Enhancing MELD Scores for Mortality Predictions in Liver Patients

Alex Gregory [1], Henry Lambson [2], Mike Wisniewski [3]

[#] *Lyle School of Engineering, Southern Methodist University*

[1] alexgregory@mail.smu.edu

[2] hlambson@mail.smu.edu

[3] mwisniewski@mail.smu.edu

Abstract— In the medical field, there is an ever-growing need for more effective approaches at diagnosing and preventing disease. When assessing patients that qualify for liver transplantation, this is an even more important requirement. To achieve better approaches for diagnosing and assessing end stage liver disease, we explore alternative ways to calculate MELD scores to achieve better mortality predictions in liver patients. The aim is to provide a series of models that explore what may or may not be achievable in further research. We view that enhancement of MELD scores is crucial to accurately assist liver patients.

Our group takes the approach of "simpler is better". For this exercise, we view the simpler solution may be the best solution - not because solving medical problems is easy, but because the data surrounding this topic is limited in number of records (patients). Therefore, with this limitation, we believe there is no need for overly complex neural networks. But we do think that neural networks still play an important role in determining MELD scores.

We found evidence to suggest that neural networks are a viable tool for use on this problem and that the simpler models perform better than complex models. Additionally, these models reclassify patient MELD scores and more accurately predict mortality likeliness per MELD bracket. Although further analysis will be required to definitively prove our model's viability, our analysis shows that there is opportunity for MELD score improvement using our neural network approach.

I. Introduction

A. Motivation

The MELD score, originally Mayo End Stage Liver Disease and later renamed Model for End-stage Liver Disease [1], was originally introduced in 2001 by researchers from the Mayo Clinic as a means of estimating survivability for transjugular intrahepatic portosystemic shunt (TIPS) placement [2]. The accuracy of this metric led the United Network for Organ Sharing (UNOS) to adopt this score the primary metric for patient prioritization for liver transplants in the United States [4]. The score gives a likelihood from 0 to 50, but traditionally capped at 6 to 40, that rates the survivability of a patient with cirrhosis.

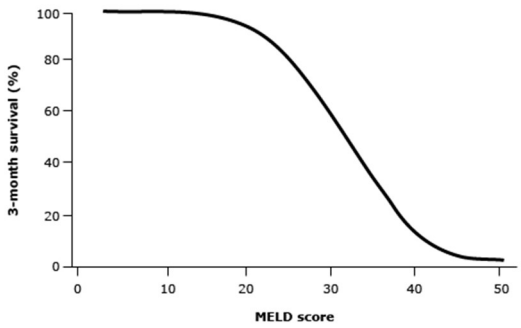

Fig. 1 Estimated 3-month survival as a function of the MELD score in patients with cirrhosis [5]

This metric is an invaluable tool for comparing patients in a standardized manner, but the metric has room for improvement. "In March 2007, Kamath et al reported that, though MELD has the ability to rank cirrhotic patients according to short term mortality, it is not a perfectly universal score, as survival is not accurately predicted in 15-20% of patients, and users must be aware of limitations of its application."[1][3] As a result of this study and others, UNOS has a list of standard exceptions for the MELD and MELD-NA score that allow for exception points to be added to a patient's score.

Standard MELD exceptions include [5][6]:

- Hepatocellular carcinoma
- Hepatopulmonary syndrome
- Portopulmonary hypertension
- Familial amyloid polyneuropathy
- Primary hyperoxaluria
- Cystic fibrosis
- Hilar/Perihilar cholangiocarcinoma (the liver transplant center must have a UNOS approved protocol for the work-up and management of patients with cholangiocarcinoma undergoing transplantation)
- Hepatic artery thrombosis (occurring within the 14 days after liver transplant surgery, but not meeting criteria for status 1A)

In January of 2016, OPTN adopted a new standard for MELD scoring that includes serum sodium, MELD-NA [7]. A new limitation introduced with this calculation was a possible skew from IV fluids or diuretics effect on patient's serum sodium [5]. With various exceptions and situations that were not recognized, OPTN established a National Liver Review Board (NLRB) to provide additional MELD exception points for patients that are not qualified under the traditional exception metrics.

In 2022, MELD 3.0 was introduced and incorporates several additional variables as a way of accounting for sex disparity in waitlist patients and a better classification of severity [8]. This metric is still be evaluated and has not been adopted by UNOS.

