Delta-XAI: A Unified Framework for Explaining Prediction Changes in Online Time Series Monitoring






Changhun Kim* Yechan Mun* Hyeongwon Jang Eunseo Lee Sangchul Hahn† Eunho Yang†

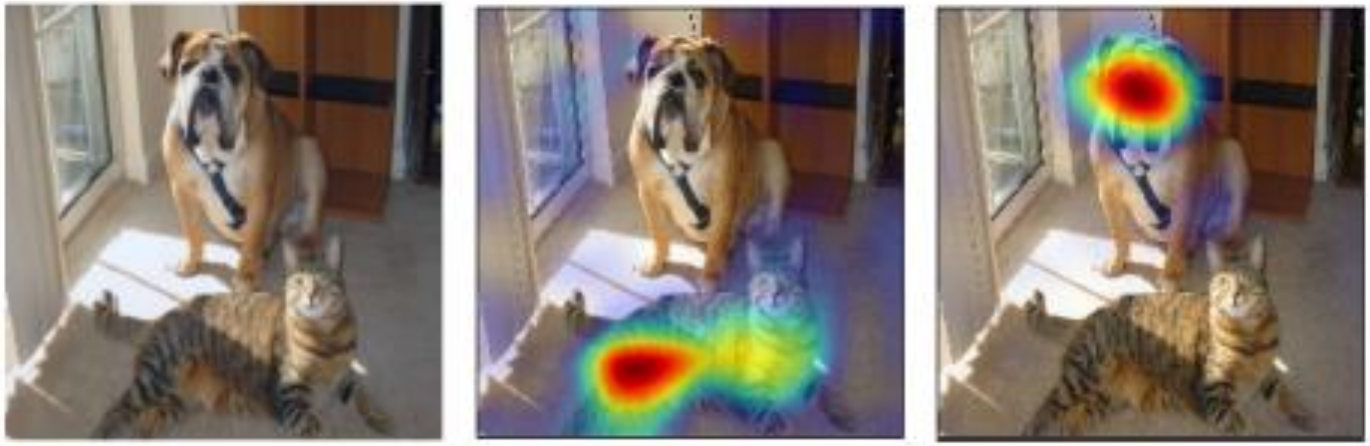
*Equal Contribution †Equal Advising

aitrics KAIST

ICLR

Motivation: Explainable Artificial Intelligence (XAI)

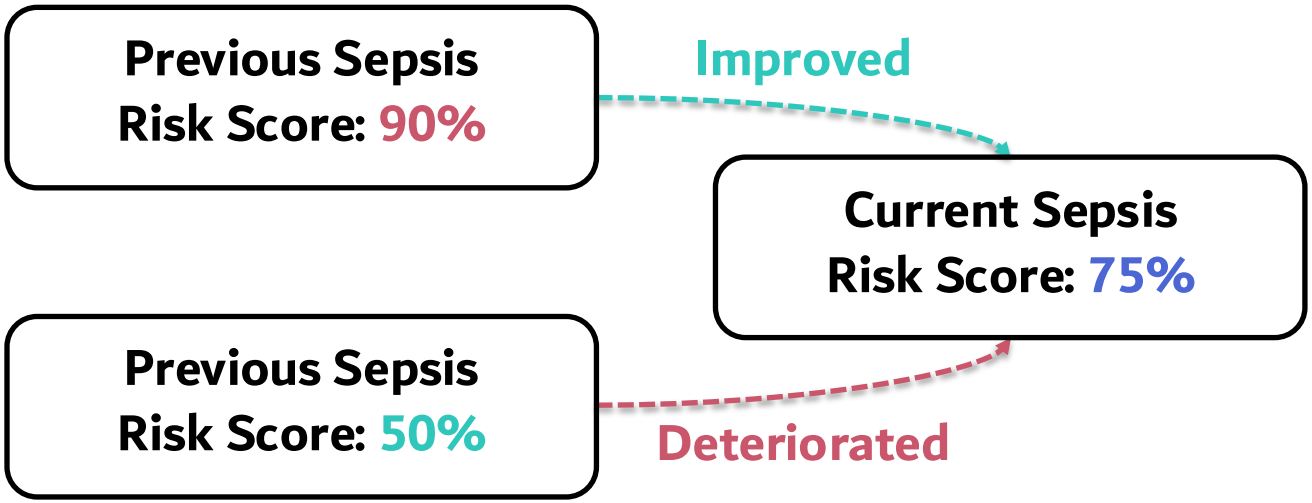
- Modern deep learning models achieve strong predictive performance, but their decision-making processes are often **difficult to interpret**.
- **Explainable artificial intelligence (XAI)** methods such as SHAP, LIME, and Grad-CAM have been proposed to provide human-understandable explanations for model predictions.
- Trustworthy AI systems are increasingly important in real-world applications, especially in **high-stakes domains** such as healthcare, and autonomous driving.



