5-2020

A Machine Learning Method for Predicting Liver Transplant Survival Outcomes

Brandon C. Revels

The University of Southern Mississippi

A Machine Learning Method for Predicting Liver Transplant Survival Outcomes

by

Brandon Revels

A Thesis
Submitted to the Honors College of
The University of Southern Mississippi
in Partial Fulfillment
of Honors Requirements

May 2020

Approved by:

_____

Chaoyang Zhang, Ph.D., Thesis Adviser
Professor of Computer Science

_____

Andrew Sung, Ph.D., Director
School of Computing Sciences and
Computer Engineering

_____

Ellen Weinauer, Ph.D., Dean
Honors College

Abstract

For years, doctors have utilized the Model for End-stage Liver Disease (MELD) score to aid in the allocation of organs for liver transplants (LT). A major issue with using the MELD score to allocate organs for transplantation is that the MELD score does not accurately predict post-transplant survival. This research project aims to investigate the use of machine learning (ML) methods to predict LT survival using the newer Scientific Registry of Transplant Recipients (SRTR) dataset. For this project, death and nonfatal graft failure were treated equally as both cases result in a loss of a donated organ. The ML algorithms used in this project were provided by both the Weka and Orange software packages. Initial trials investigated a binary classification of patients based on whether they survived for three years post-transplant and primarily utilized a random forest algorithm. Later trials moved to a multi-class classification using both random forest and other classifier algorithms. Initial results from the three-year binary classification seemed promising but performance metrics failed to improve with continued work. All multi-class trials performed similarly using various classifier algorithms. Unexpectedly, the class for 12-year survival showed a promising increase in its area under the receiver operating characteristic curve. The results of this project help to create a baseline for future ML studies utilizing the SRTR dataset and will hopefully spur further research into liver transplant survival prediction.

Keywords: machine learning, random forest, liver transplantation, Scientific Registry of Transplant Recipients, survival prediction, medical data

Dedication

I would like to dedicate this work to my friends and family for all their support throughout my research, to my grandfathers Charles Jackson and Mickey Revels Sr., and in memory of Barbara Revels and Patricia Jackson.

Acknowledgements

I would like to thank Dr. Chaoyang Zhang, Dr. Jacob Feagans, Joseph Luttrell, and Meng Song for their assistance and guidance throughout my research. Without them this project would have likely never come to fruition, and I thank each one of them greatly. I would also like to thank Mr. Donald Drapeau for his generous support through the Donald Drapeau Summer Research grant which helped to fund this research project.

Table of Contents

List of Tables

List of Illustrations

List of Abbreviations

ANN	Artificial Neural Networks
AUC	Area Under ROC Curve
INR	International Normalized Ratio
LR	Logistic Regression
LT	Liver Transplant(ation)
MELD	Model for End-Stage Liver Disease
ML	Machine Learning
MLP	Multi-Layer Perceptron
P-SOFT	Preallocation Score to Predict Survival Following Liver Transplantation
PSSP	Patient-Specific Survival Prediction
ROC	Receiver Operating Characteristic
SOFT	Survival Outcomes Following Liver Transplant
SRTR	Scientific Registry of Transplant Recipients
TRR_ID	SRTR Transplant ID
TX	Transplant
TXF	Transplant Follow-up
WEKA	Waikato Environment for Knowledge Analysis

Chapter 1: Introduction

Since the Model for End-Stage Liver Disease (MELD) score was contrived as a method of prioritizing liver transplant recipients, its use as a predictor of liver transplant outcomes has been debated. Studies investigating the merit of liver transplant (LT) survival predictions using MELD scores are often conflicting and nearly always highlight the need for a better predictive model. [1], [2] Using the data from the Scientific Registry of Transplant Recipient (SRTR) database, it is trivial to check for correlation between intake MELD scores, pre-transplant MELD scores, and the number of years survived after transplant (TX). As seen by the Pearson correlation matrix in *Table 1*, the initial MELD scores and pre-TX MELD scores correlate strongly while neither MELD score shows a correlation with the number of years survived post-TX.

