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
mrahman6@umbc.edu, psanjay@umbc.edu

Motivation

- Survival analysis aims to predict the survival probability or risk of an event over time.
- Existing statistical models and ML models, such as Cox Proportional Hazards [5] and Random Survival Forests (RSF) [4], are interpretable but less accurate, while deep learning-based survival models are accurate but not interpretable.
- **Goal:** To develop accurate yet interpretable deep survival models in the presence of censoring.

Our Proposed Solution

Key Idea: Use jackknife pseudo values to handle censoring [2] and to obtain a subject-specific prediction of survival probability. Interpretability is achieved by using Neural Additive Models (NAM) [1] or RSF [4].

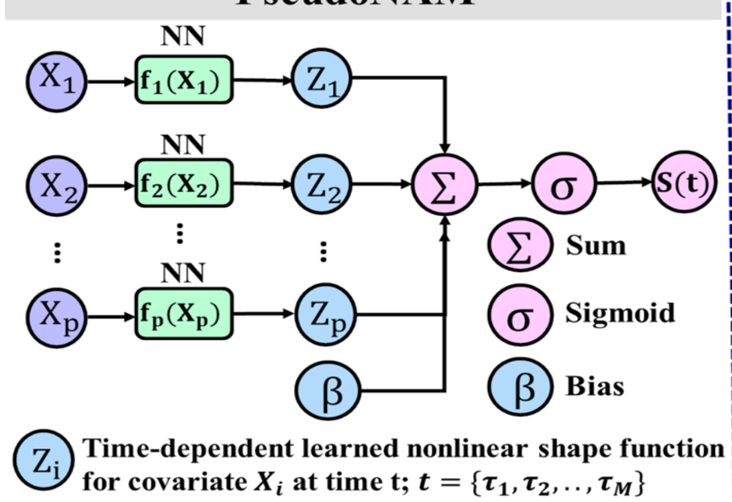
Pseudo values: For subject i, pseudo values, $\hat{y}_i(t)$, for survival probability [S(t)] at time t^* are defined as:

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

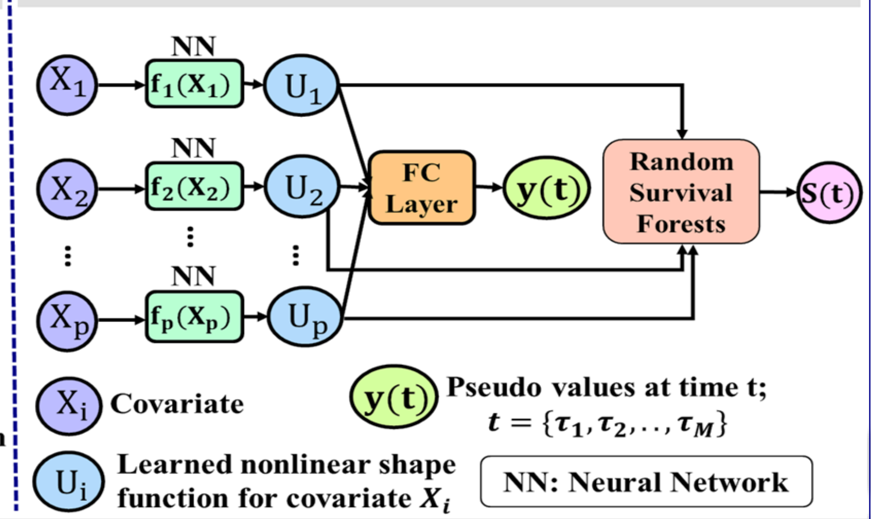
Proposed Models: **PseudoNAM** is an interpretable deep survival model, which uses the NAM to learn non-linear individual covariate effects on the survival probabilities. To improve the performance of PseudoNAM, we propose **PseudoNRSF,** which uses the learnt non-linear shape functions as input to an interpretable RSF.

