Interactive Explainable Deep Survival Analysis

Lu Wang[1,2], Xinyu Qin[1], Jingyan Jiang[4], Yan Li[2,3] and Winston Liaw[2]

***Abstract*— Being able to accurately predict the time to event of interest, commonly known as *survival analysis*, is extremely beneficial in healthcare for modeling disease progression, identifying prognostic factors, assessing risk of health by building survival models in health aging, precision medicine, supporting clinical decision making. In order to be usable by healthcare providers, survival analysis models need to be accurate, interpretable, and trustable. Efficient interaction between human stakeholders (e.g., developers, domain experts and/or end-users) and clear model interpretation not only improve the model performance but also enhance human trust. The primary goal of this paper is to develop algorithm and method that support implementation of trustworthy and time-efficient data-driven decision making for prevention and early intervention. Our experimental results on one public cancer datasets demonstrate the algorithm efficiency for predicting survival time of cancer patients.**

***Index Terms*— Survival Analysis, Deep Learning, interactive Machine Learning (iML), eXplainable Artificial Intelligence (XAI).**

I. Introduction

Survival analysis plays an essential role in healthcare science, such as prediction of patient death and time to progression in oncology, estimation of treatment cost, economic evaluations alongside clinical trials, treatment allocation of carcinoma in the precision medicine era, analysis of lifespan in aging research, and prediction of relapse in cognitive impairments.

80% of aging population have at least one chronic condition and the leading causes of death among older adults in the U.S. are chronic diseases including heart disease, cancer, stroke, Alzheimer's disease, and diabetes [1]. Most chronic diseases can be prevented to be worse if we can implement early intervention [2]. Therefor, building a robust prediction model for the early stage of chronic diseases is significant. For instance, we can estimate the probability that a patient will die by cancer or have heart failure during the upcoming months, and present the explainable results to physicians, who can develop an early intervention plan in order to help that patient extend their life expectancy and ensure health.

Recent advances in deep learning have led to its widespread application in many domains including survival analysis [3], [4], [5], [6], [7]. Deep survival analysis takes into account both *uncensored instances* (i.e., the instances whose target events are observed) and *censored instances* (i.e., those that are not observed). In addition, non-linearity relationship between features and target, and auto feature representation learning for unstructured data [8], [9] can be handled by deep survival analysis. While a deep neural network approach to survival analysis can improve accuracy, it creates a kind of "black box" decision making where stakeholders cannot tell how the model operates and how it comes to its decisions. Thus, uninterpretable deep survival analysis prediction, is challenging to use because human stakeholders don't have the information that they need to trust the way that the model makes decisions, nor can they determine when the model may be brittle versus when it is performing well.

An example of this phenomenon is a model that performed very well in distinguishing high-risk patients from non-high-risk patients based on x-ray images collected from Mount Sinai Hospital [10]. However, when the model was applied outside of Mount Sinai, the performance plummeted. This lack of generalizability happened because the model did not learn clinically relevant information from the images, but instead its predictions were based on specific characteristics of the x-ray machine that was used to image the high-risk ICU patients at Mount Sinai [10]. The model actually distinguished which machine was used instead of the risk of the patients. We can see that the model fails to capture the intrinsic patients' risk information when both model interpretation and encoding of stakeholder knowledge are missing. Thus, it lacks generalizability and fails to adapt to diverse situations.

Interactive explainable methods need to involve human experts in the model's development. Human experts need to understand how models come to their decisions so that they can identify inaccurate models before they go into the clinical validation and the certification process. Thus the development of interactive explainable models will save considerable time and effort by allowing clinicians to influence the model development with their knowledge and prevent faulty models from being fully implemented before their flaws are recognized.

Accurate diagnosis is critical in clinical decision making. However, 'prevention is better than cure' as prevention and early intervention will prevent the aging population from suffering more diseases and/or more extensive treatments. Although numbers of data driven healthcare prediction models have been proposed, most of them were merely focusing on computer-aided diagnosis instead of building an efficient prediction model in the early stage. This paper cares more

[1]Dept. of Biomedical Engineering, University of Houston, Houston, TX 77204, USA. lwang71, xqin5@central.uh.edu

[2]Dept. of Health Systems and Population Health Sciences, University of Houston, Houston, TX 77204, USA. yli159, wliaw@central.uh.edu

[3]Dept. of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S 3G8, Canada. yanrock.li@mail.utoronto.ca

[3]Argonne National Laboratory. jingyanjcmu@gmail.com

about how to support clinical decision making to provide patients effective prevention and timely intervention healthcare service via interactive explainable deep survival analysis (IEDSA).

