DeltaSHAP: Explaining Prediction Evolutions in Online Patient Monitoring with Shapley Values

Changhun Kim[1,2*] Yechan Mun[1*] Sangchul Hahn[1] Eunho Yang[1,2]

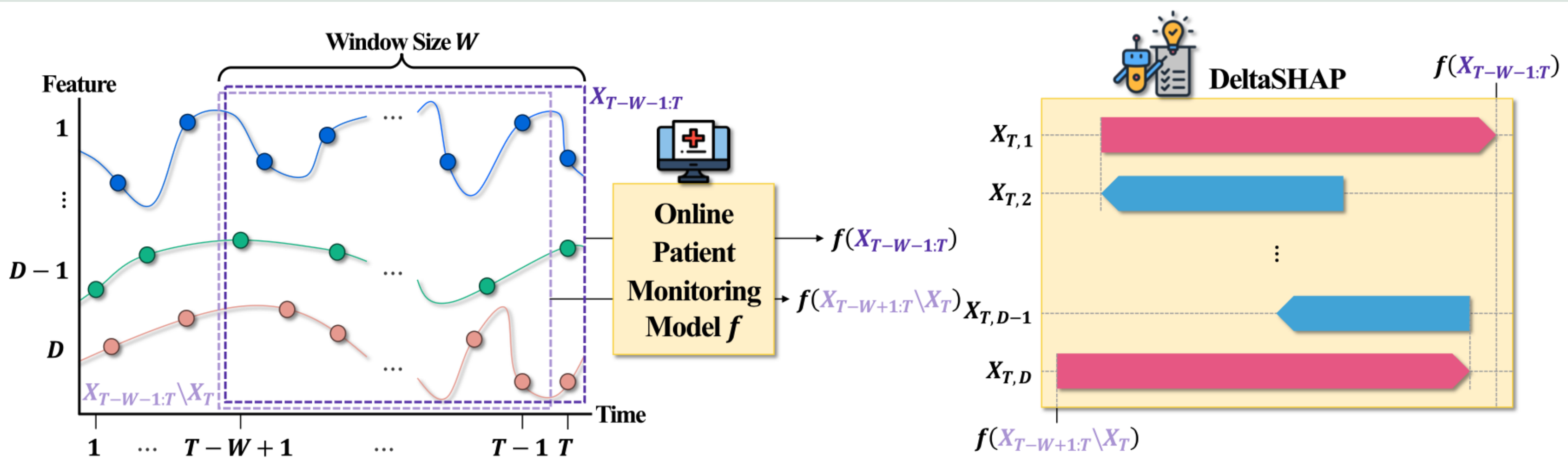
[1]AITRICS [2]KAIST *Equal Contribution

TL;DR: We propose a novel SHAP-based XAI algorithm tailored for online patient monitoring.

