FAILURE MODES OF TIME SERIES INTERPRETABILITY ALGORITHMS FOR CRITICAL CARE APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability is a crucial aspect of deploying deep learning models in critical care, especially in constantly evolving conditions that influence patient survival. However, common interpretability algorithms face unique challenges when applied to dynamic prediction tasks, where patient trajectories evolve over time. Gradient, Occlusion, and Permutation-based methods often struggle with time-varying target dependency and temporal smoothness. This paper systematically analyzes these failure modes and supports learnable mask-based interpretability frameworks as alternatives, which can incorporate temporal continuity and label consistency constraints to learn feature importance over time. We argue that learnable mask-based approaches for dynamic time-series prediction problems provide more reliable and consistent interpretations for applications in critical care and similar domains.

1 INTRODUCTION

Interpretability techniques are essential in high-stakes, resource-constrained environments such as critical care medicine to ensure model-derived interpretations align with temporal dynamics of the clinical trajectories. While deep learning models detect subtle temporal patterns, their clinical utility depends on interpretations that map predictions to pathophysiology (Dey et al., 2022). Time-series interpretability techniques must not only attribute feature importance but also contextualize when and why specific features matter, such as a blood pressure drop preceding cardiac arrest. Conventional interpretability approaches may oversimplify the temporal granularity required in critical care, attributing importance to isolated time points rather than evolving physiological contexts. This mismatch risks confusing treatment artifacts with genuine deterioration in patient health (Stiglic et al., 2020), highlighting the need for temporally coherent interpretability methods.

Among conventional methods, Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017) and GradientSHAP (Lundberg et al., 2018) are foundational approaches that leverage gradients to attribute importance to features. However, gradient-based methods often struggle to capture temporal dependencies inherent in sequential data (Srinivas & Fleuret, 2020), leading to adaptations like Temporal Integrated Gradients (TIG) (Enguehard, 2023c) and Sequential Integrated Gradients (Enguehard, 2023b).

Complementing gradient-based methods are perturbation-based approaches, which assess feature importance by altering parts of the input and observing the corresponding impact on predictions. Feature Occlusion (FO) (Suresh et al., 2017) and Augmented Temporal FO (Tonekaboni et al., 2020) are perturbation-based methods that rely on non-learnable input masking strategies or feature removal without dynamically adjusting to the data. In contrast, learnable mask-based methods such as DynaMask (Crabbé & Van Der Schaar, 2021) and ExtremalMask (Enguehard, 2023a) optimize the masking process to either preserve or perturb key features while optimizing temporal continuity and label consistency constraints. Feature Importance in Time (FIT) (Tonekaboni et al., 2020), a method specifically developed for time-series interpretability offers a robust framework for quantifying shifts in predictive distributions and understanding temporal feature importance.

Furthermore, certain model architectures enable inherent interpretability by identifying feature saliency, such as RETAIN (Choi et al., 2016), by directly learning attention weights to provide feature attributions. However, reliance on attention mechanisms can lead to inconsistencies in interpretations as supported by natural language processing literature (Jain & Wallace, 2019).

Existing methods for evaluating time-series interpretability have focused on static classification tasks, where a prediction is made only after observing the entire sequence, typically resulting in binary or multi-class output (e.g., EEG Classification, Patient Mortality Prediction). In contrast, this work extends time-series interpretability methods to dynamic prediction tasks, such as predicting acute organ failure, where predictions are generated at each time step, enabling insights into varying patient states during their stay in Intensive Care Units (ICU). In this paper, we present:

i. A systematic analysis of failure modes in time-series interpretability algorithms for dynamic prediction, focusing on target selection, attribution aggregation, and temporal smoothness.

ii. Empirical evidence supporting learnable mask-based frameworks to address the failure modes through optimization with time-series specific constraints.

