DynSHAP: Towards Explainable Dynamic Survival Analysis

Nastasya Anokhina, Jonas Jürß, Pietro Liò

UNIVERSITY OF CAMBRIDGE

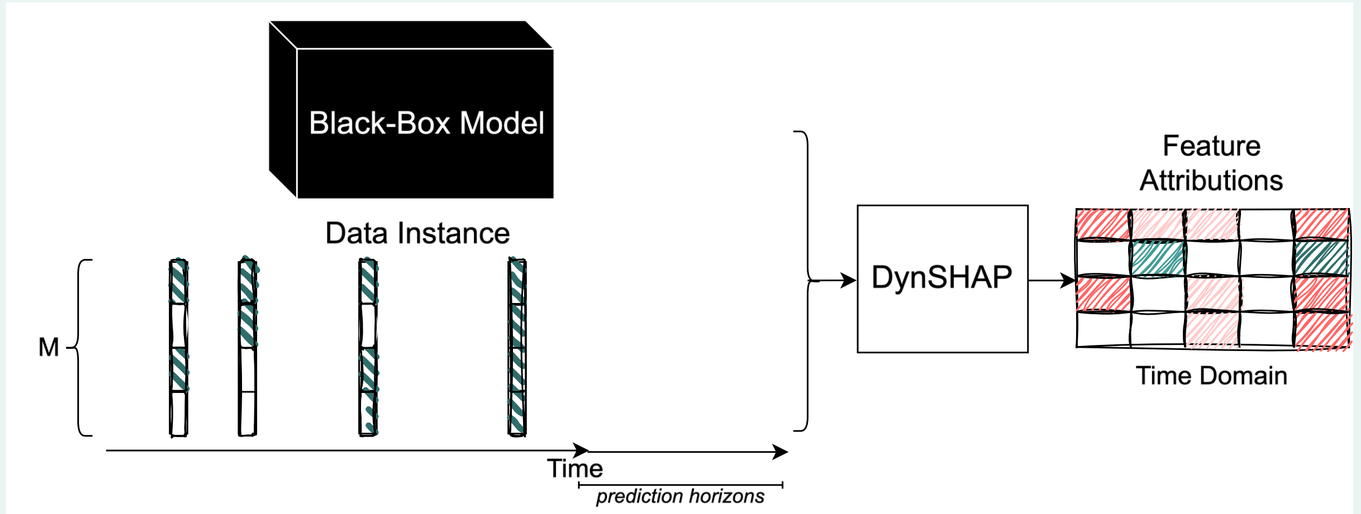
NeurIPS RCMLR Workshop

Motivation

Existing SHAP methods are **not directly applicable to Dynamic Survival Analysis** models:

- They fail to jointly handle **longitudinal input and functional survival output**
- They assume **feature independence** across time, but in practice this assumption does not hold.

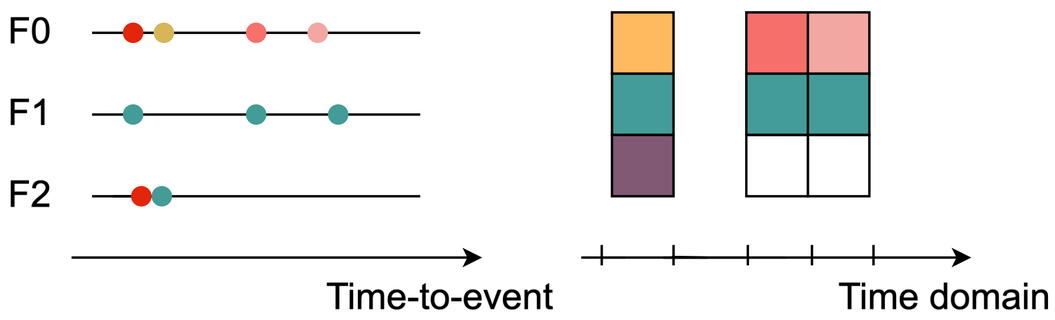
In addition to estimating the importance of every feature, DynSHAP answers the question: ***at around what time did this feature most influence the predicted survival?***



