The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Multi-Task Survival Analysis of Liver Transplantation Using Deep Learning

Atefeh (Anna) Farzindar, Anirudh Kashi
Department of Computer Science
Viterbi School of Engineering
University of Southern California
Los Angeles, California 90089
farzinda@usc.edu, kashia@usc.edu

Abstract

In this paper, we present the application of deep learning techniques to develop a modern model for the prediction of graft failure and survival analysis in liver transplant patients. We trained our model using the United Network for Organ Sharing (UNOS) dataset consisting of 59,115 patients from year 2002 to 2016 with around 150 features each. We also compare our model against another dataset — Scientific Registry of Transplant Recipients (SRTR) including 87,334 patients from year 2002 to 2018 – after selecting features by mapping them from UNOS data. Some of the most important features common to both datasets are Model for End-stage Liver Disease (MELD) score, patient body mass index (BMI), donor and patient age, cold ischemia time, and levels of various chemicals within the patient. To provide an additional tool to clinical practitioners in the allocation of a scarce resource, we developed a multi-task model to learn the survival function of a donor-recipient pair and hence predict the exact time of failure which outperforms the traditional cox hazard models. The multi-task model produces very promising C-index results of 0.82 and 0.57 on the SRTR and UNOS datasets respectively.

Introduction

Orthotopic liver transplantation (OLT) is currently the last-resort for patients with severe liver cirrhosis due to Hepatitis B, Hepatitis C, alcoholism, or hepatocellular carcinoma. During the procedure, the host liver is removed, then replaced with the graft after the anhepatic stage. Currently, Organs are procured from deceased patients and can remain in cold ischemia with Viaspan for up to 24 hours before transplantation. A shortage of grafts for small children has resulted in the development of the split liver transplantation technique, which allows one graft to be utilized in two operations. However, complications are common, with rejection occurring in up to 50% of patients. Adult grafts are also in short supply: in terms of market size, approximately 14,000 patients are listed and waiting for an OLT procedure while only 7,000 OLTs are performed annually. About 3,000 patients died or experienced progression of their disease to a point where they were no longer viable candidates while awaiting transplantation in 2015 (Kim et al. 2017). Since 2002, the MELD (Model for End-Stage Liver Disease) system has been used to prioritize patients waiting for OLT. MELD score indicates increased hepatic dysfunction and mortality risk, with a higher MELD score corresponding to an increased risk of mortality without OLT. Because waitlisted patients are prioritized by MELD score, patients must often wait until they are very sick prior to being considered for liver transplantation. However, higher MELD and more severe liver disease is also associated with inferior post-transplant survival outcomes (Schlansky et al. 2014) (Schlegel et al. 2016). Thus, the current prioritization system sometimes leads to the use of precious, scarce organs in recipients that are too sick to tolerate OLT or benefit fully from the donated organ. Furthermore, transplant surgeons have limited ability to predict which patients will do poorly after OLT because the predictive models available have low positive predictive values for post-operative mortality. Given the uncertainty of outcome, it is difficult to deny patients lifesaving OLT. Consequently, many physicians rely primarily on clinical assessment of disease burden and resilience to determine which patients are appropriate OLT candidates.

Not only is there a shortage of liver donors in general, there is a paucity of optimal donors whose organs are more likely to result in a successful transplantation. Under guidelines set by UNOS, potential donors are evaluated based on a number of features. ABO blood type, height and weight are among the most important factors taken into consideration as donor-recipient matching is done primarily through matching blood type and organ size. Old age, donor-recipient sex mismatch, donation after cardiac death as opposed to the more common donation after brain death, and use of split grafts which are some donor characteristics that are believed to have potential negative effects on the outcome of the procedure. A prolonged cold ischemia time and ABO blood type mismatch are technical factors believed to negatively affect the procedure's outcome. The cold ischemia time is a concern in particular as it places a strict time constraint on organ procurement organization (OPO) representatives to evaluate and find a potential recipient for an organ. With the extreme shortage of donor organs and the even smaller supply of optimal donor organs, it is clear that OPO representatives need all possible information in order to make the best decision (Cotler 2015).

