# Explaining Temporal Effects in Sepsis Prediction

**Anonymous Author(s)**
Affiliation
Address
`email`

## Abstract

Sepsis prediction models remain opaque to clinicians which hinder clinician adoption: without understanding why a patient is flagged as high-risk, accurate predictions may be ignored, delaying critical intervention. Existing explainability methods focus on feature importance and often overlook timing, thus failing to capture the temporal influences inherent in time-series data. We propose Positional Explanation, which separates attributions into feature content and it's position to highlight temporal effects, enabling clinicians to identify early warning indicators and monitor for specific physiological changes at critical time windows before sepsis develops. Applied to GPT-2 and Mamba models finetuned for sepsis prediction on PhysioNet and MC-MED benchmarks, our method achieves higher faithfulness scores and reveals temporal patterns in sepsis progression that existing techniques miss, potentially enabling earlier detection and improved patient outcomes.

## 1 Introduction

Sepsis is a leading cause of hospital mortality, primarily because it is often detected after irreversible organ damage [Seymour et al., 2016]. While deep learning models can predict its onset with high accuracy, they typically only signal that the risk of sepsis is high, not why [Yuan et al., 2020, Bomrah et al., 2024]. This leaves a 'lab-to-bedside' gap: without understanding the subtle physiological patterns that precede overt signs, clinicians cannot act on predictions early enough to save lives.

Explainable AI (XAI) methods have the potential to bridge this gap. Beyond fostering trust, these methods can turn predictive models into tools for clinical discovery [Wong et al., 2021, Shashikumar et al., 2021, Adams et al., 2022]. However, existing explanation methods are fundamentally misaligned with the temporal nature of diseases like sepsis.

Sepsis is a disease of trajectory; a patient's physiological trend over time—the when—is often more diagnostically significant than any single measurement—the what [Zhu et al., 2023]. An elevated heart rate, for instance, may signal danger when it appears early and persists, yet prove benign if transient. Despite this temporal criticality, existing explanation methods like LIME [Ribeiro et al., 2016] and Integrated Gradients [Sundararajan et al., 2017] only quantify feature importance, leaving temporal dynamics unexplained.

This blind spot reflects a broader challenge in machine learning. Recent studies have shown that modern deep learning architectures are highly sensitive to input order; even reordering elements in a sequence can substantially change a model's output [Liu et al., 2024, Wang et al., 2024]. This positional sensitivity in general sequence modeling directly parallels the temporal sensitivity in time-series applications like sepsis prediction. Yet current explanation methods cannot address the fundamental question underlying temporal diagnosis: "Is this feature important because of its value, or because of its timing?" Based on this, we argue that to bridge the trust gap, a clinically adequate explanation must be able to separate the importance of 'what' (the feature itself) from 'when' (its temporal effect).

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

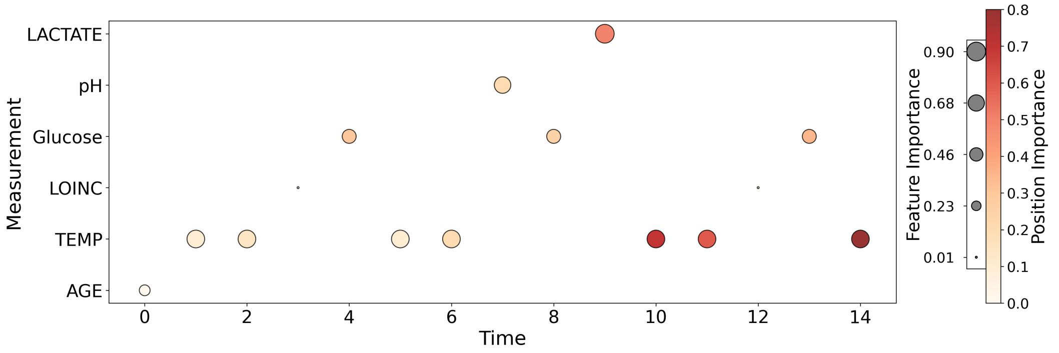


Figure 1: Feature content attribution score $\alpha^{(\text{feature})}$ and absolute position attribution score $\alpha^{(\text{position})}$ for a representative example from the PhysioNet dataset. The visualization demonstrates that feature importance and positional importance differ substantially: while TEMP measurements maintain consistent feature content attribution across time steps, their positional importance increases over time, indicating that there is a temporal effect of TEMP measurement importance for sepsis prediction.

To address this, we introduce **Positional Explanation**, a framework that decomposes the standard attribution score into two distinct components: (1) a **feature content score** reflecting its intrinsic clinical value, and (2) a **position score** that quantifies the importance of the temporal effect. We apply our framework to Mamba [Gu and Dao, 2024] and GPT-2 [Radford et al., 2019] models finetuned for early sepsis prediction, using the EHR data from PhysioNet [Reyna et al., 2020] and MC-MED [Kansal et al., 2025]. To summarize, our contributions are:

- We formalize a framework called Positional Explanation that decomposes attribution scores into feature and position effects for time-series data.
- We demonstrate through quantitative experiments that our decomposition provides more faithful explanations than existing explanation methods.
- We show that our framework identifies clinically relevant, time-dependent biomarkers missed by existing methods, offering more actionable insights for clinicians.

## 2 Related Work

The drive to deploy predictive models in high-stakes clinical settings has led to a surge in research on explainable AI (XAI) for medical time series data [Tonekaboni et al., 2019, Topol, 2019]. The primary goal is to move beyond black-box predictions and provide clinicians with transparent, trustworthy, and actionable insights, thereby fostering adoption and facilitating model auditing. This need is particularly acute in sepsis prediction, where timely and interpretable predictions can directly impact patient outcomes.

The dominant paradigm for explaining time-series models relies on post-hoc feature attribution methods that generate saliency maps. Foundational techniques like LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017], and Integrated Gradients [Sundararajan et al., 2017] are commonly adapted to clinical time series including sepsis prediction, assigning an importance score to each feature at each timestep [Shickel et al., 2017, Lauritsen et al., 2020]. More recent work has sought to create methods tailored specifically for time series, such as TimeSHAP [Bento et al., 2021] or Dynamask [Crabbé and van der Schaar, 2020], which aim to produce more faithful explanations by considering the temporal nature of the data. Other approaches generate explanations through counterfactuals—identifying what minimal changes to an input sequence would alter the model's prediction [Goyal et al., 2021, Ismail and Günnemann, 2021].

However, a critical and unaddressed limitation unites these methods: they treat each feature-timepoint observation as an atomic unit. Consequently, the resulting attribution score—whether from a saliency map or a counterfactual—fundamentally conflates the importance of a feature's content (the 'what') with the importance of its temporal position (the 'when'). For instance, in sepsis prediction, a standard explanation cannot distinguish whether an elevated lactate reading is flagged because lactate is a clinically significant marker of sepsis or because the model has learned a spurious recency bias where any observation in the final timestep is overweighted [Jain and Wallace, 2019, Ismail and Günnemann, 2021]. This entanglement prevents a deeper audit of the model's temporal reasoning, which is crucial for sepsis where the timing of physiological changes carries diagnostic significance.