Table 1: Pearson Correlation Matrix

	Initial MELD Score	**Pre-TX MELD Score**	**Years Survived Post-TX**
Initial MELD Score	1	0.896703	-0.046088
Pre-TX MELD Score	0.896703	1	-0.016356
Years Survived Post-TX	-0.046088	-0.016356	1

While there have been advances in the use of pre-TX patient data to predict short-term LT survival, there is still a need for a method capable of producing more long-term LT survival predictions. [3] The goal of this project is to apply machine learning (ML) methods to clinically obtainable donor and patient data to generate an accurate model to predict post-LT survival. At the time of writing this paper, only a handful of other studies have applied ML methods to liver transplant survival predictions. Most studies investigating transplant survival prediction have focused on utilizing artificial neural networks (ANN) to make predictions on transplant outcomes. [4]–[6] This project utilizes

random forests as a baseline ML method and attempts to utilize other ML methods based on the initial random forest results. The use of ML techniques in transplant survival prediction is just beginning to grow beyond its infancy and provides this project with a strong opportunity to expand upon current findings.

Chapter 2: Literature Review

The use of machine learning in medicine is by no means a new development nor is the need for a method of accurately predicting post-LT survival. Before discussing the specifics of this project, it is important to look at the resources and methods currently available pertaining to LT survival prediction. There have been a few attempts at using machine learning to predict long-term LT survival at time spans greater than 3 months post-TX; however, these studies typically focus on a particular liver disorder rather than TX recipient survival of all LT types. [7], [8] Other studies that investigate more generalized LT survival prediction use either the smaller United Network for Organ Sharing (UNOS) dataset or a locally obtained dataset. [5], [6]

A. MELD and Other Conventional LT Survival Predictors

Prior to the creation of the MELD score the Child-Pugh score, blood type compatibility, and overall wait time were used to prioritize LT candidates on the LT waitlist. [9] However, a study by Michael Malinchoc et al. found that the Child-Pugh score was not a good estimator of a patient's 3-month waitlist mortality, and in turn the MELD score was developed as a far superior waitlist mortality prediction model. [10] The original version of the MELD score accepted by UNOS uses only creatinine, bilirubin, and international normalized ratio (INR) in its calculation.

$$\text{MELD score} = 9.57 * \log_e(\text{creatinine mg/dL}) + 3.78 * \log_e(\text{bilirubin mg/dL}) + 11.20 * \log_e(\text{INR}) + 6.43 \quad (1)$$

The development of the MELD score led to better allocation of donor livers to LT candidates that would receive the most immediate benefit from transplantation. One of the largest flaws in the previous allocation system that was addressed by the creation of

the MELD score was the bias given to candidates with longer wait times. In the old system, wait time was directly factored into the decision to allocate an organ for LT and was primarily used as a tiebreaker between patients of the same score. [9] Therefore, waitlist time is purposefully excluded from the calculation of the MELD score.

Shortly after the acceptance of the MELD score for clinical use, researchers began investigating if the MELD score would prove equally as useful for post-TX survival prediction as it was for predicting pre-TX waitlist mortality. A 2003 study by Paul Hayashi et al. found that the pre-TX MELD scores of LT recipients did not have any correlation with 1-year or 2-year survival post-TX. [2] The results of that study are reinforced by the data presented in Table 1 which shows no correlation between MELD and post-TX survival at 1-year and beyond. In an analysis of using only the MELD score for 3-month post-TX survival, the area under the ROC curve (AUC)[1] was calculated at only 0.54. [3] Furthermore, both the previous Child-Pugh scoring system and the more recently developed donor risk index performed similarly poorly at post-TX survival prediction. [3]

In 2008, a study by Rana et al. contrived the survival outcomes following liver transplant (SOFT) score. [3] In their study, the SOFT score was found to be able to more accurately predict the 3-month post-TX survival of LT recipients. However, the SOFT score is only able to be calculated once an allograft has been allocated to the LT recipient. To account for this, a variant of the SOFT score called the preallocation score to predict survival following liver transplantation (P-SOFT) was created. The P-SOFT can be applied while a LT candidate is still on the waitlist since it excludes donor risk

[1] Note: From this point and forward, sources which state their performance metric as c-statistic are instead referred to as AUC in this paper as the two terms are generally interchangeable. See [11] for further details.

factors from the score calculation. When used to predict 3-month post-TX survival, the SOFT score was found to have an AUC of 0.70, and the P-SOFT was found to have an AUC of 0.69. [3] This makes both the SOFT score and P-SOFT vastly superior at predicting post-TX survival than using MELD alone.

