Monte Carlo ExtremalMask: Uncertainty-aware Time-series Model Interpretability for Critical Care Applications

Shashank Yadav SHASHANK@ARIZONA.EDU
Department of Biomedical Engineering
University of Arizona
Tucson, AZ, USA 85721

Vignesh Subbian VSUBBIAN@ARIZONA.EDU
Department of Biomedical Engineering
University of Arizona
Tucson, AZ, USA 85721

Abstract

Model interpretability for biomedical time-series contexts (e.g., critical care medicine) remains a significant challenge where interactions between pathophysiological signals obscure clinical interpretations. Traditional feature-time attribution methods for time series generate static, deterministic saliency masks, which fail to account for the temporal uncertainty and probabilistic nature of model-inferred feature importance in dynamic physiological systems such as acute organ failure. We address this limitation by proposing a probabilistic framework that leverages Monte Carlo Dropout to quantify model-centric epistemic uncertainty in attribution masks. We capture the stochastic variability through iterative sampling, though the inherent randomness introduces inconsistency in mask outputs across sampling iterations. We implement a dual optimization strategy that incorporates both entropy minimization and spatiotemporal variance regularization during training to ensure the convergence of attribution masks toward higher informativeness and lower entropy while preserving uncertainty quantification. This approach provides a systematic way to prioritize feature-time pairs by balancing high attribution scores with low uncertainty estimates, allowing end users to discover potential digital biomarkers associated with time-dependent pathophysiological deterioration. Our work advances the field of healthcare machine learning by formalizing uncertainty-aware interpretability for temporal models while bridging the gap between probabilistic attributions and actionable interpretations for challenges in critical care.

1. Introduction

Model interpretability plays a key role in promoting the effective adoption and use of machine learning-based decision support systems in acute care settings. In critical care medicine, for example, interpretations must align with the temporal dynamics of patient trajectories. While deep learning models for time-series data excel in detecting subtle temporal patterns in multivariate physiological data, their clinical utility fundamentally depends on their ability to map predictions to underlying pathophysiological mechanisms (Dey et al., 2022). Effective time-series interpretability must transcend univariate feature attribution to capture the spatiotemporal context of predictive relevance, such as hemo-

dynamic deterioration preceding cardiovascular decompensation. Foundational advances in generalized model interpretability include gradient-based techniques such as Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), GradientSHAP (Lundberg et al., 2018) and their variants (Wang et al., 2024). In addition to gradient-based, there are static perturbation-based methods such as Feature Occlusion (Suresh et al., 2017), Augmented Temporal Feature Occlusion Tonekaboni et al. (2020), Feature Ablation and Feature Permutation (Kokhlikyan et al., 2020). However, these methods often struggle to capture temporal dependencies in time-series data because they typically attribute significance to discrete temporal instances rather than identifying evolving temporal signals (Srinivas and Fleuret, 2020; Yadav and Subbian, 2025b). Contrary to generalized attribution methods, time-series-specific approaches such as Feature Importance in Time (Tonekaboni et al., 2020) have been developed, yet they remain inadequate for intensive care applications because they tend to capture importance for isolated individual time points rather than clinically meaningful and temporally coherent attribution masks.

Recently, differentiable mask-based methodologies have demonstrated superior performance for feature-time attribution for time-series data relative to conventional approaches. Parametric mask-based frameworks such as DynaMask (Crabbé and Van Der Schaar, 2021), ExtremalMask (Enguehard, 2023a), TimeX (Queen et al., 2023), TimeX++ (Liu et al., 2024a), and ContralLSP (Liu et al., 2024b) optimize feature-time attribution masks through end-to-end differentiable objectives that simultaneously preserve the consistency of model output while enforcing mask sparsity and temporal coherence. However, these methods generate deterministic attribution masks without quantifying epistemic or aleatoric uncertainty inherent in both the underlying data and the attribution process itself. In high-stake resource-constrained clinical environments such as Intensive Care Units (ICUs), quantifying algorithmic uncertainty constitutes an essential component of decision support systems (Ruhe et al., 2019). Since the mask-based model interpretability methods neglect uncertainty quantification, they can yield attribution scores with false or missing confidence estimates that may adversely influence clinical decision-making.

In this work, we propose Monte Carlo ExtremalMask, a Bayesian extension to the ExtremalMask framework that incorporates uncertainty quantification through stochastic variational inference implemented via Monte Carlo dropout. Our methodology quantifies epistemic uncertainty in feature-time attributions, providing not only importance magnitudes for feature-time pairs but also confidence intervals for these attributions. Monte Carlo dropout, while valuable for quantifying uncertainty, inherently introduces variability in the mask due to its stochastic sampling process. This stochasticity results in the mask values hovering around intermediate values (e.g., around 0.5) instead of committing to clear decisions (close to 0 or 1). This is because the network, when faced with dropout noise, averages its outputs rather than producing confident (bimodal) decisions, as we demonstrate here.

Furthermore, we introduce explicit mask regularization with mask entropy and mask variance as additional loss functions that optimize the mask to overcome the indecisiveness introduced by Monte Carlo dropout. This integrated approach combines stochastic sampling of Monte Carlo dropout with penalties that drive the mask values toward extremes and reduces ambiguity in the attribution. To demonstrate the ability to produce stable, clinically interpretable feature-time attributions, we evaluate our approach on the dynamic circulatory failure prediction task using data from the HiRID ICU benchmark (Yèche et al.,

2021). Our methodology produces clear, decisive masks that not only provide stable feature-time attributions with minimum variance across inference runs but also retain uncertainty quantification.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Epistemic uncertainty enhances interpretability: Quantifying model uncertainty alongside feature attributions provides a more reliable foundation for downstream evaluation and comparison of XAI methods in critical care applications.
- Temporal coherence matters: Enforcing smooth, time-aware attributions that align computational outputs with the sequential nature of physiological data improves the relevance and stability of interpretations in temporal contexts.
- Data characteristics should guide hyperparameter selection: The choice and tuning of interpretability techniques must reflect sampling frequency and feature characteristics. Densely sampled time-series allows for aggressive regularization, whereas sparse time-series may require a less aggressive setting.

2. Related Work

2.1. Mask-based Time-series Interpretability Methods

Several perturbation-based methods have recently been developed to leverage learnable masks[1] and selectively alter input time-series for extracting feature-time attributions. These masks, which are continuously valued between 0 and 1 or strictly binary, are applied element-wise to the input data to determine which segments can be perturbed. For example, Dynamask (Crabbé and Van Der Schaar, 2021) generates instance-specific importance scores by optimizing a perturbation mask. It uses localized perturbation operators (e.g., windowed moving average, gaussian) on a given time step considering adjacent temporal values to effectively capture local temporal dependencies. Building upon this idea, ExtremalMask[2] (Enguehard, 2023a) learns not only mask placements but also perturbation values themselves using a bidirectional recurrent neural network, which captures long-range feature-time interactions inherently present in time-series data.

An alternative approach involves training surrogate models to approximate black-box behaviors and generate attribution scores. TimeX (Queen et al., 2023), for example, jointly trains an explanation generator and an encoder to produce discrete attribution masks that reflect the original model's latent structure. Expanding on this, TimeX++ (Liu et al., 2024a) incorporates the Information Bottleneck principle to create in-distribution, label-preserving explanations while mitigating issues of distribution shifts through additional regularization losses. Together, these surrogate-based frameworks mark a shift towards learnable mask-based interpretability methods for time-series contexts.

1. Traditional methods such as SHAP, LIME, Integrated Gradients and its variants perform poorly on long multivariate time-series, producing fragmented, high-entropy saliency maps that lack temporal coherence (Yadav and Subbian, 2025b).
2. A primer is provided in Appendix C.

2.2. Uncertainty Estimation in Model Interpretations

Uncertainty estimation assesses the reliability of model predictions and offers insights into the model's confidence. Uncertainty is classified into two main categories: (i) Aleatoric Uncertainty, representing the intrinsic randomness or inherent noise within the data generation process, and (ii) Epistemic Uncertainty, arising from uncertainty in the model parameters themselves. While estimation of uncertainty for model predictions is a mature field with numerous established methods, including Bayesian, ensemble, sampling/dropout-based, post-hoc and auxiliary networks, generative methods, quantile regression/predicted intervals and conformal predictions (Gawlikowski et al., 2023), relatively less work has focused on quantifying the uncertainty associated with the interpretability methods used to explain those predictions.

