A Pseudo Value Based Interpretable Neural Additive Model for Survival Analysis

Md Mahmudur Rahman, Sanjay Purushotham

Department of Information Systems, University of Maryland Baltimore County, Baltimore, Maryland, USA
mrahman6@umbc.edu, psanjay@umbc.edu

Abstract

Deep learning models have achieved the start-of-the-art performance in survival analysis as they can handle censoring while learning complex nonlinear hidden representations directly from the raw data. However, the covariate effects on survival probabilities are difficult to explain using deep learning models. To address this challenge, we propose `PseudoNAM` - an interpretable model which uses pseudo values to efficiently handle censoring and uses neural additive networks to capture the nonlinearity in the covariates of the survival data. In particular, `PseudoNAM` uses neural additive models to jointly learn a linear combination of neural networks corresponding to each covariate and identifies the effect of the individual covariate on the output, and thus, is inherently interpretable. We show that our `PseudoNAM` outputs can be used in other survival models such as random survival forests to obtain improved survival prediction performance. Our experiments on three real-world survival analysis datasets demonstrate that our proposed models achieve similar or better performance (in terms of C-index and Brier scores) than the state-of-the-art survival methods. We showcase that `PseudoNAM` provides overall feature importance scores and feature-level interpretations (covariate effect on survival risk) for survival predictions at different time points.

Introduction

Survival analysis (Kleinbaum and Klein 2010), a well-studied problem, aims to estimate the risk of a subject's failure from an event, such as death due to breast cancer at a particular time point. One key challenge in survival analysis is the presence of censored subjects for whom the actual survival times remain unknown. A good survival analysis model should handle censoring, accurately discriminate the predicted risks, and should be interpretable. Traditional statistical survival analysis models such as Cox Proportional Hazard models (Cox 1972), regression models based on pseudo-observations (Andersen, Klein, and Rosthøj 2003; Andersen and Pohar Perme 2010) are interpretable but are less accurate and limited by strong assumptions on the underlying stochastic process, such as linearity, parametric, and proportional hazards assumptions. Recent survival approaches based on machine learning and deep learning models (Rahman et al. 2021; Zhao and Feng 2020; Ishwaran et al. 2008; Katzman et al. 2018) can make more accurate predictions but may require specialized objective functions to handle censoring (Lee et al. 2018). Moreover, they are black-box approaches and are not interpretable or explainable, which makes them opaque and unsuitable for medical applications.

To address the limitations of the existing methods, we propose a pseudo value based neural additive model, called **PseudoNAM**, which directly models the complex nonlinear time-varying effect of the covariate on the survival function. Our `PseudoNAM` uses neural additive models (NAM) (Agarwal et al. 2020) to jointly learn a linear combination of neural networks corresponding to each covariate and determines the magnitude of covariates' effect on the survival outcome, and thus, is inherently interpretable. The neural networks for each feature in the `PseudoNAM` are independent, and thus, can provide the individual feature contribution towards output survival prediction. Like in NAM, we sum up the individual feature contributions (neural network outputs), followed by a `logit` transformation using `sigmoid` activation function to predict the survival probability at different time points. We show different types of interpretations from `PseudoNAM`, including 1) the mean feature contributions to the survival probability predictions at different time points (overall feature importance scores) and 2) feature-level interpretations which show the time-varying covariate effect on the survival predictions.

Our experiments on three real-world datasets demonstrate that `PseudoNAM` performs similar or better than the state-of-the-art survival analysis models while providing interpretable results. We further improve the performance of `PseudoNAM` by proposing **PseudoNRSF**, a random survival forest approach that takes as input the learned outputs of individual neural networks from `PseudoNAM` and predicts the survival probabilities. We show that `PseudoNRSF` achieves state-of-the-art results in terms of c-index and Brier scores.

Related Works

Cox-based statistical and deep learning survival models (Cox 1972; Faraggi and Simon 1995; Katzman et al. 2018; Kvamme and Borgan 2019) are widely studied for analyzing time-to-event data. However, these models make strong proportional hazard and linearity assumptions that may not hold

for real data, thus leading to less accurate survival results.