B. Existing ML Methods for LT Survival Prediction

One of the earliest studies on using ML to predict LT survival was conducted by Vicente Ibáñez et al. in 2008. This study utilized logistic regression (LR) and a multi-layer perceptron (MLP) network to predict early LT failure at 90 days. The study used locally obtained data from the Liver Surgery and Transplant Unit of La Fe University Hospital in Valencia, Spain rather than the more common UNOS or SRTR datasets. The local dataset used in their study contained 701 patients that met their inclusion criteria, and for each LT record 19 features were considered. In the study's evaluation of both the LR model and MLP network on validation sets, their models showed that the difference between the two AUCs were not statistically different at 0.78 and 0.81 respectively. However, increasing the number of patients in their validation cohort from 170 to 246 brought the AUC to 0.69 for the MLP network and 0.68 for the LR model.[6]

Another study of interest was conducted in 2017, by Raji and Chandra. Their study focuses on using an MLP ANN for long-term LT survival prediction and utilizes the UNOS dataset which contained 65535 records and 389 attributes at the time of their paper's publication. Out of those records, Raji and Chandra selected 383 patient records and 27 attributes based on their filtering criteria. Their model was trained on various LT survival time spans ranging from 6 months to 13 years. While a table of exact AUC values is not given in their paper, a figure plotting the AUC for the various time periods

is included. Based on that figure, the average AUC for their ANN is approximately 0.92 between 0.5 to 10 years. For years 11, 12, and 13 the graph shows a steady decrease in AUC to 0.75, 0.55, and 0.45 respectively. Their study is also particularly interesting in that for some year values their ANN yields a near perfect accuracy. [5] However, those results may simply be an artifact of the small dataset used in their study.

Lastly, a more recent study conducted by Andres et al. in 2018, utilizes ML to predict survival after LT for primary sclerosing cholangitis patients. The study by Andres et al. is different from those previously mentioned in the fact that it is the first to utilize the SRTR dataset. Furthermore, the study uses a learning algorithm called Patient-Specific Survival Prediction (PSSP) which consists of several LR functions over a set of time points. According to their specifications, the algorithm will always produce a survival probability that decreases monotonically over time for each patient. The PSSP model was trained to provide probabilities of survival at times ranging from <1 year to 11 years. Their PSSP model used a dataset of 2769 eligible records with 5-fold cross validation for training and evaluation. [8] The authors decided to use a Hosmer-Lemeshow test as a validation metric, so the performance of the PSSP model cannot be directly compared to the previously discussed methods. Regardless, the PSSP model scored highest on the 10-year time point at scores of 0.678 for the PSSP model with donor information and 0.409 for the PSSP model without donor information on their Hosmer-Lemeshow test (values closer to 1.0 represent better calibration). [8]

Chapter 3. Materials and Methods

A. Dataset Description

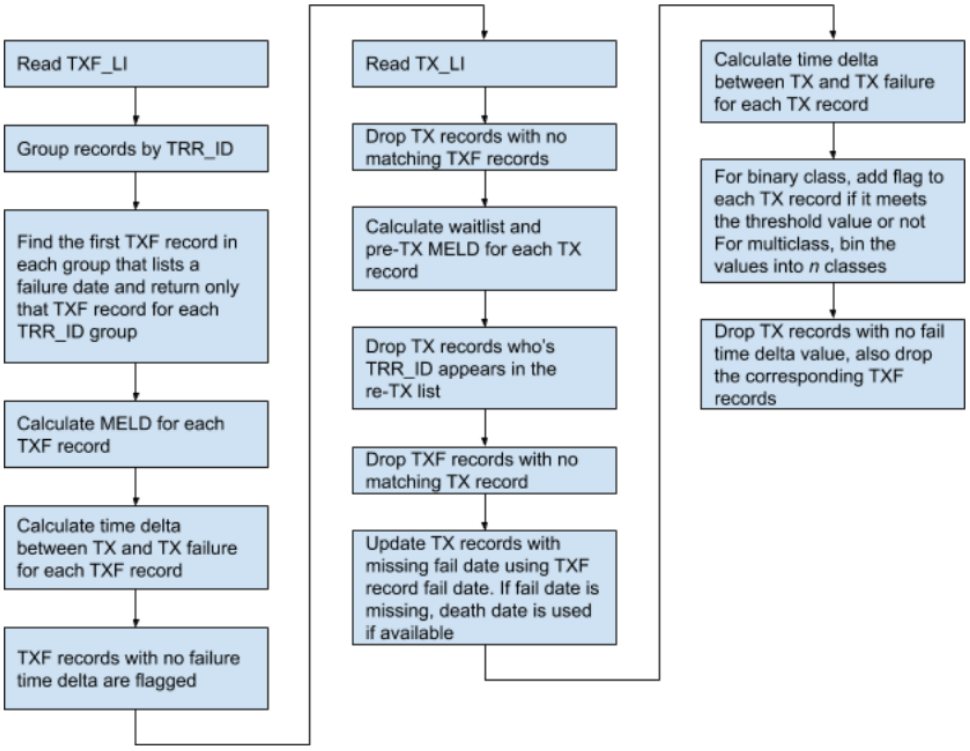
For this study, data was sourced from the SRTR database. The SRTR database is operated by The Chronic Disease Research Group, which is a division of the Hennepin Healthcare Research Institute. [12] The SRTR database contains several datasets for various organ transplant recipients, candidates, and donors. The tables of interest to this project are the TX_LI and TXF_LI tables. TX_LI is the primary dataset used in this project and contains one record per transplant as well as summarized organ donor and transplant follow-up (TXF) information. [13] There are currently 163728 entries with 309 attributes in the unprocessed TX_LI dataset. Out of the TX_LI records there are 137133 unique LT recipients. The TXF_LI dataset is the secondary dataset used in this project and contains the complete TXF records collected at 6 months, 1 year, and then annually until death/TX-failure or the patient ceases to follow-up. [13] The unprocessed TXF_LI dataset contains 1178862 records with 97 attributes. Of the records in TXF_LI, there are 143211 unique transplants recorded based on SRTR transplant IDs (TRR_ID). There appears to be a discrepancy of 20517 TX records with no matching TXF records, and those TX records are excluded from the model. Out of the 137133 unique LT recipients, 89790 did not experience graft failure or death and were also excluded from the model. These records were excluded because the event of interest has either not yet occurred in the LT recipient or could be missing due to a lack of record keeping. Therefore, only patient records which include a failure date can provide reliable LT survival data. Records from the TX_LI and TXF_LI tables are linked by their TRR_IDs.