Deep learning techniques have great potential in predicting graft survival to better allocate donor livers but there are still many challenges to be tackled. In (Yu et al. 2017) the

Multilayer Perceptron Neural Network was applied to estimate missing values and predict the degree of post-operative anemia status. In 2016, Lau et al. applied machine learning models to predict graft failure or primary non-function within specific short periods of time after the transplant procedure in order to aid the organ allocation process. The models were tested and used to determine the characteristics that contributed most in the prediction task. Also in 2016, (Raji and Chandra 2016) used a Multilayer Perceptron model to predict the mortality rate of liver transplantation patients up to 3 months after transplantation with good accuracy but poor precision and recall, possibly due to training on imbalanced datasets. In 2017, Luck et al. studied kidney graft predictive model with a Concordance-index of 0.655, higher than state of the art traditional Cox model using Efron's method.

The objective of this paper was to develop a prediction model for the survival rate of a patient post liver transplantation to help support the clinical decision on how to best allocate available donor livers to the proper recipients. In our previous study, various machine learning techniques were tested for effectiveness, with the best models being further developed to comprise the prediction model. Initially, we built models, using Random Forest and Deep Neural Network, to predict an expanded graft failure prediction range of 3 months, 6 months, 1 year and 3 years and more comprehensive performance metrics unaffected by data skewness (Farzindar et al. 2019). This approach has limited ability to provide significant results in many cases because the different models for every duration are not mutually exclusive. This means that if graft failure has been predicted in 3 months models, it does not imply that no graft failure is predicted in 1 year model and so on. The shortcomings of this initial attempt guided us into using deep survival analysis models based on the principle of multi-task learning. In this paper we present our research to develop a single model to predict the survival of the graft for any given time in the future. We perform experiments with both UNOS and SRTR datasets. Our contribution will allow physicians to compare recipients more accurately, especially in cases when the patients have similar MELD scores.

Dataset

The initial analysis for the model was done using UNOS dataset, a tax-exempt, medical, scientific and educational organization which controls the national Organ Procurement and Transplantation Network under agreement to the Division of Organ Transplantation of the Department of Health and Human Services. Since this study is based on MELD score introduced in 2002, we consider records only after this year. The data collected is a multi-organ dataset containing 59,115 patients from year 2002 to 2016.

We also performed analysis using another common organ dataset – Scientific Registry of Transplant Recipients (SRTR) including 87,334 patients from year 2002 to 2018.

Pre-Processing

We extracted the UNOS dataset section on liver transplant, which consists of 263219 rows of donor, recipient and transplant information. This dataset is highly imbalanced, containing many more entries of patients whose grafts survived. To prepare the data for analysis, we removed all records without MELD score or from living donors. We also performed analysis using SRTR dataset including 87,334 patients from year 2002 to 2018.

To handle missing data, we imputed the categorical features by mode and numerical features by mean. Rows with more than a threshold of 20% missing data were removed from the analysis. To identify inconsistency in the data, we compared categorical values with all possible values from the category and replaced them with null if any of the values were not part of their category set. Finally, we removed post-transplant features and other unimportant features as decided by feature importance in a Random Forest model, by selecting the top few features, after which the importance fell drastically.

Feature name in UNOS Data	Feature name in SRTR Data
END_STAT	REC_PX_STAT
AGE_DON	DON_AGE
ABO	CAN_ABO
ABO_DON	DON_ABO
ALBUMIN_TX	REC_PRETX_ALBUMIN
INIT_AGE	CAN_AGE_AT_LISTING
DON_TY	DON_TY
FINAL_INR	CAN_LAST_INR
DEATH_CIRCUM_DON	DON_DEATH_CIRCUM
AGE	REC_AGE_AT_TX

Table 1: An excerpt of UNOS to SRTR feature mapping

UNOS to SRTR feature mapping

We manually did the feature mapping for SRTR dataset and were able to use the same features which we used earlier for our prediction using UNOS for the fair comparison between results generated by machine learning models. Table 1 depicts some of the feature mappings between both the datasets. After our pre-processing stage for SRTR data, we were able to infer that the dataset had a lot less number of gaps in all the features compared to UNOS dataset.