Our Proposed Models

PseudoNAM



PseudoNRSF



Experimental Results

	Algorithm	Time-dependent C-index			Integrated Brier Score		
		METABRIC	SUPPORT	WHAS	METABRIC	SUPPORT	WHAS
Deep Learning Based Models	**PseudoNRSF**	0.645	0.619	**0.865**	0.171	0.196	**0.099**
	PseudoNAM	0.616	0.613	0.740	0.245	0.207	0.267
	DNNSurv [3]	0.617	0.581	0.721	0.243	0.221	0.290
	DeepSurv [5]	0.641	0.589	0.787	**0.165**	0.198	0.132
	DeepHit [6]	0.655	0.593	0.851	0.178	0.211	0.140
	DSM [7]	0.616	0.595	0.739	0.249	0.212	0.201
	CoxTime [8]	**0.660**	0.616	0.783	0.168	0.192	0.136
	PCHazard [8]	0.614	0.589	0.685	0.201	0.225	0.141
ML Based Models	**MTLR [9]**	0.550	0.550	0.618	0.225	0.263	0.162
	RSF [4]	0.616	**0.638**	0.768	0.296	**0.190**	0.206
Statistical Model	**CoxPH [5]**	0.622	0.568	0.739	0.313	0.206	0.234

Summary

- PseudoNAM models obtain good **predictive, discriminative, and interpretable results.**
  - ✓ Utilizes pseudo values to **handle censoring.**
  - ✓ capture **non-linear shape functions** through neural additive model, which is not possible in other non-deep models.
  - ✓ A step towards **transparency** in the deep learning models - through **global** and **feature-level interpretations.**
  - ✓ **Visualize** and quantify covariates' contribution and impact on the survival predictions.

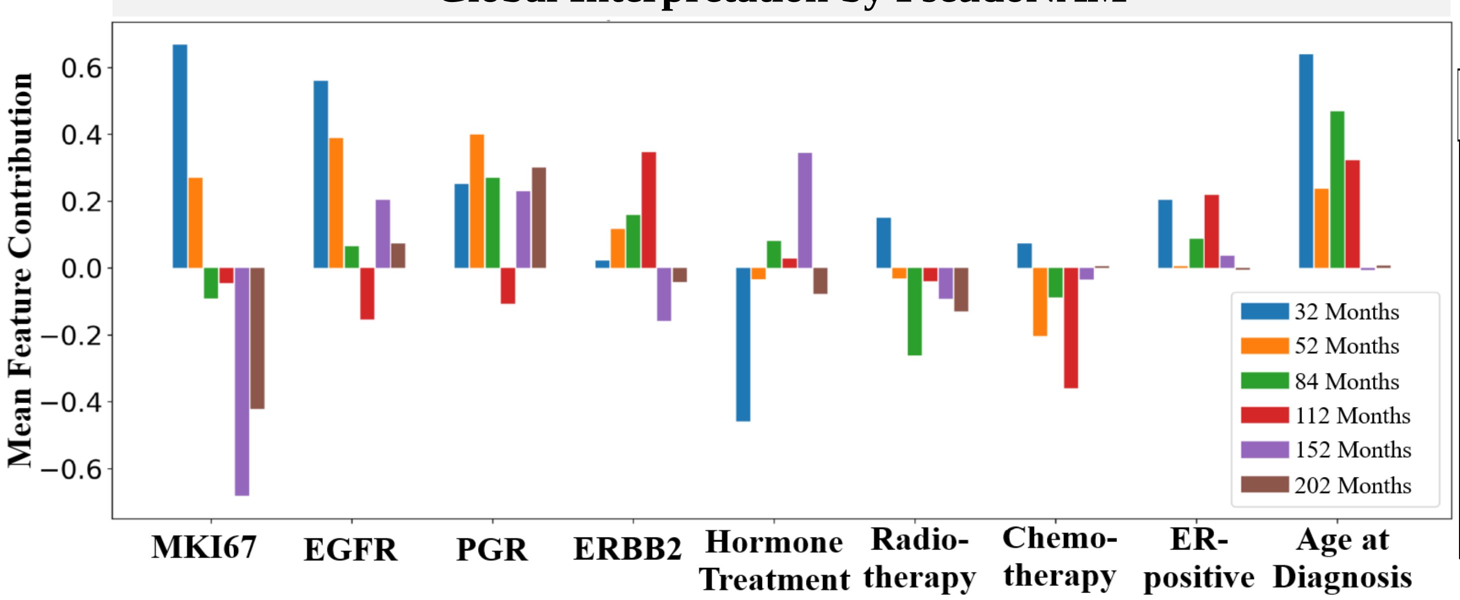
References

1. Agarwal et al. 2020. Neural additive models: Interpretable machine learning with neural nets. arXiv:2004.13912 .
2. Rahman et al. 2021. DeepPseudo: Pseudo Value Based Deep Learning Models for Competing Risk Analysis. AAAI, volume 35, 479–487.
3. Zhao et al. 2020. Deep neural networks for survival analysis using pseudo values. IEEE JBHI.
4. Ishwaran et al. 2008. Random survival forests. The annals of applied statistics
5. Katzman et al. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC.
6. Lee et al. 2018. Deephit: A deep learning approach to survival analysis with competing risks. AAAI.
7. Nagpal et al. 2021. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. IEEE JBHI.
8. Kvamme et al. 2019. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal*
9. Yu et al. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. NIPS.