This limitation is particularly striking given that modern sequence models, like the Transformer and Mamba, explicitly separate content and position through distinct token and positional embeddings [Vaswani et al., 2017, Gu and Dao, 2024]. While the model's architecture maintains this separation—enabling it to learn both what features matter and when they matter—the explanation methods used to interpret them do not. This is especially problematic for sepsis onset prediction, which is fundamentally a temporal problem where understanding both the clinical markers and their temporal evolution is essential for meaningful interpretation.

## 3 Positional Explanation

Feature attribution is the dominant paradigm for interpreting model behavior, assigning an importance score to each input feature [Doshi-Velez and Kim, 2017]. Existing methods answer the question: "Which features contributed most to the model's prediction?" However, they conflate feature content and positional effects, making it impossible to separate a feature's semantic contribution from the effect of its position.

Formally, consider a model $f : \mathcal{X} \to \mathcal{Y}$ and a single input instance $x \in \mathcal{X}$. Each component $x_i$ of $x$ represents a specific feature of that input. An **explainer**, $g$, is a function that maps the model and input instance to an attribution vector:

$$\alpha = g(f, x) \in \mathbb{R}^d \tag{1}$$

where $d$ is the dimensionality of $x$. The entry $\alpha_i$ measures the combined influence of the $i$-th feature content $x_i$ and its position on the model's prediction $f(x)$.

As shown in Equation (1), existing explainer $g$ only requires $f, x$ as input, with no positional information. Consequently, existing methods cannot reveal positional effects. Perturbation-based (LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017]) and gradient-based methods (Integrated Gradients [Sundararajan et al., 2017]) attribute importance solely to feature values at fixed positions, while decomposition-based approaches (FullGrad [Srinivas and Fleuret, 2019]) assign relevance to features at their original locations. In all cases, attributions reflect feature influence only.

**Positional Explanation Framework.** We propose *Positional Explanation*, a framework to separate feature content and positional contributions. It is general and compatible with any attribution method.

Given feature $x \in \mathcal{X}$ and position $p \in \mathcal{P}$, the framework outputs

$$\alpha = g(f, x, p) \in \mathbb{R}^{2d}, \tag{2}$$

which decomposes as

$$\alpha = (\alpha^{(\text{feature})}, \alpha^{(\text{position})}), \quad \alpha^{(\text{feature})} \in \mathbb{R}^d, \alpha^{(\text{position})} \in \mathbb{R}^d. \tag{3}$$

As shown in Equation (3), our framework explainer $g$ requires $f, x, p$ as input, meaning we are also using $p$ to show the positional influence for the prediction. Figure 1 shows the example of highlighting $\alpha^{(\text{feature})}$ and $\alpha^{(\text{position})}$ for one example across timestamps. The interpretation of $\alpha^{(\text{feature})}$ and $\alpha^{(\text{position})}$ are as follows:

1. **Feature Content Attribution** ($\alpha_i^{(\textbf{feature})}$)**:** Measures the effect of perturbing $x_i$ while keeping $p_i$ fixed. Answers: *How important is the feature content itself, given its location?*
2. **Absolute Position Attribution** ($\alpha_i^{(\textbf{position})}$)**:** Measures the intrinsic value of $p_i$ by comparing contributions of $x_i$ at its original versus random positions. Answers: *How important is this location, independent of the feature content?*

**Positional-LIME as an Example.** To illustrate, consider LIME [Ribeiro et al., 2016]. Standard LIME generates perturbed samples

$$z = m \odot x \in \mathbb{R}^d, \quad m_i \sim \text{Bernoulli}(0.5), \tag{4}$$

Table 1: Performance of GPT2 and Mamba on the MC-MED and Physionet datasets. The models achieve sufficiently high predictive performance on sepsis prediction tasks, making them suitable for subsequent analysis and explanation.

| Dataset | Finetuned Model | Accuracy | F1 | AUC | AUPRC |
|---|---|---|---|---|---|
| PhysioNet [Reyna et al., 2020] | GPT-2 [Radford et al., 2019] | 0.8680 | 0.2048 | 0.7069 | 0.1802 |
| | Mamba [Gu and Dao, 2024] | 0.8930 | 0.0531 | 0.3509 | 0.0403 |
| MC-MED [Kansal et al., 2025] | GPT-2 [Radford et al., 2019] | 0.9490 | 0.1053 | 0.3536 | 0.0900 |
| | Mamba [Gu and Dao, 2024] | 0.8940 | 0.0536 | 0.3743 | 0.0443 |

where $m_i = 0$ zeros out $x_i$ and $m_i = 1$ retains it. LIME then fits a weighted linear model

$$\alpha = g(f, x) = w \in \mathbb{R}^d, \tag{5}$$

so that each $\alpha_i$ reflects the local effect of $x_i$ on $f(x)$.

In **Positional-LIME**, positions are treated as additional features. To avoid out-of-distribution issues from zeroing positional embeddings, we instead randomize them:

$$z = m \odot (x, p) \in \mathbb{R}^{2d}, \quad m_i \sim \text{Bernoulli}(0.5), \tag{6}$$

where $m_i = 0$ indicates that the feature $x_i$ is masked and the corresponding position $p_i$ is replaced with random positional embedding.

The resulting attributions

$$\alpha = g(f, x, p) = w \in \mathbb{R}^{2d} \tag{7}$$

can then be seprated into feature and positional contributions:

$$\alpha = (\alpha^{(\text{feature})}, \alpha^{(\text{position})}), \quad \alpha^{(\text{feature})} \in \mathbb{R}^d, \alpha^{(\text{position})} \in \mathbb{R}^d. \tag{8}$$

**Generalization to Other Explainers.** More generally, this framework extends to any attribution method (e.g., SHAP [Lundberg and Lee, 2017], Integrated Gradients [Sundararajan et al., 2017], FullGrad [Srinivas and Fleuret, 2019], MFABA [Zhu et al., 2024]). By computing $\alpha^{(\text{feature})}$ and $\alpha^{(\text{position})}$ separately, we separate feature content and positional contributions, providing a more fine-grained understanding of model predictions.

# 4 Experiments

We evaluated GPT-2 small (124M) [Radford et al., 2019] and Mamba-130M [Gu and Dao, 2024] on sepsis prediction tasks using the MC-MED [Kansal et al., 2025] and Physionet [Reyna et al., 2020] datasets. For each model, we used pre-trained, fine-tuned checkpoints provided by the CareBench benchmark [Choi et al., 2025] and assessed performance directly on the corresponding test sets.

Physionet is a widely used publicly available sepsis dataset containing only tabular EHR data, whereas MC-MED provides more comprehensive information, including ECG and respiratory waveforms, ventilator settings, medications, and per-minute vitals. Following the CareBench evaluation protocol [Choi et al., 2025], we adopted the benchmark's sepsis labeling criteria and cohort selection methodology, ensuring consistent preprocessing and evaluation conditions across both models and datasets.

Table 1 summarizes model performance across four metrics: Accuracy (Acc), F1-score (F1), Area Under the Receiver Operating Characteristic curve (AUC), and Area Under the Precision-Recall Curve (AUPRC). Both models achieved strong predictive performance, establishing them as suitable candidates for subsequent explanation analyses.