Machine learning models such as Random survival forests (Ishwaran et al. 2008) and multi-task logistic regression (MTLR) (Yu et al. 2011; Fotso 2018) relax some of these assumptions and outperform statistical-based methods. Recently proposed deep learning models (Lee et al. 2018; Nagpal, Li, and Dubrawski 2021) and conditional generative adversarial networks (Chapfuwa et al. 2018) achieve state-of-the-art results for time-to-event analysis. However, these methods require either making assumptions on the underlying stochastic process or design a specialized objective function to handle censoring. Moreover, these methods lack interpretability which is required in the medical domain. To address the censoring challenge, (Zhao and Feng 2020; Rahman et al. 2021) have respectively proposed pseudo value based deep learning models for survival and competing risk analysis. However, even these methods are not directly interpretable and rely on off-the-shelf explainable AI methods such as LRP (Montavon et al. 2019) for providing explanations.

Our Proposed Models

To address the censoring and interpretability challenges of existing survival analysis approaches, in this work, we propose two interpretable pseudo value based deep learning models, **PseudoNAM** and **PseudoNRSF**. Before describing our models in detail, we will briefly introduce pseudo values.

What are Pseudo values? Pseudo values for the survival probability are derived from the non-parametric population-based Kaplan-Meier (KM) estimator, an approximately unbiased estimator of the survival probability under independent censoring (Andersen et al. 2012). For the i^{th} subject, a Jackknife pseudo value, based on the KM estimate of the survival probability (Klein et al. 2008), is computed at time horizon t^* as

$$\hat{S}_i(t^*) = n\hat{S}(t^*) - (n-1)\hat{S}^{-i}(t^*)$$

where, $\hat{S}(t^*)$ is the Kaplan-Meier estimate of the survival probability at time t^* based on a sample with n subjects and $\hat{S}^{-i}(t^*)$ is the Kaplan-Meier estimate of the survival probability at time t^* based on a leave-one-out sample with $(n-1)$ subjects, obtained by omitting the i^{th} subject. Pseudo values are calculated for both uncensored subjects and censored subjects (incompletely observed) at a specified time point.

PseudoNAM: Inspired by the success of pseudo value based deep models, DNNSurv (Zhao and Feng 2020) and DeepPseudo (Rahman et al. 2021) to handle censoring, we propose `PseudoNAM` - a multi-output neural additive model which predicts pseudo values for survival risk analysis. `PseudoNAM`, shown in Figure 1, learns non-linear representations in the data and uses pseudo values to handle censoring efficiently. `PseudoNAM` has the following form:

$$g(E[y(t|X)]) = \beta + f_1(X_1) + f_2(X_2) + \ + f_p(X_p) \quad (1)$$

Here, $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$ is a p-dimensional covariate vector for i^{th} individual; $i = 1, 2, .., n$. $g(.)$ is the link function (e.g., logit link function), β is the bias and each $f_i(.)$ is parametrized by a neural network. $y(t|X)$ is the pseudo values for survival probability at time t in the presence of covariates. Each of the networks learn the complex shape function of a specific covariate, and all the networks are trained jointly. A $n \times p$ matrix of p baseline covariates with n individuals are used as input in the input layer. Output layer returns the survival probabilities at M evaluation time points. `PseudoNAM` model provides interpretability because it jointly trains a set of neural networks corresponding to each individual covariate and returns the covariates' contribution scores to the output (i.e., the output of the neural networks) for all each covariate and for the M evaluation time points. Then we sum up the contribution scores of all covariates followed by applying a `sigmoid` activation function to get the final output, i.e., the survival probabilities at the M time points. The non-overlapping neural networks for individual covariates allow to identify the individual covariate effect on the survival probabilities.

