Supplementary Materials for DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks

Changhee Lee,[1] William R. Zame, [2] Jinsung Yoon, [1] Mihaela van der Schaar [3, 1, 4]
[1] Department of Electrical and Computer Engineering, University of California, Los Angeles, USA
[2] Department of Economics, University of California, Los Angeles, USA
[3] Department of Engineering Science, University of Oxford, UK
[4] Alan Turing Institute, London, UK
chl8856@ucla.edu, zame@econ.ucla.edu, jsyoon0823@ucla.edu, mihaela.vanderschaar@oxford-man.ox.ac.uk

Description on MP-based Survival Models

For the comparison of DeepHit with conventional machine learning algorithms, we modified survival data to perform survival analysis based on the mortality prediction using machine learning algorithms. For every time interval m where $m = 0, \cdots, T_{\text{max}}$, the cause-specific data is updated with a new dataset of patients who are not censored nor died from other causes until the m-th time interval. Now, let's focus on modifying the survival data for the event k, which can be easily generalized to other causes.

The number of patients who are not censored nor died from other causes (i.e. other than cause k) until the m-th time interval is denoted as N_m^k. Then, a new label, $\tilde{l}_k^{(i)}$, which indicates whether the i-th patient is dead ($\tilde{l}_k^{(i)} = 1$) or alive ($\tilde{l}_k^{(i)} = 0$) at the m-th time interval, is assigned for every patient. Using the updated dataset $\tilde{\mathcal{D}}_m^k = \{\mathbf{x}^{(i)}, \tilde{l}_k^{(i)}\}_{i=1}^{N_m^k}$, we train conventional machine learning (ML) algorithms (e.g. random forest, logistic regression and AdaBoost) in order to predict the new label. From this, it is possible to obtain ML classifiers independently trained at every time interval. Then, the risk score of a patient at each time interval can be assessed by using ML classifiers trained at the corresponding time. The pseudo-code for training the ML-based survival models for event cause k is described in Algorithm 1.

Algorithm 1 Pseudo code for ML-based survival model for event cause k

Initialize: $\tilde{\mathcal{D}}_m^k = \phi$ for $m = 0, \cdots, T_{\text{max}}$
for $m = 0, \cdots, T_{\text{max}}$ **do**
  for $i = 1, \cdots, N$ **do**
    if $k^{(i)} == k \ \ \& \ \ s^{(i)} \leq m$ **then**
      $\tilde{l}_k^{(i)} \leftarrow 1$
      $\tilde{\mathcal{D}}_m^k \leftarrow \tilde{\mathcal{D}}_m^k + \{\mathbf{x}^{(i)}, \tilde{l}_k^{(i)}\}$
    else if $s^{(i)} > m$ **then**
      $\tilde{l}_k^{(i)} \leftarrow 0$
      $\tilde{\mathcal{D}}_m^k \leftarrow \tilde{\mathcal{D}}_m^k + \{\mathbf{x}^{(i)}, \tilde{l}_k^{(i)}\}$
    end if
  end for
  Train ML predictor $\mathcal{H}_{\text{ML},k}^{(m)}(\mathbf{x})$ with $\tilde{\mathcal{D}}_m^k = \{\mathbf{x}^{(i)}, \tilde{l}_k^{(i)}\}_{i=1}^{N_m^k}$
end for

Additional Results for Discriminative Performance

In this section, we provide additional results on the performance benefits in terms of cause-specific C^{td}-index, comparing with the cause-specific version of survival models for the SEER and the SYNTHETIC datasets, respectively in Table 1 and 2.

For the SEER dataset, DeepHit provided consistent performance improvements over conventional benchmarks where the improvements were statistically significant ($p < 0.05$ and $p < 0.001$) except for *cs*-MP-AdaBoost and *cs*-MP-LogitR in CVD prognosis, and statistically significant ($p < 0.05$ and often $p < 0.001$) for all benchmarks in breast cancer prognosis. For the SYNTHET dataset, DeepHit outperformed all the benchmarks and the performance improvements were all statistically significant ($p < 0.001$) for both Event 1 and Event 2.

Table 1: Additional comparison of cause-specific C^{td}-index performance (mean and 95% confidence interval) tested on the SEER dataset

Algorithms	CVD	Breast Cancer
cs-RSF	0.280* (0.262 - 0.298)	0.584* (0.574 - 0.594)
cs-ThresReg	0.664‡ (0.657 - 0.671)	0.645* (0.628 - 0.662)
cs-MP-RForest	0.281* (0.263 - 0.299)	0.584* (0.574 - 0.594)
cs-MP-AdaBoost	0.671 (0.665 - 0.677)	0.741‡ (0.735 - 0.747)
cs-MP-LogitR	0.665 (0.645 - 0.685)	0.657* (0.648 - 0.666)
DeepHit	**0.684 (0.674 - 0.694)**	**0.752 (0.748 - 0.756)**

$*$ indicates p-value < 0.001
‡ indicates p-value < 0.05

Table 2: Additional comparison of cause-specific C^{td}-index performance (mean and 95% confidence interval) tested on the SYNTHETIC dataset