## 4.1 Faithfulness Test

We examine whether decomposing attributions into feature and positional components using our *Positional Explanation* framework improves explanation faithfulness in clinical settings. This decomposition enables differentiation between patients whose high risk stems from chronically abnormal lab values and those whose risk arises from sudden, recent changes, supporting more targeted clinical review.

Table 2: Insertion and deletion test results on the MC-MED and PhysioNet datasets using Positional-LIME for finetuned GPT-2 and Mamba models. The table reports Area Under the Curve (AUC) averaged over all examples. Using the feature component of Positional-LIME consistently outperforms feature-only attributions, and using the positional component consistently outperforms position-only attributions. This demonstrates that separating attributions into feature and positional components with our framework produces more faithful explanations.

(a) Insertion AUC (higher is better).

| Dataset | Model | Feature-only | Position-only | PE-Feature | PE-Position | PE-Full | Random |
|---|---|---|---|---|---|---|---|
| PhysioNet | GPT-2 | 0.354 | 0.323 | **0.419** | **0.396** | **0.465** | 0.214 |
| | Mamba | 0.347 | 0.331 | **0.392** | **0.401** | **0.454** | 0.213 |
| MC-MED | GPT-2 | 0.313 | 0.301 | **0.381** | **0.392** | **0.434** | 0.192 |
| | Mamba | 0.319 | 0.311 | **0.393** | **0.403** | **0.442** | 0.201 |

(b) Deletion AUC (lower is better).

| Dataset | Model | Feature-only | Position-only | PE-Feature | PE-Position | PE-Full | Random |
|---|---|---|---|---|---|---|---|
| PhysioNet | GPT-2 | 0.020 | 0.016 | **0.008** | **0.007** | **0.002** | 0.110 |
| | Mamba | 0.021 | 0.019 | **0.011** | **0.007** | **0.001** | 0.102 |
| MC-MED | GPT-2 | 0.007 | 0.032 | **0.006** | **0.011** | **0.005** | 0.226 |
| | Mamba | 0.072 | 0.113 | **0.066** | **0.053** | **0.045** | 0.199 |

To evaluate faithfulness, we conduct insertion and deletion tests and report average AUC scores. We compare six conditions: feature-only baseline, position-only baseline, PE-Feature (feature component from *Positional Explanation*), PE-Position (positional component from *Positional Explanation*), PE-Combined (both components from *Positional Explanation*), and a random baseline. Detailed descriptions of each approach are provided in Appendix B.

**Across Datasets and Models** Across datasets and models (further details on datasets and model setups are provided in Appendix A), PE-Feature consistently achieves higher insertion scores and lower deletion scores than Feature-only, while PE-Position achieves higher insertion and lower deletion scores than Position-only. Full insertion results are reported in Table 4a, and full deletion scores are reported in Table 4b. This demonstrates that separating feature and positional components results in more faithful attributions.

**Across Explainability Methods** We evaluate faithfulness across several explainability methods on the MC-MED dataset with GPT-2, comparing Feature-only (traditional perturbation), Position-only (position perturbation), PE-Feature (feature component of our *Positional Explanation*), and PE-Position (positional component of our *Positional Explanation*). The methods considered include LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017], Integrated Gradients [Sundararajan et al., 2017], FullGrad [Srinivas and Fleuret, 2019], and MFABA [Zhu et al., 2024] (see Appendix A.3 for details).

Although we show results here only for PhysioNet with GPT-2, the trend is consistent across all methods: PE-Feature achieves higher insertion and lower deletion scores than Feature-only, and PE-Position achieves higher insertion and lower deletion scores than Position-only. These results indicate that separating feature and positional components consistently produces more faithful explanations, independent of the underlying attribution method.

## 4.2 Independence Test

We assessed whether feature ($\alpha^{(\text{feature})}$) and positional ($\alpha^{(\text{position})}$) attributions are linearly related per measurement using the Pearson correlation coefficient. A high correlation magnitude indicates a strong linear relationship, whereas a low magnitude suggests independence. Statistical significance was evaluated using $p$-values, representing the likelihood that an observed correlation occurred by chance (see Appendix C.1.1 for computation details).

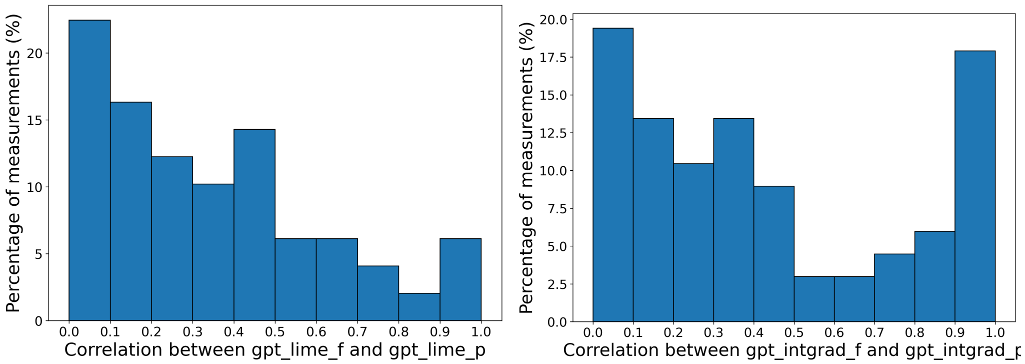
Table 3: Faithfulness comparison across explainability methods on PhysioNet using GPT-2. We report AUC for both insertion and deletion tests. Across methods, PE-Feature consistently outperforms Feature-only and PE-Position outperforms Position-only, showing that separating attributions into feature and positional components using our framework leads to more faithful explanations.

(a) Insertion AUC (higher is better).

| Explanation Method | Feature-only | Position-only | PE-Feature | PE-Position |
|---|---|---|---|---|
| LIME [Ribeiro et al., 2016] | 0.354 | 0.323 | **0.419** | **0.396** |
| SHAP [Lundberg and Lee, 2017] | 0.342 | 0.337 | **0.403** | **0.401** |
| Integrated Gradients [Sundararajan et al., 2017] | 0.361 | 0.346 | **0.427** | **0.412** |
| FullGrad [Srinivas and Fleuret, 2019] | 0.336 | 0.314 | **0.384** | **0.393** |
| MFABA [Zhu et al., 2024] | 0.351 | 0.325 | **0.417** | **0.402** |

(b) Deletion AUC (lower is better).

| Explanation Method | Feature-only | Position-only | PE-Feature | PE-Position |
|---|---|---|---|---|
| LIME [Ribeiro et al., 2016] | 0.020 | 0.016 | **0.008** | **0.007** |
| SHAP [Lundberg and Lee, 2017] | 0.019 | 0.018 | **0.007** | **0.008** |
| Integrated Gradients [Sundararajan et al., 2017] | 0.019 | 0.021 | **0.009** | **0.011** |
| FullGrad [Srinivas and Fleuret, 2019] | 0.017 | 0.019 | **0.010** | **0.010** |
| MFABA [Zhu et al., 2024] | 0.018 | 0.015 | **0.007** | **0.006** |



(a) Histogram of Correlations for measurements appreading more than 5 times using LIME

(b) Histogram of Correlations for measurements appreading more than 5 times using Integrated Gradients

Figure 2: Histogram of absolute correlation between feature ($f$) and positional ($p$) attribution per measurement. From these two histograms of LIME and Integrated Gradients, we observe that some measurements are inherently time-correlated while others are not, and these patterns differ across explanation methods.

