Time-dependent explanations of neural networks for survival analysis

Jakub Bednarz JB406103@STUDENTS.MIMUW.EDU.PL
Kamil Grudzień KP.GRUDZIEN@STUDENT.UW.EDU.PL
Krystian Sztenderski K.SZTENDERSKI@STUDENT.UW.EDU.PL
University of Warsaw, Poland

Abstract

We investigate the topic of time dependent explanations of neural networks for survival analysis. As a part of this paper, we verify how well model-specific explanations can perform on deep neural network in comparison to much more complex and sophisticated model-agnostic explanation methods such as SurvSHAP. Most of experiments are performed on broadly known Metabric dataset consisting of medical data of taxonomy of breast cancer. The code used to obtain the results is available at `https://github.com/vitreusx/surv`.

1. Introduction

Model prediction explanation techniques are emerging faster than ever. It is crucial to know the limitations of each solution and have the general idea which explanation technique suits best considered machine learning model. There are multiple well performing model agnostic and model specific methods but it is not clear how well those methods perform in comparison with each other. As our main contribution, we compare explanation results achieved with SurvSHAP (Krzyziński et al., 2022), which we consider as a baseline, with model specific methods DeepLift (Shrikumar et al., 2017) and Integrated Gradients (Sundararajan et al., 2017). Besides the quality of explanation, we also measure other metrics such as computation time, which might be crucial in some use cases.

2. Methodology

2.1 Data format

Survival analysis measures time to occurrence of an event of interest. Real world datasets contain incomplete information. For some individuals during the study no event of interest occurs and instead we are given the time of censorship - time when the subject stopped participation in the study.

Each instance in the dataset is represented as a triplet (x_i, y_i, δ_i), where $x_i \in \mathbb{R}^p$ is the variable vector, δ_i indicates the event of interest, y_i is time duration (to censorship when $\delta_i = 0$ or an event otherwise). If time to only a single event is studied, δ_i takes form of a binary variable, e.g. $\delta_i = 1$ might indicate death of the subject and $\delta_i = 0$ that the subject did not die during the study.

2.2 Dataset

We conducted our experiments on The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset (Curtis et al., 2012). It contains information about gene and protein expression profile for 1904 patients of which 1103 (57.9%) have an observed death due to breast cancer and the remaining 801 (42.1%) were right censored.

Additionally we used the synthetic dataset EXP1 introduced by (Krzyziński et al., 2022) to test our implementation. Appendix B contains results for EXP1 dataset.

2.3 Model

For our tests we used the DeepHit neural network model (Lee et al., 2018). Which for a given variable vector x produces a probability distribution $y = [y_1, y_2, ..., y_{T_{MAX}}]$, where y_k indicates that the patient will experience the event at time k (DeepHit also supports competing events, but for our case we only used it in a single event scenario).

To verify DeepHit's predictions and explanations generated for it by SurvSHAP, we also trained Cox Proportional Hazard (CPH) and Random Survival Forest (RSF) models. [1]

2.4 Explanations

We compare SurvSHAP with neural network specific explanation method DeepLift. To perform such comparison we have created an adapter for DeepLift to acquire survival function explanations similar to explanations of SurvSHAP. To achieve this for a given observation we generate variable attributions at all discrete time points.

We have also performed local explanations with DeepLiftShap (Lundberg and Lee, 2017) and Integrated Gradients using similar adapters as for DeepLift.

3. Experimental results

3.1 Evaluation measure

We have evaluated explanations obtained with SurvSHAP and DeepLift using Maximum Sensitivity metric (Yeh et al., 2019). Figure 1 shows the results for 10 test samples from the dataset. We can observe that the DeepLift's explanations change more drastically when the input is varied infinitesimally with on average about 2 times higher sensitivity than SurvSHAP's explanations for the same samples. This suggests that DeepLifts explanations are less accurate, but direct comparisons between these methods show that it does not have a significant effect on explanations.

3.2 Per sample explanations

Selecting random sample from the dataset and analyzing explanations created with SurvSHAP and DeepLift separately, makes it clear that results are more or less similar no matter which explanation technique we use. Comparing explanations per feature (presented in figures in appendix C) it can be concluded that direction and dynamics of changes are similar for both methods. There are individual features for which explanations present strongly different.

1. We used scikit-survival's CPH and RSF models (https://github.com/sebp/scikit-survival).

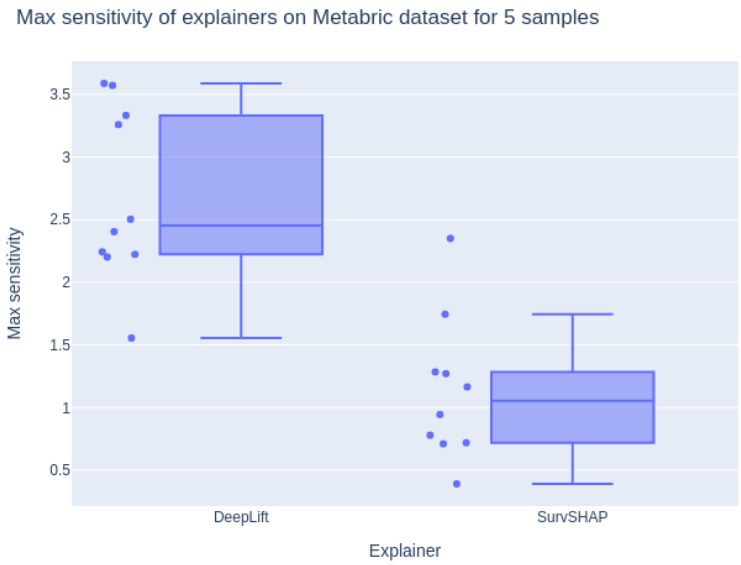

Figure 1: Maximum Sensitivity of DeepLift and SurvSHAP explanations of DeepHit model predictions for 10 samples from Metabric dataset.