This work is partially supported by grant IIS–1948399 from the US National Science Foundation

Interpretability: Global and Covariate level interpretations on METABRIC Data

Global Interpretation by PseudoNAM

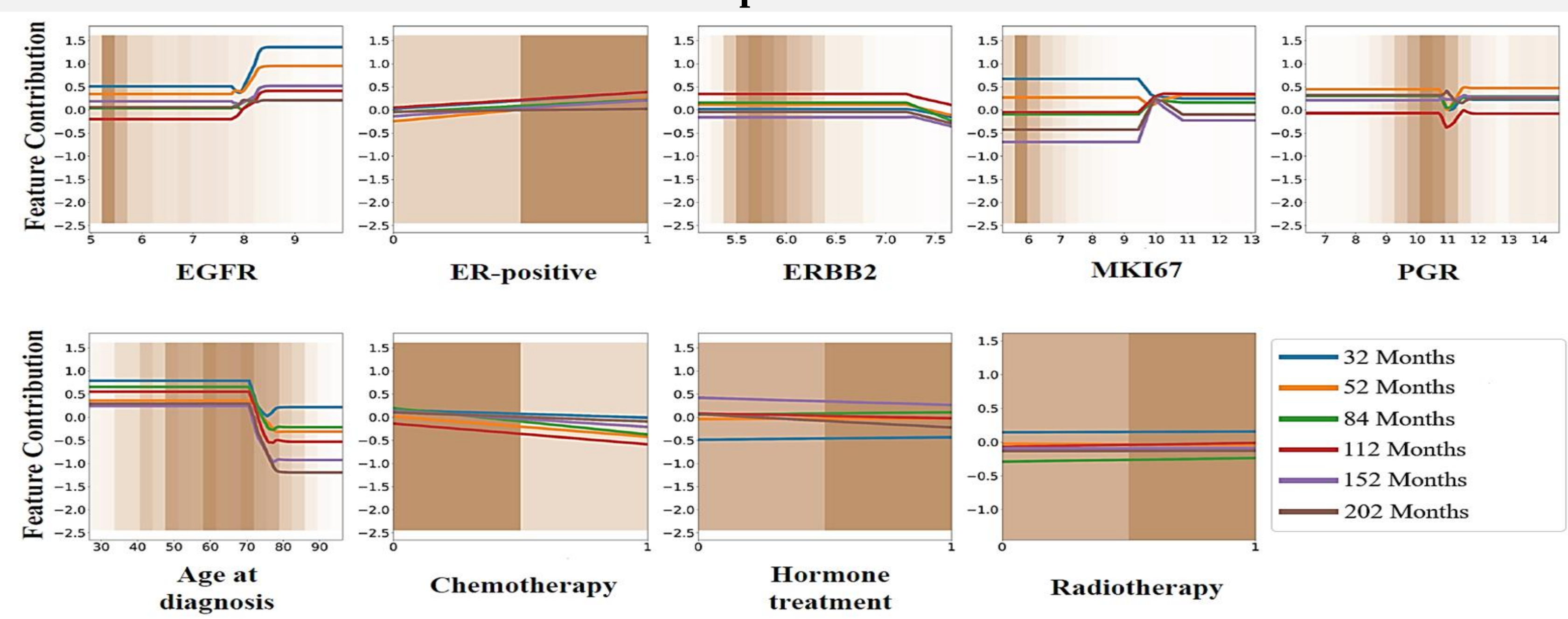


Global Interpretation by PseudoNRSF

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

Global interpretation plots provide overall feature importance scores and interpretations about the positive and negative impact of covariates on predictions. For example: Covariates such as **MKI67, radiotherapy, and chemotherapy** have positive feature contributions at earlier time points (32 months), which means that they influence better survival outcomes. However, at later time points (such as 112 months), these features have negative feature contributions - meaning they result in mortality.

Covariate Level Interpretation from PseudoNAM



Covariate level interpretation plots show individual feature time-dependent impact on survival predictions. For example: Survival probability for **age at diagnosis** at all prediction times starts decreasing after 65 years, and **Chemotherapy** is biased to the patients who did not receive chemotherapy since the density is much higher for this group (darker brown bar).