Bayesian methods naturally offer a framework for modeling uncertainty in explanations. Slack et al. (2021) developed a Bayesian framework to generate local explanations along with credible intervals that capture the associated uncertainty. Conformal predictions have been utilized to estimate uncertainty in model interpretations. Instead of a single explanation, conformal methods provide a set of explanations, or probability estimates with a guaranteed confidence level. Fast Calibrated Explanations (Löfström et al., 2024) aimed towards the generation of uncertainty-aware explanations with uncertainty intervals for both prediction and feature contributions. Similarly, Folgado et al. (2023) suggested using uncertainty quantification to reduce the complexity of feature-based explanations, especially in multimodal scenarios, through uncertainty-weighted late model fusion.

Recently, several perspectives have positioned uncertainty estimation itself as an interpretability technique that can provide local and model-specific interpretations by communicating when a model's output should or should not be trusted. Thuy and Benoit (2024) argue that uncertainty contributes to trust, actionability (through classification with rejection), and robustness under distribution shifts. Salvi et al. (2025) argues that interpretations alone do not guarantee reliability and proposes integrating uncertainty quantification to improve model trustworthiness, especially in healthcare, where models can learn spurious correlations. Such integration can promote transparency by conveying uncertainty rather than concealing it. While some of these approaches to quantify and incorporate uncertainty into interpretability have been applied to diverse time-series data, including wearables sensor data and sleep monitoring (Heremans and De Vos, 2023), a comprehensive investigation of uncertainty quantification for the interpretability of time-series models in critical care remains relatively underexplored in the current literature. Our work addresses this gap by exploring uncertainty estimation in feature-time interpretations for high-dimensional ICU time-series data and aims to enhance the interpretability and reliability of time-series models in critical care.

3. Methods

3.1. Dataset and Prediction Task

In this study, we utilize the HiRID ICU benchmark (Yèche et al., 2021), a large dataset containing high-dimensional time-series data on vital signs, laboratory values, and treatment records from critical care patients. We focus on the Dynamic Circulatory Failure Prediction

task, formulated as a binary prediction of acute circulatory failure throughout the patient's ICU stay.

3.2. Causal Crossformer Encoder as the Black Box Deep Learning Model

First, we replicated the results of the HiRID-ICU Benchmark study (Yèche et al., 2021), which established the standard transformer encoder architecture (1.64 million parameters) (Vaswani, 2017) with causal attention as the leading model for dynamic circulatory failure prediction. Next, we substantially improved this approach by implementing a causal encoder-only variant of the CrossFormer model (28.6 thousand parameters) (Zhang and Yan, 2023). Our selection of CrossFormer was driven by its state-of-the-art performance on time-series forecasting tasks and parameter efficiency. The significant reduction (98%) in parameter count facilitates faster forward-pass computations in perturbation-based methods where multiple passes are required for mask optimization. For clinical time-series data, maintaining temporal causality is also important, as predictions at any timestep should never be influenced by future data. To address this, we developed a causal variant of CrossFormer-Encoder by integrating a causal attention mask. This mask, implemented using a triangular mask, sets attention scores for future timesteps to negative infinity before the softmax operation. Hence, our model strictly adheres to temporal causality, with predictions at each timestep depending exclusively on information available at or before that specific moment, effectively preventing any future data leakage. Once the encoder is pre-trained, we freeze its parameters during mask-generator training. This ensures that the interpretability objectives cannot alter the encoder's learned representations and therefore, the calibration remain unchanged from the baseline.

3.3. Enabling Uncertainty Estimation in ExtremalMask with Monte Carlo Dropout

In this study, we adapt the ExtremalMask (EM) method (Enguehard, 2023a), a dynamic mask learning framework that identifies important feature-time pairs in multivariate time series data given an input time series $X \in \mathbb{R}^{T \times d}$, where T represents the number of time steps and d the number of clinical features. The Monte-Carlo ExtremalMask (MC-EM) method is illustrated in Figure 1. The EM part of the method learns a corresponding mask $M \in [0,1]^{T \times d}$, which selectively perturbs the input features. It jointly learns (1) a mask M and (2) a context-aware perturbation function $P_\theta(X)$ (shown as $PLNN(x)$ in Figure 1), parameterized as an autoregressive neural network. EM perturbations are optimized to preserve global temporal dynamics while masking salient regions:

$$X_{\text{masked}} = M \odot X + (1 - M) \odot P_\theta(X). \tag{1}$$

We capture model uncertainty by applying a dropout layer over the mask during both training and inference inspired by Monte-Carlo Dropout[3] (Gal and Ghahramani, 2016). This procedure enables us to generate multiple stochastic samples of the learned mask, providing a distribution over possible attributions rather than a single deterministic attribution. For a given input X, we perform K stochastic forward passes through the network with dropout

3. A primer is provided in Appendix C.

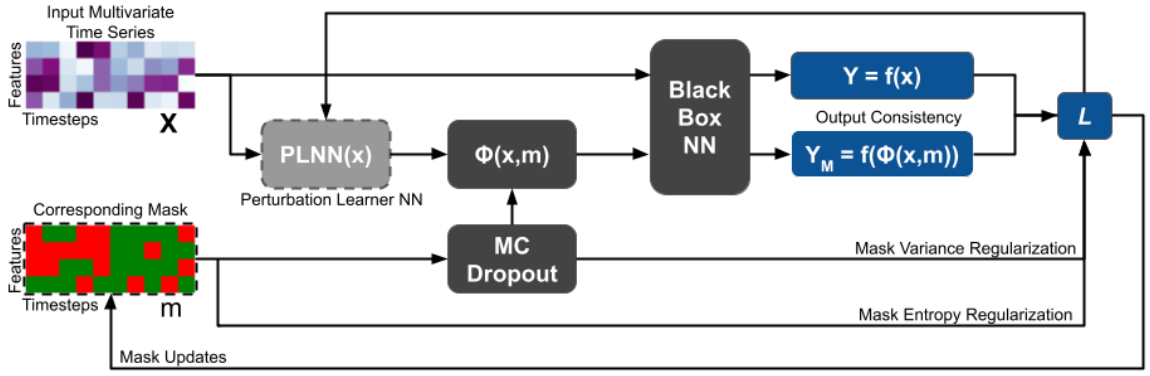

Figure 1: Illustration of the MC ExtremalMask method. A multivariate time-series input $X \in \mathbb{R}^{T \times D}$ is fed into a perturbation learner neural network (PLNN), yielding PLNN(x). A learnable mask M balances how much of each feature is replaced by the perturbation versus retained in its original form, producing $\Phi(x, m)$. Both $\Phi(x, m)$ and X are then passed to a fixed "black-box" model (causal cross-former encoder) f for explanation. The learnable parts of the framework m and $PLNN(x)$ (dashed borders) are optimized to (i) keep the perturbed outputs $f(\Phi(x, m))$ close to the original outputs $f(x)$, (ii) encourage the mask to hide as many features as possible, and (iii) keep the perturbation $PLNN(x)$ sparse. A Monte Carlo (MC) dropout layer samples different masks at each training epoch; we compute mask variance as part of the loss to enforce variance minimization and also include a mask entropy term to promote decisive (low-entropy) masking. Once trained, MC dropout remains active at inference to sample both the mask and its associated uncertainty, helping identify which features are most critical for preserving the model's original predictions.

enabled(with a dropout rate p), resulting in K samples of the mask $\{M_1, M_2, ..., M_K\}$. From these samples, we compute:

1. Mean Mask: $\bar{M} = \frac{1}{K} \sum_{i=1}^{K} M_i$, which serves as an approximate deterministic mask, representing the expected feature importance across stochastic variations.

2. Variance: $\text{Var}(M) = \frac{1}{K} \sum_{i=1}^{K} (M_i - \bar{M})^2$, which quantifies the model's epistemic uncertainty for each feature-time pair.

3. Mask Entropy: $H(M) = -\bar{M} \log \bar{M} - (1 - \bar{M}) \log(1 - \bar{M})$. The entropy of the mean mask quantifies the decisiveness of the attribution, with lower values indicating more confident decisions about feature-time attributions.