One observed case is when explanation is opposite between methods (fig. 16) and another one is when DeepLift evaluates explanation for PGR as flat line and SurvSHAP adds much different dynamic that strongly differs in time (fig. 17).

3.3 Aggregated explanations

The function explanations can be aggregated to calculate a single value variable importance as described by (Krzyziński et al., 2022). Figures 2, 3, 4, 5, 6, 7 show the aggregated explanations for SurvSHAP and DeepLift. We can observe that all of the aggregated explanations obtained are very similar between methods.

3.4 Comparison of explanations

To illustrate similarity of explanations generated with SurvSHAP and DeepLift, we use mean absolute difference of explanations per feature. In figure 20, we can see that for the initial timestamps and those at the end, metric value is increasing and decreasing respectively. However, between those episodes, value metric is stable for most of the features present in Metabric dataset. This means that the direction and dynamic of change in time for both explainers is very similar. We can expect that for bigger samples the average absolute difference is even more flat. Even more impressing is that the metric value for most of the features is not greater than about 0.007.

Method	Model	avg time (s)
SurvSHAP	Random Survival Forest	185.59
SurvSHAP	Cox Proportional Hazard	82.39
SurvSHAP	DeepHit	22.79
DeepLiftShap	DeepHit	2.91
Integrated Gradients	DeepHit	0.85
DeepLift	DeepHit	0.47

Table 1: Average (from 10 experiments each) of execution times of explanation methods for different models on a single sample from the Metabric dataset.

3.5 Baseline importance

We experimented with different baseline types used by explainers as reference samples that are compared with the inputs. Most common in the literature are baselines in form of zero scalars or scalars with mean values. Each scalar corresponds to each input tensor. Additionally, median values were tested. For most of the features there was no clear difference on Metabric dataset, though this hyperparameter was not considered as important in this case. However, in general case, for medical data it might be better to use minimal, maximal or average expected or acceptable norm of certain substance or marker. For markers present in the Metabric dataset, it is difficult to verify expected values without a general medical knowledge. Due to limited expertise on the topic, mean baselines are applied in all experiments.

3.6 Execution time

The biggest disadvantage of SurvSHAP is its computational time. Table 1 and Figure 24 present execution times of different explanation methods. For each method and model we ran the method 10 times on a single sample from the Metabric dataset. We have observed that SurvSHAP is on average 48 times slower than DeepLift when explaining neural network DeepHit.

4. Conclusion

In this work, we compared different explanation methods which can be used for explanation of deep neural network results, obtained on survival analysis. The key aspects, that were taken into consideration during evaluation of SurvSHAP vs. DeepLift were how well were we able to explain the obtained results using simpler methods and how can one benefit from using them. It appears that DeepLift method achieved comparable results to SurvSHAP which is considered a baseline method. At the same time we were able to decrease computation time by a factor of 48 on average, which is a huge improvement, especially in applications where many explanations need to be done often.

We provide a code implementation with results presented in this paper. The solution can be easily used for further experiments with use of other real-life datasets to ensure obtained results and generalization on different fields.

References

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. URL `https://EconPapers.repec.org/RePEc:nat:nature:v:486:y:2012:i:7403:d:10.1038_nature10983`.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. Survshap(t): Time-dependent explanations of machine learning survival models, 2022. URL `https://arxiv.org/abs/2208.11080`.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI Conference on Artificial Intelligence*, 2018.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL `https://arxiv.org/abs/1705.07874`.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. 2017. doi: 10.48550/ARXIV.1704.02685. URL `https://arxiv.org/abs/1704.02685`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL `https://arxiv.org/abs/1703.01365`.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity for explanations, 2019. URL `https://arxiv.org/abs/1901.09392`.

Appendix A. Aggregated variable explanations

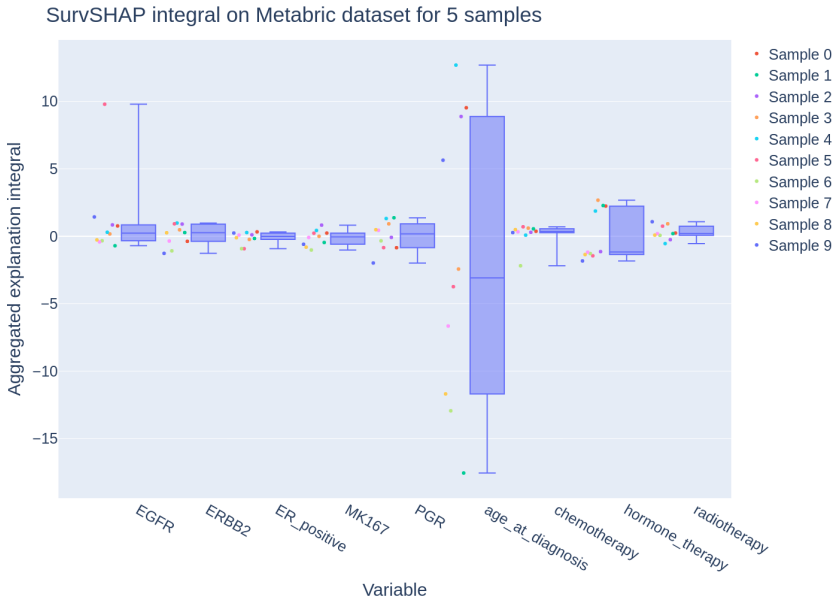

Figure 2: Aggregated explanations from SurvSHAP on DeepHit predictions for 10 samples. Aggregated using integrals.

