5. **Time-dependent explanations of neural networks for survival analysis

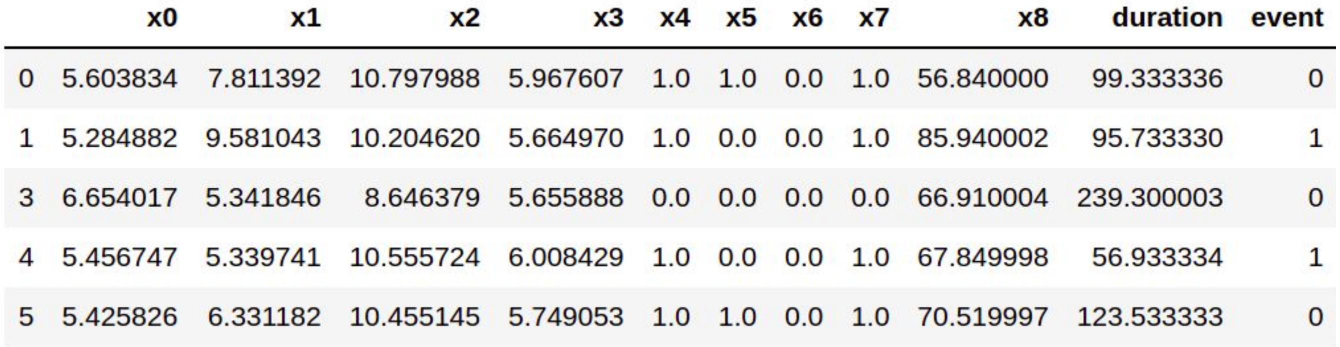
Keywords: tabular data, neural networks, experiments, pytorch

Goal: Compare SurvSHAP(t) model-agnostic explanation for survival models to explanations specific to neural networks, e.g. DeepLift. *Hopefully* model-specific explanations are comparable to SurvSHAP(t), but a lot faster to compute.

Measure:

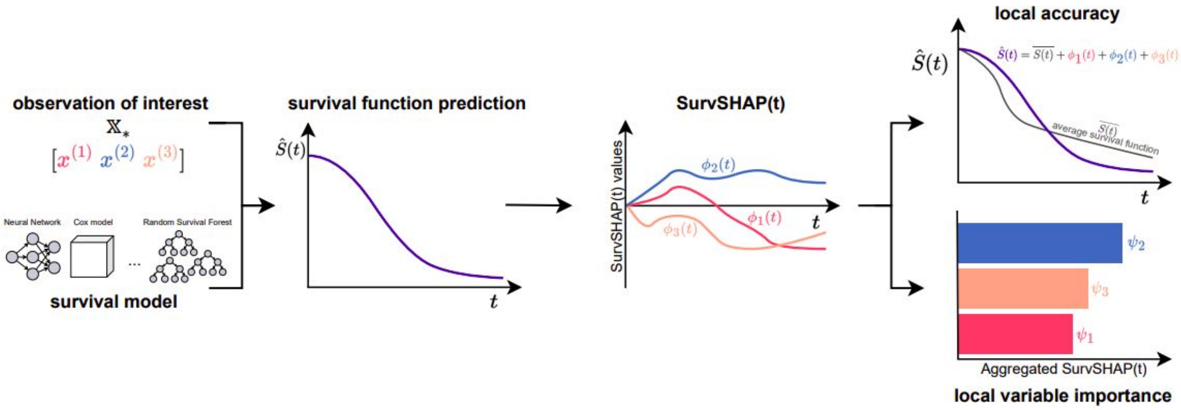
1. Computation time
2. Estimation error, i.e. how close is the faster method to SurvSHAP(t)
3. Evaluation measures, e.g. Avg-Sensitivity, Faithfulness Correlation

SURVIVAL ANALYSIS



	x0	x1	x2	x3	x4	x5	x6	x7	x8	duration	event
0	5.603834	7.811392	10.797988	5.967607	1.0	1.0	0.0	1.0	56.840000	99.333336	0
1	5.284882	9.581043	10.204620	5.664970	1.0	0.0	0.0	1.0	85.940002	95.733330	1
3	6.654017	5.341846	8.646379	5.655888	0.0	0.0	0.0	0.0	66.910004	239.300003	0
4	5.456747	5.339741	10.555724	6.008429	1.0	0.0	0.0	1.0	67.849998	56.933334	1
5	5.425826	6.331182	10.455145	5.749053	1.0	1.0	0.0	1.0	70.519997	123.533333	0

SurvSHAP(t)


observation of interest
$\mathbb{X}_*$
$[x^{(1)}\ x^{(2)}\ x^{(3)}]$
Neural Network
Cox model
Random Survival Forest
...
survival model
survival function prediction
$\hat{S}(t)$
t
SurvSHAP(t)
SurvSHAP(t) values
$\phi_2(t)$
$\phi_1(t)$
$\phi_3(t)$
t
local accuracy
$\hat{S}(t)$
$\hat{S}(t) = \overline{S(t)} + \phi_1(t) + \phi_2(t) + \phi_3(t)$
$\overline{S(t)}$
average survival function
t
ψ_2
ψ_3
ψ_1
Aggregated SurvSHAP(t)
local variable importance