We optimize the mask M with respect to a target function $f(X)$ representing the model's prediction:

$$M^* = \arg\min_M \mathcal{L}(M, X, f)$$

where $\mathcal{L}$ is a composite loss function incorporating prediction consistency, variance minimization, entropy minimization, and temporal regularization. We ran both methods on our cohort (refer to section 4.1) using identical hyperparameters: 500 optimization epochs and a learning rate of $5e^{-4}$.

3.4. Loss Functions for Mask Optimization

We incorporate several specialized loss functions to improve the mask decisiveness and lower mask variance.

1. Prediction Consistency Loss: This loss ensures that the masked input preserves the model's original prediction:

$$\mathcal{L}_{\text{consistency}} = \|f(X_{\text{perturbed}}) - f(X)\|_2^2$$

where f represents the black box prediction model (causal crossformer encoder). This term encourages the mask to retain information that is essential for the model's prediction while potentially excluding irrelevant features.

2. Entropy Minimization Loss: It drives the mask values toward 0 or 1 by penalizing intermediate values:

$$\mathcal{L}_{\text{entropy}} = -\frac{1}{T \times D} \sum_{i,j} [M_{i,j} \log M_{i,j} + (1 - M_{i,j}) \log(1 - M_{i,j})]$$

A mask with relatively low entropy implies a more decisive feature-time attribution map, which is crucial for clinical applications where ambiguity is undesirable. Mask entropy regularization was originally introduced in Dynamask (Crabbé and Van Der Schaar, 2021), while it was not incorporated into the default EM implementation. We include it in our approach to further reduce ambiguity in the attributions, thereby yielding a more decisive mask.

3. Mask Variance Minimization Loss: Monte Carlo dropout produces K stochastic mask samples $\{M^{(k)}\}_{k=1}^{K}$. We minimize the variance of the mask to promote consistency across these K samples.

$$\mathcal{L}_{\text{variance}} = \frac{1}{T \times D} \sum_{i=1}^{T} \sum_{j=1}^{D} \text{Var}\Big(\{M_{i,j}^{(k)}\}_{k=1}^{K}\Big).$$

4. Temporal Regularization: We incorporate a temporal regularization term inspired by the Dynamask (Crabbé and Van Der Schaar, 2021) procedure to enforce mask smoothness across the temporal dimension. Similar to mask entropy, it was not originally a part of the EM procedure; hence we have incorporated it into our approach. It is defined as:

$$\mathcal{L}_{\text{temporal}} = \frac{1}{T-1} \sum_{i=1}^{T-1} \|M_{i+1,\cdot} - M_{i,\cdot}\|_2^2,$$

The final loss function is a weighted combination of these individual terms:

$$\mathcal{L} = \lambda_{\text{consistency}}\,\mathcal{L}_{\text{consistency}} + \lambda_{\text{entropy}}\,\mathcal{L}_{\text{entropy}} + \lambda_{\text{variance}}\,\mathcal{L}_{\text{variance}} + \lambda_{\text{temporal}}\,\mathcal{L}_{\text{temporal}},$$

where $\lambda_{\text{consistency}}$, λ_{entropy}, $\lambda_{\text{variance}}$, and $\lambda_{\text{temporal}}$ are hyperparameters[4] that balance the contribution of each loss term.

3.5. Evaluating the Quality of Masks

We define several metrics that capture different aspects of mask quality.

1. Mask Entropy (S_M): This metric quantifies the overall decisiveness of the mask:

$$S_M = -\frac{1}{T \times D} \sum_{i,j} [M_{i,j} \log M_{i,j} + (1 - M_{i,j}) \log(1 - M_{i,j})]$$

   Lower entropy indicates clearer, more decisive attributions. A perfect mask would have all values at either 0 or 1, resulting in zero entropy.

2. Mask Information (I_M): This metric measures the amount of information retained by the mask:

$$I_M(A) = -\sum_{(t,i)\in A} \ln\bigl(1 - m_{t,i}\bigr).$$

   It represents the fraction of the original input that is preserved by the mask, balancing the trade-off between information loss and interpretability. While lower values indicate more selective masks, extremely low values might suggest that the mask is filtering out too much information.

4. Study Sample

4.1. Cohort Selection

We focused on Circulatory Failure as the primary use case for this work because it is one of the leading causes of morbidity and mortality in critical care settings (Bozkurt et al., 2023). For our experiments, we employed the Dynamic Circulatory Failure Prediction task(refer to footnote[5]) from the HiRID-ICU benchmark study (Yèche et al., 2021). Specifically, the task continuously predicts the onset of circulatory failure within the next 12 hours, provided the patient is not already in organ failure. The HiRID benchmark comprises 33,784 ICU stays, with the following split used for methods comparison and demonstration:

4. Loss weights fixed as: $\lambda_{\text{consistency}}$=1.0, λ_{entropy}=1.0, $\lambda_{\text{variance}}$=20.0, $\lambda_{\text{temporal}}$=5.0. Results from ablation experiments over these hyperparameters are provided in Appendix A2.
5. 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: 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%)

4.2. Data Extraction and Feature Choices

We followed the standardized data extraction pipeline of the HiRID-ICU benchmark, which produced a hierarchical data format (HDF) version 5 file containing data from the 33784 ICU stays. Each record in the dataset is a multivariate time series spanning 2016 time steps at 5-minute intervals—corresponding to a continuous 7-day period. For every ICU stay, 231 clinical time series features were extracted, including vital signs, hemodynamic data, treatments administered, pathological laboratory values, and ventilation parameters integral to critical care management. We filtered the test set based on two criteria: (1) the number of circulatory failure events experienced by the patient and (2) the availability of 7 days of complete data, which is the maximum duration provided in the benchmark. This filtering resulted in a subset of 319 ICU stays, of which 190 exhibited no circulatory failure events, and 129 had one or more failure events. The 190 ICU stays without failure events act as a negative control, allowing us to assess whether our framework over-attributes importance when no adverse event is present.

5. Results

In this section, we evaluate MC-EM on its ability to produce low entropy feature-time attributions, along with reliable uncertainty estimates, for dynamic circulatory failure prediction. We evaluate mask quality and assess mask uncertainty by comparing cases with and without failure events, while model performance is provided in Appendix A1.

5.1. Uncertainty Estimation and Comparison to ExtremalMask

We compared our proposed MC-EM approach to the EM method. Figure 2 illustrates the results of this comparison. In Figure 2a, we observed that our MC-EM method produced a more decisive attribution mask compared to the EM method as a bimodal distribution is observed pushing attribution scores towards extremes (0 and 1). As illustrated in Figure 2b, the inclusion of mask entropy regularization in MC-EM helps reduce mask entropy across the cohort, reflecting lesser ambiguity in the attribution masks, which is a desired outcome. Figure 2c further demonstrates that MC-EM yields more informative masks, as indicated by relatively higher mean mask information values. Finally, Figure 2d represents an example of a circulatory failure ICU stay with three distinct failure events. The mask values produced by MC-EM are predominantly near the extremes, which indicates both decisiveness and high informativeness compared to the standard EM method. Moreover, our method provides uncertainty estimates via the mask standard deviation derived from Monte Carlo