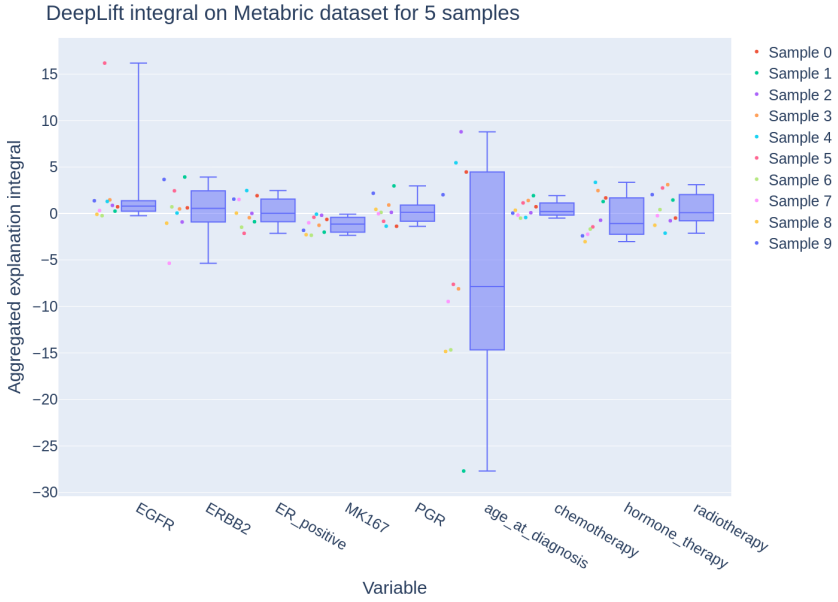

Figure 3: Aggregated explanations from DeepLift on DeepHit predictions for 10 samples. Aggregated using integrals.

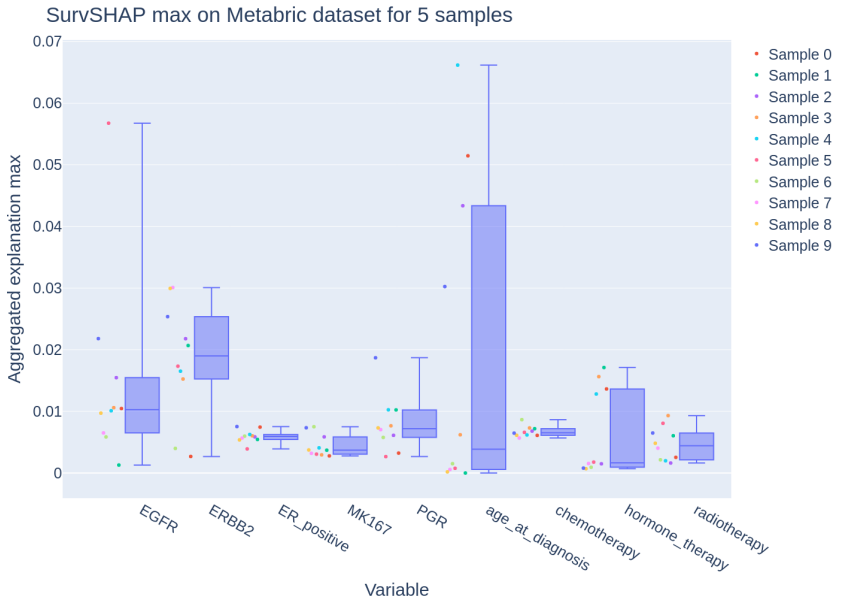

Figure 4: Aggregated explanations from SurvSHAP on DeepHit predictions for 10 samples. Aggregated using max.

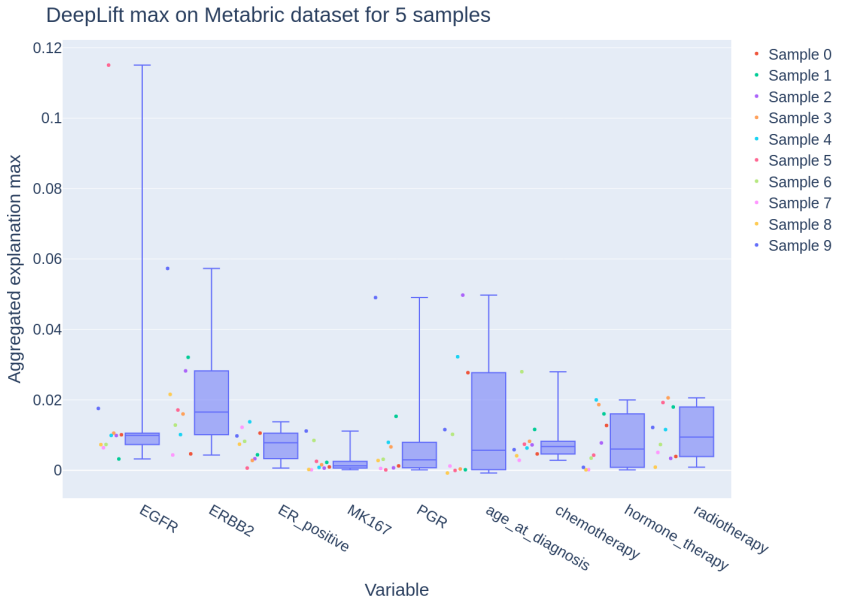

Figure 5: Aggregated explanations from DeepLift on DeepHit predictions for 10 samples. Aggregated using max.

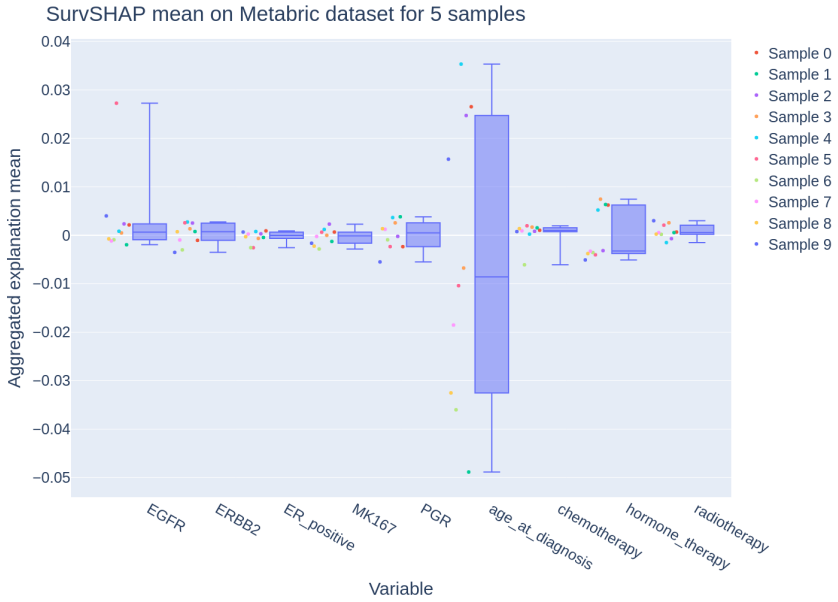

Figure 6: Aggregated explanations from SurvSHAP on DeepHit predictions for 10 samples. Aggregated using mean.