2 METHODS

2.1 DATASET DESCRIPTION AND PREDICTION TASK

We focused on Circulatory Failure, a leading cause of morbidity and mortality in critical care settings, as the use case for this work (MEMBERS et al., 2023). We utilized the Dynamic Circulatory Failure[1] Prediction task from the HiRID-ICU benchmark study (Yèche et al., 2021), which involves dynamic binary prediction throughout the patient's ICU stay. Specifically, it continuously predicts the onset of circulatory failure within the next 12 hours, provided the patient is not already in organ failure. Detailed information on the distribution of the data set, including the number of ICU stay records in the training, validation, and test sets, and the number of prediction samples, is provided in Appendix A.1. This multivariate time-series dataset spans 2016 time steps at 5-minute intervals, totaling 7 days, and includes 231 clinical features such as vital signs, hemodynamic data, treatments, pathological laboratory values, and ventilation parameters for critical care management.

2.2 MODEL ARCHITECTURES AND TRAINING

We replicated the study by (Yèche et al., 2021), which used the standard transformer encoder architecture (1.64 Million params) (Vaswani, 2017) and extended their approach by utilizing an encoder-only variant of the CrossFormer model (28.6 Thousand params) (Zhang & Yan, 2023). This choice was motivated by CrossFormer's state-of-the-art performance on time-series forecasting tasks and parameter efficiency. The reduced parameter count not only aligns with the need for efficient gradient calculations but also enables faster computation in the case of the perturbation-based method where multiple forward passes of the model are required. CrossFormer significantly outperformed the standard Transformer architecture, and test-set evaluation metrics are provided in Appendix A.2.

2.3 MODEL INTERPRETABILITY METHODS

We applied 14 time-series interpretability methods to the dynamic circulatory failure prediction task and extracted corresponding attribution maps using `captum` (Kokhlikyan et al., 2020) and `time-interpret` (Enguehard, 2023c) Python libraries. A complete list of methods is provided in Table 1, and the code to reproduce the results is available at: https://anonymous.4open.science/r/tsaifailuremodes-DB0C/.

3 RESULTS

3.1 FAILURE MODE 1: TIME-VARYING MULTI-OUTPUT MODELS

We observed that GradientSHAP, DeepLift, Integrated Gradients, and their variants are not inherently optimized for time-varying multi-output models. This becomes particularly challenging in dynamic prediction tasks, where predictions are generated at every time step, resulting in extended attribution outputs. Specifically, these methods produce attributions of size $T \times F$ for each time step T and class C, which results in a full attribution set of $T \times T \times F \times C$, where F is the number of

[1]Circulatory failure is defined as lactate levels exceeding 2 mmol / L combined with mean arterial blood pressure below 65 mmHg or administration of any vasoactive drug.

Table 1: Comparison of Model Interpretability Methods. T denotes the number of time steps, F denotes the number of features, C denotes the number of classes or outputs. Output attributions are extracted for a single ICU stay are shown.

Index	Method	Library	Output Attribution Shape
1	GradientSHAP	`captum`	$T \times T \times F \times C$
2	DeepLift	`captum`	$T \times T \times F \times C$
3	DeepLiftSHAP	`captum`	$T \times T \times F \times C$
4	Integrated Gradients	`captum`	$T \times T \times F \times C$
5	Temporal Integrated Gradients	`time-interpret`	$1 \times T \times F \times C$
6	Sequential Integrated Gradients	`time-interpret`	$T \times T \times F \times C$
7	Occlusion	`time-interpret`	$T \times T \times F \times C$
8	Augmented Occlusion	`time-interpret`	$T \times T \times F \times C$
9	Feature Ablation	`captum`	$T \times T \times F \times C$
10	Feature Permutation	`captum`	$T \times T \times F \times C$
11	RETAIN	`time-interpret`	$T \times F$
12	FIT	`time-interpret`	$T \times F$
13	DynaMask	`time-interpret`	$T \times F$
14	Extremal Mask	`time-interpret`	$T \times F$

features. This failure mode is illustrated in Figure 1 for GradientSHAP, where we analyzed a patient record with an ICU stay of 7 days and three brief periods of elevated risk of circulatory failure. We randomly selected three consecutive time steps ($T = 199, T = 200, T = 201$) during a circulatory failure event and extracted the corresponding attribution maps for the GradientSHAP method. The results showed that the attribution maps varied significantly across these adjacent time steps and contradicted the expectation that a short 5-minute interval would not result in substantial changes in the patient's state, especially when the patient is already in a state of circulatory failure. It further leads to two sub-modes of failure:

i. Temporal Aggregation: Multi-dimensional attribution maps (size: $T \times F \times C$) generated for each time step T lack interpretable aggregation across the temporal dimension, obscuring sustained pathophysiological patterns.

