OrdShap: Feature Position Importance for Sequential Black-Box Models

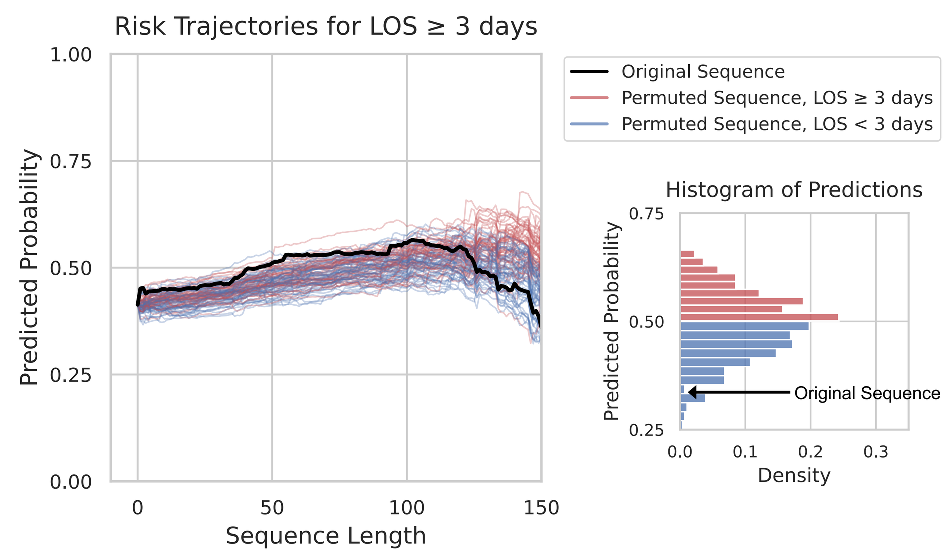
Davin Hill[1], Brian Hill[2†], Aria Masoomi[1], Vijay Nori[3], Robert Tillman[3], Jennifer Dy[1]

[1]Northeastern University [2]Age Bold [3]Optum AI

† Work done while at Optum AI

Introduction

- **Context:** Sequential deep learning classification / regression models (e.g. Transformers, RNNs) excel in domains with temporal dependencies.
- **Feature attribution** methods quantify the importance of each feature with respect to the model prediction.
- **Problem:** Current attribution methods mask / perturb feature *values* – but sequential models are also dependent on feature *order*:



Background

Shapley Value [1]

$$\phi_i(N,\nu) = \sum_{S\subseteq N\setminus\{i\}} \frac{(|N|-|S|-1)!(|S|!)}{|N|!}\left[\nu(S\cup\{i\}) - \nu(S)\right]$$

Sanchez-Bergantiños Value [2]

$$\phi_i^{\mathrm{SB}}(N,\omega) = \sum_{S\subseteq N\setminus\{i\}}\sum_{\sigma\in\mathfrak{S}_S}\frac{(|N|-|S|-1)!}{|N|!(|S|+1)}\sum_{l=1}^{|S|+1}\left[\omega(\sigma_{+i,\ell}) - \omega(\sigma)\right]$$