Survival Analysis

Survival analysis is the measure of time to an event. The event can be anything defined by the user and explicitly available in the data. Here, the event we are measuring is time for liver-graft failure. Newly transplanted liver is more susceptible to infections and sometimes can be rejected by the body. Things that determine these are called covariates. Survival analysis handles complications in the data effectively. Things like incomplete data can exist due to subjects dropping out from the clinical trial even before the trial ends, or the event not happening during the entire course of the trial. This is called *censoring*. Ignoring these kinds of data would cause generalization bias while testing. The probability of the event happening could be less and we need to capture that percentage with respect to the entire population under study.

Year	Patient count in 12-23 MELD Range
2002	2375
2003	2815
2004	3061
2005	3092
2006	3264
2007	3033
2008	2858
2009	2725
2010	2546
2011	2433
2012	2449
2013	2335
2014	2446
2015	2603
2016	1963

Table 2: Number of people in 12-23 MELD range every year

Motivation

After superseding Child-Turcotte-Pugh (calculating the severity of cirrhosis) score, MELD has become the de-facto metric for organ allocation for liver transplantation. MELD has been successful in lowering the importance of waiting time and placing more weight in liver disease severity. MELD score has also been demonstrated to be a good predictor of three-month mortality for patients awaiting liver transplantation (Bambha and Kamath 2013). However, since the 3-month mortality values are associated with MELD scores, having a small range, it causes ties to be frequent as seen above (Wiesner et al. 2003). Patients who fall into the same MELD range with the same blood type would be prioritized by waiting time (Bambha and Kamath 2013). In Table 2, we have shown the number of people who fall into the 12-23 MELD range for each year from 2002 to 2016. This range was chosen because the hazard ratio is statistically lower and most transplants occurred during this time period. MELD score at 15 represents a transition point, where the comparative hazard of undergoing transplant versus remaining on the waiting list, drops significantly. Beyond 18, substantial and progressively higher survival benefit was shown (Merion et al. 2005). It was also shown that since the introduction of MELD in the period February 27, 2002 to February 26, 2003, the average MELD score at time of transplantation is 24.

From the statistics, you can see that there are many ties within the 12-23 range. This means waiting time and other factors are still necessary for the allocation of livers. Currently, there are also many special severity conditions that are not reflected by MELD but do justify expedited liver transplants. These are called MELD exceptions and require either manual increments to MELD scores or petitioning (Bambha and Kamath 2013).

With this goal in mind, we intend to design a multi-task deep learning model for analyzing patient-specific liver graft. This model predicts both the time of graft failure and its rank in cox partial log-likelihood framework. Our model's first output, for ranking patients by survival time, can be used in prioritizing patients that fall in the same operational MELD range. The second output, for estimating graft survival times, can be used to perform survival analysis i.e to get the exact time of when the liver might fail, to help surgeons make more informed decisions.

Survival Analysis with Continuous Time

Let T be the failure or censored time. Failure time is if the event happens during the study and censored time is if the event does not happen during the study. The Survival function S(t) can be defined as the probability of event happening after t.

$$S(t) = \Pr(T \geq t)$$

A more important metric is the Hazard function λ(t). Hazard function defines the probability of the event happening at an instant in time, given that the event did not occur before.

$$\lambda(t) = \lim_{dt \to \infty} \frac{\Pr(t \leq T \leq t + dt | T \geq t)}{dt}$$

This Hazard value which indicates risk can be used to rank donor-recipient combinations.

Cox Proportional Hazards Model

The standard method for survival analysis before the popularity of Deep Learning was through Cox models. The most well known among these kinds of models is the Cox Proportional Model which makes the assumption that the rate of risk is constant throughout the study. It estimates the Risk function h(x) in the following way:

$$h(x) = e^{\beta x}$$

where β is the parameter vector and x is the feature vector.