ii. Temporal Causality: Interpretability heatmaps for time T incorporate attribution scores influenced by future observations (e.g., $T + 1$, $T + 2$, ...), violating clinical causality.

Similarly, other gradient, occlusion, ablation, and static permutation-based methods are plagued with the same failure modes as illustrated in Appendix A.3 (figs. 3 to 10), further highlighting the challenges in achieving temporal coherence in attribution results.

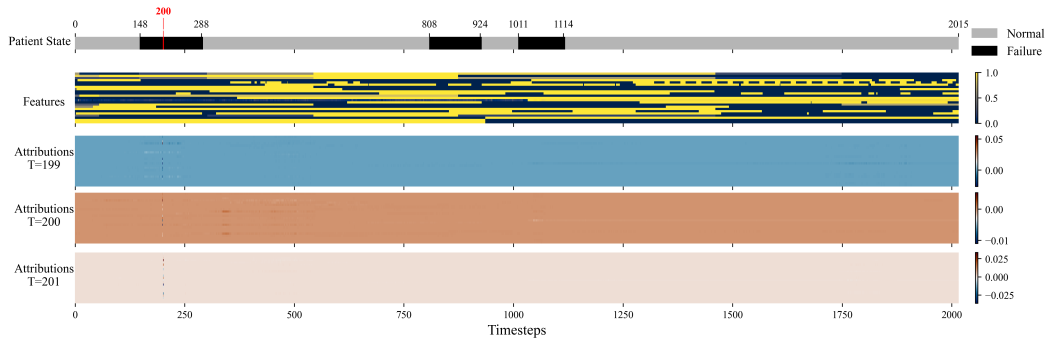

Figure 1: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **GradientSHAP** Attributions at selected time steps (T=199, T=200, and T=201).

3.2 Failure Mode 2: Issues with Temporal Smoothness

Interpretability methods specifically designed for time-series contexts, such as TIG and FIT, produce attributions that are not coherent in time. We observed that feature attributions for these methods show abrupt changes in feature importance that are inconsistent with the smooth trends typically observed in ICU time-series data. This behavior is illustrated in Figure 2 for both TIG and FIT. For FIT, this incoherence may stem from its dependence on pointwise Kullback-Leibler (KL) divergence to quantify predictive distributional shifts, which isolates individual time steps by contrasting observed features against counterfactuals where other features are masked. While theoretically rigorous, this framework risks overemphasizing transient perturbations in the predictive distribution rather than capturing the gradual, interdependent evolution of patient states.

