



# Adapting Ensemble Neural Networks to Clinical Prediction in High-Dimensional Settings

Simon de Montigny<sup>1,2(✉)</sup> and Philippe Broët<sup>1,2,3</sup>

<sup>1</sup> CHU Sainte-Justine Research Center, Montreal, QC, Canada

<sup>2</sup> School of Public Health, University of Montreal, Montreal, QC, Canada  
simon.de.montigny@umontreal.ca

<sup>3</sup> University Paris-Saclay, University Paris-Sud, CESP, INSERM, Paris, France

**Abstract.** Neural networks have been investigated as models for survival data using a training criterion similar to that of the Cox proportional hazards model, a criterion not designed for clinical prediction. In this paper, we develop a new survival learning algorithm where a neural network ensemble minimizes the integrated Brier score. We compare the results obtained with this method to a standard implementation of random survival forests in R and to an ensemble of linear units.

**Keywords:** Neural networks · Survival analysis · Predictive models

## 1 Introduction

Neural networks (NNs) have been discussed for clinical use and survival analysis starting in the mid 90s, but early works had serious shortcomings [1]. Many survival deep learning models have now been proposed [2–8], with a clear focus on regularization and validation. Predictive accuracy of these NN models are usually assessed with the C-index [9] or the Brier score [10]. Limitations remain for clinical applications: these NNs have loss functions that don't measure predictive accuracy, and they are not well suited for high-dimensional data. In this work, we propose a new survival learning algorithm which combines predictions from an ensemble of NN models minimizing the integrated Brier score, optionally with  $L_1$  penalization. We compare this procedure to the state-of-the-art ensemble approach which is the Random Survival Forest [11], and to a baseline ensemble of linear units that maximize partial likelihood under  $L_1$  penalization. To evaluate performance in the high-dimensional setting, we created different survival data sets by adding non-informative covariates to the well-known Primary Biliary Cirrhosis (PBC) dataset [12].

## 2 Probabilistic Survival Model

The health status of a patient is measured until a certain event occurs or until he is lost to follow-up. Let the random variables  $T$  and  $C$  be the time-to-event

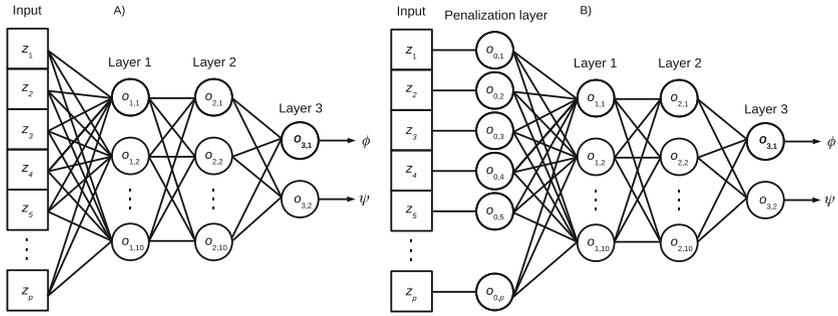
and the censoring time, respectively. We define  $X = \min(T, C)$  as the observed follow-up time and  $\delta = 1_{(X=T)}$  as the event indicator. We assume noninformative and independent censoring for  $T$  and  $C$  [13]. The survival function of  $T$  is defined by  $S(t) = P[T > t]$  ( $t \geq 0$ ), the hazard function by  $\lambda(t) = -\left(\frac{d}{dt}S(t)\right)/S(t)$ , and the cumulative hazard function by  $\Lambda(t) = \int_0^t \lambda(s)ds$ ; we have  $S(t) = \exp(-\Lambda(t))$ .

To take into account that some patients are not susceptible to the event of interest, we use an improper survival function  $S(t)$  such as  $\lim_{t \rightarrow \infty} S(t) = \epsilon$  where  $\epsilon$  ( $0 < \epsilon < 1$ ) is the tail defect; we then have  $\Lambda(t) \leq -\ln \epsilon$ . Broadly speaking, the random variable  $T$  takes the value  $\infty_+$  for non-susceptible patients. In this context, we consider an improper semi-parametric model given by  $S(t | Z) = \exp\left\{-\theta \exp[\phi(Z)] [1 - A(t)^{\exp[\psi(Z)]}]\right\}$  where  $Z = (Z_1; \dots; Z_p)$  is a  $p$ -dimensional vector of covariates, where  $A(t)$  can be any function decreasing with time from one to zero, and where  $\theta$  is a positive parameter. This type of model is a useful alternative to the standard Cox model which allows to investigate survival effects evolving in time. Here,  $\phi(Z)$  and  $\psi(Z)$  are two risk functions that correspond to the long-term effect (linked to the tail defect) and the short-term effect (linked to the time-to-event survival distribution for susceptible patients), respectively. The tail defect is given by  $\epsilon = \exp[-\theta \exp(\phi(Z))]$ . We define  $\theta$  and  $A(t)$  based on the Nelson-Aalen estimator of the cumulative hazard rate, noted  $H(t)$ , as follows. We set  $\theta = \max\{H(t)\}$  and, given  $H^-(t) = \max\{H(t)1_{(H(t)<\theta)}\}$  and  $H^*(t) = H(t)1_{(H(t)<\theta)} + H^-(t)1_{(H(t)=\theta)}$ , we set  $A(t) = 1 - \theta^{-1}H^*(t)$ . Moreover, for small values of  $\psi(Z)$ ,  $S(t|Z)$  can be re-expressed as a time-dependent proportional hazard model [14].