Since the inception of the MELD score in 2001, many iterations of the score have included more values and situational edits to values as a means of increasing the accuracy of the score. Can a dynamic model be trained to provide a more accurate MELD score? With the wealth of medical data available to analyze, we argue that it should be possible to train a model that is able to account for most of the exception situations listed above and consider unique factors in a patient's labs to generate a more accurate MELD score. To limit the scope of this paper with the limited data we have available, our goal is to provide a proof of concept that deep neural networks can generate scores that are as good or better than traditional methods of generating MELD scores.

B. Related Work

In 2023, researchers at University Hospital Leipzig evaluated 654 patient records to improve on the MELD score [9]. Their study yielded another formula for calculating MELD derived from penalized regression and random forest models to identify the highest impact lab values. They argue that machine learning can be used as means of highlighting more impactful variables for MELD iteration. The data we are training comes from their study.

II. Data Preparation and Methods

The original data contained 778 patients with potential for liver transplant during the evaluated from November 2012 to June 2015. This dataset was trimmed by the original authors as shown in Figure 2.

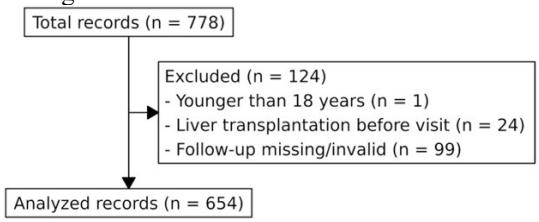

Fig. 2 Records excluded in source pre-processing [9]

We further trimmed the data due to missing columns that would not allow us to calculate our three MELD scores for accuracy evaluation. This dropped 9 more records that were missing INR, sodium, or albumin. We used MELD, MELD-NA, and MELD 3.0 to evaluate, calculating them with standard formulas from MDCALC [1]. The original dataset was also in scientific and not metric values, so lab values were converted to metric to follow the formulas on MDCALC.

Our data was divided into columns that are binary and columns that are continuous or discrete, but not binary. In our approach to scale and normalize our data, we used a standard scaling method and only scaled our non-binary columns. We view scaling binary columns as unnecessary for our models. Because of the limited amount of data, our train and test datasets were split using a 70/30 split.

Our data is further processed using dimensionality reduction algorithms; however, we explain in detail our process in the "Models" section of this paper as we view our dimensionality reduction as variations of our base models.

In depth step by step data preparation can be found in our Jupyter Notebook [13].

III. Models

We approached this problem from a neural network perspective as opposed to the traditional machine learning models found in Gibb et al [9]. We architected two models in our approach: a simple Sequential neural network and a relatively more complex ResNet-lite network. Our Sequential model contains 8 layers, with a glorot uniform distribution of weights at the input layer, ReLU activations for each layer, and a linear output layer. The Sequential model incorporates dropout and batch normalization at various parts of the model.

Our ResNet network is based off the Tensorflow ResNet architecture, however we modified the architecture to contain fewer layers and a multi-layer fully connected network after convolutions. This architecture contains an initial Conv1D layer followed by five residual Conv1D blocks which connects to a 4-layer sequential network. Like our Sequential model, we incorporated dropout and batch normalization at various points in the architecture. This network outputs to a single neuron layer with a linear activation function.

During our preprocessing of our data, we explore two dimensionality reduction methods to further enhance our models. The dimensionality reduction methods chosen are Principal Component Analysis (PCA) and UMAP. We incorporate PCA and UMAP into both base models to bring our total model count to six. For both PCA models, we reduce our dimensions down to 12 total dimensions as we found these 12 dimensions explain roughly 80% of the variance. For both UMAP models, a similar approach was used. We calculate UMAP distances using Euclidean distance. It was important for us to keep parameters and hyperparameters of all models as consistent as possible to accurately analyze the results. These models serve as a jumping off point for further analysis.

More details for both models can be found within our Jupyter Notebook [13].

IV. Results

Based on the results of these six models, we concluded that for this problem, simpler is better. The simplest model of the six ended up being the best performing, with the much more complex ResNet models performing the worst by a wide

margin. Even when utilizing the dimensionality reduction techniques of PCA and UMAP, the model's accuracy deteriorated. Figure 2 will show the results of Model 1's classifications compared to the original MELD scores of the test set, shown in Figure 3.