Figure 2 shows the distribution of absolute correlation values across measurements. The results indicate variability in temporal dependence: some measurements strongly depend on time, while others are largely independent.

Examples of temporal correlation analysis of measurements in the MC-MED dataset using GPT-2 with Position-LIME:
**High temporal correlation:** LABPTT, GLOBULIN, WAM DIFTYP, TEMP
**Low temporal correlation:** AGE, RACE, AST (SGOT), PLATELET COUNT (PLT)

These findings suggest that static variables (e.g., demographics, baseline labs) are generally position-independent, whereas dynamic variables (e.g., coagulation tests, temperature) exhibit strong temporal dependence. Full correlation values and $p$-values are reported in Appendix C.1.2.

To validate our hypothesis that separating attribution into feature and positional components is helps identifying true temporal dependencies, we conduct an evaluation using a Large Language Model

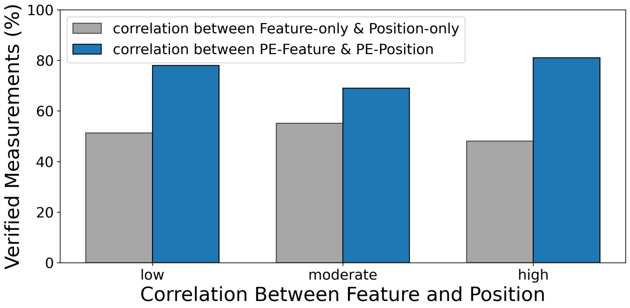


Figure 3: LLM verification accuracy for temporal correlation detection across different correlation bins. Our Positional Explanation (PE) framework, which correlates PE-Feature and PE-Position scores after separation, consistently outperforms the baseline method that directly correlates feature-only and position-only attributions. Higher verification accuracy across all bins demonstrates that decomposing attribution signals better identifies measurements with genuine temporal dependencies, helping clinicians distinguish between time-correlated and time-independent measurements.

(LLM) as a proxy for ground-truth verification. We compare two methods for measuring temporal correlations, with results presented in Figure 3.

The baseline uses correlation between feature-only and position-only attribution. We compare it to correlation between PE-Feature and PE-Position using our Positional Explanation framework. For evaluation, we group feature-position pairs into three bins based on their computed correlation scores: high correlation (correlation $> 0.7$), moderate correlation ($0.3 <$ correlation $\leq 0.7$), and low correlation (correlation $\leq 0.3$). Within each bin, we measure the LLM verification accuracy to assess how well our correlation scores align with LLM-verified temporal dependencies. The results show that our PE-based attribution consistently achieves higher verification rates across all correlation bins, demonstrating that separating the score improves the identification of features with genuine temporal effects and confirming the effectiveness of our method in detecting temporal correlations.

We also show qualitative result of what the LLM output for such correlation in Appendix C

### 4.3 Relevance Test

To evaluate the quality of feature attributions, we conducted a systematic comparison between traditional feature-only explanations and our Positional Explanation framework using LLM verification. We analyzed feature importance scores across clinical measurements to assess which method more accurately identifies clinically relevant features independent of temporal context.

For quantitative evaluation, we computed average feature importance scores across the entire dataset and organized features into three bins based on their attribution scores: high influence, moderate influence, and low influence. The top 10 most influential measurements identified using our framework include: `INFLUENZA B`, `NUR1373`, `ALBUMIN`, `POC16`, `KETONE: URINE (UA)`, `SARS-COV-2 RNA`, `MYCOPLASMA PNEUMONIAE`, `POC:POTASSIUM`, `POC:GLUCOSE BY METER`, `MAGNESIUM`.

Figure 4 presents the LLM verification results comparing feature-only attributions (original explanation method that perturbs only features) against PE-Feature scores from our Positional Explanation framework (which perturbs both features and positions before extracting the feature component). The results demonstrate that our PE-Feature approach consistently achieves higher LLM verification accuracy across all influence bins. This superior performance confirms that disentangling positional and feature effects produces more clinically meaningful feature attributions, enabling better identification of truly relevant measurements for clinical decision-making.

We also show qualitative result of what the LLM output for such feature was in Appendix D

## 5 Conclusion

We introduced the *Positional Explanation* framework, which separates standard feature attributions into *feature content* and *position* effects, enabling explanations that distinguish *what* drives a model's

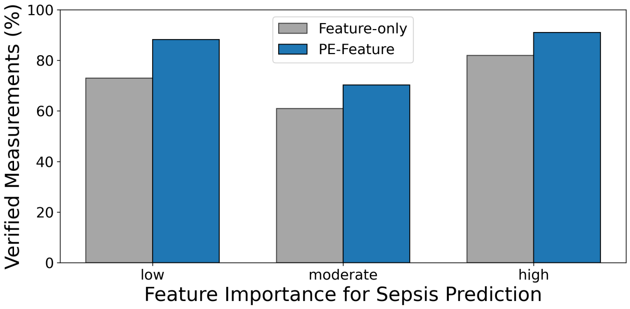


Figure 4: LLM verification accuracy for feature attribution methods. Our PE-Feature scores from the Positional Explanation framework achieve higher LLM verification accuracy compared to feature-only attributions, demonstrating improved feature attribution quality through separation.

prediction from *when* it is clinically significant. Applied to Mamba and GPT-2 models finetuned for sepsis prediction on PhysioNet and MC-MED datasets, our approach provides more faithful, temporally aware explanations than existing explanation methods, and identifies clinically relevant, time-dependent biomarkers that are otherwise overlooked. Importantly, Positional Explanation is model- and method-agnostic and can be applied to any feature attribution framework for any types of data including image, text, and time-series.

While these results are promising, broader clinical validation is necessary. Current evaluation relies primarily on LLM-based models. We will engage multiple clinicians specialized in sepsis to evaluate real-world interpretability, trust, and utility. To demonstrate generality, we plan to extend the framework to new models, develop scalable metrics for temporal effects, and integrate it into clinical decision support systems for timely, actionable alerts.

Overall, Positional Explanation provides a general, flexible framework for temporally aware explainability in clinical predictive modeling, bridging the gap between accurate prediction and actionable, time-sensitive insight.

# References

Roy Adams, Kevin E. Henry, Anoop Sridharan, Heather Soleimani, Karandeep A. Zell, Chuan S. L. Tan, Jenna N. Wiens, Craig E. V. Barton, and Karandeep A. Singh. Prospective, multi-site study of a deep learning model for early detection of sepsis. *Nature Medicine*, 28(8):1649–1654, 2022. doi: 10.1038/s41591-022-01894-0.

João Bento, Pedro Saleiro, Pedro Bizarro, and Mário A T Oliveira. Timeshap: Explaining recurrent models through time. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 336–345. IEEE, 2021.