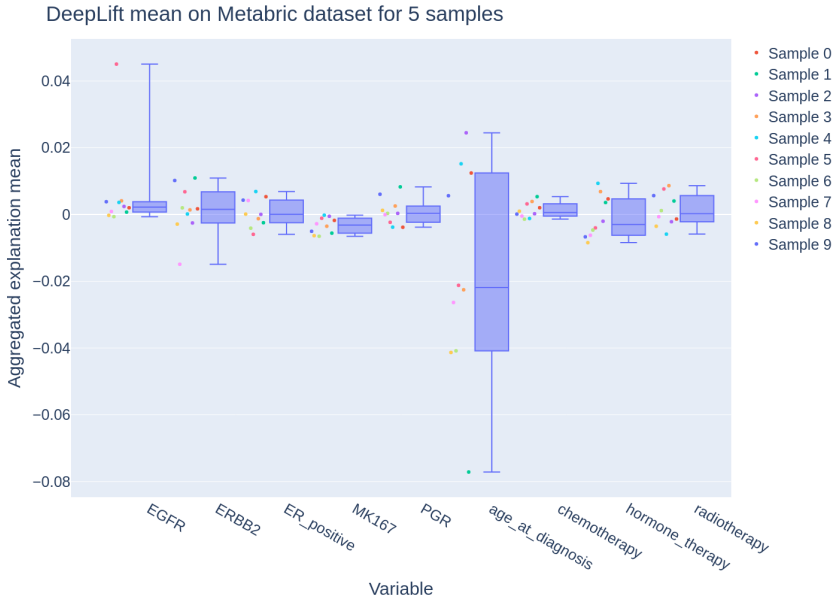

Figure 7: Aggregated explanations from DeepLift on DeepHit predictions for 10 samples. Aggregated using mean.

Appendix B. Synthetic dataset EXP1

Here we present explanations and execution times of various methods for the synthetic dataset EXP1 from (Krzyziński et al., 2022).

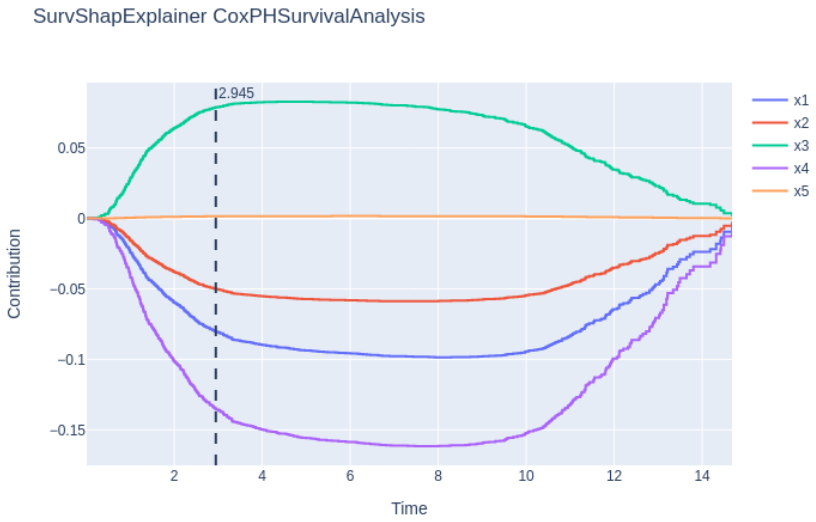

Figure 8: Sample explanations of SurvSHAP for Cox Proportional Hazard model on the EXP1 dataset.

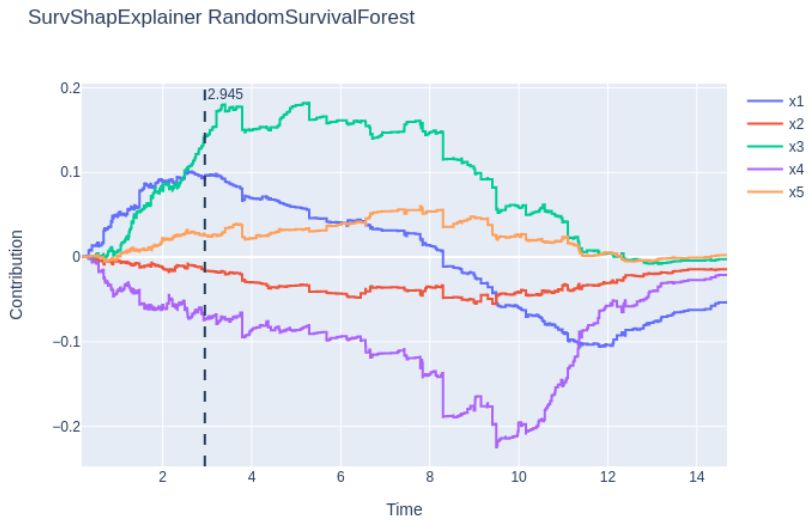

Figure 9: Sample explanations of SurvSHAP for Random Survival Forest model on the EXP1 dataset.