Original Image | Grad-CAM 'Cat' | Grad-CAM 'Dog'

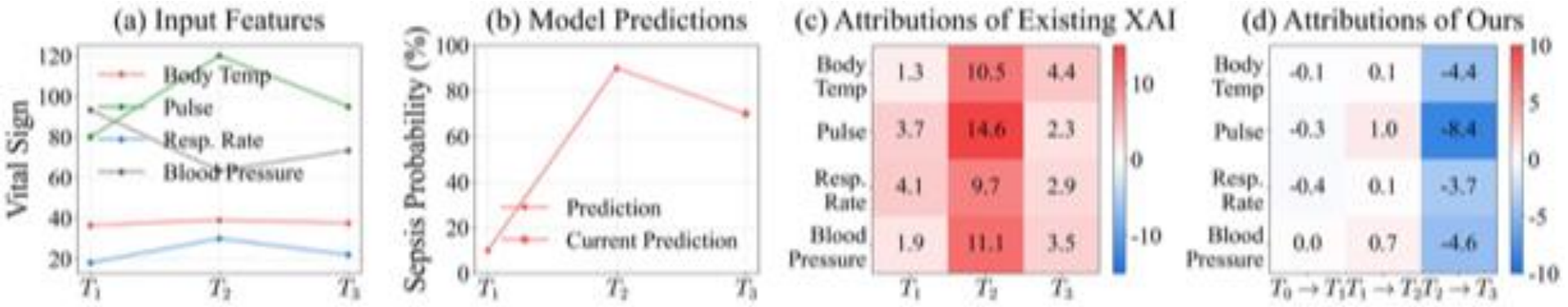
Motivation: XAI for Online Time Series Monitoring

- In online time series monitoring, **explaining prediction changes** between nearby time steps is more important than **explaining a single prediction**.
- For example, in sepsis monitoring, clinicians require explanations for both **improvement** (90% → 75%) and **deterioration** (50% → 75%), even if the **final prediction is the same** (75%).
- Existing XAI methods are mostly designed for **single-time predictions** and cannot adequately explain **temporal prediction changes**.



Motivation: XAI for Online Time Series Monitoring

- A patient's risk may rise from 10% at T_1 to 90% at T_2, then decrease to 70% at T_3. Clinicians want to explain the **recovery** from $T_2 \rightarrow T_3$.
- Existing XAI methods explain only the prediction at T_3, often still highlighting features responsible for the earlier deterioration at T_2.
- Our Delta-XAI framework instead explains prediction **changes** between consecutive time steps, such as $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_3$.



Proposed Method: Adopting Static XAI to Prediction Changes

- We formalize **explanations for prediction change** in online time series monitoring by introducing a **prediction difference wrapper** that **unifies existing 14 single-time XAI methods** and enables them to directly explain temporal prediction changes.

$$g : \mathbb{R}^{(T_2-T_1+W)\times D} \to [0,1]^C,\ g(\mathbf{X}_{T_1-W+1:T_2}) := f(\mathbf{X}_{T_2-W+1:T_2}) - f(\mathbf{X}_{T_1-W+1:T_1}),$$