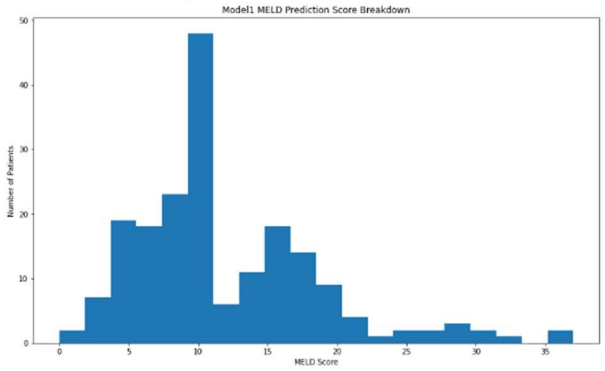

Fig. 2 Distribution of MELD Scores-Model 1

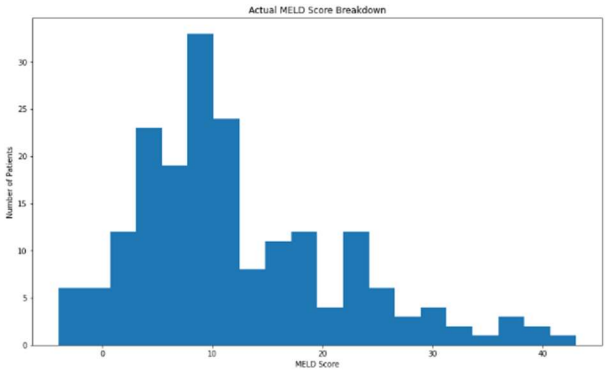

Fig. 3 Distribution of MELD Scores-Original

From these figures, we can see that Model 1 predicted lower MELD scores on average than the traditional method of calculating the scores. While the original scores had a maximum of just over 40, Model 1 had a maximum score of just over 35. Model 1 classified much fewer patients as having a MELD score of over 20, with most of the patients being classified around the score of 10.

Table 1 will show how Model 1 classified the patients in the test dataset into MELD brackets. From this table, we will be able to see how Model 1 is able to make better classifications than the traditional method of calculating MELD. In the notebook [13], there is a full version of this table which includes the classifications and ratios from all six models.

TABLE 1 EXPECTED VS OBSERVED DEATHS FOR MODEL 1 (ROUNDED)

MELD Bracket	Expected Deaths, Total	Expected Mortality Ratio, Total	Expected Deaths, Test Set	Expected Mortality Ratio, Test Set	Observed Deaths, Model 1	Observed Mortality Ratio, Model 1
[6,9]	12	0.04	3	0.03	1	0.01
[10,20]	41	0.17	9	0.13	18	0.17
[20,30]	33	0.52	10	0.40	9	0.64
[30,40]	12	0.71	8	0.80	3	1.00
[40+]	3	0.60	1	1.00	0	0.00

The table shows that Model 1 rebalanced the MELD scores of many of the patients into brackets the suit them better. Looking first at the [6,9] bracket, three of the patients that were originally classified into this bracket died, while only one patient from Model 1's classifications died in this bracket. This means that the model reclassified the patients that were originally in this bracket into higher brackets where their likelihood of dying would be seen as higher. The original classifications saw a mortality ratio of 0.033708, while Model 1's classifications saw a much lower ratio at 0.014493. This lower ratio for this bracket means that Model 1's classifications were more accurate that the traditional method of calculating MELD.

Model 1 classified most patients into the [10,20] bracket, with 18 of the test patients appearing here, whereas the original classifications only saw 9 patients in this bracket. Model 1 had a higher mortality ratio in this bracket, but only slightly even though it classified double the number of patients into this range. Both mortality ratios are low, which is to be expected as this bracket is for middling MELD values.

The [20,30] bracket is where the MELD values start to get high, leading to these patients being seen as much more likely to die. The original classifications only had one more patient in this bracket than Model 1's classification, however the mortality ratios of the two are quite different. The original had a ratio of 0.4, while Model 1 had a ratio of 0.642857, meaning that many of the patients that were originally in this bracket were placed in different brackets by Model 1. This data shows that patients who were originally in this bracket that did not die were placed into the [10,20] bracket rather than this one. The patients that were placed into this bracket are most likely made up of some of the patients that were originally in the [30,40] range, as Model 1 only classified three patients into the [30,40] bracket, and all three ended up dying. This gave the [30,40] bracket from Model 1 a mortality ratio of 1.0, while the original scores had ten patients classified in this range, with eight of them dying.

From these results we can see that Model 1 is much stricter on which patients it gives this high of a MELD score to, and those that it did give this score to ended up dying. In the [40+] bracket, the original classifications had one patient who died, and Model 1 had zero patients.