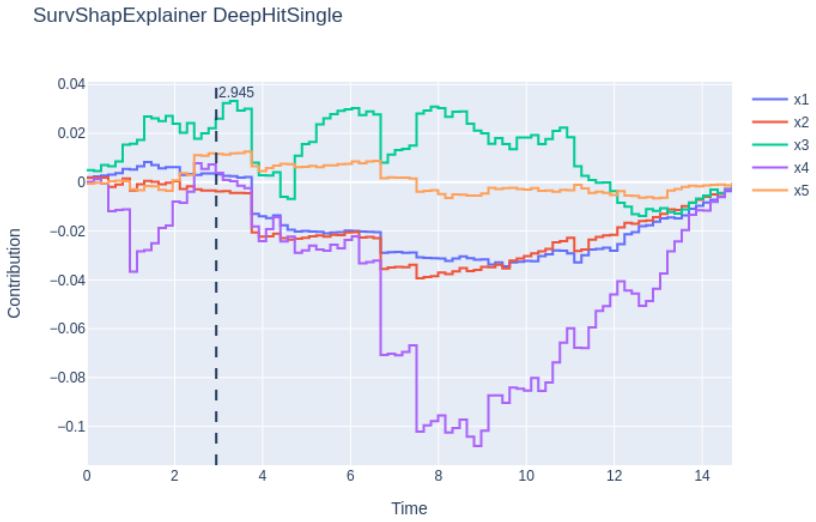

Figure 10: Sample explanations of SurvSHAP for DeepHit model on the EXP1 dataset.

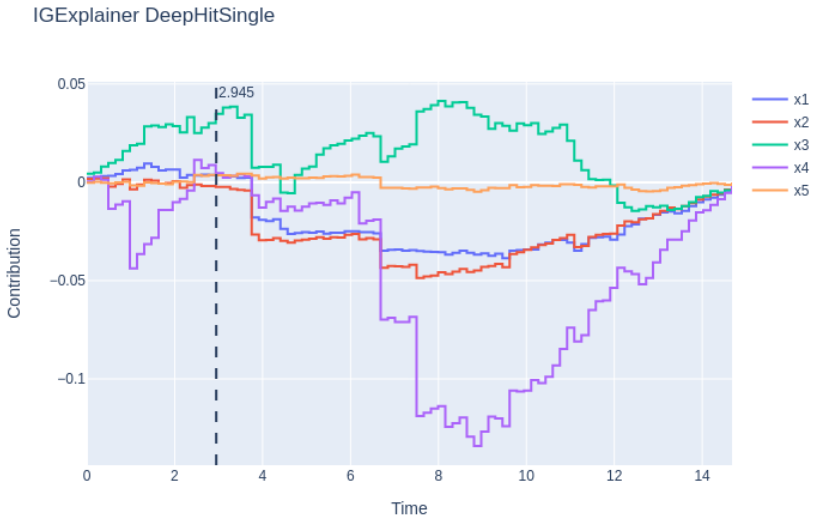

Figure 11: Sample explanations of Integrated Gradients for DeepHit model on the EXP1 dataset.

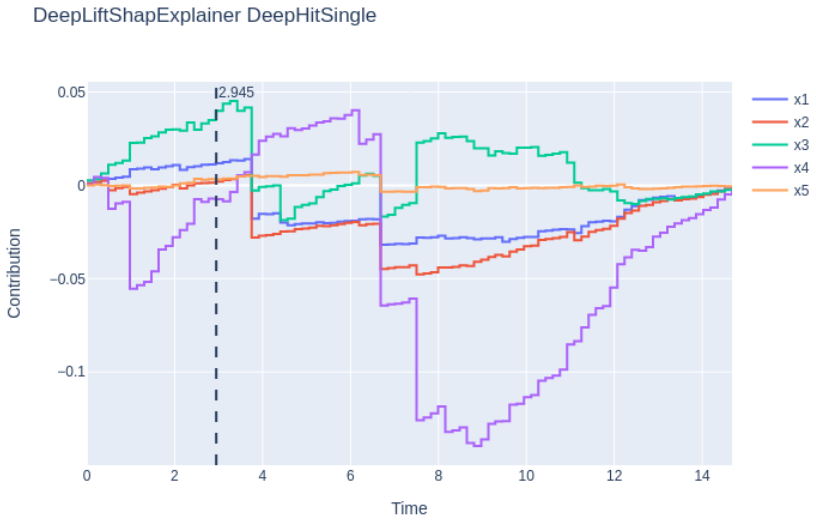

Figure 12: Sample explanations of DeepLiftShap for DeepHit model on the EXP1 dataset.

Appendix C. Explanations per feature on METABRIC dataset

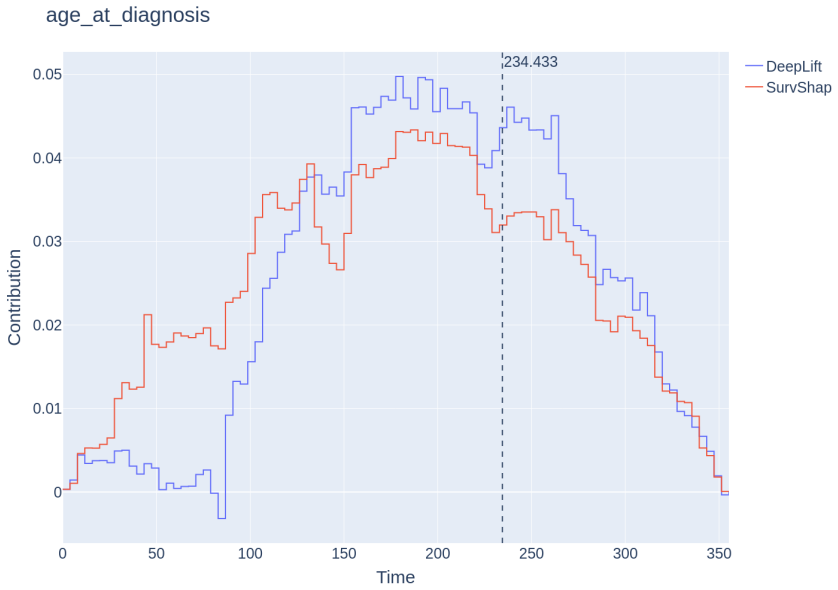

Figure 13: Chosen sample explanation of 'age at diagnosis' with DeepLiftShap and SurvSHAP for DeepHit model on the METABRIC dataset.