Algorithms	Event 1	Event 2
cs-RSF	0.669* (0.664 - 0.674)	0.657* (0.652 - 0.662)
cs-ThresReg	0.579* (0.574 - 0.584)	0.588* (0.585 - 0.591)
cs-MP-RForest	0.620* (0.611 - 0.629)	0.610* (0.603 - 0.617)
cs-MP-AdaBoost	0.607* (0.600 - 0.614)	0.607* (0.601 - 0.613)
cs-MP-LogitR	0.579* (0.572 - 0.586)	0.586* (0.583 - 0.589)
DeepHit	**0.755 (0.749 - 0.761)**	**0.755 (0.748 - 0.762)**

$*$ indicates p-value < 0.001

Weighted Average of AUC Curve

Through out the section, we focus on the single risk/event case for further investigation on loss functions, and, thus, we omit cause-specific indicator k for notational simplicity. To highlight the gain of introducing $\mathcal{L}_2$, we adopt the analysis that the C^{td}-index is equivalent to the weighted average of the area under time-specific ROC curve (AUC) (Antolini, Boracchi, and Biganzoli 2005),

$$C^{td} = \sum_{m=0}^{T_{\max}} w(m) \cdot AUC(m), \tag{1}$$

In this equation, $AUC(t)$ is based on the incident/dynamic definition of sensitivity and specificity in (Heagerty and Zheng 2005) and $w(t)$ indicates the weight for $AUC(t)$ which is proportional to having acceptable pairs. $AUC(t)$ and $w(t)$ are defined as

$$AUC(t) = \frac{\sum_{i \neq j} \mathbb{1}\big(F(t|\mathbf{x}^{(i)}) > F(t|\mathbf{x}^{(j)})\big) \cdot \mathbb{1}(s^{(i)} = t, s^{(j)} > t)}{\sum_{i \neq j} \mathbb{1}(s^{(i)} = t, s^{(j)} > t)},$$

$$w(t) = \frac{\sum_{i \neq j} \mathbb{1}(s^{(i)} = t, s^{(j)} > t)}{\sum_{m=0}^{T_{\max}} \sum_{i \neq j} \mathbb{1}(s^{(i)} = m, s^{(j)} > m)},$$

It is worth to stress that the same analysis on C-index is available for survival models with the PH assumption (Heagerty and Zheng 2005).

In Figure 1 and 2, we depicted $AUC(t)$ of DeepHit over time with corresponding $w(t)$ for the UNOS and METABRIC dataset, respectively. An averaging window with the window size of 10 time intervals is applied for $AUC(t)$ in order to mitigate fluctuations since $AUC(t)$ is only available at times where death event occurs. DeepHit trained with both $\mathcal{L}_1$ and $\mathcal{L}_2$ (i.e. DeepHit) outperforms the same network trained with only $\mathcal{L}_1$ (i.e. DeepHit w/ $\alpha = 0$) consistently over time. Moreover, compared to conventional survival models, DeepHit provides performance gain on time intervals where $w(t)$ is relative high. In other words, $\mathcal{L}_2$ helps the proposed network to focus its discriminative performance on time intervals where there a large number of patients who face the death event. Therefore, DeepHit is able to achieve the performance gain in terms of C^{td}-index.

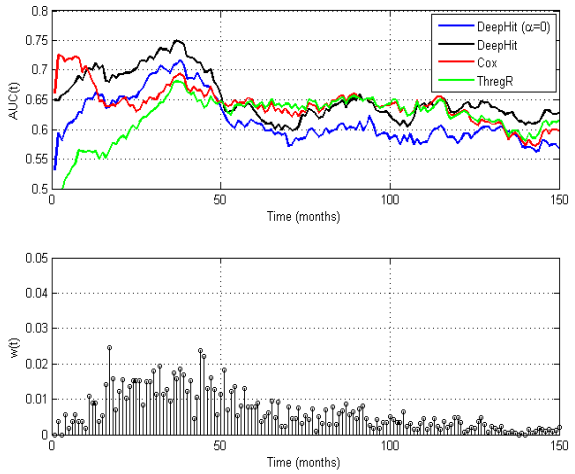

Figure 1: The performance in $AUC(t)$ over time with corresponding $w(t)$ for the UNOS dataset.

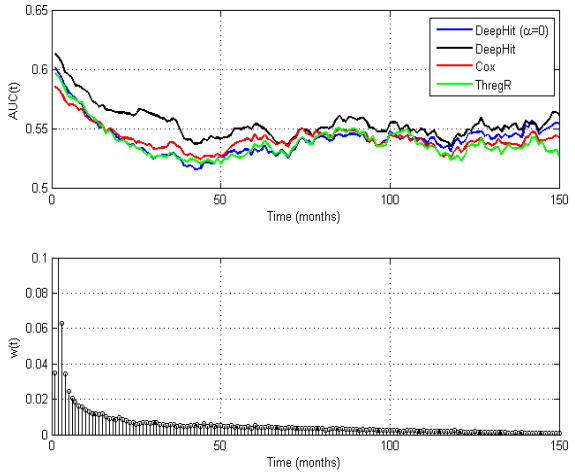

Figure 2: The performance in $AUC(t)$ over time with corresponding $w(t)$ for the METABRIC dataset.

References

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

Heagerty, P. J., and Zheng, Y. 2005. Survival model predictive accuracy and roc curves. *Biometrics* 61:92–105.