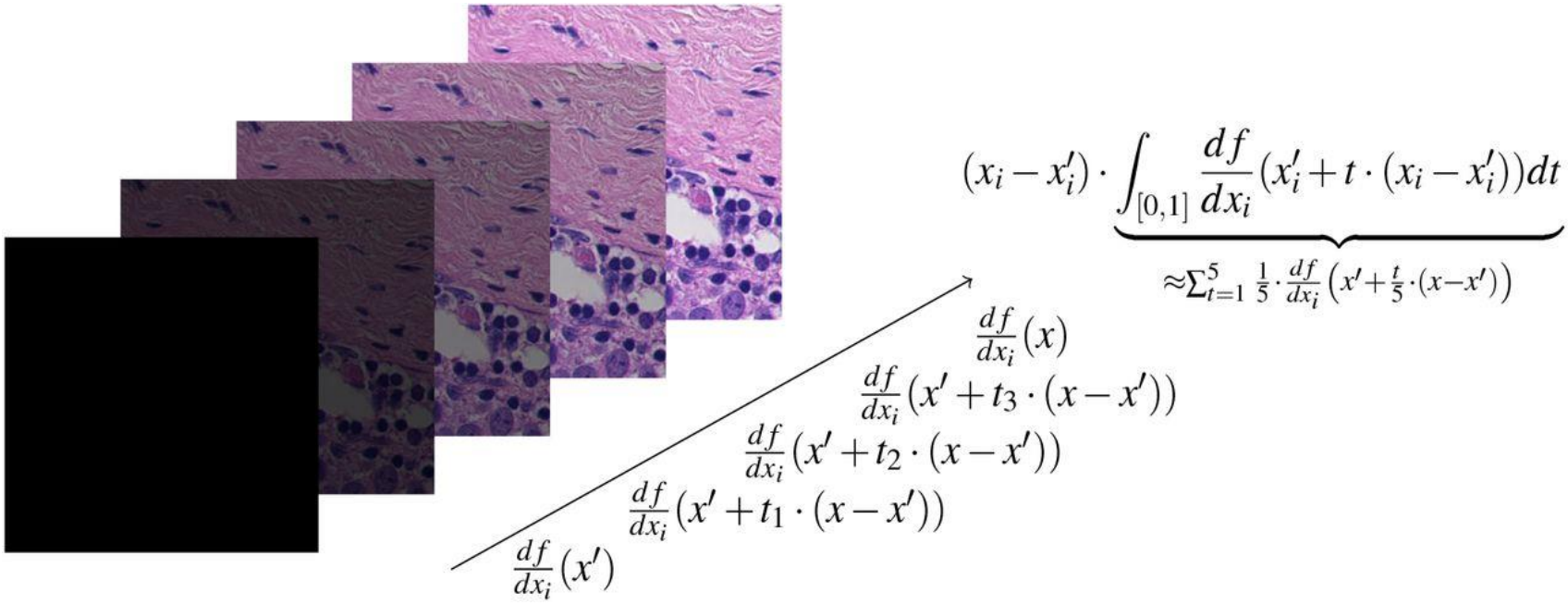
- For **linear and complete methods** such as SHAP and Integrated Gradients, explanations for prediction change reduce to differences of single-time attributions.

Theorem 1 (Attribution Decomposition Theorem for Online Completeness)**.** *Given a linear and complete attribution method φ with a* fixed *baseline, the following decomposition holds:*

$$f(\boldsymbol{X}_{T_2-W+1:T_2})_{\hat{c}} - f(\boldsymbol{X}_{T_1-W+1:T_1})_{\hat{c}} = \underbrace{\sum_{t=T_1+1}^{T_2}\sum_{d=1}^{D}\varphi(f,\boldsymbol{X}_{t,d} \mid T_2)}_{\textit{Addition of newest features}} + \underbrace{\sum_{t=T_2-W+1}^{T_1}\sum_{d=1}^{D}[\varphi(f,\boldsymbol{X}_{t,d} \mid T_2) - \varphi(f,\boldsymbol{X}_{t,d} \mid T_1)]}_{\textit{Delayed effect of intermediate features}} - \underbrace{\sum_{t=T_1-W+1}^{T_2-W}\sum_{d=1}^{D}\varphi(f,\boldsymbol{X}_{t,d} \mid T_1)}_{\textit{Removal of oldest features}}. \quad (4)$$

Proposed Method: Background of Integrated Gradients (IG)