Main Contributions

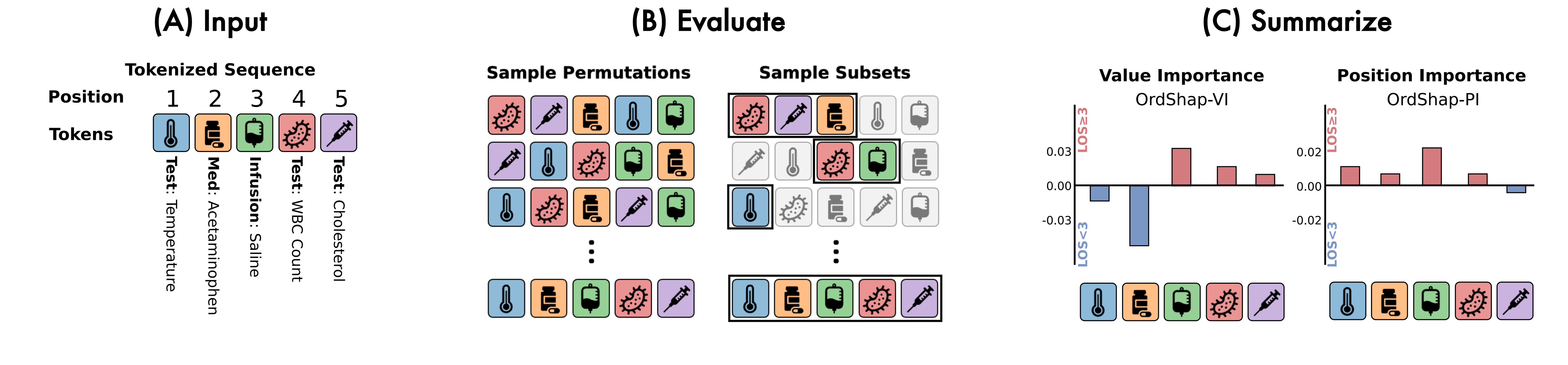
- **OrdShap framework:** First method to disentangle value importance (VI) from position importance (PI) in sequential models
- **Theoretical grounding**: Established game-theoretic connection to Sanchez-Bergantiños values with desirable axiomatic properties
- **Algorithms**: Two efficient methods to approximate OrdShap

Acknowledgements / References

[1] Lloyd S. Shapley. A value for n-person games. 1953.
[2] Estela Sanchez and Gustavo Bergantiños. On Values for Generalized Characteristic Functions. 1997.

This work was supported in part by the National Heart, Lung, and Blood Institute

Overview



Method

1. OrdShap

Novel characteristic function for sequential models:

$$\tilde{\omega}_{f,x}(S,\sigma) = \mathbb{E}_{x'\sim\mathcal{X}}\left[f(x')\,\middle|\, x'_{\sigma^{-1}(i)} = \begin{cases} x_i & \forall i \in S \\ x'_i & \forall i \in N\setminus S\end{cases}\right]$$

OrdShap: Attributions conditioned on feature position

$$\gamma_{i,\ell}(N,\tilde{\omega}) = \sum_{\substack{S\subseteq N\\ i\in S}}\sum_{\substack{\sigma\in\mathfrak{S}_N\\ \sigma^{-1}(i)=\ell}}\frac{(|S|-1)!(|N|-|S|)!}{(|N|-1)!|N|!}\left[\tilde{\omega}(S,\sigma) - \tilde{\omega}(S\setminus\{i\},\sigma)\right]$$