### 2.1 Neural Network Architecture Proposal

We propose to model the risk functions  $\phi(Z)$  and  $\psi(Z)$  with a NN having a  $p$ -dimensional input and a two-dimensional output  $(o_{3,1}; o_{3,2})$ . The network, shown in Fig. 1A, is described by  $o_{a,b} = h_a\left(w_{a,b,0} + \sum_{j=1}^{10} w_{a,b,j}o_{a-1,j}\right)$  for layers  $a = 2, 3$ , and by  $o_{1,b} = h_{1,b}\left(w_{1,b,0} + \sum_{j=1}^p w_{1,b,j}z_j\right)$  for layer 1. We use  $h_1(x) = h_2(x) = \text{selu}(x)$ , a scaled exponential linear unit [15], and  $h_3(x) = 5 \tanh(x)$ , a scaled hyperbolic tangent. The resulting survival function is noted  $\hat{S}(t|Z)$ . A variant of the network, where input variables are subjected to  $L_1$  penalization, is described in Fig. 1B. In this case, the equation for the first layer is given by  $o_{1,b} = \phi_1\left(w_{1,b,0} + \sum_{j=1}^p w_{1,b,j}o_{0,j}\right)$  with  $o_{0,j} = w_{0,j}z_j$ , where  $w_{0,j}$  is the weight of the  $j$ th unit of the penalization layer (note that these units have no bias term).

We base the loss function of the network on the integrated Brier score [16], defined by  $\text{IBS} = \frac{1}{\tau} \int_0^\tau \text{BS}(t) dt$  where  $\tau = \max(X_i \delta_i)$  is the time of the last uncensored event, and where  $\text{BS}(t)$  is the Brier score at time  $t$ , a pointwise mean square error between  $\hat{S}(t|Z)$  and what is observed. The observation variable takes value 1 if the event did not occur up to time  $t$ , value 0 if the event did occur, and it does not exist in case of censoring. To account for this third case, the



**Fig. 1.** A) Three layered NN. B) Modified NN with penalization layer.

error is weighted by the inverse probability of censoring. Thus, we have  $BS(t) = \frac{1}{n} \sum_{i=1}^n \left\{ \left[ \hat{S}(t|Z_i) \right]^2 \hat{G}^{-1}(X_i) 1_{(X_i \leq t, \delta_i=1)} + \left[ 1 - \hat{S}(t|Z_i) \right]^2 \hat{G}^{-1}(t) 1_{(X_i > t)} \right\}$ . The function  $\hat{G}(t)$  is the nonparametric Kaplan-Meier estimate of the censoring distribution. The square root  $\sqrt{BS(t)}$  represents the deviation between the predicted outcome and the true event status. In the modified network, a penalization term  $\lambda_1 \sum_{j=1}^p |w_{0,j}|$  is added to the IBS, where  $\lambda_1$  is the penalization parameter.

### 2.2 Classical Approaches

The baseline model (ensemble of linear units) that we use in our experiments is derived from the hazard  $\lambda(t|Z) = \nu(t)e^{\phi(Z)}$ , with  $\nu(t)$  a baseline hazard, and from the partial likelihood function  $L = \prod_{i=1}^n e^{\phi(Z_i)\delta_i} / \left( \sum_{j=1}^n e^{\phi(Z_j)} 1_{(X_j \geq X_i)} \right)$ . Model parameters in  $\phi(Z)$  are adjusted to maximize  $L$ . Equivalently, we can minimize  $\ell = -\sum_{i=1}^n \left( \phi(Z_i)\delta_i - \sum_{j=1}^n \phi(Z_j) 1_{(X_j \geq X_i)} \right)$ , that is the negative partial log-likelihood. We use  $\ell$  as the loss for each unit of the ensemble. Applications of NNs to survival analysis have also focused on minimizing  $\ell$  or its variants.