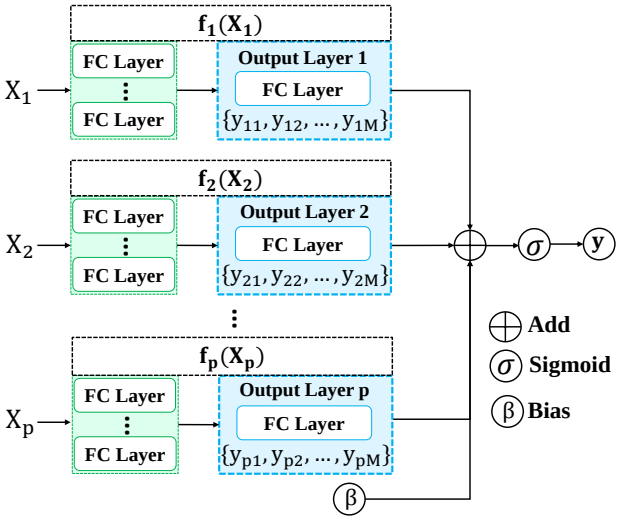

Figure 1: Architecture of `PseudoNAM`. $\mathbf{X} = \{X_1, X_2, .., X_p\}$ is a p dimensional vector of covariates. $f_i(X_i)$ is the neural network corresponding to covariate X_i and β is the bias. Sigmoid (σ) is worked as inverse logit link function. y is the output, i.e., survival probability at M time points. FC Layer means Fully Connected Layer.

PseudoNRSF: While `PseudoNAM` provides interpretable predictions; its performance is limited by the NAM model architecture. To improve the performance of `PseudoNAM` model and to obtain global and local interpretations like Random survival forests (RSF) (Ishwaran et al. 2008), we propose `PseudoNRSF` - a two-stage deep learning model. In the first stage, `PseudoNAM` model is used to learn the individual feature contribution scores for predicting pseudo values. In the second stage, these learned feature contribution scores are input to an RSF with the goal of directly predicting survival probabilities. Thus, `PseudoNRSF` returns the subject-specific survival proba-

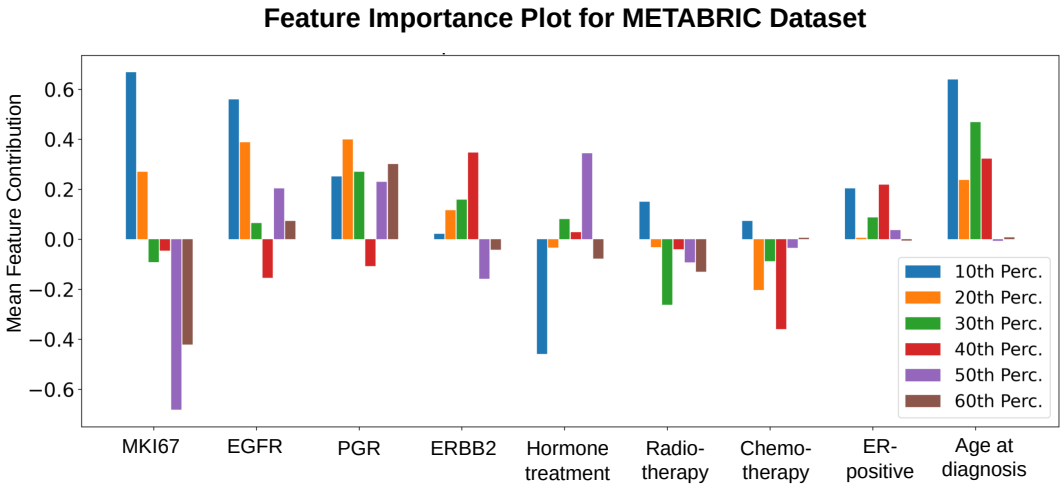

Figure 2: Mean individual feature contributions on survival probabilities at different time points on METABRIC dataset. Here, 10^{th} Perc., 20^{th} Perc. are representing 10^{th} percentile and 20^{th} percentile of the survival time distribution at which we get the `PseudoNAM` model predictions.

Weight	Feature
0.1664 ± 0.0235	Age at diagnosis
0.0624 ± 0.0089	MKI67
0.0485 ± 0.0062	EGFR
0.0431 ± 0.0062	ERBB2
0.0387 ± 0.0027	PGR
0.0182 ± 0.0028	Hormone treatment
0.0149 ± 0.0020	Chemotherapy
0.0147 ± 0.0020	Radiotherapy
0.0046 ± 0.0011	ER-positive

Figure 3: Importance of feature (mean weight $\pm$ sd) on the survival probability predictions measured by `PseudoNRSF` on METABRIC dataset.

bilities at different time points as output, and the predictions are interpretable due to the use of RSF. `PseudoNRSF` has the interpretation property of the RSF. We can easily get the effect of each covariate (importance score) on the overall survival probability. However, using the `PseudoNAM`, we can see the change of covariate effect on survival probabilities at different time points.