B. Data Preprocessing

In order to achieve a better performance from the ML models, the data must be preprocessed before training. The following steps were taken to process the SRTR dataset, and a visual summary of this process can be found in Illustration 1. All data is first imported from the TX_LI and TXF_LI datasets. All entries in the TXF_LI dataset are then grouped by their TRR_IDs. In each group, the first entry with a date of TX failure is found and the failure date is copied to all entries in the group. Once the date has been copied, the earliest TXF record in each group is returned. The returned records are then merged back into one table for further processing. For each record MELD scores are calculated based on the given bilirubin, INR, and creatinine using the standard MELD score equation (1). The MELD score is recalculated here despite a MELD score attribute existing in the dataset due to the fact that the existing MELD score attribute contains incompatible data such as the deprecated Child-Pugh score. The next value that is calculated is the time delta between the date of LT and date of LT failure for each patient. Records with missing dates or invalid times are flagged. Next, entries in the TX_LI table without corresponding follow-ups in the TXF_LI table are removed. As with the TXF_LI dataset, TX_LI record waitlist and pre-TX MELD scores are recalculated with the data from each record. Since this project is only interested in predicting the survival of first-time LT recipients, all TX_LI records with TRR_IDs in the re-TX list are removed from the dataset. Furthermore, all entries in the TXF table without a matching TX entry are once again dropped. In order to ensure that both deaths and graft failures were included in the dataset TX_LI records with missing failure dates are updated with the failure date from the corresponding TXF_LI entry. This is accomplished by attempting to update the

value with the TXF graft failure date first and then the TXF death date second, whichever the first available is. If no values are available, then the date remains NaN and the record will be removed later. Next, the time delta between LT and LT failure is calculated again for the TX_LI records. For the binary class models, each record is flagged according to whether the time delta value surpasses a given threshold or not. Likewise, for the multiclass version of the dataset each record is labelled according to what bin the time delta value falls into. Lastly, any TX_LI records which do not have a valid time delta value are dropped from the TX_LI dataset, and the corresponding record in the TXF_LI dataset is also dropped.

Illustration 1: Preprocessing Stage I



After the previous steps, the dataset is nearly ready for use. However, there are still a few adjustments needed to ensure improved ML model performance. A