Random Survival Forest (RSF) is one of the most effective machine learning approaches for survival prediction. Broadly speaking, the RSF builds a series of binary decision trees from which a final prediction is obtained by combining the predictions from each individual tree. These latter tree-based learners are non-parametric approaches that partition recursively the predictor space into disjoint sub-regions that are homogeneous according to the outcome of interest. These partitions are obtained from a splitting criterion, usually the logrank statistic, that can be expressed as a score test from the partial likelihood function.

## 3 Experiment

### 3.1 Simulated Dataset

The PBC dataset has  $n = 312$  observations and  $p = 17$  covariates. To test the capacity of the models to select relevant covariates, we generated two modified

versions of the PBC dataset. For the second version, we added 500 uninformative variables (each of them, for every patient, generated randomly following a uniform distribution on the interval 0–1), resulting in a dataset with  $p = 517$  covariates. For the third version, we added 5000 uninformative variables in the same manner instead of 500, resulting in a dataset with  $p = 5017$  covariates.

### 3.2 Models

We tested four models on the dataset: a survival NN ensemble (SNNE), a SNNE with  $L_1$  penalization (SNNE- $L_1$ ), a RSF, and an ensemble of linear units (baseline). The survival random forest model is generated with the `rfsrc` function (with default values) from the R package `randomForestSRC`. We implemented the three other models in Python with Keras and TensorFlow. The ensemble method comprises bagging with 1000 bootstrap samples for all four models.

The prediction of NN ensembles for a patient is the average of the survival curves  $\hat{S}(t|Z)$  from every network where the patient was out-of-bag. Note that  $H(t)$ ,  $\theta$ ,  $A(t)$ ,  $\hat{G}(t)$  and  $\tau$  are computed in-bag. The process is similar for the baseline model: the survival estimate for each bootstrap sample is given by  $\hat{S}(t|Z) = [K(t)]^{\exp[h(w_{1,0} + \sum_{j=1}^p w_{1,j} w_{0,j} z_j)]}$ , where  $w_{1,j}$  for  $j = 0, \dots, p$  are the weights of the linear unit, where  $w_{0,j}$  are the penalization weights, and where  $K(t) = \exp[-H(t)]$  is the Fleming-Harrington estimator.

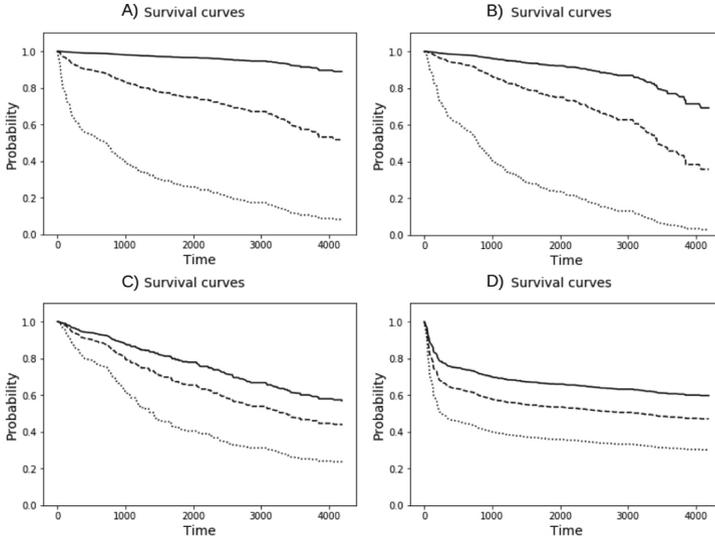
For the SNNE model, we normalized the inputs (in-bag) and we used the Glorot uniform initializer. We then trained each NN for 200 epochs with mini-batches (size 32) with the default Adam optimizer, and we selected the best weights with 15% in-bag validation. In addition, for the SNNE- $L_1$  model, we used  $\lambda_1 = 0.01$  and we initialized the penalization layer with a uniform distribution (0.95–1.05 interval). For the baseline model, we used the same training setup (with  $\lambda_1 = 0.01$  for penalization), except that we used the batch mode of training (no validation set), because  $\ell$  is not a sum of individual error terms (mini-batches with validation have not been studied in the literature for partial likelihood).

**Table 1.** Out-of-bag prediction error, computed with  $\tau = 4191$  (time of the last uncensored event). SNNE- $L_1$  shows best performance (values highlighted in bold). These values do not include the penalization term for the SNNE- $L_1$  and baseline models.