Methods

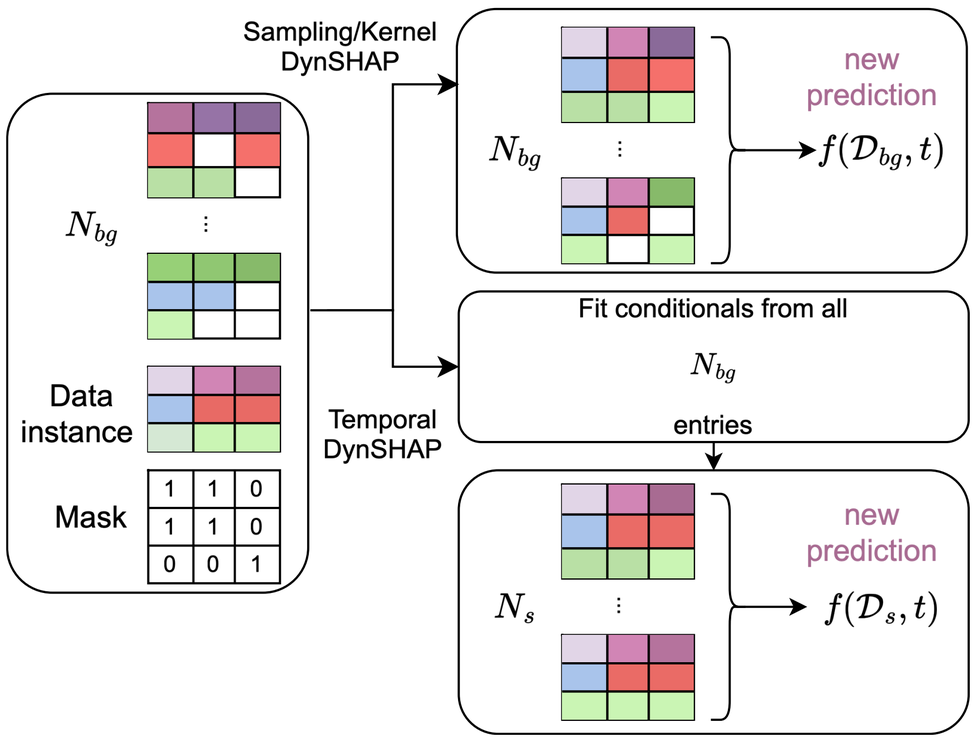
Discretising the Time Domain

Each **time-feature pair is a player** in the Shapley game

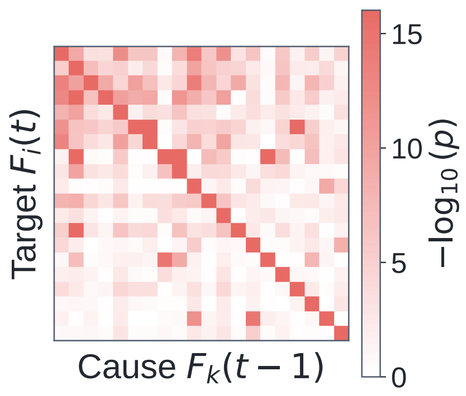


DynSHAP Estimators

- **Sampling/Kernel** DynSHAP: marginal sampling → *"what would the model predict if we intervened on player X?"*
- **Temporal** DynSHAP: conditional sampling → *"what would the model predict given a **realistic** sample with the intervened player value X?"*



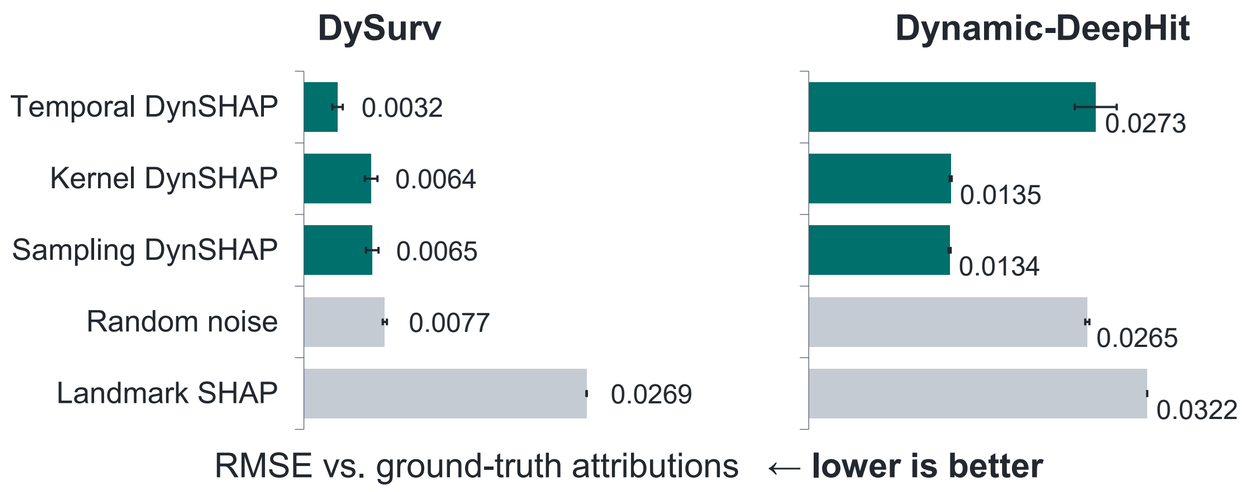
"Realistic" sample: true data distribution unknown, so estimate it with linear-gaussian models, motivated by Granger causality tests.