Experiments

We conducted experiments on three real-world datasets to answer the following questions: a) how well our proposed models perform compared to the state-of-the-art survival models? b) how well can our `PseudoNAM` explain their predictions?

Datasets: Table 1 shows statistics of the following datasets.

METABRIC: This data (Katzman et al. 2018)[1] contains patients' gene expressions and clinical variables for breast cancer survival prediction.

SUPPORT: This dataset (Knaus et al. 1995) is from the Vanderbilt University study to estimate survival of 9,105 seriously ill hospitalized patients.

WHAS: This dataset was collected to examine the effects of a patient's factors on acute myocardial infraction (MI) survival (Hosmer and Lemeshow 2002).

Implementation Details: The (ground-truth) pseudo values for survival probabilities are obtained using the `jackknife` function of `R` package `prodlim` at each evaluation time point (separately for training and validation sets). We performed stratified 5-fold cross-validation so that the ratio of censored and uncensored subjects remained the same in each fold. We jointly train our PseudoNAM's feature networks based on an early stopping criterion and choose the best model based on the model's performance on validation data. Each feature network consists of 3 hidden layers with a number of units [128, 64, 32]. We used `relu` activation function in the hidden layer of each covariate's neural network and `tanh` activation function in the output layer of the neural networks. In the final output layer, we sum up the output of individual feature neural networks and use the `sigmoid` activation function to get the survival probability at $10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th}$ percentile of the maximum survival time of the training data. We did not perform hyperparameter tuning. We set the learning rate 0.0001, output penalty coefficient 0.001, weight decay coefficient 0.000001, dropout rate