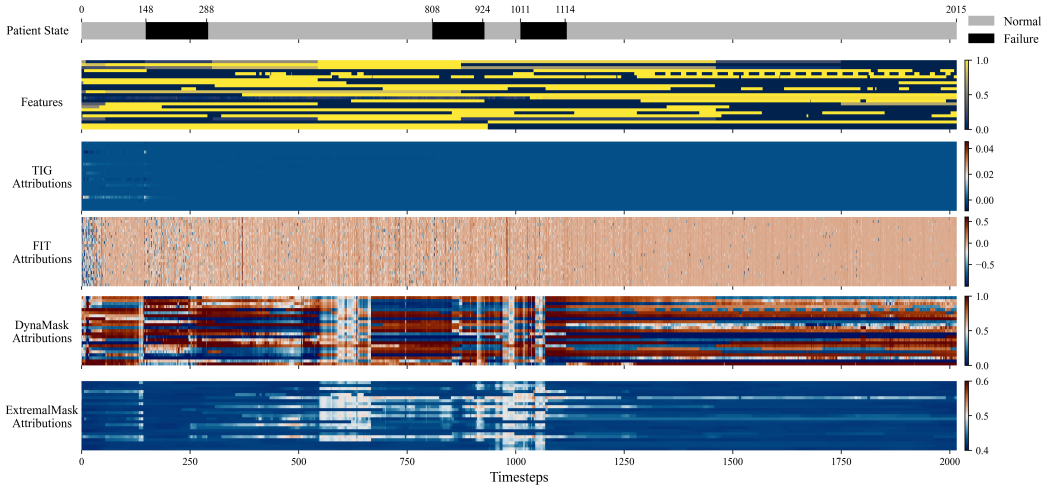

Figure 2: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows patient state. The first heatmap shows normalized feature values over time. The next four heatmaps show attributions for **Temporal Integrated Gradients (TIG)**, **Feature Importance in Time (FIT)**, **DynaMask**, and **ExtremalMask**, respectively, for all times.

3.3 Alternative Approaches

Methods such as DynaMask and ExtremalMask dynamically learn masks to selectively perturb features in a way that optimizes the relevance and consistency of attributions (Figure 2). A major benefit of these frameworks is their ability to achieve significantly lower variance in attributions across time steps, ensuring better temporal continuity. Furthermore, the optimization problem in learnable mask-based approaches is framed to ensure that model predictions on perturbed data remain consistent with the original predictions, often achieved by incorporating losses to enforce label consistency. We argue that the potential of learnable mask-based approaches is immense and they can offer consistent and meaningful interpretations for dynamic prediction tasks.

4 Discussion

Mask-based approaches provide a significant advantage for interpretability in time-series tasks by enabling the integration of task-specific constraints such as temporal continuity and label consistency. These constraints ensure that the feature importance scores evolve smoothly across adjacent time steps, preserving the inherent dependencies in temporal data. In contrast, traditional interpretability methods often fail to account for such temporal structures, leading to fragmented and inconsistent attributions that are unsuitable for dynamic prediction tasks. Mask-based methods offer a robust solution for generating coherent model interpretability attribution scores, particularly in high-stakes applications such as acute organ failure prediction, by tailoring the interpretability framework to the unique characteristics of time-series datasets.

REFERENCES

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, pp. 2166–2177. PMLR, 2021.

Sanjoy Dey, Prithwish Chakraborty, Bum Chul Kwon, Amit Dhurandhar, Mohamed Ghalwash, Fernando J Suarez Saiz, Kenney Ng, Daby Sow, Kush R Varshney, and Pablo Meyer. Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, 3(5), 2022.

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

Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*, 2023b.

Joseph Enguehard. Time interpret: a unified model interpretability library for time series. *arXiv preprint arXiv:2306.02968*, 2023c.

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

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

WRITING COMMITTEE MEMBERS, Biykem Bozkurt, Tariq Ahmad, Kevin M Alexander, William L Baker, Kelly Bosak, Khadijah Breathett, Gregg C Fonarow, Paul Heidenreich, Jennifer E Ho, et al. Heart failure epidemiology and outcomes statistics: a report of the heart failure society of america. *Journal of cardiac failure*, 29(10):1412, 2023.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Suraj Srinivas and François Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. *arXiv preprint arXiv:2006.09128*, 2020.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. *Advances in Neural Information Processing Systems*, 33:799–809, 2020.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

A Appendix

A.1 Dataset description

Table 2: Dataset Description for the Dynamic Circulatory Failure Prediction Task. M: Million

Set	ICU Stays	Predictions(% positive)
Train	23643	11.56M (4.51%)
Validation	5072	2.42M (4.22%)
Test	5069	2.44M (4.67%)

A.2 Model Evaluation

Table 3: Model evaluation metrics on the test set. Mean and standard deviation are calculated over ten runs. M: Million, K: Thousand