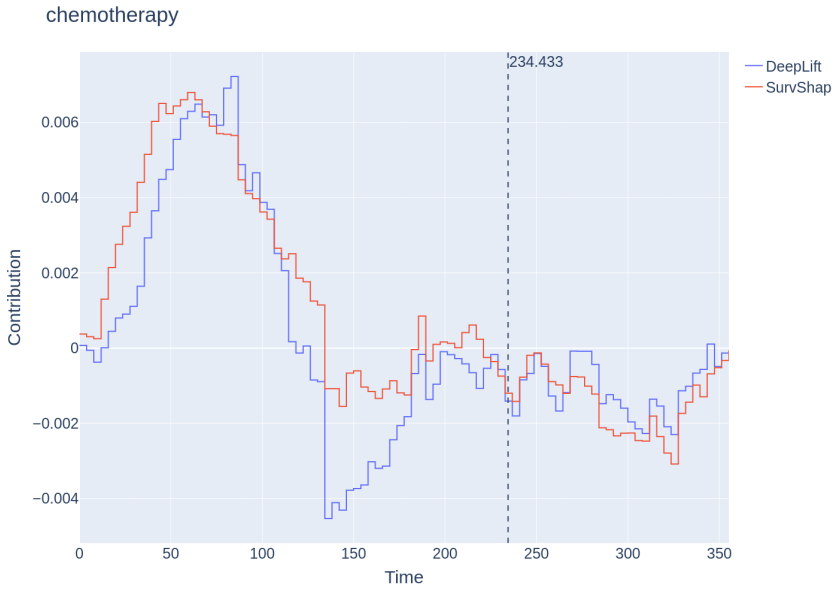

Figure 14: Chosen sample explanation of 'chemotherapy' with DeepLiftShap and SurvSHAP for DeepHit model on the METABRIC dataset.

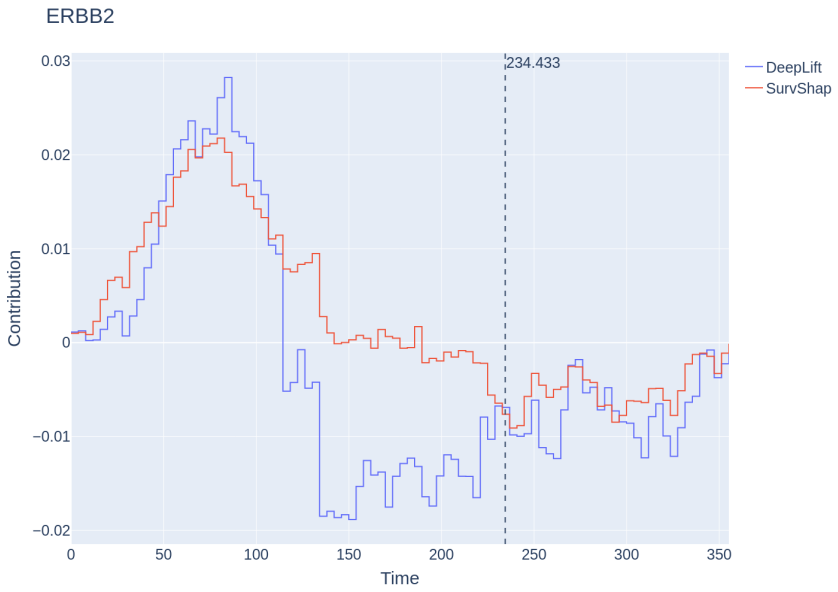

Figure 15: Chosen sample explanation of 'ERBB2' with DeepLiftShap and SurvSHAP for DeepHit model on the METABRIC dataset.

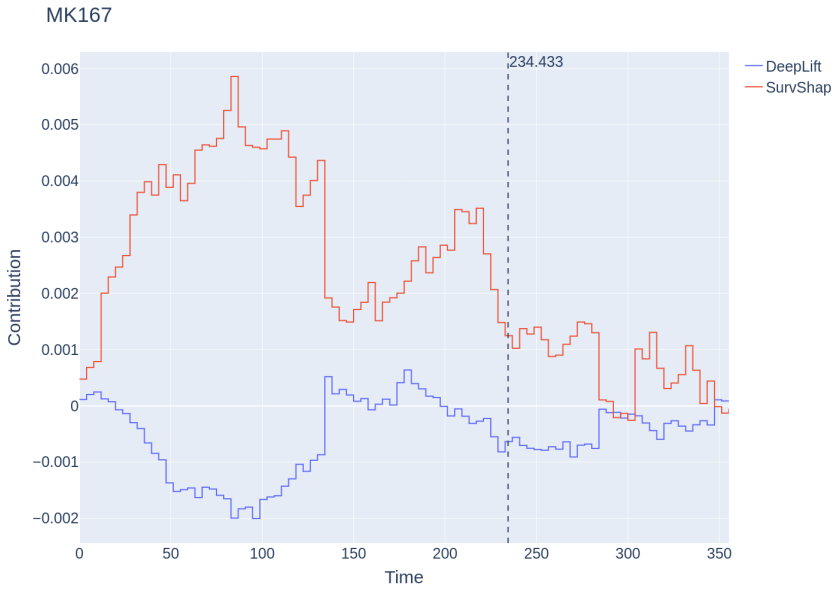

Figure 16: Chosen sample explanation of 'MK167' with DeepLiftShap and SurvSHAP for DeepHit model on the METABRIC dataset.