[1]https://github.com/jaredleekatzman/DeepSurv

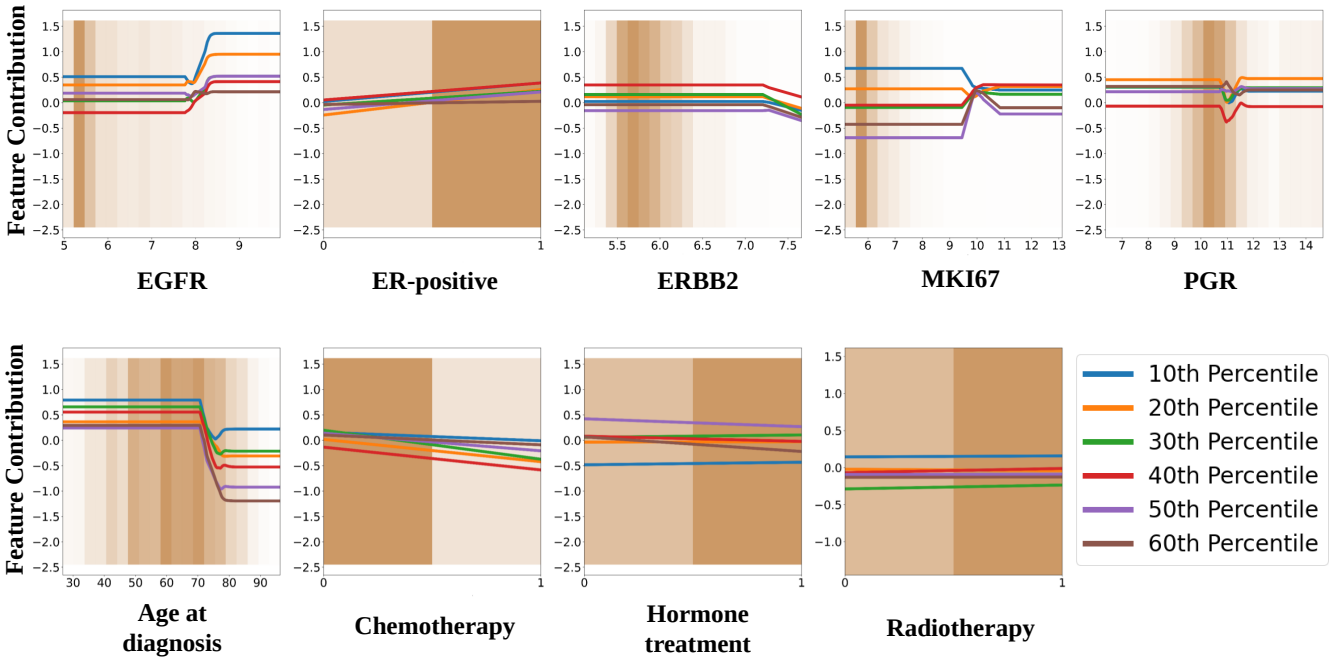

Figure 4: Feature-level contributions to survival probability at different time points (10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th} percentile of survival time distribution) for individual features in METABRIC dataset. The darker the brown bars indicates higher density of the data. Age at diagnosis, EGFR, ERBB2, MKI67, PGR are continuous valued features, while others are discrete valued.

Table 1: Descriptive Statistics of the three Real-World Survival Datasets

2*Dataset	2* No. of Observation	2* No. of uncensored (%)	2* No. of censored (%)	2* No. of features	Event Time				Censoring Time			
					Min	Max	Mean	Median	Min	Max	Mean	Median
METABRIC	1904	1103(57.9)	801(42.1)	9	0.1	355.2	100.0	85.9	0.1	337.0	159.6	158.0
SUPPORT	8873	6036 (68.0)	2837 (32.0)	14	3.0	1944.0	205.5	57.0	344.0	2029.0	1059.9	918.0
WHAS	1638	690 (42.1)	948 (57.9)	5	0.1	1965.9	696.7	515.5	371.0	1999.0	1298.9	1347.5

0.0, and feature dropout rate 0. We used Adam optimizer with batch size 128 during training and used Penalized Mean Squared Error loss function to train our models. For performance metrics, we used (a) time-dependent concordance index (Antolini, Boracchi, and Biganzoli 2005) adjusted with an inverse propensity of censoring estimate for evaluating the discriminative-ability, and (b) Integrated IPCW Brier Score (denoted as Brier Score) (Gerds and Schumacher 2006) metric for evaluating the predictive-ability. To encourage reproducibility, the source codes for our proposed models will be released to the public.

Model Comparisons: We compared the following survival analysis models:
nolistsep

[noitemsep]**Statistical models:** Cox Proportional Hazard Model [`CoxPH`] (Cox 1972)
Machine learning models: Random Survival Forest [`RSF`] (Ishwaran et al. 2008), Multi-task Logistic Regression [`MTLR`] (Yu et al. 2011)
Deep Learning models: `DNNSurv` (Zhao and Feng 2020), `DeepHit` (Lee et al. 2018), `DeepSurv` (Katzman et al. 2018), `CoxTime` (Kvamme, Borgan, and Scheel 2019), Deep Survival Machine [`DSM`] (Nagpal, Li, and Dubrawski 2021), Piecewise Constant Hazard [`PCHazard`] (Kvamme and Borgan 2019), and our proposed models: `PseudoNAM` and `PseudoNRSF`.

Results and Discussion

Table 2 shows the performance comparison of the survival models based on time dependent concordance index and Brier scores. From this table, we see that our `PseudoNAM` obtains similar or comparable performance to other survival analysis models, while `PseudoNRSF` outperforms all the survival models on the WHAS dataset, and obtains similar performance as the state-of-the-art models on the other two datasets. We notice that the independent neural networks for individual covariates limits `PseudoNAM` to learn the shared

Table 2: Model comparisons of the performance metrics (mean and 95% confidence interval) evaluated on survival datasets