2. Value Importance (OrdShap-VI)

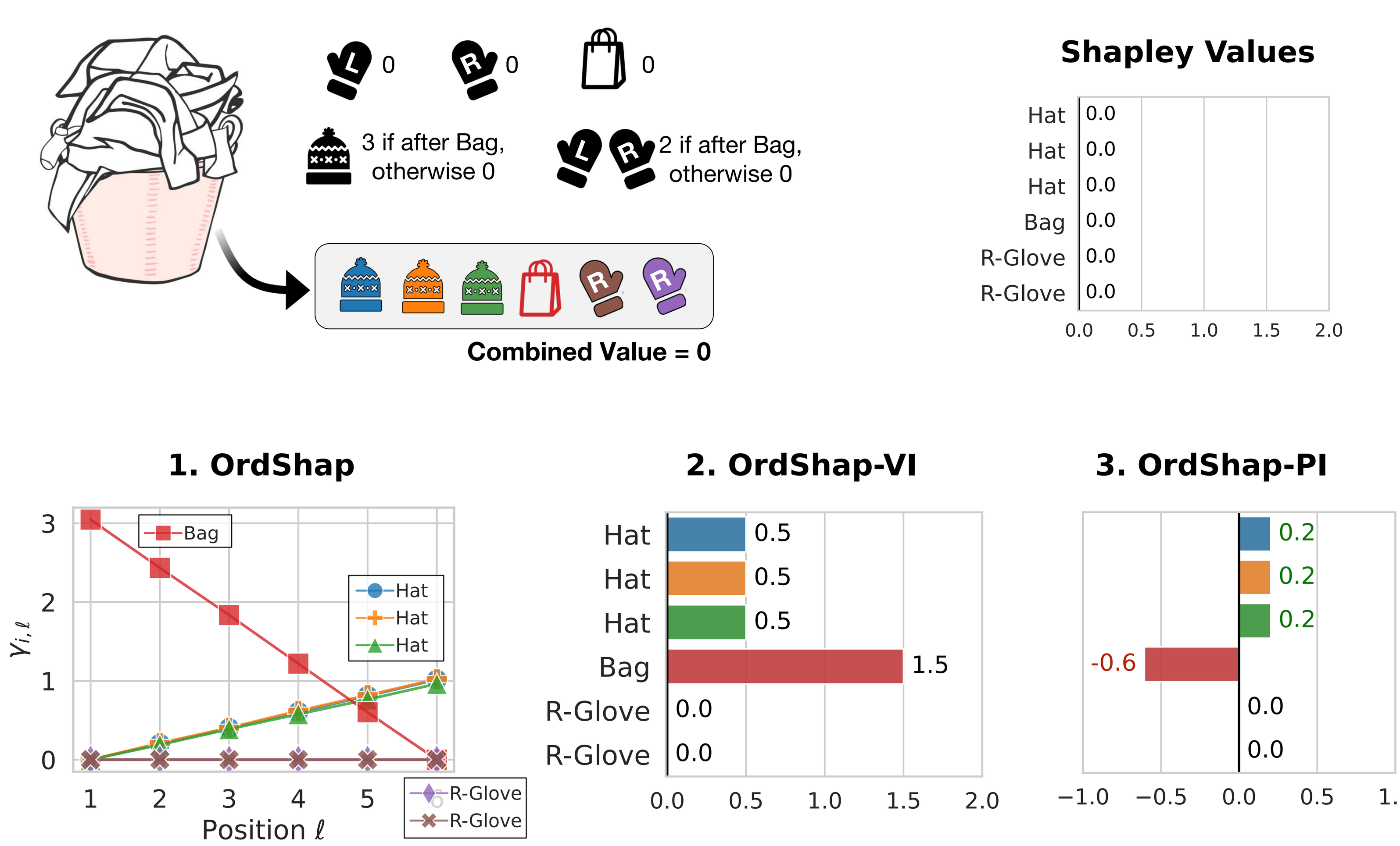
$$\text{OrdShap-VI}_i:\ \bar{\gamma}_i(N,\tilde{\omega}) = \frac{1}{|N|}\sum_{\ell\in N}\gamma_{i,\ell}(N,\tilde{\omega})$$

Theorem 1: OrdShap-VI satisfies the SB axioms. Given $\tilde{\omega}$, there exists a corresponding function ω representing $\tilde{\omega}$ averaged over all $\sigma\in\mathfrak{S}_N$ which contain a given $\pi\in\mathfrak{S}_S\ \forall S\subseteq N$. Then $\bar{\gamma}_i(N,\tilde{\omega}) = \phi^{\mathrm{SB}}(N,\omega)$.