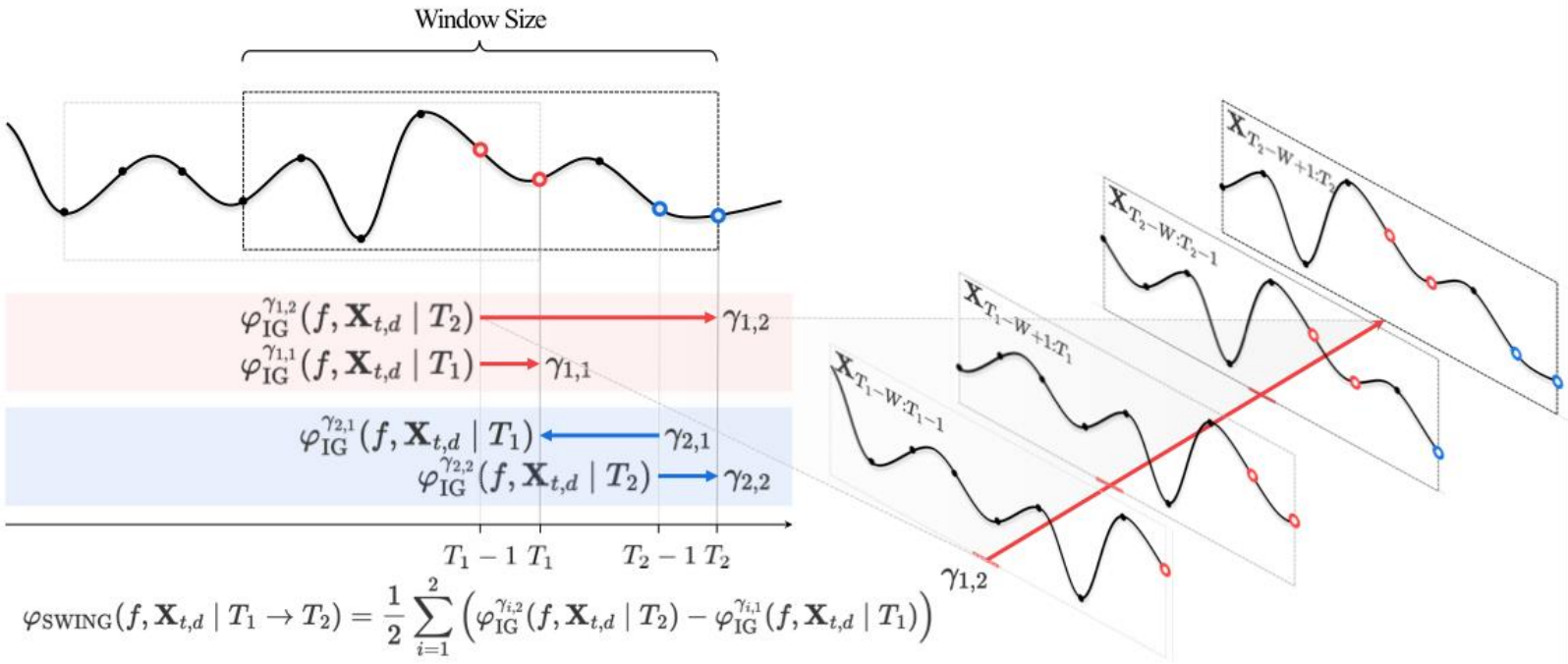
- **Integrated Gradients (IG)** uses linear interpolation from baseline; intermediate points can be **Out-of-Distribution (OOD) region**.
- IG also fails to consider **temporal dependencies;** it merely scales the input.



Rahnfeld, Jens, et al. "A Comparative Study of Explainability Methods for Whole Slide Classification of Lymph Node Metastases using Vision Transformers." PLOS Digital Health, 2025.

Proposed Method: Shifted Window Integrated Gradients (SWING)

- We extend **Integrated Gradients (IG)** to explain prediction changes by integrating along temporally shifted windows rather than static input-baseline pairs.
- SWING uses **recent past observations** as **realistic baselines** and follows **piecewise-linear historical paths** to reduce **off-manifold artifacts**.
- By combining **retrospective baseline selection**, **dual-path integration**, and **historical integration**, SWING produces temporally consistent explanations while preserving key theoretical properties.