Time-dependent Concordance Index

	PseudoNRSF	PseudoNAM	DNNSurv	CoxPH	CoxTime	DeepHit	DeepSurv	DSM	MTLR	PCHazard	RSF
METABRIC	0.645±0.038	0.616±0.025	0.617±0.014	0.622±0.013	**0.660±0.055**	0.655±0.045	0.641±0.017	0.616±0.040	0.550±0.043	0.614±0.041	0.616±0.058
SUPPORT	0.619±0.019	0.613±0.017	0.581±0.009	0.568±0.016	0.616±0.012	0.593±0.012	0.589±0.009	0.595±0.005	0.550±0.024	0.589±0.022	**0.638±0.010**
WHAS	**0.865±0.038**	0.740±0.022	0.721±0.018	0.739±0.013	0.783±0.027	0.851±0.038	0.787±0.030	0.739±0.013	0.618±0.104	0.685±0.038	0.768±0.041

Brier Score

	PseudoNRSF	PseudoNAM	DNNSurv	CoxPH	CoxTime	DeepHit	DeepSurv	DSM	MTLR	PCHazard	RSF
METABRIC	0.171±0.005	0.245±0.013	0.243±0.010	0.313±0.020	0.168±0.011	0.178±0.022	**0.165±0.013**	0.249±0.020	0.225±0.024	0.201±0.014	0.296±0.016
SUPPORT	0.196±0.01	0.207±0.007	0.221±0.002	0.206±0.005	0.192±0.008	0.211±0.008	0.198±0.006	0.212±0.003	0.263±0.019	0.225±0.018	**0.190±0.007**
WHAS	**0.099±0.010**	0.267±0.022	0.290±0.029	0.234±0.029	0.136±0.018	0.140±0.054	0.132±0.018	0.201±0.005	0.162±0.028	0.141±0.009	0.206±0.033

effect on the survival probability at different times, thus resulting in comparable but not the best results.

Model Intepretations

The main advantage of our `PseudoNAM` models is that they can provide interpretations. Here, we discuss the two ways of interpreting the `PseudoNAM` model predictions: overall feature importance scores and feature-level interpretations.

`PseudoNAM` first learns the individual feature contributions for pre-specified time points (here, we choose $10^{th}, 20^{th}, ..., 60^{th}$ percentile of the event horizon). Then we sum up these feature contributions followed by the `sigmoid` transformation to get the final survival probabilities at the pre-specified time points. Figure 2 shows the **overall feature importance scores** measured as mean individual feature contributions on the survival probability at the pre-specified time points for the METABRIC dataset. We see that the features can have a positive or negative impact (overall effect) on survival probability predictions at different time points for breast cancer patients. For example, the covariates such as `MKI67, radiotherapy, and chemotherapy` have `positive` feature contributions at the initial time points (10^{th} percentile), which means that they influence better survival outcomes (higher survival probabilities). However, at later time points (such as 60^{th} percentile), these features have `negative` feature contributions - meaning they result in mortality. This is expected because the survival probability remains higher at initial time points, and it decreases over time. Therefore, the treatment like `chemotherapy` fails to reduce the risk of death at later time points, and the older people (age at diagnosis) are at greater mortality risk.

Figure 3 shows the permutation feature importance, which is measured by observing how random re-shuffling of each covariate influences model performance. We use `eli5`[2] library to compute the feature importance for `PseudoNRSF` model. We observe that `age at diagnosis` has the highest importance on the survival probability predictions and `ER-positive` has the lowest feature importance.

Figure 4 shows the individual feature contribution (i.e., the outputs of the individual neural networks of `PseudoNAM`) on survival probability at different time points (i.e., time-varying covariate effect on survival predictions) for the METABRIC dataset. Here x-axis shows the feature values, and the y-axis shows their contributions. In other words, this plot provides **feature-level interpretations**. For example, the survival probability for the feature `age at diagnosis` at all the time points starts decreasing after 65 years; and we see that the feature `chemotherapy` is biased to the patients who did not receive chemotherapy since the density is much higher for this group (darker brown bar). The plot also shows that the model predicted a decrease in survival probability for a few patients who received chemotherapy, especially at later time points.

Why PseudoNAM is suited for healthcare domain?