Model (# Parameters)	AUC-ROC	AUC-PR	F1	MCC
Transformer-Encoder (1.64M)	90.26±0.42	34.84±0.69	26.32±2.56	28.72±1.62
Crossformer-Encoder (28.6K)	96.33±0.14	62.87±0.80	54.82±0.81	53.94±0.65

A.3 Additonal feature-time attribution maps

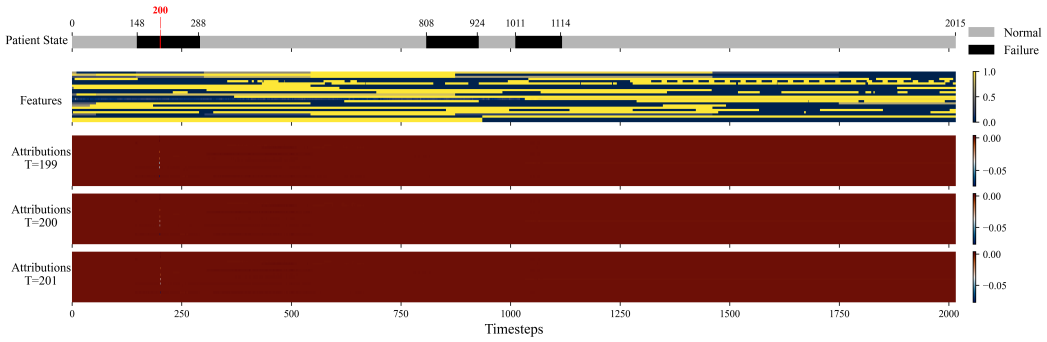

Figure 3: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **Integrated Gradients (IG)** Attributions at selected time steps (T=199, T=200, and T=201).

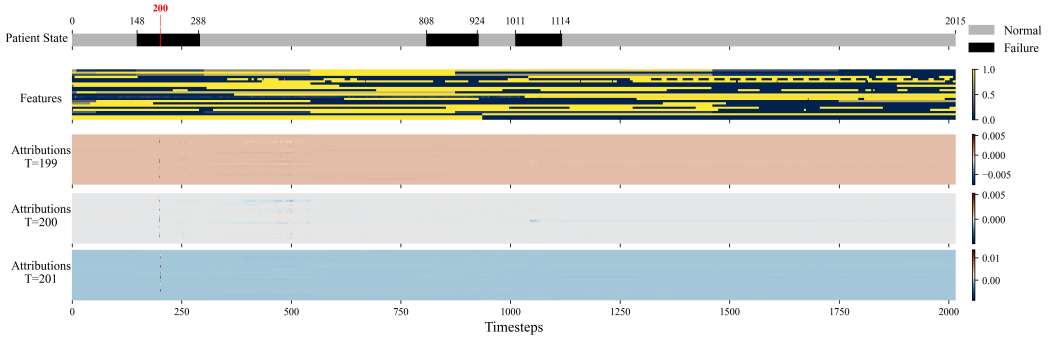

Figure 4: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **DeepLift** Attributions at selected time steps (T=199, T=200, and T=201).

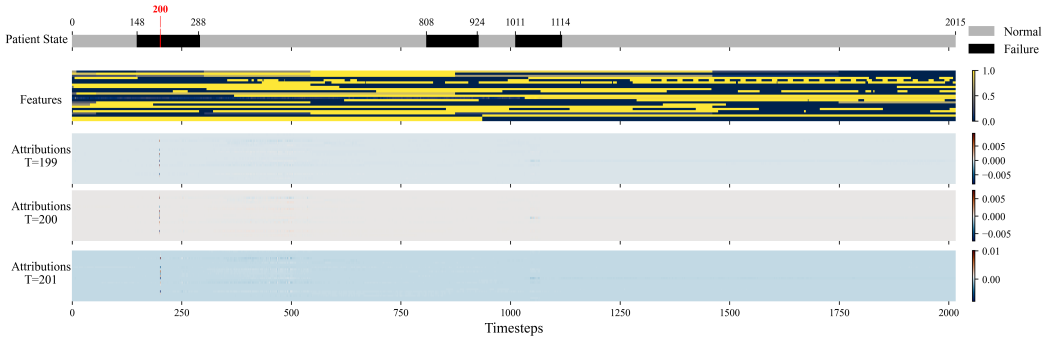

Figure 5: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **DeepLiftSHAP** Attributions at selected time steps (T=199, T=200, and T=201).