Proposed Evaluation Metrics

- **Area-based ranking quality:** AUPD/AUPP aggregate CPD/CPP over all prefixes, evaluating whether highly ranked features consistently affect predictions.

$$\mathrm{AUPD}(g, \mathbf{X}, K) = \frac{1}{2K} \sum_{k=1}^{K} [\mathrm{CPD}(g, \mathbf{X}, k) + \mathrm{CPD}(g, \mathbf{X}, k-1)]$$

- **Macro-level temporal dynamics:** MPD/MPP first aggregate attribution scores across nearby online windows to capture evolving explanations.

$$\varphi(\mathbf{X}_{t,d} \mid T) = \frac{1}{2W-1} \sum_{T'=t-W+1}^{t+W-1} \varphi(\mathbf{X}_{t,d} \mid T')$$

- **Magnitude-aware attribution quality:** Corr. measures whether attribution magnitudes align with actual prediction changes after feature removal.

$$\mathrm{Corr.} = \mathrm{Corr}\left([\varphi^{(1)}, \ldots, \varphi^{(K)}], [|g_1^{\uparrow} - g_0^{\uparrow}|, \ldots, |g_K^{\uparrow} - g_{K-1}^{\uparrow}|]\right)$$

Experiments: Main Table

- SWING consistently outperforms existing XAI methods.

Table 2: Performance comparison of XAI methods on clinical prediction tasks: MIMIC-III decompensation benchmark using LSTM as backbone architecture. Evaluation is performed by removing the most or least salient 50 feature points per time step, using forward-fill substitution.

Algorithm	Removal of Most Salient 50 Points				Removal of Least Salient 50 Points				Corr. ↑
	CPD ↑	AUPD ↑	MPD ↑	AUMPD ↑	CPP ↓	AUPP ↓	MPP ↓	AUMPP ↓	
LIME (Ribeiro et al., 2016)	2.26±0.04	1.72±0.03	13.78±0.08	7.70±0.02	32.46±0.16	14.26±0.10	33.45±0.14	15.30±0.10	0.02±0.00
GradSHAP (Lundberg & Lee, 2017)	13.73±0.06	9.05±0.04	16.68±0.08	11.19±0.05	32.97±0.13	13.96±0.06	30.13±0.17	11.95±0.09	0.14±0.00
IG (Sundararajan et al., 2017)	13.42±0.06	9.10±0.05	16.14±0.07	11.31±0.04	33.55±0.12	13.97±0.04	29.46±0.17	10.85±0.05	0.17±0.00
DeepLIFT (Shrikumar et al., 2017)	13.58±0.06	9.42±0.04	16.03±0.08	11.25±0.06	35.96±0.16	14.61±0.07	31.53±0.15	11.41±0.06	0.19±0.00
FO (Suresh et al., 2017)	13.14±0.10	9.92±0.07	17.79±0.10	12.92±0.06	44.02±0.19	22.32±0.09	16.92±0.09	6.24±0.03	0.26±0.00
AFO (Tonekaboni et al., 2020)	13.24±0.07	9.30±0.05	17.16±0.07	11.95±0.03	36.13±0.22	16.64±0.09	24.13±0.16	9.32±0.07	0.28±0.00
FIT (Tonekaboni et al., 2020)	3.40±0.04	2.70±0.03	7.11±0.04	6.14±0.04	35.52±0.11	17.55±0.08	**12.07±0.05**	10.19±0.05	0.06±0.00
WinIT (Leung et al., 2023)	19.64±0.07	12.25±0.04	**24.87±0.13**	15.45±0.08	29.22±0.07	13.05±0.05	26.11±0.12	11.92±0.06	0.21±0.00
Dynamask (Crabbé & Van Der Schaar, 2021)	11.72±0.08	7.56±0.04	13.15±0.08	8.25±0.04	53.07±0.24	26.22±0.08	49.80±0.16	24.26±0.06	0.04±0.00
Extrmask (Enguehard, 2023)	16.66±0.11	10.47±0.06	17.51±0.12	10.63±0.05	29.91±0.17	15.13±0.09	29.64±0.17	14.84±0.12	0.08±0.00
ContraLSP (Liu et al., 2024b)	12.88±0.36	8.69±0.26	18.00±0.16	11.11±0.18	41.62±0.30	21.17±0.10	42.67±0.29	21.94±0.10	0.03±0.00
TimeX (Queen et al., 2024)	16.99±0.09	11.45±0.06	19.45±0.10	12.45±0.06	50.34±0.10	24.11±0.04	51.06±0.09	24.50±0.05	0.03±0.00
TimeX++ (Liu et al., 2024a)	11.12±0.05	7.00±0.04	13.14±0.02	7.76±0.02	34.21±0.17	13.72±0.07	32.34±0.11	13.08±0.05	0.03±0.00
TIMING (Jang et al., 2025)	14.99±0.07	9.71±0.05	16.50±0.08	11.53±0.04	31.22±0.16	13.36±0.05	27.19±0.19	10.24±0.07	0.19±0.00
SWING	**23.87±0.16**	**16.23±0.10**	22.27±0.19	**15.52±0.12**	**17.76±0.04**	**5.85±0.04**	18.20±0.06	**6.06±0.05**	**0.40±0.00**