From the distribution of patients in Model 1's classification, overall, Model 1 is lowering the MELD score of the patients. From these classifications, we can see that Model 1 classifies very high-risk patients into the [30,40] bracket, and seems to reserve a score of over 40 for extreme outliers, seeing as there are none classified in that bracket for this test set. From Model 1's classifications, the brackets can be reworked to the following:

- [6,9] - Extremely low chance of death
- [10,20] - Low chance of death
- [20,30] - High chance of death
- [30,40] - Extremely high chance of death

Because of the large difference in mortality ratio between brackets [10,20] and [20,30] it may be worthwhile to divide these two brackets into three or four smaller brackets so that more precision when assessing MELD scores for chance of death can be achieved.

V. Final Word

Given our analysis and comparison of our neural network with traditional MELD scores, we draw two conclusions: there is evidence to suggest simpler models perform stronger than more complex architectures and there is evidence to suggest that our model approach classifies and predicts mortality better than MELD.

We note that further analysis and research will be required to bring a definitive case of using neural network architectures over current MELD score calculations and models. We believe our study presents strong evidence that neural network architectures can be a viable solution for MELD score predictions. We are hesitant to suggest that MELD scores antiquated or inaccurate forms of measurement for patients, rather we believe that model usage to reclassify, or more accurately identify and bucket patients into correct MELD brackets would create stronger cases for patients to receive correct medical attention.

We highly encourage the medical community to leverage neural network architectures into their own practices for potential benefits. Although there is no simple solution to complex problems in medicine, there are ways to simplify the road to solving these complex problems. With good data governance, we also encourage the medical community to open their datasets to the public to create an open-source community to help solve these problems.

Lastly, we want to thank our Professor Dr. Eric Larson who has encouraged us and allowed us to make creative decisions during our research into this topic. Without his guidance and support, we don't believe we would have a good foundation for further research.

References

[1] (2023) MDcalc website. [Online] Available: https://www.mdcalc.com/calc/78/meld-score-model-end-stage-liver-disease-12-older#evidence

[2] Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, D'Amico G, Dickson ER, Kim WR. A model to predict survival in patients with end-stage liver disease. Hepatology. 2001 Feb;33(2):464-70. doi: 10.1053/jhep.2001.22172. PMID: 11172350.

[3] Kamath PS, Kim WR; Advanced Liver Disease Study Group. The model for end-stage liver disease (MELD). Hepatology. 2007 Mar;45(3):797-805. doi: 10.1002/hep.21563. PMID: 17326206.

[4] Freeman RB Jr, Wiesner RH, Harper A, McDiarmid SV, Lake J, Edwards E, Merion R, Wolfe R, Turcotte J, Teperman L; UNOS/OPTN Liver Disease Severity Score, UNOS/OPTN Liver and Intestine, and UNOS/OPTN Pediatric Transplantation Committees. The new liver allocation system: moving toward evidence-based transplantation policy. Liver Transpl. 2002 Sep;8(9):851-8. doi: 10.1053/jlts.2002.35927. PMID: 12200791.

[5] (2023) UpToDate website. [Online]. Available: https://www.uptodate.com/contents/model-for-end-stage-liver-disease-meld#H938189995

[6] (2023) Organ Procurement and Transplant Network website. [Online]. Available: https://optn.transplant.hrsa.gov/policies-bylaws/policies/

[7] (2023) Organ Procurement and Transplant Network website. [Online]. Available: https://optn.transplant.hrsa.gov/news/meld-serum-sodium-policy-changes/

[8] Kim WR, Mannalithara A, Heimbach JK, Kamath PS, Asrani SK, Biggins SW, Wood NL, Gentry SE, Kwong AJ. MELD 3.0: The Model for End-Stage Liver Disease Updated for the Modern Era. Gastroenterology. 2021 Dec;161(6):1887-1895.e4. doi: 10.1053/j.gastro.2021.08.050. Epub 2021 Sep 3. PMID: 34481845; PMCID: PMC8608337.

[9] Gibb, Sebastian, Berg, Thomas, Herber, Adam, Isermann, Berend and Kaiser, Thorsten. "A new machine-learning-based prediction of survival in patients with end-stage liver disease" Journal of Laboratory Medicine, vol. 47, no. 1, 2023, pp. 13-21. https://doi.org/10.1515/labmed-2022-0162

[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.

[13] Gregory A, Lambson H, Wisniewski M, "Enhancing MELD Scores for Mortality Predictions in Liver Patients", Unpublished Jupyter Notebook