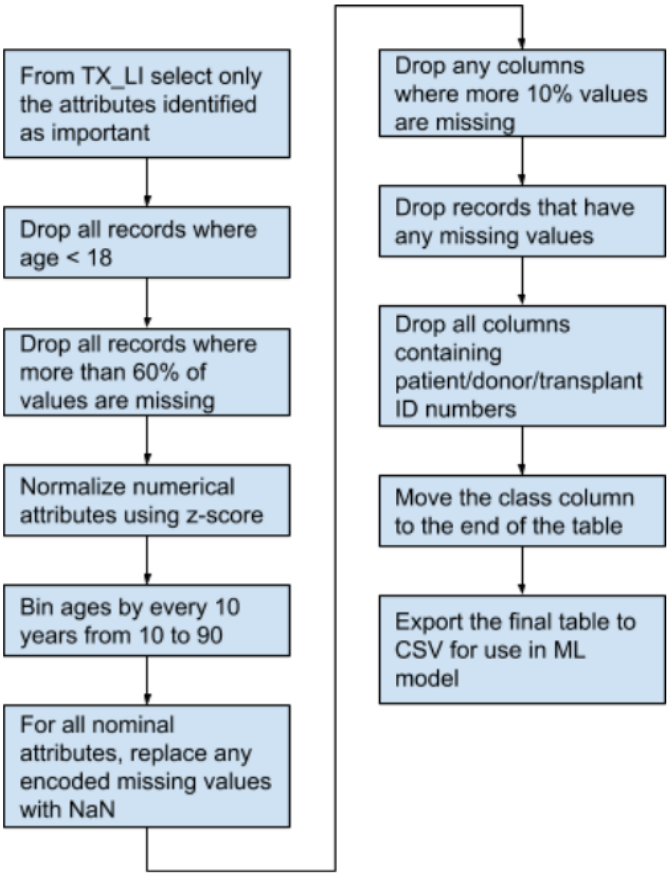
summarized version of this process can be found in Illustration 2. First, a subset of 96 attributes are chosen based on professional input from the original 309 TX_LI attributes. Next, all records with LT recipients under 18 are dropped along with any records that are missing > 60% of all values. Some ML algorithms require numerical values to be normalized during preprocessing; for this project, z-score normalization (2) was used on all numeric data aside from ages.

$$z - score = \frac{x - avg(attrib)}{stdev(attrib)} \quad (2)$$

Ages were instead classified into bins ranging from 10 to 90 years. A significant portion of the selected attributes are nominal, and most of the nominal attributes utilize encoded missing values. The encoded missing values artificially increases the completeness of the dataset, and to correct this all encoded missing values were replaced with NaN values. In order to reduce the dimensionality and sparsity of the dataset further, attributes which were missing more than 10% of values across all records were dropped. Next, records with any missing values were dropped resulting in a final dataset containing 24763 LT records. The last items dropped from the dataset before beginning the ML training were all attributes containing identification numbers. The final number of attributes used for training the models totaled to 69, including the class attribute. The first binary class dataset was based on 3-year survival, and the latter binary class dataset was based on 12-year survival. Three years was chosen initially as it was the point in the dataset which evenly divided the number of class members; however, after multiclass trials 12-years was found to provide superior model performance. The multiclass dataset utilized bins from 0 to 4 months, 4 to 8 months, 8 months to 1 year, 1 year to 2 years, 2 to 4 years, 4 to

6 years, 6 to 8 years, 8 to 12 years, and 12+ years right exclusive. The intervals of the multiclass bins were chosen to keep the number of class members as even as possible.

Illustration 2: Preprocessing Stage II



C. Machine Learning Software and Algorithms

The machine learning algorithms used in this project were provided by the Waikato Environment for Knowledge Analysis (WEKA) 3 and Orange software packages. [14], [15] The main algorithm used in this project is the random forest classifier algorithm provided by both software packages. The random forest algorithm was chosen as it proved to be the most performant in initial trials using the processed dataset. Other algorithms utilized from the WEKA software include the naive Bayes classifier and the MLP classifier. The MLP classifier algorithm was chosen since it has

previously demonstrated favorable performance in LT survival prediction in other studies. [4], [5] Furthermore, the naive Bayes classifier was chosen as it also demonstrated performance similar to random forest in some initial trials. The same three algorithms used in WEKA were also utilized in the Orange package, and each algorithm did not show a significant difference in performance between the software packages. Trials carried out in WEKA utilizing the random forest algorithm use custom settings to calculate attribute importance and extend the number of iterations to 200. All other trials in WEKA and Orange use default settings for all algorithms. Unless otherwise noted, all trials also used 5-fold cross validation in the training and evaluation of the model. Performance statistics are taken from both WEKA and Orange, whichever showed superior performance of the model being tested. Illustrations 3 and 4 were generated using the Orange software.

Chapter 4: Results

The AUC of a model represents a general performance across classification thresholds where values closer to 1.0 indicate a perfect model, and values closer to 0.5 indicate a model that is no better than random chance. The F1 score of a model also gives an estimate of classifier performance ranging from 0.0 to 1.0 where numbers closer to 1.0 denote superior performance. In this project, the F1 score of a classifier refers to the weighted average F1 score across all classes unless otherwise noted, and all scores were recorded directly from the toolkits' data output. Given the results of the studies discussed in Chapter 2, the goal for this project is for the model to significantly surpass an AUC of at least 0.7. The F1 scores were used primarily for comparison between the models within this project and as a supplement to the other performance metrics when comparing against LT survival models outside of this project.