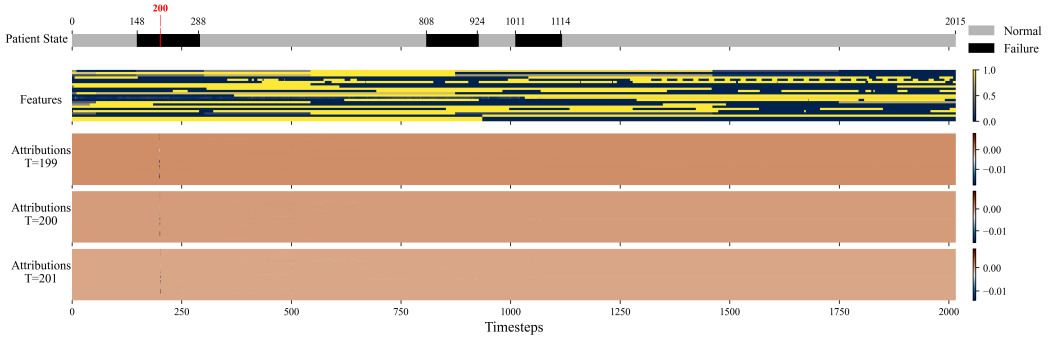

Figure 6: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **Sequential Integrated Gradients (SIG)** Attributions at selected time steps (T=199, T=200, and T=201).

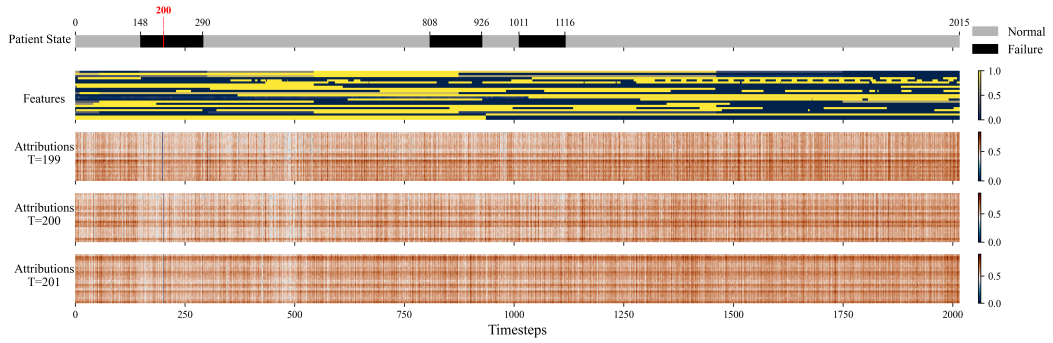

Figure 7: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **Occlusion** attributions at selected time steps (T=199, T=200, and T=201).

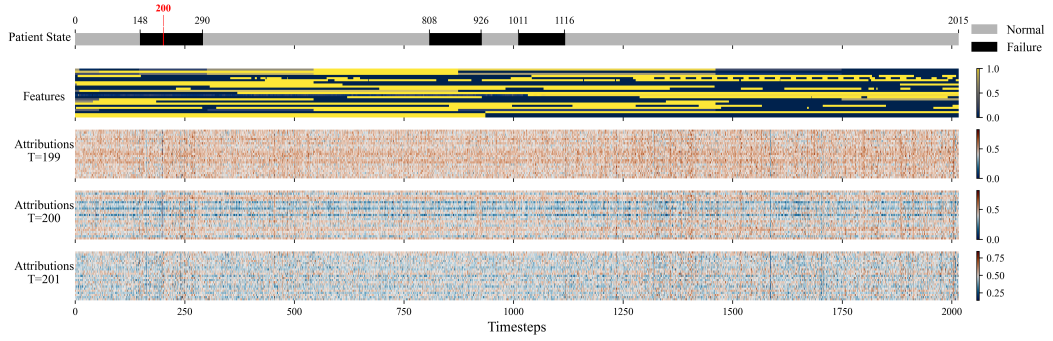

Figure 8: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **Augmented Occlusion** attributions at selected time steps (T=199, T=200, and T=201).