II. Method

Survival analysis aims to estimate time-dependent survival probability in longitudinal studies, and hence time-dependent interpretation is desired. To achieve this, given an original input sample, we define time-dependent counterfactual example based on the properties of survival analysis. We further propose a time-dependent gradient integration to interpret the survival model.

A. The Brazilian National Cancer Institute (INCA) cancer dataset description

The public dataset we used in our experiments was collected by the Brazilian National Cancer Institute (INCA), which includes recorded information related to cancer cases among the Brazilian population. In this study, our primary focus is on predicting time to death for cancer patients. This dataset contains almost 40 variables including for 1,048,576 instances. We finally selected 14,338 patients with 26 variables to exclude the patients with majority of missing data.

B. Interactive Explainable Deep Survival Analysis Algorithm

In survival analysis, we can only observe either a *survival time* (O_i) or a *censored time* (C_i) for each instance. The dataset is said to be right-censored iff $T_i = \min(O_i, C_i)$ is observed during the study, where T_i is named as *observed time*. Usually an instance is represented as a triplet (X_i, T_i, δ_i), where $X_i \in \mathbb{R}^{1\times p}$ is the feature vector and δ_i is the censoring indicator which equals to 1 for an uncensored instance, and 0 for a censored instance. The primary goal of survival analysis is to learn a predictive function $f_\Theta(\cdot)$ parameterized by Θ, such that the predicted survival time is as close as possible to the true survival time. The learning process is to estimate the parameter Θ by minimizing the empirical expectation of a loss $\mathcal{L}(\Theta; X, T, \delta) = \sum_{i=1}^{N} \ell_{sur}(f_\Theta(X_i), T_i, \delta_i))$, where N is the number of training instances. $\ell_{sur}(\cdot)$ is a designed loss function for survival analysis, which leverages both uncensored instances (whose true survival time are known) and censored instances (whose true survival time are unknown but should be greater than the corresponding censored time).

This paper focuses on explaining deep discrete-time survival analysis, which is appropriate here for the following reasons: 1) Most of the existing deep continuous-time survival analysis models are extended from parametric censored regressions or Cox model [11], [12], [3], [4], and hence inherit shortcomings from their corresponding base models. Discrete-time survival analysis models require no assumption of the underlying distribution w.r.t. survival time nor survival function. Therefore, they are more generalized. 2) Deep continuous-time survival analysis models formulate the survival prediction as a regression problem; therefore, a post-hoc interpretation method for a general deep neural network can be used to interpret a deep continuous-time survival analysis model. So that a model-specific interpretation method is not needed for deep continuous-time survival models.

In the following, a formulation for discrete-time survival analysis is considered in this paper. Assume the maximal observed time is divided into T_{max} intervals $(t_0, t_1], \cdots (t_{T_{max}-1}, t_{T_{max}}]$, where $t_0 = 0$. Therefore, the survival time of the i-th instance can be estimated via predicting whether it is still survival $(T_i > t_j)$ or not within each time interval $(t_{j-1}, t_j]$ for $j = 1, \cdots, T_{max}$. For convenience, we denote $\mathrm{I}(T_i > t_j)$ as $Y_{i,j}$, where $\mathrm{I}(T_i > t_j) = 1$ if $T_i > t_j$, and 0 otherwise. Therefore, the discrete time survival analysis problem is formulated as the following optimization problem:

$$\min_{\Theta} \sum_{j=1}^{T_{max}} \sum_{i\in U_j} \ell\left(S_\Theta^{(j)}(X_i), Y_{i,j}\right), \tag{1}$$

$$s.t.\ \ 1 \geq S_\Theta^{(j)}(X_i) \geq S_\Theta^{(j+1)}(X_i) \geq 0, \tag{2}$$

where $U_j = \{i \mid \delta_i = 1 \cup C_i > t_j\}$ represents the set of uncensored instances and the instances that are censored after the j-th time interval; thus, leverages all instances including handling the censored instances. ℓ is an empirical loss for classification, e.g., logistic loss. $S_\Theta^{(j)}(X_i)$ is the estimated survival probability of the i-th instance at the j-th time interval. The constraint term in Eq. (1) preserves the natural property of survival probability, i.e., remains in $[0, 1]$ and monotonically non-increasing over time.