Experiments: Ablation Study

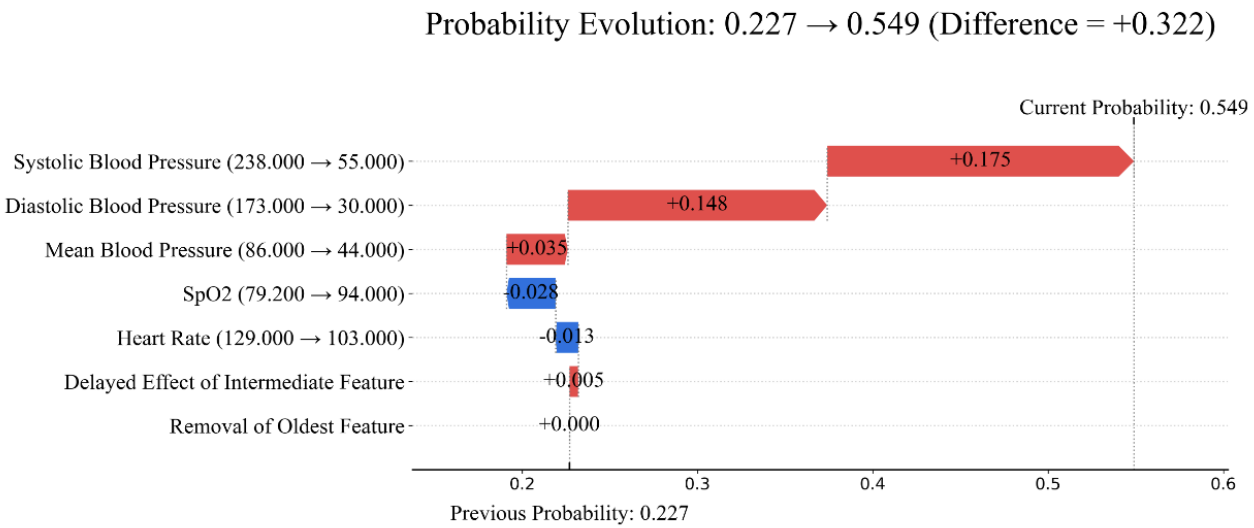
- Each component contributes to reliable prediction-change explanations.