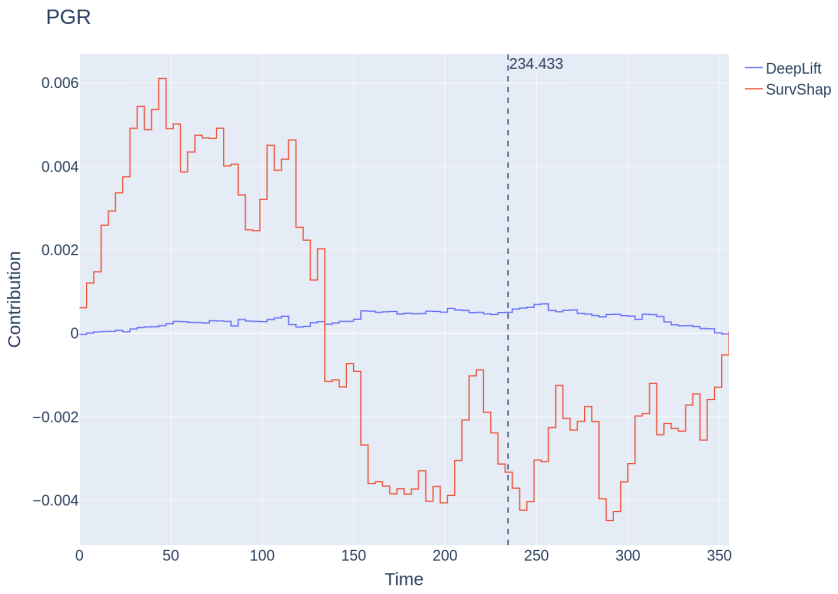

Figure 17: Chosen sample explanation of 'PGR' with DeepLiftShap and SurvSHAP for DeepHit model on the METABRIC dataset.

Appendix D. SurvSHAP and DeepLift sample explanations

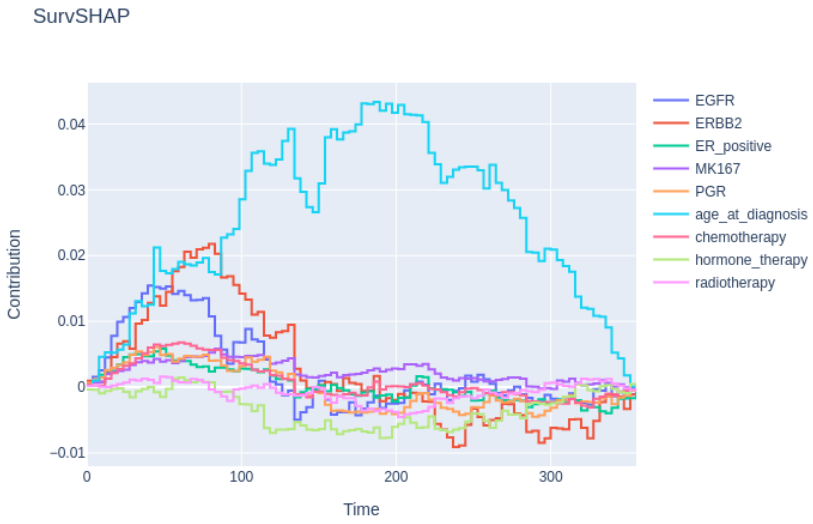

Figure 18: Sample explanations of SurvSHAP for DeepHit model on the METABRIC dataset.

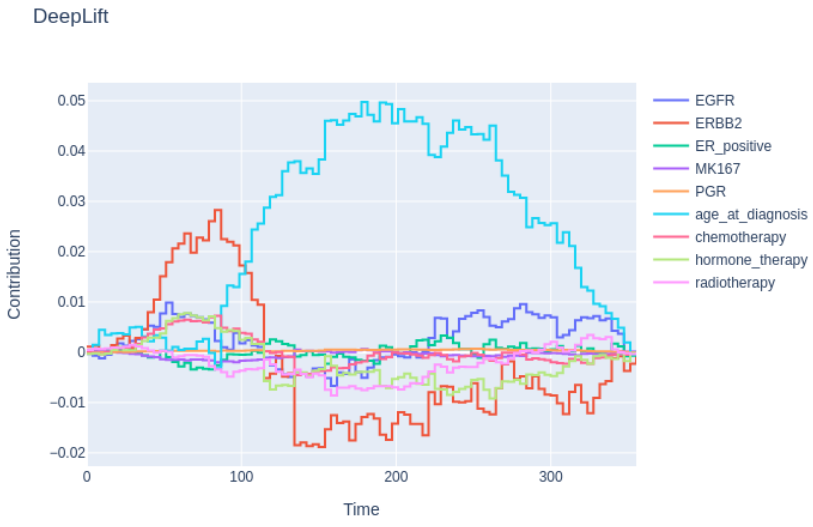

Figure 19: Sample explanations of DeepLift for DeepHit model on the METABRIC dataset.

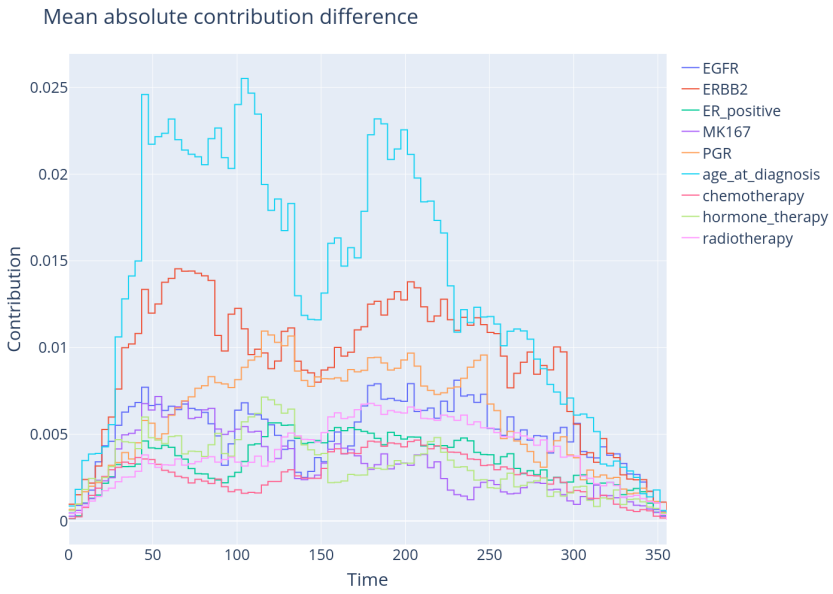

Figure 20: Mean absolute difference between explanations of SurvSHAP and DeepLift for DeepHit model on the METABRIC dataset.