The initial models based on the 3-year binary class dataset attained maximum performance using the random forest algorithm. Although the AUC reached 0.709, the predictive ability of the model was rather average with an F1 score of only 0.685. The naive Bayes and MLP classifiers performed similarly average as can be seen in Table 2.

Table 2: 3-Year Binary Class Model Performances

	AUC	**F1**	**Precision**	**Recall**
Random Forest	0.709	0.658	0.660	0.663
Naive Bayes	0.701	0.659	0.658	0.659
MLP	0.645	0.606	0.607	0.606

The next dataset used to train the models was the multiclass dataset. Models using the multiclass dataset showed comparatively poor performance across all classes when looking at their F1 scores and AUCs versus those of the binary classification models. Again, the random forest algorithm performed best for the multiclass dataset with an

average AUC of 0.624 and average F1 score of 0.172 as seen in Table 3. However, performance for the 12+ years class showed a considerable increase in AUC to 0.874 and an F1 score of 0.394. This spike in performance prompted investigation of a 12-year binary classification.

Table 3: Multiclass Model Performances

	AUC	F1	Precision	Recall
Random Forest	0.624	0.172	0.201	0.221
Naive Bayes	0.616	0.172	0.176	0.197
MLP	0.563	0.159	0.159	0.159

Given the increased performance for the 12+ year class in the multiclass models, a dataset with a 12-year binary class was created and then used to train a random forest model. As seen in Table 4, the performance of this model was much higher than the previous models. The 12-year random forest model uses a separate training and validation set as opposed to cross-validation in order to more accurately gauge effectiveness on unseen data. The training set used a randomly selected subset of LT patient data with the true and false classes balanced, and the validation set utilized the remaining data that was not included in the training set. The data in Table 4 is based on the model validation results.

Table 4: 12-Year Random Forest Model Performance

	AUC	F1	Precision	Recall
True	0.870	0.248	0.144	0.905
False	0.870	0.836	0.993	0.721
Weighted Average Over Classes	0.870	0.807	0.951	0.730

Illustration 3: 12-Year Model True Class ROC Curve

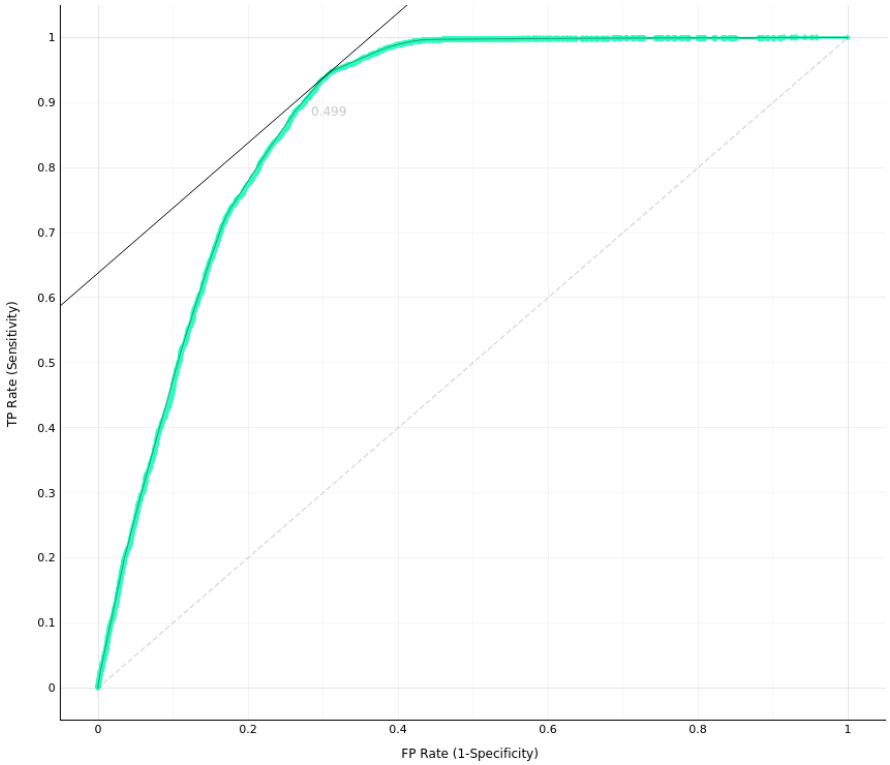
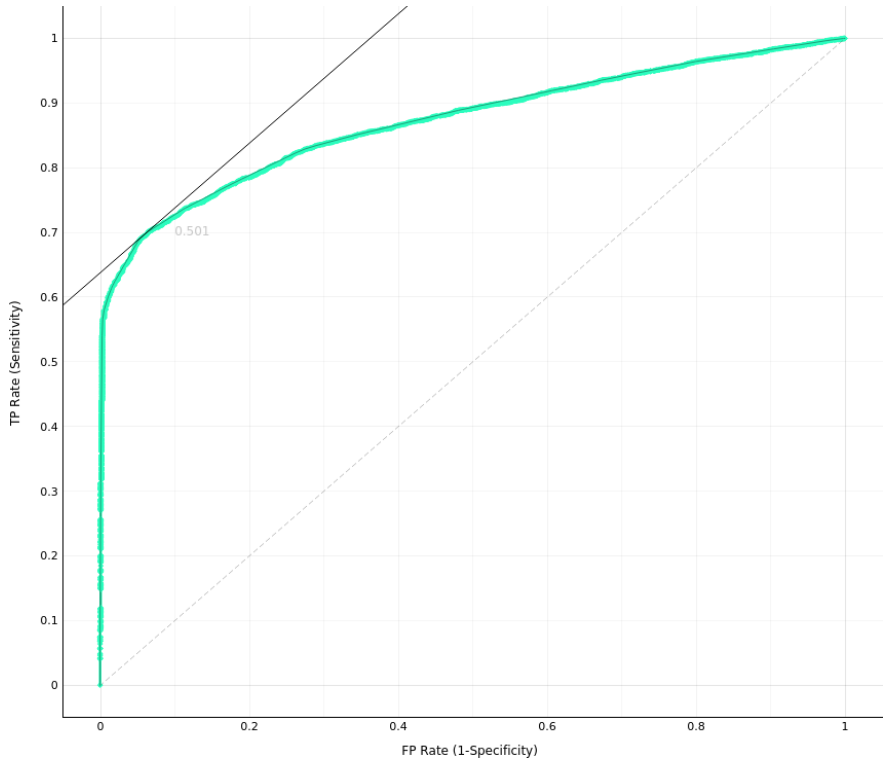