This risk function is optimized by minimizing the cox partial likelihood loss function:

$$L(\beta) = \prod_{i:E_i=1} \frac{exp(h(x_i))}{\sum_{j \epsilon R(T_i)} exp(h(x_j))}$$

where T_i is the event time, E_i is a binary value indicating whether the event happened or not and $R(T_i)$ is the risk set $\{i : T_i \geq t\}$ indicating all the recipients who are still at risk at time t. As number of data points increases, the possibility of multiple events happening at the same time increases. These are called *tied* events. In the below sections we define a modified cox partial likelihood function, with Effon's approximation (Menon et al. 2012) similar to (Luck et al. 2017).

Deep Survival Model

In Deep Survival Model, Deep Learning techniques are used directly to learn the hazard function. These models overcome many of the restrictions of cox models like the proportionality assumption. In this paper, we analyze multi-task methods of achieving our desired goal of improving donor-recipient selections for transplantation. Deep Survival Analysis is one of the ways which efficiently captures the

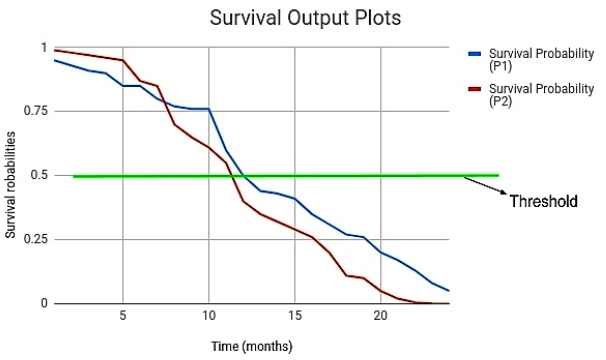

Figure 1: Survival values of two donor-recipient pairs

time-line of clinical trial, and helps rank recipients for every donor. Below sections explain details on the model and optimization loss functions used.

Figure 1 is a plot of the probability values of two patients, which are the outputs (second outputs) of two donor-recipient inputs. We can clearly say that, after the threshold, survival probability of P1 is higher than that of P2, so P1 is a better candidate over P2. So, our model helps to make these kinds of decisions by predicting the survival timelines.

Model Details

The model we built is a five layered network with three hidden layers: 512, 256 and 64 respectively, one input layer and two output layers as shown in Figure 2. The model is a multi-task Deep Neural Network with the following as outputs:

- A single output proportional to Hazard value, trained by implicitly ranking it before all data points which have their events happening after the current point.
- Multiple sigmoid units trained using isotonic regression to predict the probability of graft failure at time t, where $t \in [0, T]$. The granularity of the time-line (T) is decided based on the data and user's requirements.

The hidden layers comprises of ReLU activation, batch normalization and dropout. There is no activation function for the first output unit and sigmoid activation is used for the second output units.

Loss Functions

The first loss function is a cox partial likelihood loss combined with an Effron's approximation to handle ties. Let Y_i be the observed time (either censored time or event time) of patient i. C_i is 1 if event is non-censored, 0 if event is censored.

$$l_1(s^{(1)}) = \sum_j \left(\sum_{i \epsilon H_j} \log s_j^{(1)} - \sum_{l=0}^{m-1} \log \left(\sum_{i:Y_i \geq t_j} s_i^{(1)} - \frac{l}{m} \sum_{i \epsilon H_j} s_i^{(1)} \right) \right)$$

where t_j denotes unique times, H_j the set of indices i such that $Y_i = t_j$ and $C_i = 1$ and $n_j = |H_j|$. m is the number