In this paper, we introduce a sophisticated designed network structure to tackle the constraint in Eq.(1). Based on the Markov assumption, survival probability at the j-th time interval is:

$$S_\Theta^{(j)}(X_i) = f_{\theta_j}(X_i) \cdot S_\Theta^{(j-1)}(X_i) = \prod_{1}^{j} f_{\theta_j}(X_i), \tag{3}$$

where $f_{\theta_j}(X_i)$ denotes the predicted probability that the event of interest is **not** happened during the j-th time interval, and $S_\Theta^{(j-1)}(X_i)$ is the survival probability at the $(j-1)$-th time interval. Since $S_\Theta^{(0)}(X_i) = 1$, i.e., for all observed instances that event of interest is not happened prior the starting time, the second equality holds in Eq. (3). θ_j is the set of model parameters associated with the j-th time interval, and note that parameters for different time intervals, i.e., θ_j and θ_k, may share same components based on the structure of the neural network. Eq.(3) ensures that the $S_\Theta^{(j)}(X_i)$ obeys the constraint in Eq.(1) and the corresponding network structure is illustrated in Figure 1. To interpret the aforementioned deep discrete-time survival analysis model, we propose an *Integrated Gradients (IG)* based attribution method. *IG* is one of the most commonly used interpretation algorithm for deep learning models [13]. In *IG* the attribution score of the l-th feature of a given sample (x) is mathematically defined as $IG_l(x) := (x_l - x'_l) \times \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_l} d\alpha$, which is a path integration of the gradients from a *reference/baseline* (i.e., x') to the given sample along a straight line. Therefore,

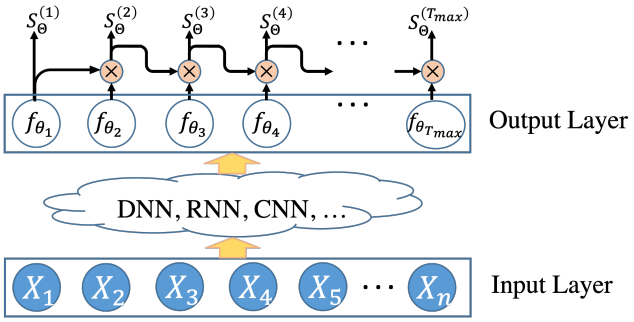

Fig. 1: Conceptual representation of the designed network structure that ensures the monotonically non-increasing property of survival probability. A specific network structure will be designed to meet the property of input data, e.g., CNN for image data and DNN for tabular data, and hybrid structure can be used to handle multimedia data.

the attribution score is highly dependent on the selection of reference.

The most common type of feedback from physicians is that particular features are important, and in practice they only care about the top influential features. Let $\Phi_l = \{\phi_{i,l}^{(j)} | i = 1, \cdots, N, \text{ and } j = 1, \cdots, T_{max}\}$ denotes the set of attributions related to the l-th feature, and $\mathcal{I} \subset \{1, \cdots, p\}$ be the set of important features selected by physicians. To encode this feedback, we encourage $\{||\Phi_{l'}||_2 \geq \frac{1}{p}\sum_{l=1}^{p}||\Phi_l||_2 \mid \forall l' \in \mathcal{I}\}$, i.e., for each selected feature the l_2 norm of its correlated attributions should be greater than or equal to the average of l_2 norm of the attribution across all features. In addition, to encourage the model to focus on the top influential features, we propose the following regularization term:

$$\Omega(\Phi(\Theta, X)) = \lambda_1 \sum_{l' \in \mathcal{I}} \text{ReLU}(\frac{1}{p}\sum_{l=1}^{p}||\Phi_l||_2 - ||\Phi_{l'}||_2) \quad (4)$$
$$+\lambda_2 \sum_{l'' \notin \mathcal{I}} ||\Phi_{l''}||_2,$$