Illustration 4: 12-Year Model False Class ROC Curve



Chapter 5: Discussion

As can be seen by the results in tables 2 and 3, the models for LT survival prediction at time frames less than twelve years show average performances that match the SOFT score or the model trained by Ibáñez et al. at best. [3], [6] Furthermore, none of the models matched the performance of the MLP ANN reported by Raji and Chandra which used the UNOS dataset. [4], [5] This difference in performance is likely due to a difference in skill between researchers; given that this is an undergraduate level project, inefficiencies and flaws in the methods used are to be expected. In retrospect, a large flaw exists with the processed dataset because there is still nominal data that uses numbers to represent encoded attributes. After investigating the algorithms used by WEKA and Orange, most classifier algorithms will preprocess data internally when the dataset is loaded for training. This likely causes inaccuracy since the nominal attributes represented by numeric values will erroneously be normalized by the software. A solution to this issue would be to convert all numerically encoded nominal values to use unique strings; however, many attributes have over a hundred possible values that would require unique replacement strings to be assigned. Due to the time constraints of this project and the ability to override individual attribute types in Orange, nominal constraints were left numerically encoded. Similarly, it was most likely unnecessary to normalize any of the numeric attributes in the dataset since most classifier algorithms automatically normalize any numeric attributes.

Despite the poor performance for predicting transplant survival at less than 12 years, the model for predicting 12-year survival shows some promise for very long-term LT survival prediction. Although the 12-year binary class model does not have

exceptional sensitivity, the model does possess a reasonable amount of specificity. This may prove useful for the purpose of organ allocation as the model is relatively good at determining if a LT recipient will not survive for at least 12 years. When combined with more accurate short-term LT survival predictors and professional input, the 12-year model presented in this project could help to reduce the number of organs lost to transplant failure.