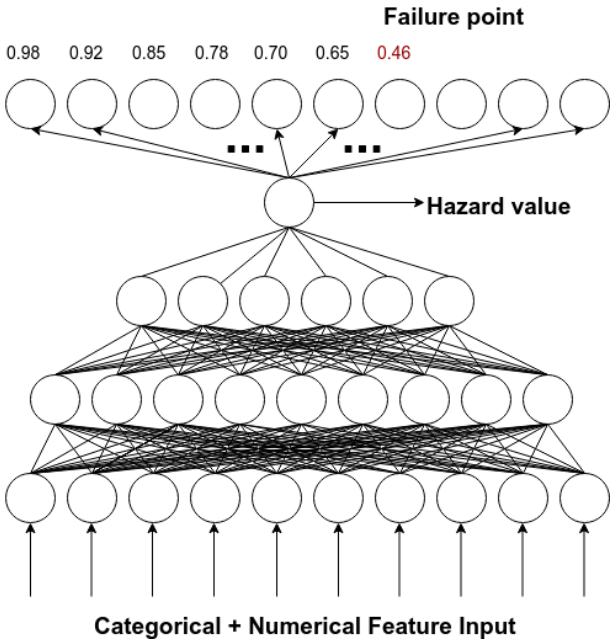

Figure 2: Compressed model structure

of tied events at t_j. Note that in the case of the Cox model, $s^{(1)} = e^{\theta.X}$

Second loss function is a combination of isotonic regression and ranking loss as derived in(Menon et al. 2012), modified as in(Luck et al. 2017) to handle censored data.

$$l_2(w) = \sum_{Acc(i,j)} l\Big(s^{(2)}(x_j; w) - s^{(2)}(x_i; w), 1\Big)\, y_i(1 - y_j)$$

where $Acc(i, j)$ are the set of acceptable pairs, that is i is not censored and at the time of $i's$ event, j is not censored. $l(.,.)$ is some convex loss function, here l_2 distance and $s_{(2)}$ is the output from the second layer.

The result of the first output layer is just a value, proportional to the Hazard value. We can infer the ranking of patients by comparing their Hazard values. Lesser the Hazard value, lesser the risk the patient is in and hence will be positioned lower in the rank. Output of the second layer on the other hand identifies meaningful result from the Hazard value and provides time-lines of when exactly the event might happen.

Implementation

We run our experiments on USC's High Performance Computing servers. The compute system has 12GB of Tesla K40 GPU memory. The HPC system has secure servers to hold sensitive medical data, which can be accessed and used only by certified members. Our pre-processing is a separate module which takes about 10 minutes to run and it generates three files - train, validation and test with 70%, 15% and 15% as the split. The model takes about 4 - 5 hours to train and the hyperparameters are tuned using validation set. We use random search to set the hyperparameters. These are the ranges of hyperparameters that we tried:

Dataset	Method	C-index
UNOS	Single Loss function	0.62
	Double loss function	0.57
SRTR	Single Loss function	0.76
	Double loss function	**0.82**

Table 3: C-index obtained for the two models tested on UNOS and SRTR datasets

1. 2 to 5 hidden layers
2. Learning rate from $1e-6$ to 0.01
3. Empirical weights ranging from 0.3 to 0.6 to combine the loss functions.
4. Activation functions: tanh, relu

Evaluation and Results

Metric: C-index

For Survival Analysis we use a metric called C-index (Concordance index) (Harrell Jr et al. 1982). It measures how good the ranking system is by finding the probability of correctly ordered pairs. For example, if $(T_1, E_1), (T_2, E_2), ..., (T_j, E_j)$ are the event times and occurrences in our dataset, C-index starts by counting the number of pairs which are correctly ordered by the model. C-index is the ratio of this value by the total number of admissible pairs. A pair $(T_i, E_i), (T_j, E_j)$ are considered admissible if i is not censored and during $i's$ event, j is still under risk.

Results

Table 3 contains the results for both UNOS and SRTR datasets. Single loss function corresponds to using just cox-partial likelihood and Double loss function corresponds to using both losses and hence jointly learning likelihood and ranks. The C-index values for survival analysis are acceptable if it is in between 0.6 and 0.7 and excellent above 0.7. C-index of 0.5 is considered random (Luck et al. 2017).

Intuitively, the double loss function should be more effective than the single loss function since the training is more constrained. The results of the two loss functions should be coherent. We believe the bad UNOS results are due to missing and inconsistent data. SRTR is a much clearer dataset and we get better results.