where $\text{ReLU}(\frac{1}{p}\sum_{l=1}^{p}||\Phi_l||_2 - ||\Phi_{l'}||_2)$ is greater than 0 iff $\frac{1}{p}\sum_{l=1}^{p}||\Phi_l||_2 > ||\Phi_{l'}||_2$. Therefore, the first part of regularization term encourages the selected features to have above average contribution to the output. The second part of regularization ($\sum_{l'' \notin \mathcal{I}} ||\Phi_{l''}||_2$) encourages the group sparsity among none-selected features, i.e., if the l''-th feature is unimportant then all elements of its feature attribution ($\Phi_{l''}$) are encouraged to be close to zero, where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ adjust the regularization strength of each part.

III. Experiments and Results

A. Interactive Human Expertise Integration

We integrate the feedback from one physician (i.e., Dr. Winston Liaw, M.D., M.P.H.) in our team into our algorithm to improve the prediction performance to achieve interactive Machine Learning (iML). Dr. Liaw recognized several features such as Extension (localized or metastasis), and Illness.Code (Brazilian ICD-10 codes, specifically cancer types as some cancers, e.g., C61–ovarian cancer, are more aggressive than others, e.g., C509–female breast cancer [14]) that are more important in the INCA dataset to predict the death time of cancer patients.

B. Comparison methods

We demonstrate the performance of our proposed method by comparing with several other commonly used methods:

- **The Cox proportional hazards model (Cox)**: The Cox model is the most commonly used method in survival analysis [15], and it is trained by using the *coxph* function in the *survival* R package[1] [16].
- **DeepSurv**: DeepSurv is an extension of Cox proportional hazards model that employs a deep neural network to replace the linear regression in standard Cox model thereby handling the non-linearity [11]. A Pytorch implementation of DeepSurv can be found in the *pycox* package[2].
- **CoxTime**: CoxTime is a relative risk model that extends Cox regression beyond the proportional hazards [12], and its Pytorch implementation is available in the *pycox* package[3].
- **DeepHit**: DeepHit is a deep learning model designed for discrete-time survival analysis, where a feed forward DNN that incorporates both ranking loss and binary loss (log-likelihood) at each discrete time point is used to predict the probability density values at each time point [5]. A Pytorch implementation of Deephit can also be found in the *pycox* package[4].

C. Evaluation metric: the concordance index (C-index)

The concordance index (C-index), is a general performance measure of prediction models that generates continuous, ordinal and dichotomous outcomes [17], which quantifies the quality of predicted rankings. Let us consider a pair of 2-tuples $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$, where y_i is the true target value, and $\hat{y}_i$ is the predicted outcome. The concordance probability is defined as:

$$c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 > y_2). \quad (5)$$

By definition, the C-index has the same scale as the area under the ROC Curve (AUC) in binary classification, and if y_i is binary, then the C-index is same as the AUC. Therefore, similar to the AUC, $c = 1$ indicates perfect prediction and $c = 0.5$ indicates the prediction is as good as a random guess.

As censored data can be easily taken into account, the C-index is the most commonly used evaluation metric in survival analysis [18]. In 1982, Harrell et al. proposed the first definition and computational formulation of C-index for time-to-event data [18], i.e., the proportion of concordant pairs divided by the total number of possible evaluation pairs. Based on the types of learning targets, the existing survival prediction methods can be divided into two categories: risk score orientated and survival time orientated. The risk score orientated methods, e.g. the Cox proportional hazard model and its extensions, aim at learning a risk score for each instance. Note that, the instance with a low risk score should survive longer.

[1] https://cran.r-project.org/web/packages/survival/index.html

D. Results

We compare our proposed method interactive explainable deep survival analysis (IEDSA) with several other commonly used methods in survival analysis mentioned in the subsection III-B in terms of C-index defined in the subsection III-C using the Brazilian National Cancer Institute (INCA) cancer dataset described in the subsection II-A. To test the model performance, we splited the INCA cancer dataset into 80% training and 20% testing sets. The comparison results are presented in Table I and we can see that our proposed IEDSA method achieved the best prediction performance in the testing set.