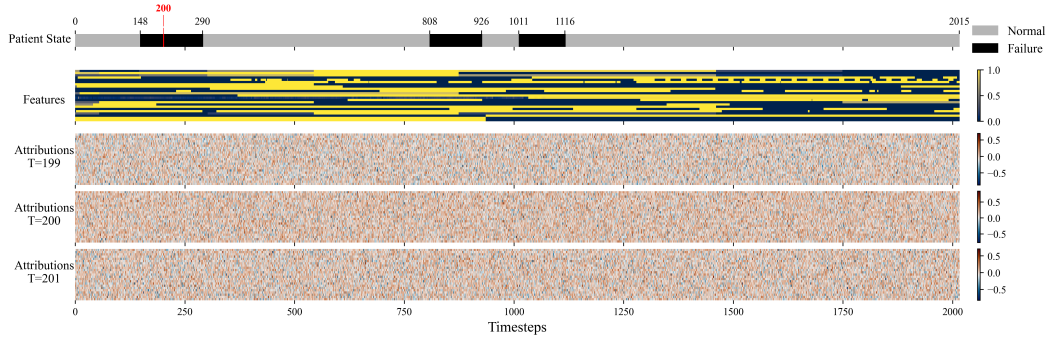

Figure 9: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **Feature Ablation** attributions at selected time steps (T=199, T=200, and T=201).

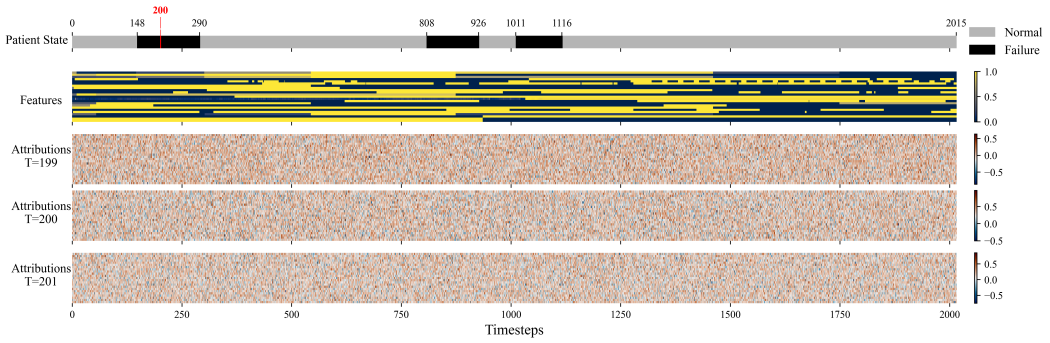

Figure 10: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time, and the three lower heatmaps show **Feature Permutation** attributions at selected time steps (T=199, T=200, and T=201).

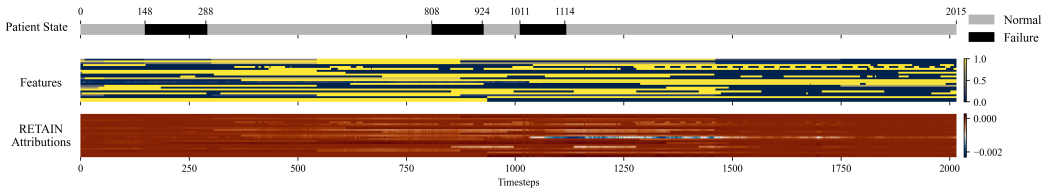

Figure 11: Illustration of a patient's timeline from T=0 to T=2015 for the dynamic circulatory failure prediction task. The top bar shows the patient's state. The first heatmap shows normalized feature values over time. The second heatmap shows attributions obtained from the **RETAIN** method for the whole timeline.