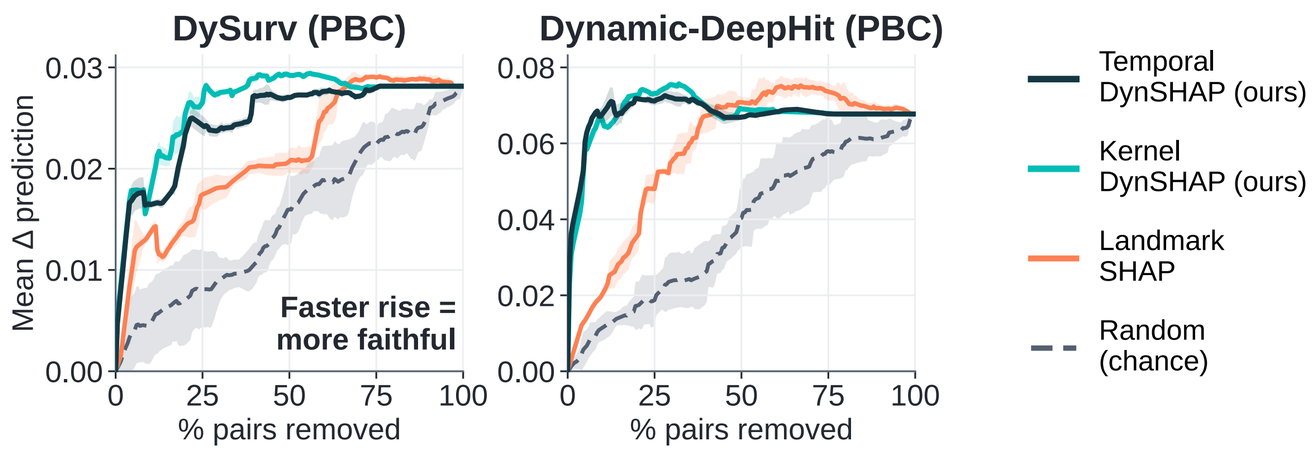
Experiments

Quantitative results

- DynSHAP **recovers ground truth** better than the baselines in synthetic data
- Temporal DynSHAP **recovers temporally dependent features better** for one of the state-of-the-art DSA models but not the other



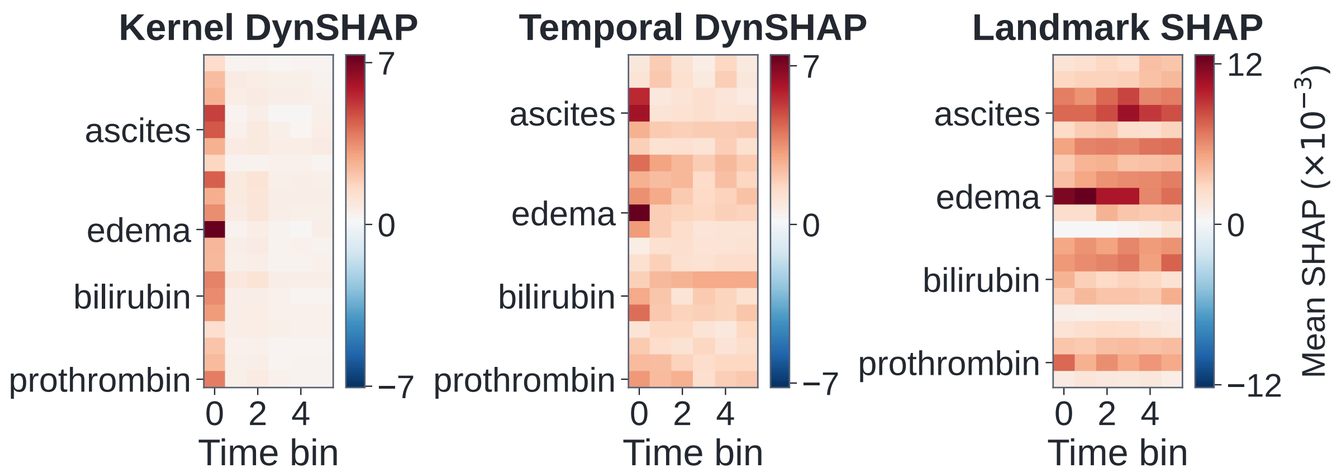
- DynSHAP is **more faithful to model learning** than the baselines



Qualitative results

Dynamic-DeepHit on PBC

- DynSHAP's explanations are **consistent with medical findings**



CODE

PAPER

Takeaway

DynSHAP attributes each dynamic survival prediction to individual time–feature pairs, showing which patient measurements mattered and when.