Appendix E. Other explanation methods on METABRIC dataset

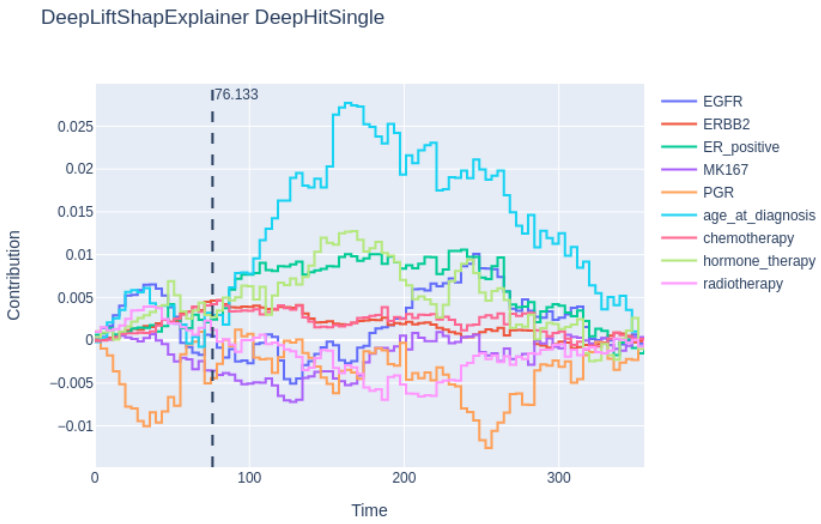

Figure 21: Sample explanations of DeepLiftShap for DeepHit model on the METABRIC dataset.

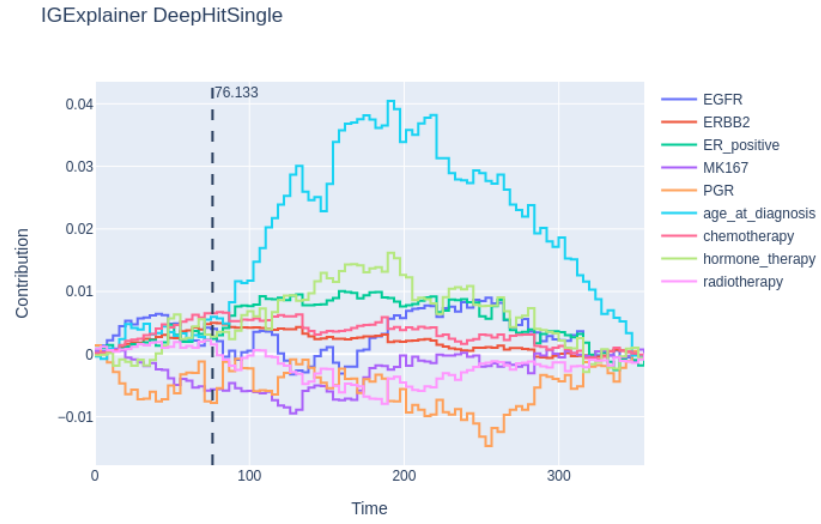

Figure 22: Sample explanations of Integrated Gradients for DeepHit model on the METABRIC dataset.

Appendix F. Execution times

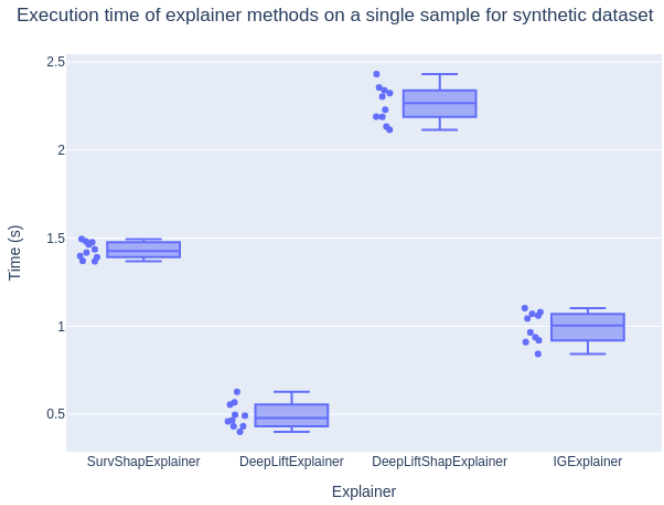

Figure 23: Execution times (in seconds) of different explanation methods on the DeepHit model prediction for a single sample from the EXP1 dataset. For each method the experiment was repeated 10 times.

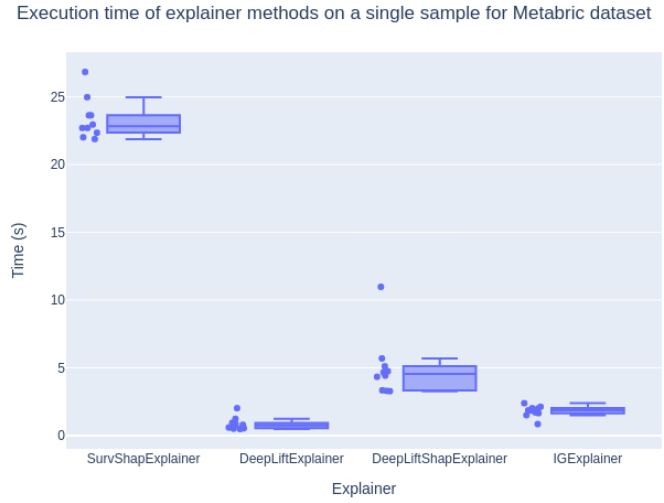

Figure 24: Execution times (in seconds) of different explanation methods on the DeepHit model prediction for a single sample from the Metabric dataset. For each method the experiment was repeated 10 times.