We also present the Fig. 2, plotted based on the calculated expected gradient values, which explains IEDSA prediction model in terms of feature importance. As one of the most important eXplainable Artificial Intelligence (XAI) methods, feature importance is measured as a way to explain the identification results generated from ML models. Based on this plot it can be seen that four features have stronger predictive power in the trained model, which are Indicator of Rare Case (i.e., is the disease rare), Morphology Description (i.e., carcinoma information including type and sites), Description of Disease (i.e., the distribution of cancer within the anatomical structures of a biological organism) and Status Address (i.e., the state where the patient is living now). Note that, in Fig. 2, a positive value influences the model to predict alive, whereas the negative ones influence the model toward prediction of death.

TABLE I: **Performance comparison of the proposed method interactive explainable deep survival analysis (IEDSA) and other existing related methods using C-index.**

	Cox	Deepsurv	Coxtime	Deephit	IEDSA
Train dataset	**0.819**	0.730	0.741	0.740	0.817
Test dataset	0.629	0.725	0.738	0.729	**0.813**

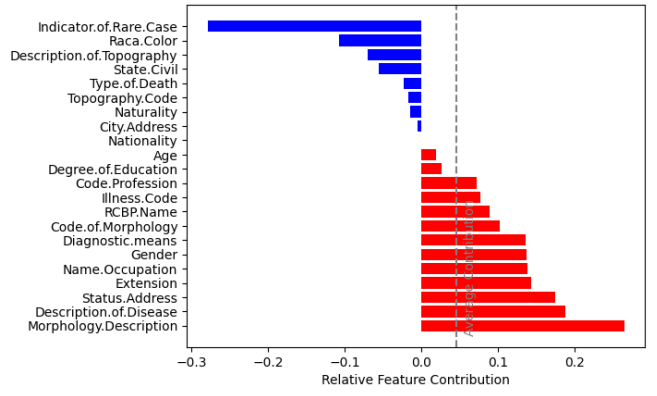

Fig. 2: ***Feature importance generated from our proposed IEDSA method.***

IV. Conclusion

In this paper, we proposed a time-dependent interpretation method for survival analysis to improve transparency and trustworthiness of deep discrete-time survival analysis models achieved the targets of both iML and XAI. Moreover, the proposed interpretation method can be used to encode domain knowledge and expert feedback in an interactive way to achieve better prediction performance. In the future, we plan to develop a sufficient system to keep human in the loop that allows human experts to provide feedback to the algorithm in order to optimize it via a human-computer interaction (HCI) interface.

Acknowledgement

This work is supported by US National Science Foundation grant IIS-2245739.

References

[1] N. C. on Aging, "Healthy aging," 2018.

[2] W. H. Organization, P. H. A. of Canada, C. P. H. A. of Canada *et al.*, *Preventing chronic diseases: a vital investment*. World Health Organization, 2005.

[3] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 544–547.

[4] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei, "Deep survival analysis," in *Machine Learning for Healthcare Conference*, 2016, pp. 101–114.

[5] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] E. Giunchiglia, A. Nemchenko, and M. van der Schaar, "Rnn-surv: A deep recurrent model for survival analysis," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 23–32.

[7] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu, "Deep recurrent survival analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4798–4805.

[8] P. Deepa and C. Gunavathi, "A systematic review on machine learning and deep learning techniques in cancer survival prediction," *Progress in Biophysics and Molecular Biology*, 2022.

[9] R. W. Pettit, R. Fullem, C. Cheng, and C. I. Amos, "Artificial intelligence, machine learning, and deep learning for clinical outcome prediction," *Emerging topics in life sciences*, vol. 5, no. 6, pp. 729–745, 2021.

[10] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[11] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, p. 24, 2018.

[12] H. Kvamme, Ørnulf Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *Journal of Machine Learning Research*, vol. 20, no. 129, pp. 1–30, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-424.html

[13] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[14] A. Yoneda, M. E. Lendorf, J. R. Couchman, and H. A. Multhaupt, "Breast and ovarian cancers: a survey and possible roles for the cell surface heparan sulfate proteoglycans," *Journal of Histochemistry & Cytochemistry*, vol. 60, no. 1, pp. 9–21, 2012.

[15] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220, 1972.

[16] T. Therneau, "A package for survival analysis in S. R package version 2.37-4," *URL http://CRAN. R-project. org/package= survival. Box*, vol. 980032, pp. 23 298–0032, 2013.

[17] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[18] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.