As shown in Table 2, our `PseudoNAM` models obtain good predictive and discriminative performance on all the survival datasets. Moreover, using our models, one can visualize each covariates' contribution to the survival probability. Therefore, `PseudoNAM` helps to identify the potential risk factors for an event, such as death due to breast cancer. The visualization of the feature-level interpretations can be a step towards transparency in the deep learning models, which can inform clinical decision-making and perhaps lead to trust in the model. Thus, `PseudoNAM` models with high predictiveability and inherent interpretability could be well-suited for survival analysis in the healthcare domain.

Conclusion

In this paper, we proposed interpretable pseudo value-based deep learning approaches `PseudoNAM` and `PseudoNRSF` to model the nonlinear time-varying covariate effect on survival predictions. Our proposed models use 1) pseudo values to handle censoring and 2) neural additive networks to capture the complex nonlinear relationships and to obtain interpretable predictions. Empirical results show that our proposed models achieve similar or better performance than the state-of-the-art survival methods. Our `PseudoNAM` model provides both overall feature importance scores and feature-level interpretations of predicted survival probabilities at different time points. For future work, we study and compare the interpretability of our proposed models with other parametric survival approaches.

[2]https://github.com/eli5-org/eli5

Acknowledgements

This work is partially supported by grant IIS–1948399 from the US National Science Foundation and grant 80NSSC21M0027 from the National Aeronautics and Space Administration.

References

- Agarwal, R.; Frosst, N.; Zhang, X.; Caruana, R.; and Hinton, G. E. 2020. Neural additive models: Interpretable machine learning with neural nets. *arXiv:2004.13912*.

Andersen, P. K.; Borgan, O.; Gill, R. D.; and Keiding, N. 2012. *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K.; Klein, J. P.; and Rosthøj, S. 2003. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1).

Andersen, P. K.; and Pohar Perme, M. 2010. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1): 71–99.

Antolini, L.; Boracchi, P.; and Biganzoli, E. 2005. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24): 3927–3944.

Chapfuwa, P.; Tao, C.; Li, C.; Page, C.; Goldstein, B.; Duke, L. C.; and Henao, R. 2018. Adversarial time-to-event modeling. In *International Conference on Machine Learning*, 735–744. PMLR.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Faraggi, D.; and Simon, R. 1995. A neural network model for survival data. *Statistics in medicine*, 14(1): 73–82.

Fotso, S. 2018. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*.

Gerds, T. A.; and Schumacher, M. 2006. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6).

Hosmer, D. W.; and Lemeshow, S. 2002. *Applied survival analysis: regression modelling of time to event data*. Wiley.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; Lauer, M. S.; et al. 2008. Random survival forests. *The annals of applied statistics*, 2(3): 841–860.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1).

Klein, J. P.; Gerster, M.; Andersen, P. K.; Tarima, S.; and Perme, M. P. 2008. SAS and R functions to compute pseudo-values for censored data regression. *Computer methods and programs in biomedicine*.

Kleinbaum, D. G.; and Klein, M. 2010. *Survival analysis*. Springer.

Knaus, W. A.; Harrell, F. E.; Lynn, J.; Goldman, L.; Phillips, R. S.; Connors, A. F.; Dawson, N. V.; Fulkerson, W. J.; Califf, R. M.; Desbiens, N.; et al. 1995. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*.

Kvamme, H.; and Borgan, Ø. 2019. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*.

Kvamme, H.; Borgan, Ø.; and Scheel, I. 2019. Time-to-event prediction with neural networks and Cox regression. *arXiv preprint arXiv:1907.00825*.

Lee, C.; Zame, W.; Yoon, J.; and van der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. Layer-wise relevance propagation: an overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer.

Nagpal, C.; Li, X. R.; and Dubrawski, A. 2021. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*.

Rahman, M. M.; Matsuo, K.; Matsuzaki, S.; and Purushotham, S. 2021. DeepPseudo: Pseudo Value Based Deep Learning Models for Competing Risk Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 479–487.

Yu, C.-N.; Greiner, R.; Lin, H.-C.; and Baracos, V. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, 24: 1845–1853.

Zhao, L.; and Feng, D. 2020. Deep neural networks for survival analysis using pseudo values. *IEEE journal of biomedical and health informatics*, 24(11): 3308–3314.