Contribution

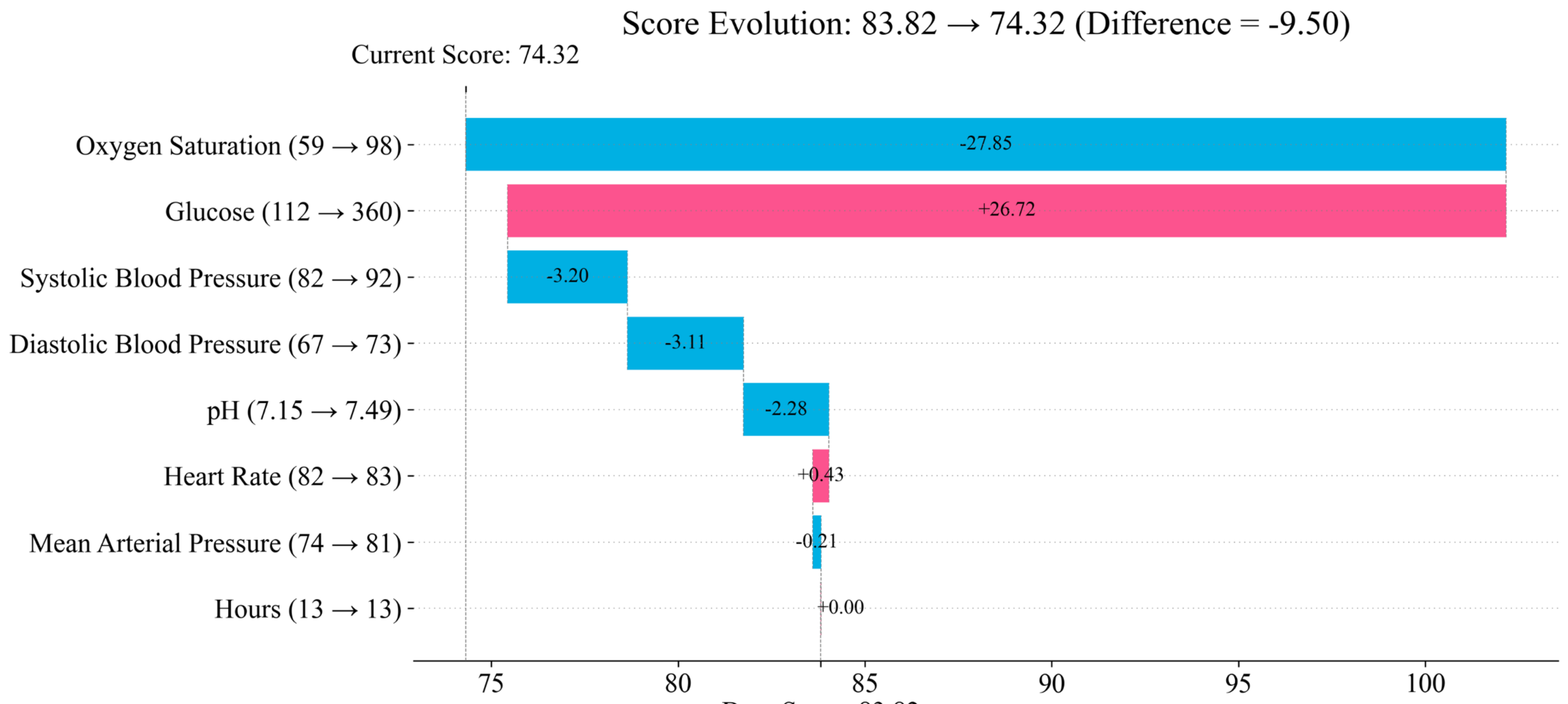
- We propose **DeltaSHAP**, a novel XAI algorithm for **online patient monitoring** that attributes **prediction changes** over time to **newly observed features**, with **directional attributions** and **real-time efficiency** via Shapley Value Sampling.
- We introduce **new evaluation metrics**—Area Under Prediction Difference (**AUPD**) and Area Under Prediction Preservation (**AUPP**)—to quantitatively evaluate the **faithfulness** and **sufficiency** of feature attributions in time series XAI.
- Extensive experiments on real-world clinical datasets demonstrate that DeltaSHAP provides more **faithful, stable, and clinically interpretable explanations** compared to existing XAI baselines.

Proposed Method: DeltaSHAP



- **DeltaSHAP** attributes the **prediction change** caused by **newly observed features** at time T by considering:
$$\Delta = f(X_{T-W+1:T}) - f(X_{T-W+1:T} \setminus X_T).$$
- **Shapley Value Sampling** estimates each feature's **marginal effect** by averaging over **sampled permutations**:
$$\hat{\phi}_j(f, X_{T-W+1:T}) = \frac{1}{N} \sum_{\pi \in \Omega} \left[v(S_{\pi,j} \cup \{j\}) - v(S_{\pi,j}) \right],$$
where $S_{\pi,j}$ is set of features before j in permutation π, and $v(S)$ is a model output when only features in S at T are observed.
- **Baseline Selection**
  - Missing features are filled with **last observations**, matching **preprocessing** and avoiding **unrealistic imputations**.
- **Why DeltaSHAP?**
  - **Intuitive**: attributes **prediction change** to **observed features** at T with **directional** explanation.
  - **Practicality**: model-agnostic and time-efficient.
  - **Normalization** with $\phi_j(f, X_{T-W+1:T}) = \hat{\phi}_j(f, X_{T-W+1:T}) \cdot \Delta / \sum_{k \in \mathcal{F}_{\text{obs}}} \hat{\phi}_k(f, X_{T-W+1:T})$ satisfies the **efficiency** property.