Model	IBS ( $p = 17$ )	IBS ( $p = 517$ )	IBS ( $p = 5017$ )
SNNE	0.1217	0.1545	0.1898
SNNE- $L_1$	<b>0.1151</b>	<b>0.1310</b>	<b>0.1316</b>
RSF	0.1252	0.1550	0.1855
Baseline	0.2270	0.1956	0.2147

The out-of-bag IBS for all models and for the three datasets is given in Table 1. The SNNE yields a slightly lower IBS value than the RSF, but this

advantage is lost in the presence of uninformative variables. The SNNE- $L_1$  has the overall best performance. The baseline model performs notably worse than the other models due to batch training without validation.



**Fig. 2.** Survival stratification for A) **SNNE** model, B) **SNNE- $L_1$**  model, C) **RSF** model, D) **baseline** model (solid curve for low-risk group, dashed curve for mid-risk group, dotted curve for high-risk group)

To highlight the differences between models, we stratified the out-of-bag survival estimates (for the second version of the PBC dataset) into three groups based on the survival probability value at the time of the last uncensored event: patients in the upper quartile form the low-risk group, patients in the interquartile range form the mid-risk group, and patients in the lower quartile form the high-risk group. The groupwise survival curves obtained with each model are shown in Fig. 2. Despite having similar performance, the SNNE and RSF models have very noticeably different survival curves, with the RSF model having more pessimistic survival for the low-risk group and more optimistic survival for the high-risk group. The SNNE- $L_1$  model makes a compromise between SNNE and RSF for the survival of the low-risk group, whereas it predicts low survival for the high-risk group, like SNNE. The baseline model generates survival curves that clearly display the proportional hazards assumption, and its predictions show a trend similar to those of RSF: survival is pessimistic in the low-risk group and optimistic in the high-risk group.

Our results indicate that there is potential in using NNs for survival prediction based on the integrated Brier score. In particular, they allow penalization strategies via modifications of the loss function. We showed that this strategy is well suited to situations where few relevant predictors are expected.

## 4 Conclusion

In this paper, We have shown that an ensemble of NNs provides a valuable tool for survival prediction in high dimensional setting. The proposed strategy shows better predictive performance than survival random forests on the PBC dataset. The originality of the proposed model lies in its choice of loss function to train an NN ensemble with regularization. Future work will evaluate the interest of such approach in ultra-high dimensional genomics datasets.

## References

1. Schwarzer, G., Vach, W., Schumacher, M.: On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat. Med.* **19**(4), 541–561 (2000)
2. Luck, M., Sylvain, T., Cardinal, H., Lodi, A., Bengio, Y.: Deep learning for patient-specific kidney graft survival analysis. *CoRR abs/1705.10245* (2017)
3. Chapfuwa, P., et al.: Adversarial time-to-event modeling. In: *ICML (2018)*
4. Fotso, S.: Deep neural networks for survival analysis based on a multi-task framework. *CoRR abs/1801.05512* (2018)
5. Giunchiglia, E., Nemchenko, A., van der Schaar, M.: RNN-SURV: a deep recurrent model for survival analysis. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) *ICANN 2018. LNCS*, vol. 11141, pp. 23–32. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01424-7\\_3](https://doi.org/10.1007/978-3-030-01424-7_3)
6. Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018)
7. Manyam, R.B., Zhang, Y., Keeling, W.B., Binongo, J., Kayatta, M., Carter, S.: Deep learning approach for predicting 30 day readmissions after coronary artery bypass graft surgery. *CoRR abs/1812.00596* (2018)
8. Nezhad, M.Z., Sadati, N., Yang, K., Zhu, D.: A deep active survival analysis approach for precision treatment recommendations: application of prostate cancer. *Expert Syst. Appl.* **115**, 16–26 (2019)
9. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. *JAMA* **247**(18), 2543–2546 (1982)
10. Brier, G.W.: Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**(1), 1–3 (1950)
11. Ishwaran, H., Kogalur, U., Blackstone, E., Lauer, M.: Random survival forests. *Ann. Appl. Stat.* **2**(3), 841–860 (2008)
12. Therneau, T., Grambsch, P.: *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, New York (2000). <https://doi.org/10.1007/978-1-4757-3294-8>
13. Fleming, T.R., Harrington, D.P.: *Counting Processes and Survival Analysis*. Wiley, Hoboken (2005)

14. Broët, P., De Rycke, Y., Tubert-Bitter, P., Lellouch, J., Asselain, B., Moreau, T.: A semiparametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics* **57**(3), 844–852 (2001)
15. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. *CoRR* abs/1706.02515 (2017)
16. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18**(17–18), 2529–2545 (1999)