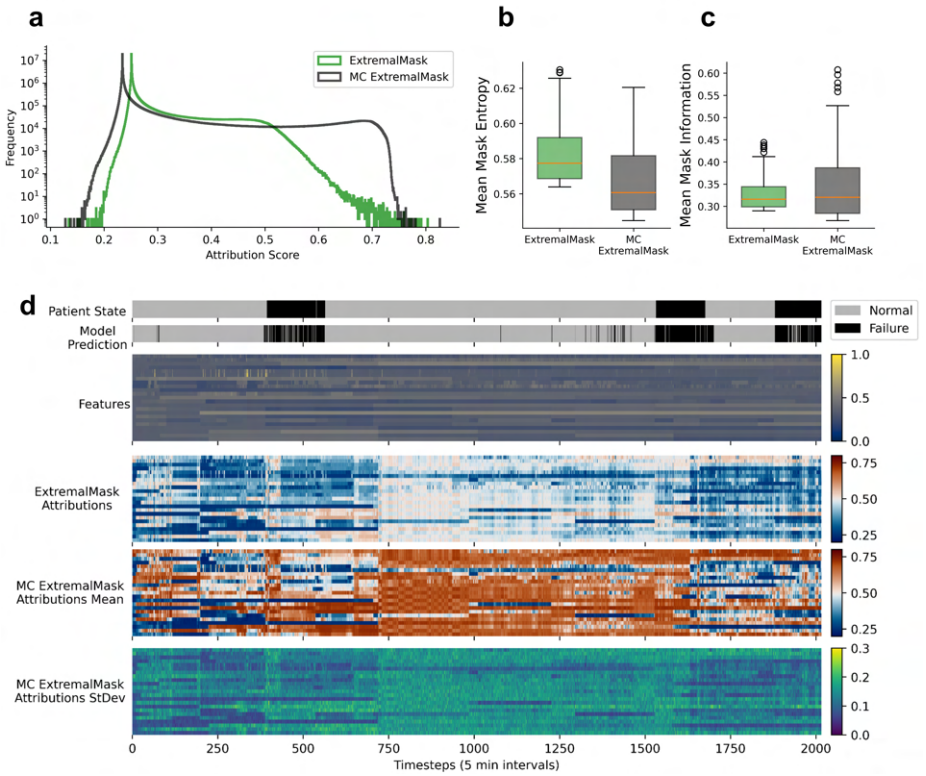

Figure 2: Comparison of MC-EM with EM. (a) Attribution scores derived from both the EM and MC-EM methods are compared for 119 ICU stays in the test set cohort, highlighting the mask decisiveness in attribution values. (b) Mask entropy values for the same ICU cohort from both methods. (c) A comparison of mask information metrics further illustrates that the MC-EM method yields more informative masks. (d) A representative case of a circulatory failure ICU stay which features three distinct events of failure alongwith model's prediction. In this panel, the corresponding time series data is displayed alongside mean attribution masks for both methods and the uncertainty estimate (mask standard deviation) provided by the MC-EM method.

dropout inference. Notably, high mean mask values are naturally accompanied by increased variance, reflecting the multiplicative effects of dropout noise and heteroscedasticity. Our method incorporates mask entropy minimization with mask variance minimization, which encourages consistency among stochastic dropout samples. These regularizations together effectively counteract the potential adverse effects of Monte Carlo dropout, resulting in clearer and more stable feature-time attributions. Extended results comparing MC-EM and EM for dynamic ARDS classification across the HiRID, MIMIC-IV (Johnson et al., 2023), and eICU (Pollard et al., 2018) datasets are provided in Appendix B (Figures 6,7,8 and Table 7), further illustrating the consistency of our findings across a different clinical task and datasets.

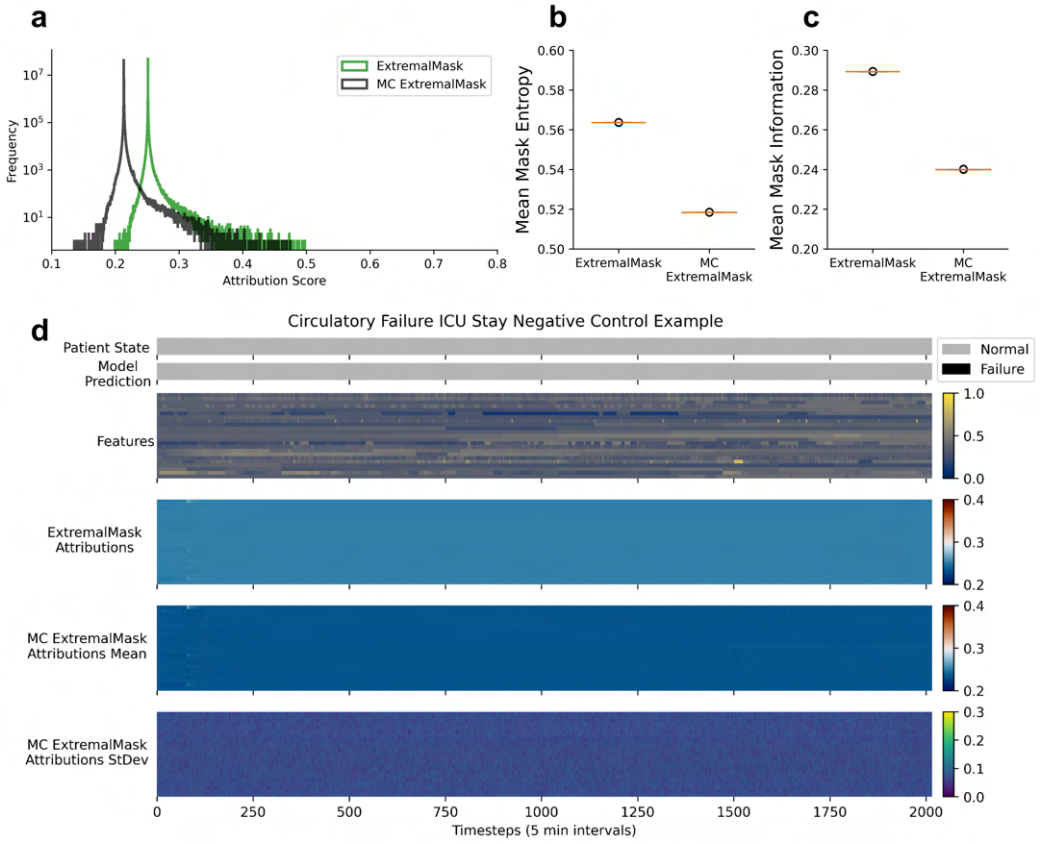

Figure 3: Comparison of MC-EM with EM for negative controls. (a) Attribution scores derived from both the EM and MC-EM methods are compared for 190 ICU stays acting as a negative control in the test set cohort (b) Mask entropy values for the same ICU cohort from both methods. (c) A comparison of mask information. (d) A representative case of a circulatory failure ICU stay negative control—featuring zero failure events is presented. In this panel, the corresponding time series data is displayed alongside close-to-zero mean attribution masks for both methods and the close-to-zero uncertainty estimate (mask standard deviation) provided by the MC-EM method.

5.2. Evaluating Over-Attribution via Negative Control Analysis

We compared MC-EM and EM on cases acting as negative controls, where no circulatory failure events occurred. Figure 3 illustrates this comparison between the EM and MC-EM methods across 190 negative-control ICU stays in the test cohort. As shown in Figure 3a, the distribution of mask attributions for MC-EM is predominantly shifted toward lower values compared to the EM method, and both methods exhibit scores below 0.5. This observation supports the fact that neither approach over-attributes importance when no actual failure events are present. Meanwhile, the mean mask entropy in Figure 3b remains consistently lower for MC-EM. The reduced entropy suggests that MC-EM produces more decisive yet

appropriately restrained attributions, even when no critical events are anticipated. Figure 3c illustrates that MC-EM also yields lower mask information in these negative-control scenarios. This lower informativeness, in contrast to the positive-control cases, is desirable because it indicates that the model is not forcing patterns of importance where none exist. Finally, Figure 3d presents a representative negative-control ICU stay, illustrating that no circulatory failure events occurred. Consistent with the aggregated findings, both EM and MC-EM methods produce near-zero mean attribution masks; however, MC-EM exhibits lower overall mask values and minimal uncertainty (as shown by the minimal mask standard deviation).

5.3. Ablation Experiments

5.3.1. Standard MC ExtremalMask vs. a PLNN-less variant

Here, we investigate the impact of incorporating a unidirectional RNN-based perturbation learner neural network (PLNN) into MC-EM, comparing it against a PLNN-less variant. As shown in Figure 4a, the attribution score histograms for MC-EM with PLNN exhibit a slightly heavier right tail (towards higher attribution scores around 0.7–0.8) distribution than those without PLNN, suggesting that the PLNN can better distinguish and highlight highly relevant features compared to the non-PLNN version. Figure 4b demonstrates that the PLNN-based method achieves lower attribution standard deviations in mask attributions, which implies that a PLNN-based MC-EM produces stable attributions across stochastic dropout inferences. However, the inclusion of PLNN leads to a modest increase in mean mask entropy while providing higher mask information as illustrated by Figures 4c and 4d. The increased entropy likely reflects the PLNN's capacity to learn long-range feature-time interactions context-aware perturbation masks. Figure 4e illustrates a representative ICU stay (corresponds to the example used in Figure 2d) in which the with-PLNN variant not only identifies critical features but also exhibits a more coherent attribution map over time. Though the mask shows a relatively higher entropy, it remains stable (lower variance) and benefits from capturing the long-range patterns by the PLNN. Given that the unidirectional RNN can effectively preserve temporal causality, its usage as a PLNN contextualizes specific timesteps with past timesteps and learns perturbations that respect changes in patient state over time.

