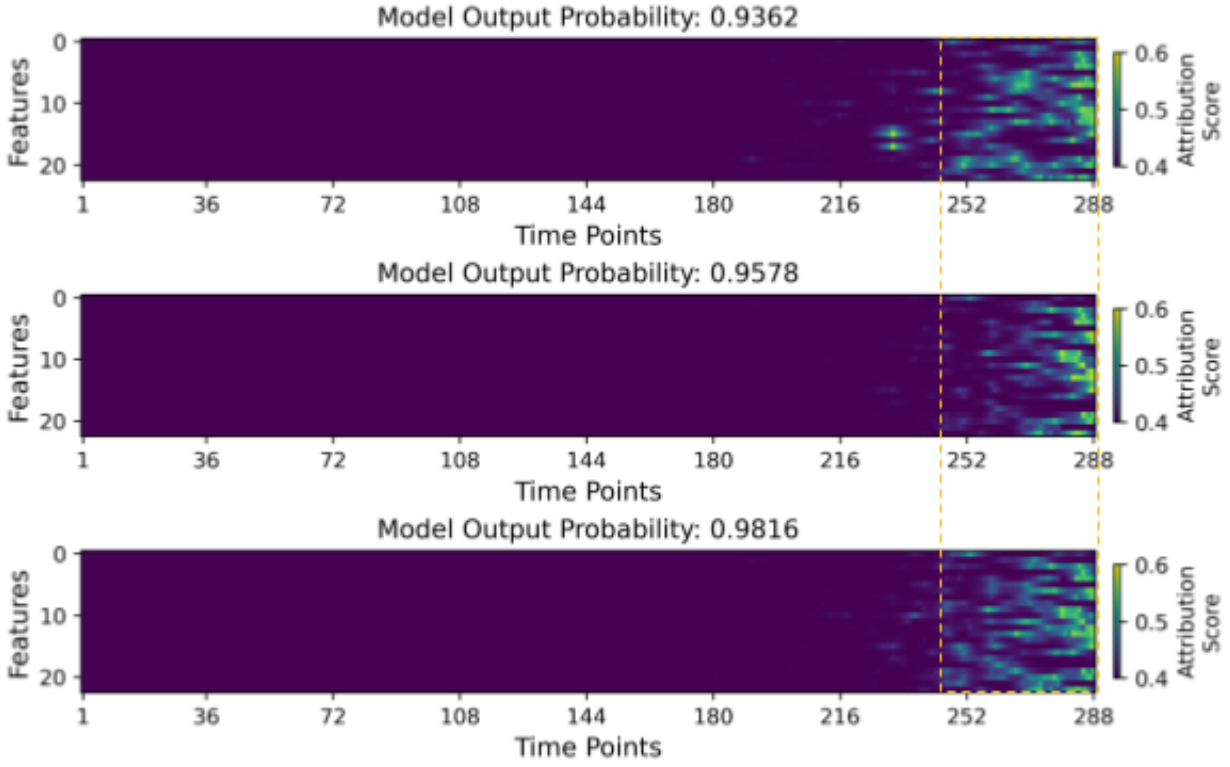



Figure 10: Comparison of model variants based on the attribution scores from ExtremalMask method. Each panel corresponds to a distinct mask solution for the same time-series from model variants A, B and C. The mask retains only those feature-time pairs that are most critical to the model's adverse-event prediction. Notably, all three masks emphasize the final timesteps (highlighted in dashed yellow), consistently reflecting the model's reliance on late-stage physiological signals for making a prediction for an adverse event.