3. Position Importance (OrdShap-PI)

$$\text{OrdShap-PI}_i:\ \underset{\beta_i}{\arg\min}\left[(\ell-\bar{\ell})\beta_i - (\gamma_{i,\ell}(N,v) - \bar{\gamma}_i)\right]^2$$

Illustrative Example



Approximation

Naïve Calculation: $\mathcal{O}(d!\,2^d\,\delta_f)$

MC Sampling: $\mathcal{O}(dKL\delta_f + d^2KL)$

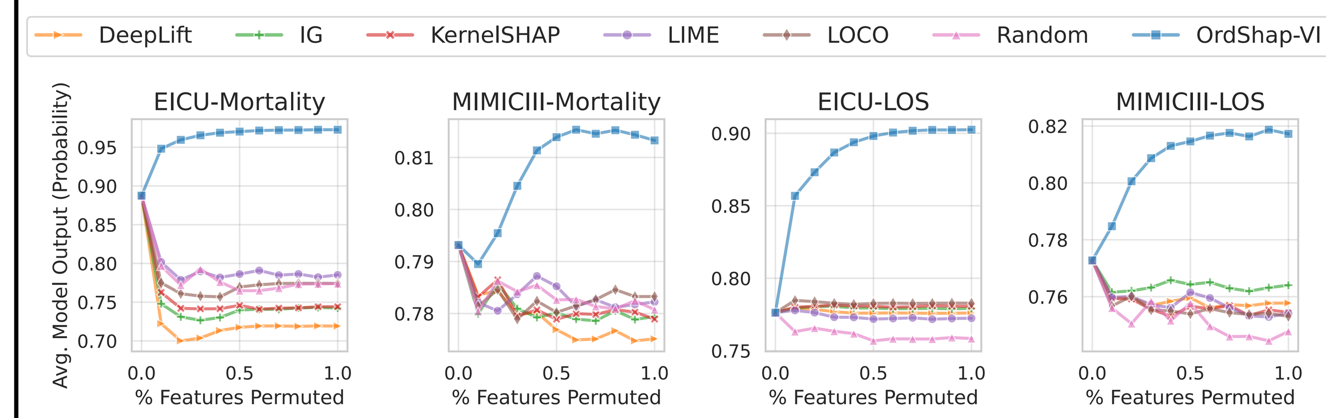
Least-Squares Algorithm with MC Sampling: $\mathcal{O}(KL\delta_f + d^2KL + d^3)$

$$\min_{\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_n}\sum_{\substack{S\subseteq N\\ S\neq\varnothing,N}}\sum_{\sigma\in\mathfrak{S}_N}\mu(|S|)\left[\sum_{i\in S}\alpha_i + \sum_{i\in S}\left[\sigma^{-1}(i)-\bar{\ell}\right]\beta_i - \left[\tilde{\omega}(S,\sigma) - \tilde{\omega}(\varnothing,\sigma_N^{id})\right]\right]^2$$

$$s.t.\ \sum_{i\in N}\alpha_i = \frac{1}{|N|!}\sum_{\sigma\in\mathfrak{S}_N}\tilde{\omega}(N,\sigma) + \tilde{\omega}(\varnothing,\sigma_N^{id})$$

Theorem 2: Optimal α, β corresponds to OrdShap

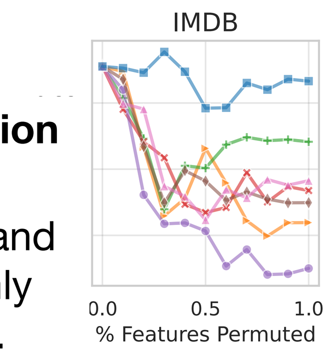
Experiments



Top: Quantitative evaluation of OrdShap-PI. Model output is evaluated while iteratively permuting tokens according to attribution magnitude.

Bottom: AUC evaluation of OrdShap-VI

We calculate IncAUC and ExcAUC while randomly permuting token order.