DeepHit

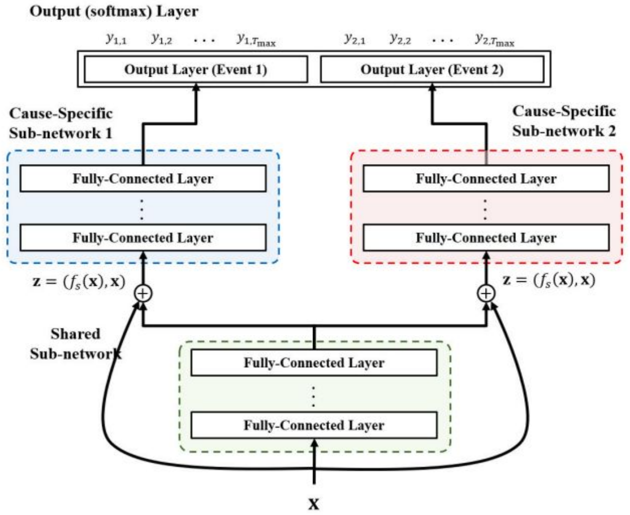
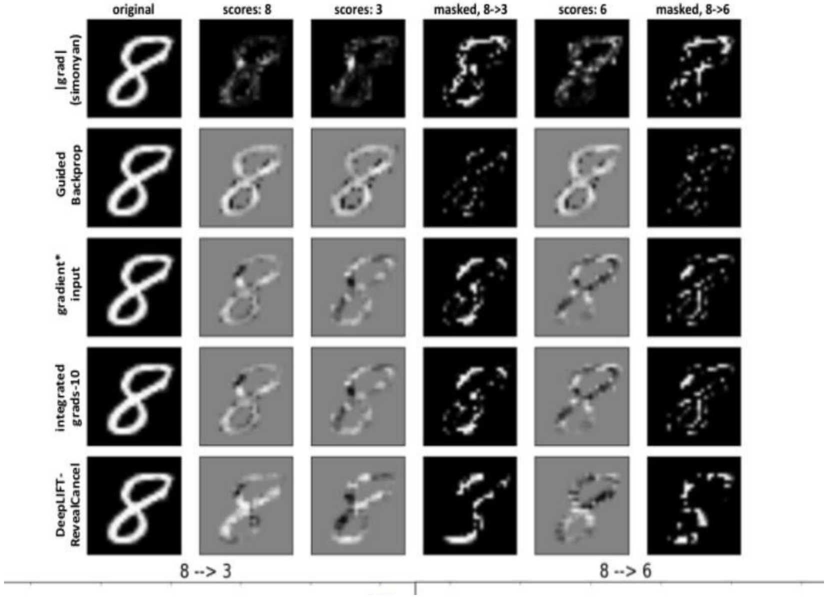
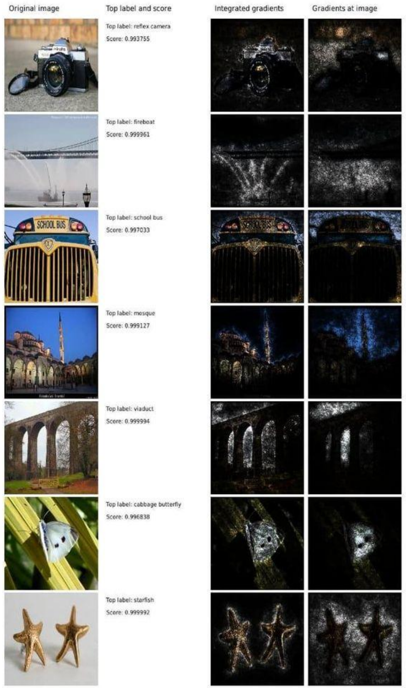

Output (softmax) Layer
$y_{1,1}$ $y_{1,2}$ ... $y_{1,T_{\max}}$
$y_{2,1}$ $y_{2,2}$ ... $y_{2,T_{\max}}$
Output Layer (Event 1)
Output Layer (Event 2)
Cause-Specific Sub-network 1
Cause-Specific Sub-network 2
Fully-Connected Layer
Fully-Connected Layer
Fully-Connected Layer
Fully-Connected Layer
$\mathbf{z} = (f_s(\mathbf{x}), \mathbf{x})$
$\mathbf{z} = (f_s(\mathbf{x}), \mathbf{x})$
Shared Sub-network
Fully-Connected Layer
Fully-Connected Layer
$\mathbf{x}$

Figure 2: The architecture of DeepHit with two competing events.

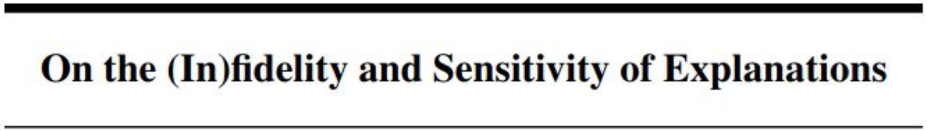
DeepLIFT


original
scores: 8
scores: 3
masked, 8->3
scores: 6
masked, 8->6
|grad| (simonyan)
Guided Backprop
gradient* input
integrated grads-10
DeepLIFT-RevealCancel
8 --> 3
8 --> 6

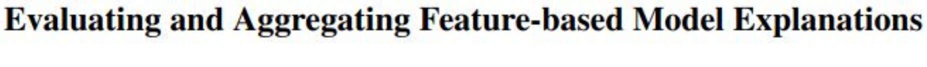
Integrated Gradients



Metrics



On the (In)fidelity and Sensitivity of Explanations



Evaluating and Aggregating Feature-based Model Explanations