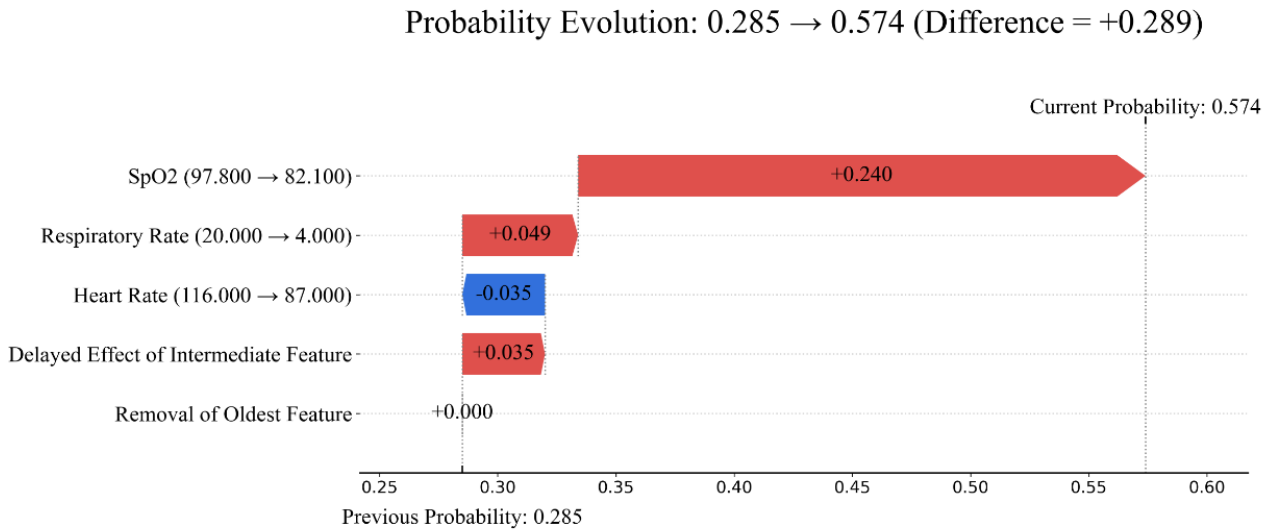
Table 3: Ablation study of SWING examining retrospective baseline selection (RBS), dual-path integration (DPI), and piecewise-linear historical integration (PHI) on the MIMIC-III decompensation benchmark (Johnson et al., 2016), with LSTM (Hochreiter & Schmidhuber, 1997) backbone and interval $T_2 - T_1 = 24$. We vary the baseline distance d (default: 1) and remove the most or least salient 50 feature points per time step, using forward-fill substitution.

Algorithm	Removal of Most Salient 50 Points				Removal of Least Salient 50 Points				Corr. ↑
	CPD ↑	AUPD ↑	MPD ↑	AUMPD ↑	CPP ↓	AUPP ↓	MPP ↓	AUMPP ↓	
w/o RBS, PHI	45.82±0.20	26.85±0.13	40.85±0.22	26.31±0.14	104.05±0.34	48.43±0.19	77.94±0.30	29.56±0.11	0.20±0.00
w/o RBS	40.25±0.22	24.39±0.12	46.71±0.35	29.10±0.21	79.70±0.30	31.29±0.07	82.59±0.35	33.33±0.11	0.16±0.00
w/o DPI ($\gamma_{1,1}, \gamma_{2,2}$)	**55.01**±0.35	**32.39**±0.21	**53.09**±0.38	**33.08**±0.21	71.01±0.30	26.12±0.10	76.26±0.34	29.29±0.09	0.19±0.00
$d = 0$	33.80±0.23	20.11±0.13	48.98±0.44	29.78±0.24	82.28±0.29	33.99±0.10	83.69±0.35	33.31±0.12	0.11±0.00
$d = 3$	42.00±0.33	26.00±0.28	47.49±0.44	29.74±0.34	68.53±0.49	25.60±0.18	71.88±0.62	27.54±0.27	0.19±0.00
$d = 5$	41.98±0.31	25.65±0.25	47.18±0.45	29.44±0.33	72.31±0.66	27.88±0.22	74.62±0.71	29.11±0.26	0.18±0.00
$d = 10$	41.02±0.28	24.94±0.18	46.89±0.39	29.23±0.25	76.55±0.48	29.96±0.16	78.11±0.44	31.04±0.16	0.17±0.00
SWING	41.07±0.22	26.46±0.18	50.58±0.28	32.29±0.20	**60.29**±0.14	**21.60**±0.08	**64.43**±0.19	**23.87**±0.10	**0.21**±0.00

Experiments: Qualitative Analysis

- SWING is not only accurate, but also provides clinically coherent explanations.





Conclusion

- We formulate **explanations for prediction change** in online time series monitoring and **adapt 14 existing XAI methods** through a **unified prediction wrapper** with **tailored evaluation metrics**.
- We propose **SWING**, which extends IG with **historical integration paths** to capture temporal dynamics, mitigate OOD effects, and preserve key theoretical properties.
- Extensive experiments show that **SWING consistently outperforms state-of-the-art time series XAI methods** across diverse benchmarks, backbones, and metrics.

Paper

Code