Sherali Bomrah, Mohy Uddin, Umashankar Upadhyay, Jyoti Priya, Eshita Dhar, Shih-Chang Hsu, and Shabir Syed-Abdul. A scoping review of machine learning for sepsis prediction- feature engineering strategies and model performance: a step towards explainability. *Critical Care*, 28: 180, 2024.

Seewon Choi, Mayank Keoliya, Rajeev Alur, Mayur Naik, and Eric Wong. Carebench: Stable prediction of adverse events in medical time-series data, 2025.

Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1236–1247. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/084a0c2053618953a0a65261394338d3-Paper.pdf`.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

Yash Goyal, Been Kim Wu, and Joachim Ernst. Counterfactual explanations for time-series models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1496–1508. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/0e64a7b12e3472038596519183 8b08cd-Paper.pdf`.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL `https://arxiv.org/abs/2312.00752`.

Aaqib Ismail and Stephan Günnemann. Benchmarking deep learning interpretability in time series predictions. *Advances in Neural Information Processing Systems*, 34:23605–23618, 2021.

Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019. URL `https://arxiv.org/abs/1902.10186`.

Anshul Kansal, Eric Chen, Billy T. Jin, et al. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12:1094, 2025. doi: 10.1038/s41597-025-05419-5.

Simon Meyer Lauritsen, Martin Kristensen, Mads Vincent Olsen, and Michael Stig Larsen. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1):3852, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 4768–4777, 2017.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48 (2):210–217, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Christopher W Seymour, Vincent X Liu, Theodore J Iwashyna, Frank M Brunkhorst, Thomas D Rea, André Scherag, Gordon Rubenfeld, Jeremy M Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S Deutschman, Gabriel J Escobar, and Derek C Angus. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):762–774, 2016. doi: 10.1001/jama.2016.0288.

Supreeth P. Shashikumar, Gabriel Wardi, Atul Malhotra, and Shamim Nemati. Artificial intelligence sepsis prediction algorithm learns to say “i don’t know”. *npj Digital Medicine*, 4(1):134, 2021. doi: 10.1038/s41746-021-00504-6.

Benjamin Shickel, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning for electronic health records. In *IEEE journal of biomedical and health informatics*, volume 22, pages 1589–1604. IEEE, 2017.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems 32*, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017.

Sana Tonekaboni, Shalmali Joshi, Michael D McCradden, and Anna Goldenberg. What clinicians want: a survey of explainable ai needs for clinical decision support. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, J. Michael McCullough, Olivia DeTroyer-Cooley, Jennifer Pestrue, M. Elizabeth Phillips, Justin Konye, Patrick J. Schulte, Mihir A. Kora, Dmitriy A. Dligach, and Majid Afshar. External validation of a widely implemented commercial sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8): 1065–1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

Kuo-Ching Yuan, Lung-Wen Tsai, Ko-Han Lee, Yi-Wei Cheng, Shou-Chieh Hsu, Yu-Sheng Lo, and Ray-Jade Chen. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International Journal of Medical Informatics*, 141:104176, 2020.

Jia-Liang Zhu, Shi-Qi Yuan, Tao Huang, Lu-Ming Zhang, Xiao-Mei Xu, Hai-Yan Yin, Jian-Rui Wei, and Jun Lyu. Influence of systolic blood pressure trajectory on in-hospital mortality in patients with sepsis. *BMC Infectious Diseases*, 23(1):90, 2023.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Minhui Xue, Dongxiao Zhu, and Kim-Kwang Raymond Choo. Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):17228–17236, 2024. doi: 10.1609/aaai.v38i15.29669.

# A Experimental Setup

## A.1 Dataset Description

### A.1.1 Datasets

We utilize sepsis prediction datasets curated by CAREBench [Choi et al., 2025], which processes two publicly available datasets: PhysioNet 2019 [Reyna et al., 2020] and MC-MED [Kansal et al., 2025].

**PhysioNet 2019** comprises over 40,000 ICU patients with up to 40 clinical variables recorded hourly, totaling 2.5 million hourly time windows. The dataset includes vital signs, laboratory values, and demographics in tabular format without physiological waveforms.

**MC-MED** contains 118,385 emergency department visits from 70,545 unique patients (2020–2022). This dataset uniquely combines minute-level vital signs and continuous physiological waveforms (ECG, photoplethysmogram, respiration) with comprehensive clinical data including demographics, medical histories, medications, and laboratory results.

### A.1.2 Sepsis Prediction Task Curation

CAREBench adapted the curation methodology to each dataset's clinical setting and available data.

**PhysioNet 2019:** Sepsis labels were pre-defined using Sepsis-3 criteria, requiring both clinical suspicion of infection (blood culture or IV antibiotic orders) and a two-point SOFA score change.

**MC-MED:** CAREBench implemented a two-stage process:

1. **At-Risk Cohort Selection** – Patients meeting all criteria:
   - Admission source of ED
   - Temperature $< 36°C$ or $> 38.5°C$ within 24 hours of admission (`Temp_time`)
   - At least one of the following within 24 hours of admission:
     - WBC Count $> 12K$ or $< 4K/\mu L$ (`WBC_time`)
     - HR $> 90$ bpm (`HR_time`)
     - RR $> 20$ (`RR_time`)
   - At least 1 of the `WBC_time`, `HR_time`, `RR_time` within 12 hours of `Temp_time`
   - No intravenous antibiotic at or before the time of the first criteria met
2. **Sepsis Labeling** – Adapted Sepsis-3 definition for ED settings with $h = 1.5$ hour prediction horizon. Positive labels assigned when emergency SOFA (eSOFA) criteria met:
   - Presumed serious infection:
     - Blood culture obtained (regardless of the results)
     - $\geq 4$ QADs starting within $\pm$ 2 days of `blood_culture_day`
   - Any 1 of below within $\pm$ 2 days of `blood_culture_day` (acute organ dysfunction):
     - Vasopressor initiation
     - Initiation of mechanical ventilation
     - Doubling in serum creatinine level or decrease by $\geq 50\%$ of eGFR (excluding patients with end-stage kidney disease [585.6])
     - Total bilirubin level $\geq$ 2.0mg/dL and doubling
     - Platelet count $<$ 100 cells/$\mu$L and $\geq$ 50% decline from baseline (excluding baseline $<$ 100 cells/$\mu$L)
     - Serum lactate $\geq$ 2.0 mmol/L

## A.2 Model Description

We employed GPT-2 (124M parameters) [Radford et al., 2019] and Mamba-130M [Gu and Dao, 2024], pre-trained language models fine-tuned for sepsis prediction using the CAREBench-curated datasets.

### A.2.1 Model Architectures

**GPT-2 Small:** A 124M parameter decoder-only transformer with 12 layers, 768 hidden dimensions, and 12 attention heads. Its autoregressive architecture with causal self-attention naturally captures temporal dependencies in patient trajectories, leveraging pre-trained sequential representations for modeling physiological progression patterns.

**Mamba-130M:** A 130M parameter state-space model addressing transformer limitations in long-sequence processing. Its selective state-space mechanism achieves linear complexity with sequence length, enabling efficient processing of extended patient histories. The architecture's continuous-time formulation aligns naturally with physiological processes, offering advantageous inductive biases for modeling sepsis dynamics.