5.3.2. MC Extremal Mask: Evaluating the Impact of Loss Functions

Here, we compare four MC-EM variants that incorporate different combinations of entropy minimization (EntMin) and variance minimization (VarMin): i) MC-EM (VarMin+EntMin), ii) MC-EM (no extra losses), iii) MC-EM (VarMin Only), and iv) MC-EM (EntMin Only). Figure 5a shows histograms of mask attribution scores across 129 ICU stays (positive samples). MC-EM variants that exclude entropy minimization (no extra losses and VarMin Only) display broader attribution distributions without pronounced bimodality, demonstrating the importance of explicitly driving masks toward near-binary extremes for more decisive attributions. Figure 5b then compares the distribution of mask standard deviations using histograms. Here, the peak standard deviation for methods lacking entropy minimization is higher, suggesting that ambiguous mid-range mask values are more susceptible to fluctuations introduced by MC Dropout. When the entropy minimization loss is added, it

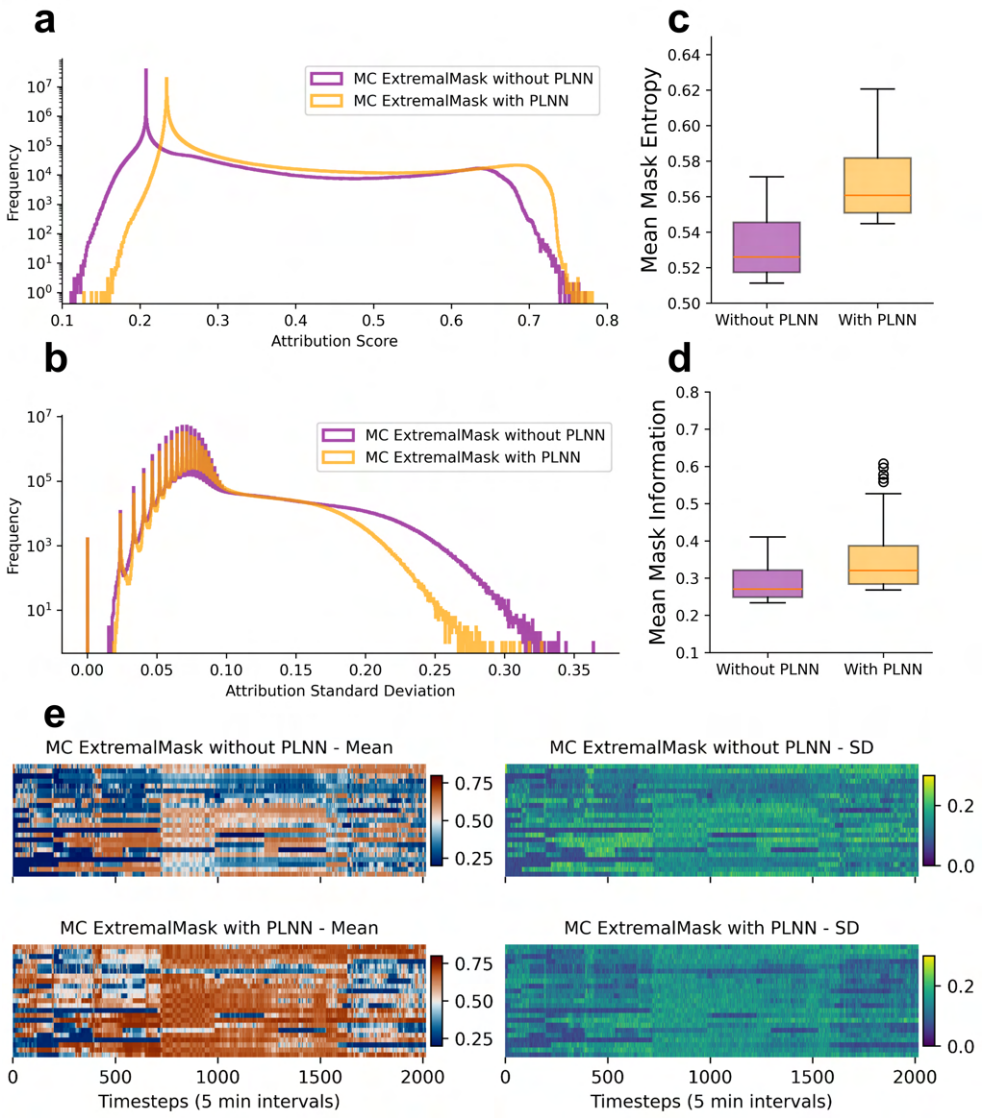

Figure 4: Comparison of MC-EM with a PLNN-less variant and derived evaluation metrics. (a) Attribution score histograms for MC-EM with and without PLNN (Perturbation Learning Neural Network) are presented. (b) Attribution standard deviation for both MC-EM variants is compared. (c) Comparison of mask entropy across 129 ICU stays in our cohort. (d) Comparison of mask information metrics for the same cohort. (e) A representative ICU stay is illustrated, and mask outputs for MC-EM with and without a PLNN are compared.

indirectly reduces variance by steering mask values away from these ambiguous regions and thus produces more consistent outputs, as observed by a lower spread in the mask standard deviations. Although entropy and variance minimization are distinct objectives, this result

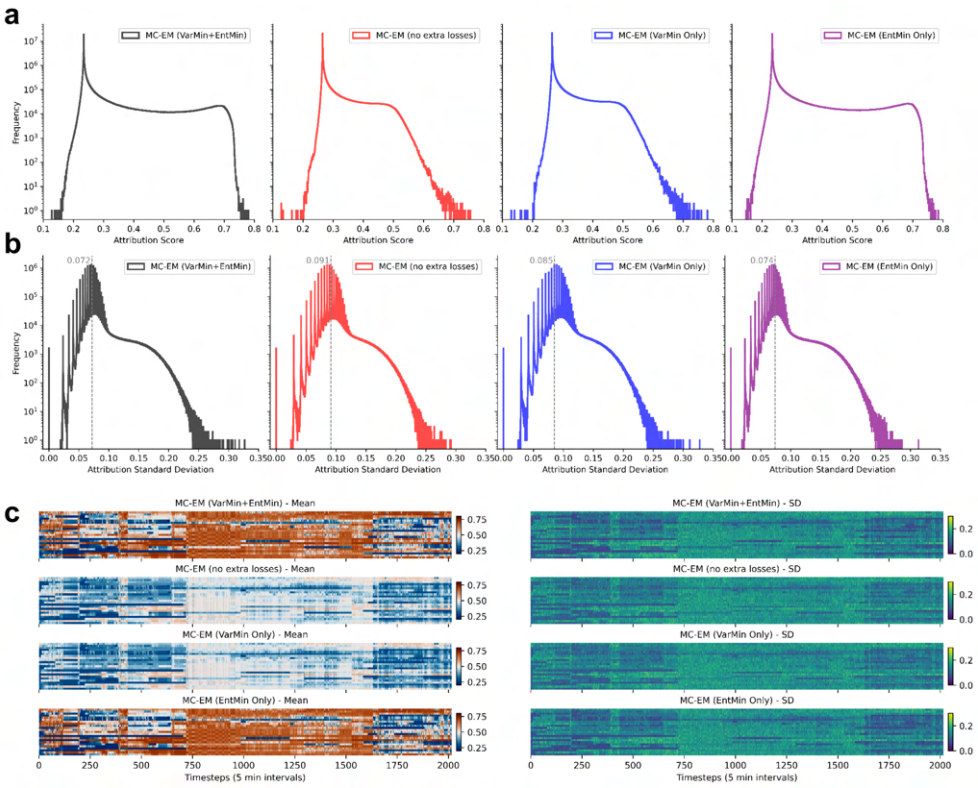

Figure 5: Evaluation MC-EM Ablation Variants. (a) Comparison of mask attribution score histograms for MC-EM and its ablation variants—MC-EM (VarMin+EntMin; standard), MC-EM (no extra losses), MC-EM (VarMin Only), and MC-EM (EntMin Only)—across 129 ICU stays (positive samples). (b) Comparison of mask standard deviation for the same variants across the 129 ICU stays. (c) A representative ICU stay is showcased, comparing the mask outputs obtained from MC-EM (VarMin+EntMin) and the different ablation variants.

indicates that reducing entropy can also help keep mask variance in check. Finally, Figure 5c illustrates a representative ICU stay(corresponds to the example used in Figure 2d), illustrating that our MC-EM method yields attribution masks that are both comparatively decisive (near-binary) and stable (lower variance). In contrast, the other variants either do not push the mask strongly toward binary extremes or exhibit higher overall uncertainty, highlighting the synergy between entropy and variance minimization in producing clear and consistent attributions.