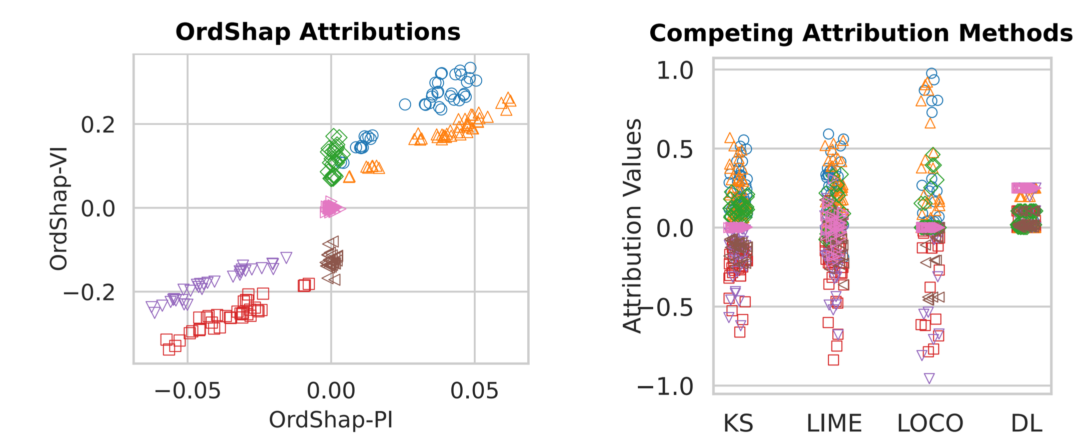
Metric	Explainer	EICU-LOS	EICU-Mort	MIMICIII-LOS	MIMICIII-Mort	IMDB
Inclusion AUC ↑	DL	0.906 (0.010)	0.804 (0.013)	0.893 (0.011)	0.854 (0.012)	0.784 (0.016)
	IG	0.903 (0.010)	0.803 (0.013)	0.896 (0.011)	0.859 (0.012)	0.791 (0.016)
	KS	0.904 (0.010)	0.801 (0.014)	0.898 (0.011)	0.855 (0.012)	0.863 (0.011)
	LIME	0.866 (0.012)	0.776 (0.016)	0.885 (0.012)	0.804 (0.015)	0.859 (0.012)
	LOCO	0.904 (0.010)	0.799 (0.014)	0.894 (0.011)	0.854 (0.012)	0.855 (0.012)
	Random	0.812 (0.014)	0.769 (0.016)	0.851 (0.014)	0.705 (0.017)	0.797 (0.016)
	OrdShap-VI	**0.913** (0.010)	**0.809** (0.013)	**0.899** (0.011)	**0.862** (0.011)	**0.866** (0.012)
Exclusion AUC ↓	DL	0.614 (0.025)	0.727 (0.022)	0.817 (0.018)	0.484 (0.025)	0.855 (0.012)
	IG	0.629 (0.024)	0.728 (0.021)	0.795 (0.019)	0.481 (0.025)	0.867 (0.013)
	KS	0.626 (0.024)	0.730 (0.021)	0.805 (0.019)	0.485 (0.025)	**0.766** (0.018)
	LIME	0.771 (0.017)	0.750 (0.018)	0.840 (0.016)	0.591 (0.022)	0.804 (0.016)
	LOCO	0.618 (0.025)	0.732 (0.021)	0.803 (0.019)	0.495 (0.025)	0.784 (0.017)
	Random	0.821 (0.014)	0.759 (0.017)	0.866 (0.013)	0.694 (0.018)	0.849 (0.014)
	OrdShap-VI	**0.573** (0.028)	**0.724** (0.022)	**0.788** (0.020)	**0.472** (0.025)	0.779 (0.021)

Right: Evaluation on nonlinear, synthetic data. Tokens are generated with positional dependency; this dependency can be detected using OrdShap.

Synthetic Data Definition

Token t_i	Value $v(t_i)$	VI	PI	Legend
A	$0.1i^2+2$	✓	✓	○
B	$0.1i^2$	✓	✓	△
C	2	✓	✗	◇
D	0	✗	✗	▷
$\bar{A},\bar{B},\bar{C}$	$-\{A,B,C\}$			□ ▽ ◁

$$f_{\text{nonlinear}}(t) = \text{sigmoid}\left(\sum_{i=1}^{10} v(t_i)\right)$$



Right: Qualitative Case Study For Length-of-Stay Prediction

We apply OrdShap and KernelShap to a sample and investigate tokens with high attribution