### A.2.2 Training Configuration

Following CAREBench methodology:

- **Custom Tokenization:** Dataset-specific tokenizers handle hospital-specific medical codes and limited vocabulary
- **Training Duration:** 100 epochs ensuring convergence on limited medical data
- **Hyperparameter Selection:** Learning rate $\in \{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$ via validation performance

This configuration enables effective adaptation from general language understanding to domain-specific temporal patterns and medical terminology in sepsis prediction.

### A.3 Explanation Methods

This section briefly describes the explanation methods employed in conjunction with our Positional Explanation approach.

- **LIME (Local Interpretable Model-agnostic Explanations)** [Ribeiro et al., 2016] generates local explanations for individual predictions by fitting an interpretable surrogate model (typically linear) within the neighborhood of the target instance. The method creates perturbations around the input sample and trains the surrogate model on these variations, with samples weighted by their proximity to the original instance.
- **SHAP (SHapley Additive exPlanations)** [Lundberg and Lee, 2017] computes feature importance scores based on cooperative game theory principles. Each feature receives an attribution value representing its marginal contribution to the prediction relative to a baseline, with the property that all attribution values sum to the difference between the model's output and the baseline prediction.
- **Integrated Gradients (IntGrad)** [Sundararajan et al., 2017] computes feature attributions by integrating gradients along a linear path from a baseline input to the target input. This path integral approach ensures satisfaction of fundamental attribution axioms, including sensitivity and implementation invariance.
- **FullGrad** [Srinivas and Fleuret, 2019] extends standard gradient-based attribution by incorporating gradient information from all network layers. The method aggregates input gradients with bias gradients across all intermediate representations, providing more comprehensive attribution maps that capture multi-layer feature interactions.
- **MFABA (More Faithful and Accelerated Boundary-based Attribution)** [Zhu et al., 2024] computes attributions by constructing paths from input samples to adversarial examples that cross the model's decision boundary. The method employs second-order Taylor approximations to better model loss function changes during gradient ascent optimization.

## B Additional Faithfulness Test Results

This section presents comprehensive results from our insertion and deletion experiments across all experimental configurations. We systematically evaluate faithfulness across two datasets (PhysioNet and MC-MED), two transformer architectures (GPT-2 and Mamba), and five explanation methods (LIME, SHAP, Integrated Gradients, FullGrad, MFABA).

### B.1 Faithfulness Test Experimental Setup

For each explanation method, we compare five attribution approaches:

- **Feature-only**: Traditional perturbation-based explanations
- **Position-only**: Positional explanation perturbing only position
- **PE-Feature**: Feature component of our *Positional Explanation* framework
- **PE-Position**: Position component of our *Positional Explanation* framework
- **PE-Full**: Both feature and position components of our *Positional Explanation* framework
- **Random**: Baseline for comparison

We employ two complementary faithfulness metrics: insertion tests (where higher AUC indicates better faithfulness) and deletion tests (where lower AUC indicates better faithfulness).

### B.2 Key Findings

The results demonstrate consistent improvements in explanation faithfulness when separating positional and feature components:

**Insertion Test Performance.** Our positional explanation components (PE-Feature and PE-Position) consistently outperform their traditional counterparts (Feature-only and Position-only) across all experimental configurations. PE-Feature achieves higher AUC scores than Feature-only, while PE-Position surpasses Position-only, indicating more faithful identification of important features.

**Deletion Test Performance.** The superiority of our approach is further confirmed in deletion tests, where PE-Feature consistently achieves lower AUC scores than Feature-only, and PE-Position outperforms Position-only. Lower scores in deletion tests indicate that removing highly-attributed features causes greater performance degradation, confirming these features are indeed more important for model predictions.

**Cross-Architecture and Cross-Method Consistency.** The improvements hold across both GPT-2 and Mamba architectures, as well as different explanation methods including gradient-based attribution, attention-based explanations, and perturbation-based approaches, demonstrating the broad generalizability of our positional explanation approach.

### B.3 Detailed Results

Tables 4a and 4b present the complete faithfulness evaluation results across all experimental configurations. The insertion test results demonstrate the ability of each method to identify truly important features, while the deletion test results show how effectively each method identifies features whose removal significantly impacts model performance. These comprehensive results validate our theoretical framework and demonstrate the practical benefits of separating positional and feature attributions in transformer explanations.

## C Additional Independence Test Results

### C.1 Independence Test Analysis

This section presents the complete results from our independence test analysis, expanding on the verification scores reported in Section 4.2.

#### C.1.1 Measurements

The correlation was measured using the Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(\alpha_i^{(\text{feature})} - \overline{\alpha^{(\text{feature})}})(\alpha_i^{(\text{position})} - \overline{\alpha^{(\text{position})}})}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(\text{feature})} - \overline{\alpha^{(\text{feature})}})^2}\sqrt{\sum_{i=1}^{n}(\alpha_i^{(\text{position})} - \overline{\alpha^{(\text{position})}})^2}}, \tag{9}$$

where $r \in [-1, 1]$, $\overline{\alpha^{(\text{feature})}}$ is the mean feature attribution, and $\overline{\alpha^{(\text{position})}}$ is the mean positional attribution. Values of $r$ close to $1$ or $-1$ indicate strong positive or negative correlation, while values near $0$ suggest little to no linear relationship.

To assess statistical significance, we tested the null hypothesis:

$$H_0 : r = 0 \quad \text{(feature and positional attributions are uncorrelated).}$$

The corresponding $p$-value quantifies the probability of observing a correlation at least as extreme as the measured $r$ under $H_0$. At the $\alpha = 0.05$ significance level,

- If $p < 0.05$: we reject $H_0$, concluding significant correlation.
- If $p \geq 0.05$: we fail to reject $H_0$, finding no clear evidence of correlation.

#### C.1.2 Temporal Correlation Patterns

Our analysis identified distinct patterns in temporal correlation across different medical measurements:

**Examples of independent features** (low correlation, high $p$-value) using Positional-LIME on the MC-MED dataset with GPT-2 included:

- AGE: correlation = 0.0244, $p$ = 0.9020
- RACE: correlation = -0.0330, $p$ = 0.8675
- AST (SGOT): correlation = -0.0082, $p$ = 0.9668
- PLATELET COUNT (PLT): correlation = 0.0301, $p$ = 0.8888

**Examples of independent features** (low correlation, high $p$-value) using Positional-LIME on the MC-MED dataset with GPT-2 included:

Table 4: Our Positional Explanation (PE) framework consistently outperforms traditional attribution methods. PE-Feature and PE-Position achieve higher insertion AUC and lower deletion AUC than their Feature-only and Position-only counterparts, confirming more faithful identification of important features. The improvements hold across both GPT-2 and Mamba architectures and multiple explanation methods. PE = Positional Explanation, Feat = Feature, Pos = Position.