6. Discussion

In this work, we presented Monte Carlo ExtremalMask, a novel framework that bridges stochastic uncertainty estimation with dynamic feature-time attributions in critical care time series. Our method leverages Monte Carlo dropout to quantify epistemic uncertainty

while incorporating loss functions that explicitly minimize mask entropy and variance. This dual regularization strategy forces the learned mask values toward near-binary extremes (closer to 0 or 1), which reduces ambiguity in the attributions and enhances consistency across stochastic samples.

Technical Implications Our approach advances learnable mask-based perturbations with epistemic uncertainty estimation. We introduced mask entropy minimization with variance minimization, which stabilizes attribution mask across dropout samples. Moreover, the integration of a unidirectional RNN as a perturbation learner captures important temporal dependencies and maintains temporal causality, ensuring that the generated perturbations respect the temporal nature of the critical care time series. These enhancements result in attribution maps that not only highlight potential clinically relevant feature-time pairs but also provide robust uncertainty estimates for applications in critical care, such as predicting dynamic organ failure. Our approach also demonstrates attribution scores indicate heteroscedasticity, where feature–time pairs with higher average attribution scores tend to exhibit greater variance in their estimated importance. Also, for cases where there is no circulatory failure, the framework produces near-zero attributions, serving as a negative control and reinforcing that our method does not unexpectedly over-attribute importance to feature-time pairs.

Clinical Implications Our method offers a significant advance by providing uncertainty estimates along with a better-learned mask, as shown by its application to dynamic circulatory failure prediction. In general, our method enables the generation of low-entropy attribution maps with accompanying uncertainty intervals that would equip end users with actionable information on features leading to organ failure. The decisiveness of the attributions would help ensure that end-users in the clinic are able to pinpoint critical features influencing patient outcomes and retrospectively validate what went wrong before the onset of failure. Furthermore, the use of negative controls (ICU stays with no failure events) validates that the model refrains from flagging irrelevant features, which potentially builds trust for its inclusion into clinical decision support systems.

Limitations Our approach, while demonstrating promising improvements in mask interpretability and uncertainty estimation, has several important limitations. First, the black-box encoder model occasionally produces rapid state transitions in its predictions, which may reflect sensitivity to short-term fluctuations and could be further improved by incorporating temporal smoothing to enhance consistency over time. Second, the performance of our method is highly sensitive to the selection of hyperparameters. Specifically, the number of mask training epochs, the learning rate, and the weights assigned to the loss function components can significantly affect the characteristics of the resulting masks. These weights must be carefully tuned for specific prediction tasks. Also, it remains unclear how robust the model would be to changes in these parameters across different datasets. Additionally, our study did not extend to clinical biomarker detection, which could potentially identify predictive signals for the onset of circulatory failure. Extracting such temporal signatures would require applying our approach extensively over a much larger set of ICU stays to discern robust predictive biomarkers, an investigation that remains a direction for future work.

7. Code and Data Availability

Our implementation builds on the `time_interpret` library (Enguehard, 2023b) and is available at https://github.com/xinformatics/mcem. The datasets used in this study are publicly accessible via PhysioNet (Goldberger et al., 2000), subject to data usage agreements.

References

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.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, pages 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*, pages 9329–9342. PMLR, 2023a.

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

Duarte Folgado, Marília Barandas, Lorenzo Famiglini, Ricardo Santos, Federico Cabitza, and Hugo Gamboa. Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series. *Information Fusion*, 100:101955, 2023.

Loic Le Folgoc, Vasileios Baltatzis, Sujal Desai, Anand Devaraj, Sam Ellis, Octavio E Martinez Manzanera, Arjun Nair, Huaqi Qiu, Julia Schnabel, and Ben Glocker. Is mc dropout bayesian? *arXiv preprint arXiv:2110.04286*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56 (Suppl 1):1513–1589, 2023.

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.

Elisabeth RM Heremans and Maarten De Vos. Explaining uncertainty in ai for clinical decision support systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 404–411. Springer, 2023.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

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.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234, 2022.

Wei Liao and Joel Voldman. A multidatabase extraction pipeline (metre) for facile cross validation in critical care research. *Journal of Biomedical Informatics*, 141:104356, 2023.

Brian Liu and Madeleine Udell. Impact of accuracy on model interpretations. *arXiv preprint arXiv:2011.09903*, 2020.

Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. Timex++: Learning time-series explanations with information bottleneck. *arXiv preprint arXiv:2405.09308*, 2024a.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, et al. Explaining time series via contrastive and locally sparse perturbations. *arXiv preprint arXiv:2401.08552*, 2024b.

Tuwe Löfström, Fatima Rabia Yapicioglu, Alessandra Stramiglio, Helena Löfström, and Fabio Vitali. Fast calibrated explanations: Efficient and uncertainty-aware explanations for machine learning models. *arXiv preprint arXiv:2410.21129*, 2024.

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.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Owen Queen, Tom Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency. *Advances in Neural Information Processing Systems*, 36:32129–32159, 2023.

David Ruhe, Giovanni Cina, Michele Tonutti, Daan de Bruin, and Paul Elbers. Bayesian modelling in practice: Using uncertainty to improve trustworthiness in medical applications. *arXiv preprint arXiv:1906.08619*, 2019.

Massimo Salvi, Silvia Seoni, Andrea Campagner, Arkadiusz Gertych, U Rajendra Acharya, Filippo Molinari, and Federico Cabitza. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, 197:105846, 2025.

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

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34:9391–9404, 2021.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 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.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020.

Arthur Thuy and Dries F Benoit. Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, 317(2):330–340, 2024.

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.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*, 2024.

Shashank Yadav and Vignesh Subbian. When attention fails: Pitfalls of attention-based model interpretability for high-dimensional clinical time-series. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 289–305. PMLR, 25–27 Jun 2025a. URL https://proceedings.mlr.press/v287/yadav25a.html.

Shashank Yadav and Vignesh Subbian. Failure modes of time series interpretability algorithms for critical care applications and potential solutions. *arXiv preprint arXiv:2506.19035*, 2025b.

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.

Appendix A.

A.1. Black Box Model performance comparison

The causal CrossFormer-Encoder model significantly outperformed the standard causal transformer model on the dynamic circulatory failure task. Specifically, our causal CrossFormer achieved higher performance across AUCROC, AUCPR, F1, and MCC scores (Table 2). This improvement highlights the effectiveness of CrossFormer's segment-based embedding and causal cross-dimensional attention in capturing temporal dependencies. Moreover, recent research has indicated that models that achieve higher performance provide better interpretations—providing clearer insights into model behavior (Liu and Udell, 2020; Li et al., 2022). Hence, the causal CrossFormer not only minimizes generalization error but also improves the quality of downstream interpretations.

Table 2: Model evaluation metrics on the test set. Mean and standard deviation are averaged over ten runs. M: Million, K: Thousand

Model (# Parameters)	**AUC-ROC**	**AUC-PR**	**F1**	**MCC**
Causal Transformer (1.64M)	90.26±0.42	34.84±0.69	26.32±2.56	28.72±1.62
Causal Crossformer (28.6K)	97.19±0.20	68.05±0.52	59.87±0.71	58.75±0.65

A.2. Hyperparameter Tuning of Loss Weights

Table 3 and Table 4 show the grid search over λ_{entropy} and $\lambda_{\text{variance}}$.