Chapter 6: Conclusion

This project has delved into various studies done to create a model that can predict LT survival as well as presented a new ML model for LT survival prediction based on the SRTR dataset. Due to the worldwide scarcity of organs for transplant there is a constant demand for ways to reduce organ waste due to failed transplants. [4] This project was started with two goals in mind, to create a model capable of accurately predicting LT survival and to create a LT survival prediction model using only attributes obtainable pre-TX or otherwise estimable pre-TX. This project began with heavy work on data analysis and preprocessing. A substantial amount of effort was put forth to obtain the largest and most complete dataset possible after preprocessing. Thanks to the large size of the SRTR dataset, many records with missing data could be dropped to avoid data imputation without reducing the size of the dataset too far. After training several models on the processed dataset, the random forest classifier was found to perform best with a binary class of 12-years post-TX survival. The final model presented its highest performance in its specificity where about 80% of false results were correctly classified as false. Given the ability of the model to determine those who will not survive at least 12 years post-TX, this model may find use as a supplement to current short-term survival prediction methods. Despite the effort put into preprocessing the dataset, there is still a large amount of room for improvement in the preprocessing steps and the resulting dataset used by the models. Hopefully this project will inspire more studies related to LT survival prediction and lead to increased use of the larger SRTR dataset in similar projects.

Works Cited

[1] S. Habib *et al.*, "MELD and prediction of post–liver transplantation survival," *Liver Transpl.*, vol. 12, no. 3, pp. 440–447, 2006, doi: 10.1002/lt.20721.

[2] P. H. Hayashi *et al.*, "Model for End-Stage Liver Disease score does not predict patient or graft survival in living donor liver transplant recipients," *Liver Transpl.*, vol. 9, no. 7, pp. 737–740, 2003, doi: 10.1053/jlts.2003.50122.

[3] A. Rana *et al.*, "Survival Outcomes Following Liver Transplantation (SOFT) Score: A Novel Method to Predict Patient Survival Following Liver Transplantation," *Am. J. Transplant.*, vol. 8, no. 12, pp. 2537–2546, 2008, doi: 10.1111/j.1600-6143.2008.02400.x.

[4] C. G. Raji and S. S. V. Chandra, "Various Medical Aspects of Liver Transplantation and its Survival Prediction using Machine Learning Techniques," *Indian J. Sci. Technol.*, vol. 10, no. 13, pp. 1–17, Apr. 2017, doi: 10.17485/ijst/2017/v10i13/94111.

[5] C. G. Raji and S. S. V. Chandra, "Long-Term Forecasting the Survival in Liver Transplantation Using Multilayer Perceptron Networks," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 47, no. 8, pp. 2318–2329, Aug. 2017, doi: 10.1109/tsmc.2017.2661996.

[6] Vicente Ibáñez *et al.*, "Predicting Early Transplant Failure: Neural Network Versus Logistic Regression Models," *Open Transplant. J.*, vol. 3, pp. 14–21, May 2009, doi: 10.2174/1874418400903010014.

[7] V. Bhat, M. Tazari, K. D. Watt, and M. Bhat, "New-Onset Diabetes and Preexisting Diabetes Are Associated With Comparable Reduction in Long-Term Survival After Liver Transplant: A Machine Learning Approach," *Mayo Clin. Proc.*, vol. 93, no. 12,

pp. 1794–1802, Dec. 2018, doi: 10.1016/j.mayocp.2018.06.020.

[8] A. Andres *et al.*, "A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis," *PLoS ONE*, vol. 13, no. 3, pp. 1–14, Mar. 2018, doi: 10.1371/journal.pone.0193523.

[9] R. Wiesner *et al.*, "Model for end-stage liver disease (MELD) and allocation of donor livers," *Gastroenterology*, vol. 124, no. 1, pp. 91–96, Jan. 2003, doi: 10.1053/gast.2003.50016.

[10] M. Malinchoc, P. S. Kamath, F. D. Gordon, C. J. Peine, J. Rank, and P. C. J. ter Borg, "A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts," *Hepatology*, vol. 31, no. 4, pp. 864–871, 2000, doi: 10.1053/he.2000.5852.

[11] L. E. Chambless and G. Diao, "Estimation of time-dependent area under the ROC curve for long-term risk prediction," *Stat. Med.*, vol. 25, no. 20, pp. 3474–3486, Oct. 2006, doi: 10.1002/sim.2299.

[12] Scientific Registry of Transplant Recipients, "Mission, Vision, and Values." https://www.srtr.org/about-srtr/mission-vision-and-values/ (accessed Oct. 11, 2019).

[13] "SAF Data Dictionary." https://www.srtr.org/requesting-srtr-data/saf-data-dictionary/ (accessed Oct. 11, 2019).

[14] F. Eibe, M. A. Hall, and I. H. Witten, "The WEKA Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques.'" Morgan Kaufmann, 2016, Accessed: Oct. 11, 2019. [Online]. Available: https://www.cs.waikato.ac.nz/~ml/weka/.

[15] University of Ljubljana Bioinformatics Laboratory, "Orange Data Mining

Toolbox.” https://orange.biolab.si/#Orange-Features (accessed Oct. 11, 2019).