(a) Insertion test results (AUC). Higher values indicate more faithful performance.

| Dataset | Model | Explanation | Feat-only | Pos-only | PE-Feat | PE-Pos | PE-Full | Random |
|---|---|---|---|---|---|---|---|---|
| PhysioNet | GPT-2 | LIME | 0.354 | 0.323 | 0.419 | 0.396 | 0.465 | 0.214 |
| | | SHAP | 0.342 | 0.337 | 0.403 | 0.401 | 0.452 | 0.209 |
| | | IntGrad | 0.361 | 0.346 | 0.427 | 0.412 | 0.478 | 0.221 |
| | | FullGrad | 0.336 | 0.314 | 0.384 | 0.393 | 0.443 | 0.215 |
| | | MFABA | 0.351 | 0.325 | 0.417 | 0.402 | 0.461 | 0.208 |
| | Mamba | LIME | 0.347 | 0.331 | 0.392 | 0.401 | 0.454 | 0.213 |
| | | SHAP | 0.352 | 0.323 | 0.415 | 0.395 | 0.463 | 0.207 |
| | | IntGrad | 0.364 | 0.348 | 0.431 | 0.416 | 0.472 | 0.226 |
| | | FullGrad | 0.338 | 0.312 | 0.393 | 0.382 | 0.445 | 0.218 |
| | | MFABA | 0.353 | 0.334 | 0.422 | 0.404 | 0.460 | 0.202 |
| MC-MED | GPT-2 | LIME | 0.313 | 0.301 | 0.381 | 0.392 | 0.434 | 0.192 |
| | | SHAP | 0.321 | 0.314 | 0.392 | 0.403 | 0.446 | 0.207 |
| | | IntGrad | 0.332 | 0.322 | 0.413 | 0.421 | 0.461 | 0.215 |
| | | FullGrad | 0.303 | 0.296 | 0.375 | 0.384 | 0.421 | 0.194 |
| | | MFABA | 0.324 | 0.312 | 0.401 | 0.395 | 0.452 | 0.203 |
| | Mamba | LIME | 0.319 | 0.311 | 0.393 | 0.403 | 0.442 | 0.201 |
| | | SHAP | 0.331 | 0.322 | 0.414 | 0.411 | 0.451 | 0.214 |
| | | IntGrad | 0.339 | 0.336 | 0.421 | 0.432 | 0.463 | 0.223 |
| | | FullGrad | 0.312 | 0.303 | 0.382 | 0.391 | 0.433 | 0.208 |
| | | MFABA | 0.330 | 0.321 | 0.412 | 0.410 | 0.450 | 0.212 |

(b) Deletion test results (AUC). Lower values indicate more faithful performance.

| Dataset | Model | Explanation | Feat-only | Pos-only | PE-Feat | PE-Pos | PE-Full | Random |
|---|---|---|---|---|---|---|---|---|
| PhysioNet | GPT-2 | LIME | 0.020 | 0.016 | 0.008 | 0.007 | 0.002 | 0.110 |
| | | SHAP | 0.019 | 0.018 | 0.007 | 0.008 | 0.003 | 0.102 |
| | | IntGrad | 0.019 | 0.021 | 0.009 | 0.011 | 0.005 | 0.112 |
| | | FullGrad | 0.017 | 0.019 | 0.010 | 0.010 | 0.004 | 0.111 |
| | | MFABA | 0.018 | 0.015 | 0.007 | 0.006 | 0.002 | 0.109 |
| | Mamba | LIME | 0.021 | 0.019 | 0.011 | 0.007 | 0.001 | 0.102 |
| | | SHAP | 0.021 | 0.018 | 0.012 | 0.007 | 0.002 | 0.103 |
| | | IntGrad | 0.011 | 0.011 | 0.008 | 0.009 | 0.003 | 0.111 |
| | | FullGrad | 0.010 | 0.012 | 0.007 | 0.008 | 0.001 | 0.113 |
| | | MFABA | 0.020 | 0.017 | 0.010 | 0.006 | 0.001 | 0.100 |
| MC-MED | GPT-2 | LIME | 0.007 | 0.032 | 0.006 | 0.011 | 0.005 | 0.226 |
| | | SHAP | 0.011 | 0.024 | 0.010 | 0.013 | 0.006 | 0.228 |
| | | IntGrad | 0.053 | 0.103 | 0.037 | 0.077 | 0.035 | 0.218 |
| | | FullGrad | 0.049 | 0.098 | 0.036 | 0.064 | 0.032 | 0.220 |
| | | MFABA | 0.022 | 0.030 | 0.014 | 0.012 | 0.015 | 0.225 |
| | Mamba | LIME | 0.072 | 0.113 | 0.066 | 0.053 | 0.045 | 0.199 |
| | | SHAP | 0.085 | 0.116 | 0.058 | 0.056 | 0.046 | 0.201 |
| | | IntGrad | 0.089 | 0.102 | 0.047 | 0.057 | 0.033 | 0.204 |
| | | FullGrad | 0.086 | 0.098 | 0.046 | 0.055 | 0.031 | 0.206 |
| | | MFABA | 0.071 | 0.111 | 0.065 | 0.052 | 0.044 | 0.200 |

- LABPTT: correlation = 0.9080, $p$ = 1.11e-04
- GLOBULIN: correlation = 0.9061, $p$ = 3.90e-05
- WAM DIFTYP: correlation = 0.8447, $p$ = 2.15e-03
- TEMP: correlation = 0.6659, $p$ = 3.55e-02

**Gemini 2.5 pro**

**Measurements with High Temporal Correlation**

- **LABPTT:** This is a measure of blood clotting time, often monitored frequently for patients on anticoagulant therapy (like heparin). The value at time t is highly predictive of the value at t+Δt because therapeutic adjustments are made based on current levels, creating a continuous, closely tracked time-series.
- **GLOBULIN:** Globulin levels in the blood change in response to underlying physiological processes (e.g., inflammation, liver function) that evolve over hours to days. This creates a relatively smooth, slow-moving signal where consecutive measurements are strongly related.
- **WAM DIFTYP:** This probably refers to a component of the white blood cell count. These counts rise and fall in response to events like infection. The biological process is not instantaneous, leading to strong temporal dependence in measurements taken hours or days apart.
- **TEMP:** Body temperature is a classic example of a high-frequency physiological signal. It has inherent patterns (circadian rhythm) and responds dynamically to illness (fever). A measurement at one moment is a very strong predictor of the temperature a few minutes or even an hour later.

**Measurements with Low Temporal Correlation**

- **AGE:** Age is a static or near-static feature within the context of most data collection periods (e.g., a hospital stay). Since it doesn't change, it has no dynamic temporal pattern for a model to learn from. From a signal processing standpoint, a constant signal has perfect autocorrelation, but from a feature importance perspective, its lack of change means it offers no evolving temporal information. Thus, classifying it as "low" makes sense in this context.
- **RACE:** This is a static demographic feature. It is constant for an individual and has no temporal variability.
- **AST (SGOT):** AST is a liver enzyme that changes in response to liver injury or stress. These changes typically occur over hours to days, creating a clinically meaningful trend. While it doesn't fluctuate as rapidly as temperature, a measurement today is certainly informative for predicting the measurement tomorrow. It's not a static value and its trend is often a critical diagnostic signal.
- **PLATELET COUNT (PLT):** Platelet levels are dynamic and change over days in response to bleeding, infection, or bone marrow conditions. The trajectory of platelet counts is a crucial clinical indicator. The value is not static and exhibits significant temporal dependence, even if it's less "high-frequency" than vital signs.