Table 3: Mask Entropy across λ_{entropy} and $\lambda_{\text{variance}}$

$\lambda_{\text{entropy}}\downarrow \setminus \lambda_{\text{variance}}\rightarrow$	0.1	1	10
0.1	0.623 (0.011)	0.624 (0.011)	0.624 (0.011)
1	0.568 (0.021)	0.568 (0.021)	0.568 (0.020)
10	0.392 (0.030)	0.392 (0.030)	0.392 (0.031)

Table 4: Mask Information across λ_{entropy} and $\lambda_{\text{variance}}$

$\lambda_{\text{entropy}}\downarrow \setminus \lambda_{\text{variance}}\rightarrow$	0.1	1	10
0.1	0.437 (0.078)	0.437 (0.079)	0.437 (0.079)
1	0.349 (0.083)	0.350 (0.084)	0.349 (0.083)
10	0.187 (0.066)	0.185 (0.064)	0.186 (0.065)

After selecting the optimal $\lambda_{\text{entropy}} = 1$ and $\lambda_{\text{variance}} = 10$, we tuned $\lambda_{\text{consistency}}$. Table 5 compares three settings:

Table 5: Ablation over $\lambda_{\text{consistency}}$ (with $\lambda_{\text{entropy}} = 1$, $\lambda_{\text{variance}} = 10$)

Configuration	$\lambda_{\text{consistency}}$	Mask Entropy	Mask Information
	0.1	0.414 (0.031)	0.199 (0.067)
MC-EM (default)	1	**0.567 (0.020)**	**0.349 (0.083)**
	10	0.608 (0.013)	0.408 (0.079)
EM (baseline)	—	0.582 (0.017)	0.328 (0.039)

High $\lambda_{\text{consistency}}$ led to overly sparse masks with poor predictive fidelity, while low values produced diffuse, less informative masks. The default setting ($\lambda_{\text{consistency}} = 1$) yielded the best tradeoff between entropy and informativeness.

A.3. Computational Cost

Training on all 319 samples (129 with circulatory failure, 190 negative controls) with batch size 1 yields the following runtimes and GPU memory usage:

Table 6: Runtime and GPU memory profiling (batch size 1)

Model Variant	Training Time	GPU Memory (MB)
EM (baseline)	1.42h	878
EM + PLNN	3.72h	912
MC-EM (no PLNN)	2.66h	878
MC-EM + PLNN	5.05h	960

The compute overhead in MC-EM primarily arises from $K = 100$ stochastic passes for MC Dropout per sample, and PLNN buffers increase GPU memory by approximately 5%.

Appendix B.

B.1. Extended Results - ARDS (Acute Respiratory Distress Syndrome)

We applied the causal-CrossFormer and MC-EM framework to dynamic ARDS prediction on three external ICU cohorts: HiRID (278 ICU stays with at least one ARDS event; Figure 6), MIMIC-IV (49 ICU stays with at least one ARDS event; Figure 7), and eICU-CRD (50 ICU stays with at least one ARDS event; Figure 8) to extend the generalization ability of MC-EM. We selected the ICU Stays with uniform inclusion criteria of at least seven days of continuous data in HiRID or thirty days in MIMIC-IV and eICU, plus at least one ARDS onset during the ICU stay. ARDS labels for MIMIC-IV and eICU-CRD were derived using the METRE pipeline Liao and Voldman (2023) and established clinical definitions Tang et al. (2020). A quantitative comparison of mask entropy and mask information values for both methods is provided in Table 7.

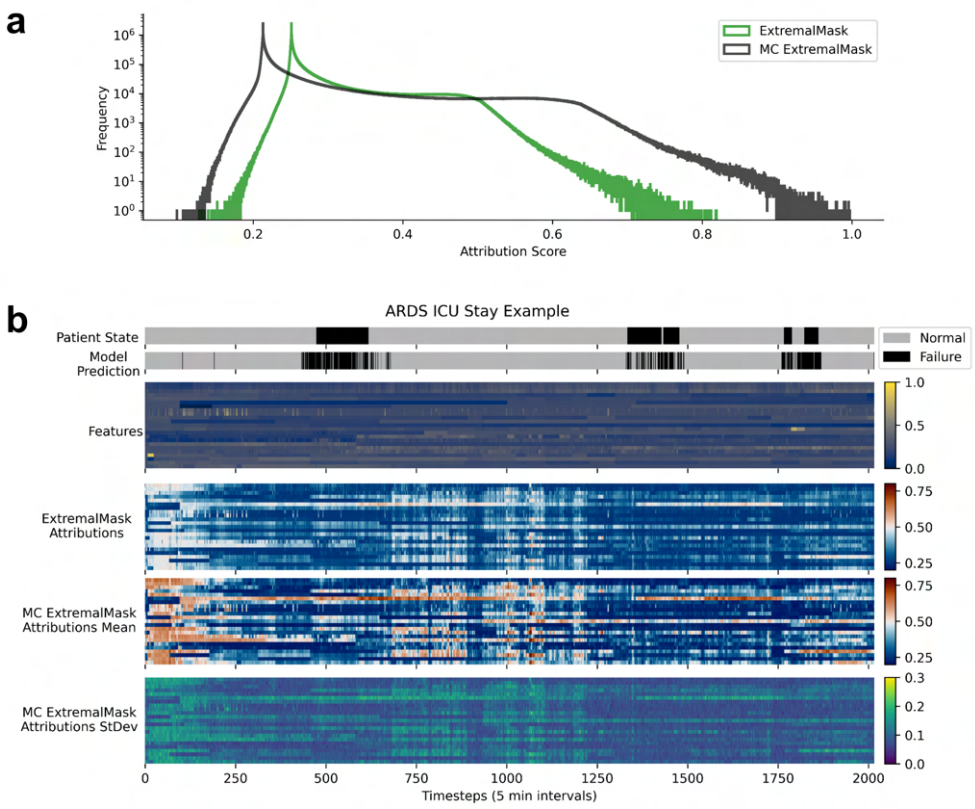

Figure 6: Comparison of MC-EM with EM on the HiRID ARDS Cohort. (a) Attribution scores from both EM and MC-EM methods are shown across 278 ICU stays in the HiRID ARDS test cohort. (b) A representative example of an ARDS ICU stay with four distinct failure events. The panel includes the corresponding ground truth, the model's prediction, time series features, mean attribution masks from both EM and MC-EM methods, as well as the uncertainty estimates (mask standard deviation) from MC-EM.

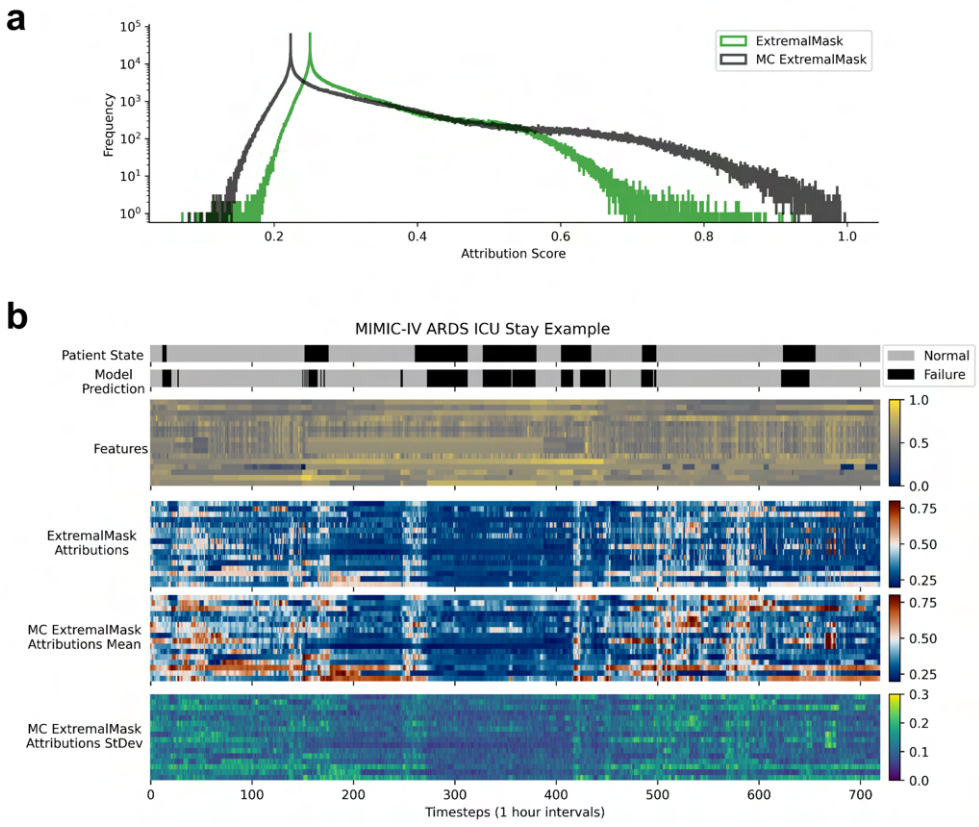

Figure 7: Comparison of MC-EM with EM on the MIMIC-IV ARDS Cohort. (a) Attribution scores from both EM and MC-EM methods are shown across 49 ICU stays in the MIMIC-IV ARDS test cohort. (b) A representative example of an ARDS ICU stay with seven distinct failure events. The panel includes the corresponding ground truth, the model's prediction, time series features, mean attribution masks from both EM and MC-EM methods, as well as the uncertainty estimates (mask standard deviation) from MC-EM.