Conclusion

In this paper, we presented our study on Multi-task learning for developing a deep learning method that directly models the survival analysis function to predict survival times for graft in liver transplant patients. In this research, we compared the results using c-index metrics for two methods: Single loss function and Double loss function. We used the two dataset of UNOS and SRTR where the features were mapped.

This study on survival analysis results in improved learning efficiency and prediction accuracy for the graft futility in liver transplantation patients, when comparing to our previous works for training the models separately.

The predictive and modeling capabilities of our multi-task Survival analysis will enable medical team to use deep neural networks as a valuable tool in their clinical decisions related to allocating organs.

Acknowledgment

The authors would like to thank Joan Brown, Manas Bhatnagar and Dr. Juliet Emamaullee our D-Health partner at the Department of Surgery of the Keck School of Medicine of the University of Southern California (USC) for their guidance that assisted the research. We would also like to expand our gratitude to Integrated Media Systems Center (IMSC) and USC's Center for High-Performance Computing (HPC) for their valuable support.

References

Bambha, K., and Kamath, P. S. 2013. Model for end-stage liver disease (meld). *UpToDate. Waltham, MA: UpToDate*.

Cotler, S. J. 2015. Liver transplantation: donor selection.

Farzindar, A.; Kashi, A.; Bhagwat, P.; and Umate, P. 2019. A deep learning-based multi-model ensemble method for prediction of graft futility in liver transplantation patients. In *5th International Conference on Communication, Management and Information Technology (ICCMIT 2019)*.

Harrell Jr, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; Rosati, R. A.; et al. 1982. Evaluating the yield of medical tests. *Jama* 247(18):2543–2546.

Kim, W.; Lake, J.; Smith, J.; Skeans, M.; Schladt, D.; Edwards, E.; Harper, A.; Wainright, J.; Snyder, J.; Israni, A.; et al. 2017. Optn/srtr 2015 annual data report: liver. *American Journal of Transplantation* 17:174–251.

Lau, L.; Kankanige, Y.; Rubinstein, B.; Jones, R.; Christophi, C.; Muralidharan, V.; and Bailey, J. 2017. Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation* 101(4):e125–e132.

Luck, M.; Sylvain, T.; Cardinal, H.; Lodi, A.; and Bengio, Y. 2017. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*.

Menon, A. K.; Jiang, X. J.; Vembu, S.; Elkan, C.; and Ohno-Machado, L. 2012. Predicting accurate probabilities with a ranking loss. In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, 703. NIH Public Access.

Merion, R. M.; Schaubel, D. E.; Dykstra, D. M.; Freeman, R. B.; Port, F. K.; and Wolfe, R. A. 2005. The survival benefit of liver transplantation. *American Journal of Transplantation* 5(2):307–313.

Raji, C., and Chandra, S. V. 2016. Artificial neural networks in prediction of patient survival after liver transplantation. *J. Health. Med. Inform* 7:1.

Schlansky, B.; Chen, Y.; Scott, D.; Austin, D.; and Naugler, W. 2014. Waiting time predicts survival after liver transplantation for hepatocellular carcinoma: a cohort study using the

united network for organ sharing registry. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society* 20(9):1045–1056.

Schlegel, A.; Linecker, M.; Kron, P.; Györi, G.; Oliveira, M. D.; Müllhaupt, B.; Clavien, P.; and Dutkowski, P. 2016. Risk assessment in high- and low-meld liver transplantation. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 17(4):1050–1063.

Wiesner, R.; Edwards, E.; Freeman, R.; Harper, A.; Kim, R.; Kamath, P.; Kremers, W.; Lake, J.; Howard, T.; Merion, R. M.; et al. 2003. Model for end-stage liver disease (meld) and allocation of donor livers. *Gastroenterology* 124(1):91–96.

Yu, C.-H.; Bhatnagar, M.; Hogen, R.; Mao, D.; Farzindar, A.; and Dhanireddy, K. 2017. Anemic status prediction using multilayer perceptron neural network model. In *GCAI*, 213–220.