Figure 5: Feature temporal correlation analysis for sepsis prediction using Gemini 2.5 pro. Features are categorized into High, Moderate, and Low time-correlated groups, reflecting progression markers, intermediate trends, and condition markers, respectively, as verified using the prompt shown in fig. 9.

These findings suggest that certain static variables (e.g., demographics, baseline labs) remain position-independent, while dynamic variables (e.g., coagulation tests, temperature) exhibit strong temporal dependence.

### C.2 LLM Verification Results

To validate our independence test findings, we employed large language models for additional verification. The detailed outputs from both Gemini and GPT models are presented below.

**LLM Analysis.** Figures 5 and 6 show the comprehensive LLM evaluation results for feature independence assessment.

## D Additional Relevance Test Results

We analyzed which clinical measurements contribute most to model predictions, independent of temporal context. For each measurement, we computed the average feature importance across the entire dataset and identified the top 10 most influential features:

`INFLUENZA B, NUR1373, ALBUMIN, POC16, KETONE: URINE (UA), SARS-COV-2 RNA, MYCOPLASMA PNEUMONIAE, POC:POTASSIUM, POC:GLUCOSE BY METER, MAGNESIUM.`

**LLM Analysis.** We show the llm outputs for feature relevance test in Figure 7 and 8.

**GPT-5**

**Measurements with High Temporal Correlation**

- **LABPTT:** Typically shows continuous adjustment (e.g., heparin titration), so strong serial dependence.
- **GLOBULIN:** Usually slow drift, limited short-term serial structure.
- **WAM DIFTYP:** Infections/inflammation often produce evolving patterns (neutrophil/lymphocyte shifts) rather than single jumps.
- **TEMP:** Shows smooth-ish rises/falls and circadian/fever cycles.

**Measurements with Low Temporal Correlation**

- **AGE:** Static.
- **RACE:** Static.
- **AST (SGOT):** Often flat until an insult, then spikes—step-like, so low smooth serial correlation.
- **PLATELET COUNT (PLT):** Frequently stable then precipitous drop; jumpy/change-point behavior. Low smooth correlation.

Figure 6: Feature temporal correlation analysis for sepsis prediction using GPT 5. Features are categorized into High, Moderate, and Low time-correlated groups, reflecting progression markers, intermediate trends, and condition markers, respectively, as verified using the prompt shown in fig. 9.

**Gemini 2.5 pro**

**Pathogens and Infectious Triggers** Sepsis is fundamentally a dysregulated host response to an infection. Therefore, identifying the causative pathogen is a powerful predictor.

- **SARS-COV-2 RNA, INFLUENZA B, and MYCOPLASMA PNEUMONIAE:** These directly identify the infectious agent. Severe viral infections (like COVID-19 and Influenza) can lead to viral sepsis or predispose a patient to a secondary bacterial infection. Atypical bacteria like *Mycoplasma pneumoniae* are common causes of pneumonia, which is a leading trigger for sepsis.

**Core Biomarkers of Systemic Illness** These lab values reflect the widespread physiological chaos caused by the septic response, such as inflammation, metabolic derangement, and organ dysfunction.

- **ALBUMIN:** Low albumin (hypoalbuminemia) is a classic and very strong predictor of poor outcomes in sepsis. It reflects the severity of the systemic inflammatory response and a state of negative protein balance.
- **POC:GLUCOSE BY METER:** High glucose (stress hyperglycemia) is extremely common in critically ill patients. It's a result of the body's stress response and is associated with worse outcomes in sepsis.
- **POC:POTASSIUM and MAGNESIUM:** Electrolyte abnormalities are hallmarks of organ dysfunction in sepsis, particularly Acute Kidney Injury (AKI).
- **KETONE: URINE (UA):** The presence of ketones can indicate metabolic acidosis, a sign of severe metabolic stress and shock.

Figure 7: Feature influence analysis on sepsis prediction using Gemini 2.5 Pro. Features are categorized into High, Moderate, and Low influence based on their impact on model predictions, as verified using the prompt shown in fig. 10.

## E Prompt

We provide the prompts used to verify our model analyses using large language models (LLMs), specifically Gemini 2.5 Pro and GPT-5 Mini.

**Independence Test Prompt.** This prompt is designed to verify the results of the independence test, assessing whether features are statistically independent. The full prompt is shown in fig. 9.

**Relevance Test Prompt.** This prompt is used to verify the results regarding the influence of each feature, evaluating their impact on the model's predictions. The full prompt is shown in fig. 10.

**GPT-5 mini**

**Features with Established Importance**

- **Influenza B:** While specific studies on Influenza B's role in sepsis prediction are limited, respiratory viral infections, including influenza, have been associated with sepsis. Differentiating between influenza and Mycoplasma pneumoniae infections using peripheral blood routine parameters has shown diagnostic value, with an AUC of 0.845.
- **Albumin:** Hypoalbuminemia is a recognized marker of systemic inflammation and poor nutritional status, both relevant in sepsis. Low albumin levels have been linked to increased mortality in sepsis patients.
- **SARS-CoV-2 RNA:** COVID-19 has been associated with sepsis in some patients. Laboratory abnormalities in COVID-19 patients, particularly in coagulation function, have been related to the occurrence of sepsis.
- **Mycoplasma pneumoniae:** This pathogen can cause respiratory infections that may progress to sepsis. Studies differentiating influenza and Mycoplasma pneumoniae infections show its diagnostic relevance.
- **POC: Potassium:** Electrolyte imbalances, including abnormal potassium levels, are common in sepsis and can impact cardiac and muscular function.
- **POC: Glucose by Meter:** Hyperglycemia is frequently observed in sepsis and is associated with worse outcomes. Elevated blood glucose levels are primary risk factors for poor outcomes in various conditions.
- **Magnesium:** Hypomagnesemia has been linked to increased mortality in sepsis patients, suggesting its potential role as a prognostic marker.

**Features with Limited or Unclear Importance**

- **Ketone: Urine (UA):** While ketonuria can indicate metabolic disturbances such as diabetic ketoacidosis, its role in sepsis prediction is not well-established in the literature.

Figure 8: Feature influence analysis on sepsis prediction using ChatGPT-5 Mini. Features are categorized into High, Moderate, and Low influence based on their impact on model predictions, as verified using the prompt shown in fig. 10.

**Prompt**

You will be provided with results from our explainability method, which categorizes features based on their temporal correlation into three groups: High Time-Correlated Features, Moderate Time-Correlated Features, and Low Time-Correlated Features.
For each feature:

- Indicate whether you agree that the feature belongs in its assigned temporal correlation group.
- Briefly justify your agreement or disagreement based on reasoning about temporal patterns.

Here are the feature groups:

Figure 9: Prompt template for verifying feature temporal correlation group assignment.

**Prompt**

You are an expert in sepsis prediction. We have categorized features based on their impact on sepsis prediction into High, Moderate, and Low influence.
For each feature:

- Indicate whether you agree with the feature's assigned impact group.
- Briefly justify your agreement or disagreement based on reasoning about its role in sepsis prediction.

Here are the features:

Figure 10: Prompt template for verifying feature influence on sepsis prediction.