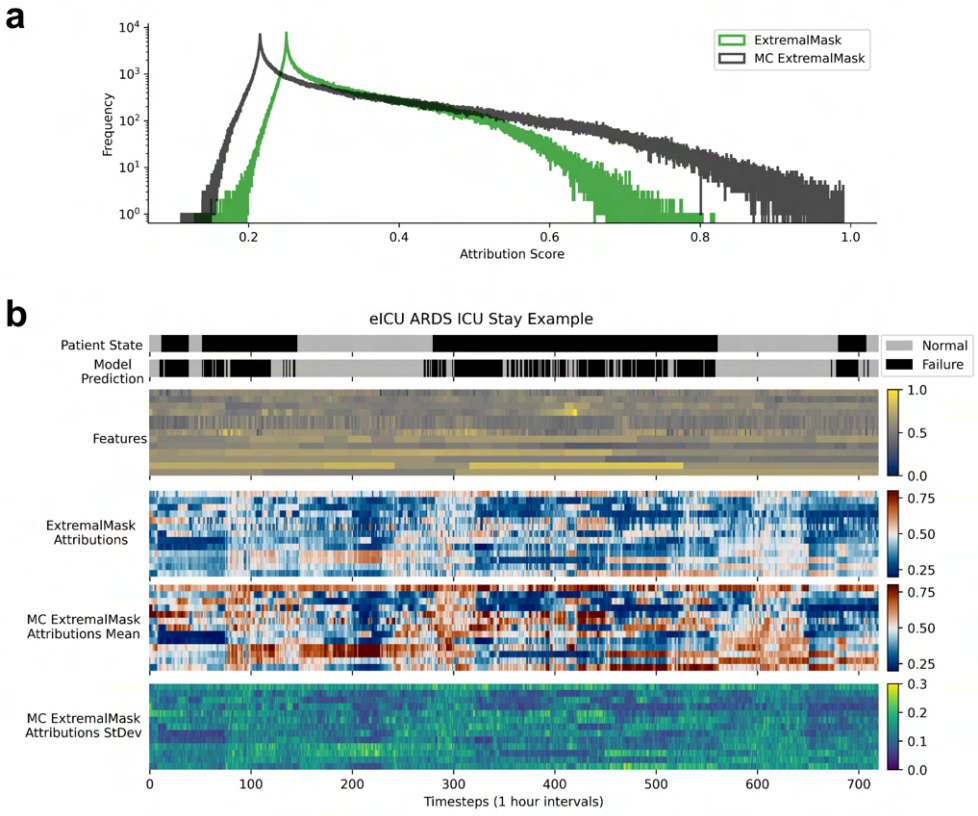

Figure 8: Comparison of MC-EM with EM on the eICU ARDS Cohort. (a) Attribution scores from both EM and MC-EM methods are shown across 50 ICU stays in the eICU ARDS test cohort. (b) A representative example of an ARDS ICU stay with four distinct failure events. The panel includes the corresponding ground truth, the model's prediction, time series features, mean attribution masks from both EM and MC-EM methods, as well as the uncertainty estimates (mask standard deviation) from MC-EM.

Table 7: Mask Entropy and Mask Information for ARDS across datasets. ExMa: Extremal Mask, MC ExMa: Monte Carlo Extremal Mask

Dataset	Model	Mask Entropy	Mask Information
HiRID	ExMa	0.596 (0.022)	0.355 (0.051)
	MC ExMa	0.567 (0.030)	0.374 (0.081)
MIMIC-IV	ExMa	0.599 (0.015)	0.366 (0.035)
	MC ExMa	0.579 (0.020)	0.378 (0.057)
eICU	ExMa	0.609 (0.018)	0.392 (0.044)
	MC ExMa	0.587 (0.024)	0.412 (0.074)

Appendix C.

C.1. ExtremalMask

ExtremalMask (Enguehard, 2023a) is a perturbation-based model agnostic interpretability method specifically developed for univariate and multi-variate time series data. It uniquely learns trainable masks alongside adaptive perturbations to identify which feature-time pairs significantly impact the model output. In contrast, methods (e.g. Dynamask (Crabbé and Van Der Schaar, 2021)) use static perturbations such as Gaussian blurs, averages, or fixed noise, which often fail to accurately capture the temporal nature. ExtremalMask overcomes these limitations by employing a dual-learning approach. It simultaneously optimizes a binary mask which indicates the importance of individual feature-time pairs, and an associated perturbation model $P_\theta(X)$. Typically, this perturbation model is implemented using a unidirectional / bidirectional recurrent neural network (RNN), enabling it to effectively handle sequential dependencies and temporal patterns. ExtremalMask operates through two primary strategies:

- **Preservation game**: Identifies the minimal subset of features required to maintain the original prediction.
- **Deletion game**: Identifies a minimal subset of features whose removal most significantly changes the original prediction.

ExtremalMask demonstrates superior performance over fixed perturbation and traditional time-series model interpretability methods. It provides temporally aware feature-time attributions which are essential for accurate risk assessment and early identification of patient deterioration. (Yadav and Subbian, 2025b,a)

C.2. Monte Carlo Dropout

Monte Carlo Dropout (MCD) applies standard dropout at inference to produce a distribution of outputs without modifying the network. During inference, MCD performs K stochastic forward passes, each sampling a binary mask $z_k \sim \text{Bernoulli}(p)$ on the weights θ,

yielding masked parameters $\theta \odot z_k$. The ensemble of predictions

$$\hat{y} \;=\; \frac{1}{K}\sum_{k=1}^{K} f\big(x;\theta\odot z_k\big), \qquad \widehat{\mathrm{Var}}(y) \;=\; \frac{1}{K}\sum_{k=1}^{K} f\big(x;\theta\odot z_k\big)^2 - \hat{y}^2$$

serves as the point estimate and an epistemic uncertainty proxy. Gal and Ghahramani (2016) showed that this formulation is equivalent to variational inference in a deep Gaussian process with

$$q(\theta) = \sum_{z} q(z)\,\delta\big(\theta - \theta\odot z\big),$$

but Folgoc et al. (2021) demonstrated that this discrete mixture never concentrates as data grows and can assign zero probability to the true model in closed-form tests. Consequently, MCD's variance is a heuristic instability score rather than a true Bayesian credible interval. In our work, we repurposed MCD to measure uncertainty in feature–time attribution masks $m(x;\theta)$ instead of model outputs. We draw K dropout masks $z_k \sim \mathrm{Bernoulli}(p)$ and compute for each

$$m_k = m\big(x;\,\theta\odot z_k\big).$$

For every feature–time pair (i,t), we then calculate

$$\overline{m}_{i,t} = \frac{1}{K}\sum_{k=1}^{K} m_k(i,t), \qquad \mathrm{Var}(m_{i,t}) = \frac{1}{K}\sum_{k=1}^{K} m_k(i,t)^2 - \overline{m}_{i,t}^2.$$

Here, $\mathrm{Var}(m_i)$ quantifies the uncertainty in feature-time pair (i,t) scores across different dropout masks. We aim to capture this uncertainty in interpretation masks rather than to recover a true Bayesian posterior. The simple variational approximation behind MCD is sufficient for our needs and the variability in the sampled masks highlights which attributions are reliable and which depend strongly on model perturbations.