Qualitative Experiments



- DeltaSHAP provides **clinically intuitive explanations**: **reduced oxygen saturation lowers** the risk score, while **increased glucose raises it**—consistent with **clinical understanding**.

Limits of Current XAI in Online Monitoring

- Explaining **prediction differences between time steps** is essential for understanding **patient risk evolution**, but existing methods **rarely capture temporal change**.
- Clinicians need **directional attributions**—how recent feature changes increase or decrease risk—rather than **unsigned importance** alone, but recent time series XAI methods provide no **directionality**.
- These attributions must be computed in **real time**, but current approaches are often **too slow for practical use, limiting their clinical utility**.

Proposed Evaluation Metrics

- Let $f(X)$ be a model prediction, and let $X_k^{\uparrow}$ and $X_k^{\downarrow}$ be the input with the top-k and bottom-k important features removed, respectively. Cumulative Prediction Difference / Preservation (**CPD / CPP**) are defined as:
$$\mathrm{CPD}(f, X, K) = \sum_{k=0}^{K-1} \left| f(X_k^{\uparrow}) - f(X_{k+1}^{\uparrow}) \right|,$$
$$\mathrm{CPP}(f, X, K) = \sum_{k=0}^{K-1} \left| f(X_k^{\downarrow}) - f(X_{k+1}^{\downarrow}) \right|.$$
- Extending these, we define Area Under Prediction Difference (**AUPD**) and Area Under Prediction Preservation (**AUPP**) as:
$$\mathrm{AUPD}(f, X, K) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{CPD}(f, X, k),$$
$$\mathrm{AUPP}(f, X, K) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{CPP}(f, X, k).$$
- **Why AUPD and AUPP?**
  - **Ranking-sensitive:** Emphasize the impact of **higher-ranked features** for more meaningful attribution evaluation.
  - **Smoothing effect:** Aggregating over multiple k values mitigates **instability** from **individual steps**.

Quantitative Experiments

Algorithm	AUPD ↑		AUPP ↓		Wall-Clock Time	
	MIMIC-III	P19	MIMIC-III	P19	MIMIC-III	P19
LIME	8.20±0.03	1.13±0.00	21.58±0.03	3.58±0.00	0.22	0.29
GradSHAP	6.20±0.02	0.96±0.00	19.68±0.03	3.09±0.00	<u>0.03</u>	0.04
IG	13.46±0.00	2.28±0.00	14.51±0.00	2.42±0.00	0.04	0.04
DeepLIFT	<u>13.95±0.00</u>	2.24±0.00	14.35±0.00	2.45±0.00	<u>0.03</u>	<u>0.03</u>
FO	13.55±0.00	2.34±0.00	<u>14.14±0.00</u>	2.37±0.00	1.43	1.14
AFO	13.08±0.05	<u>3.27±0.00</u>	15.14±0.04	<u>1.03±0.00</u>	39.62	14.18
FIT	12.60±0.00	2.15±0.00	16.16±0.00	3.08±0.00	0.12	0.11
WinIT	10.06±1.48	1.27±0.00	16.56±1.75	3.23±0.00	0.30	0.29
DeltaSHAP	**22.59±0.01**	**3.68±0.00**	**3.04±0.01**	**0.89±0.00**	**0.02**	**0.02**

- **Main Results:** DeltaSHAP achieves 62% higher faithfulness than prior methods across clinical benchmarks including MIMIC-III and PhysioNet 2019 with LSTM backbone architecture.
- **Computational Efficiency:** Runs 33% faster than existing time-series XAI methods, enabling real-time use.
- **Ablation study** shows forward-fill boosts attribution quality, normalization ensures efficiency, and sampling while $N = 25$ balances speed and accuracy.