Predicting waitlist mortality for liver transplant candidates: a comparative analysis between statistical scores and machine learning models

Abdelghani Halimi, Nesma Houmani, Sonia Garcia-Salicetti, Ilias Kounis, Audrey Coilly

The 12th International Conference on E-Health and Bioengineering - EHB 2024
Grigore T. Popa University of Medicine and Pharmacy Iasi, November 14-15, 2024, Iasi, ROMANIA

Predicting Waitlist Mortality for Liver Transplant Candidates: A Comparative Analysis between Statistical Scores and Machine Learning Models

Abdelghani Halimi[1,2], Nesma Houmani[1], Sonia Garcia-Salicetti[1], Ilias Kounis[3,4,5,6], Audrey Coilly[3,4,5,6]
[1]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, Palaiseau, 91120, France,
[2]Chaire BOPA, Rue de la Chapelle de l'Hôpital, Villejuif, France
[3]Inserm, Université Paris-Saclay, UMR-S 1193, Villejuif, France
[4]AP-HP Hôpital Paul-Brousse, Centre Hépato-Biliaire, Villejuif, France
[5]Université Paris Saclay, Inserm, Physiopathogénèse et traitement des maladies du Foie, Villejuif, France
[6]France FHU Hepatinov, Villejuif, France

***Abstract*— Accurately predicting waitlist mortality for liver transplant candidates is a critical yet challenging task. Traditional models such as MELD, MELD-Na, and MELD 3.0 have been widely used by clinicians but fall short in delivering precise mortality predictions when compared to machine learning (ML) models. In this study, we conduct a comprehensive comparative analysis of these conventional scoring systems against advanced ML models, including LDA, TabNet, Random Forest, and LightGBM. Results not only highlight the improved predictive accuracy of certain ML models over MELD-based scores but also identify the most significant variables influencing 3-month waitlist mortality. This analysis enables the proposal of new, critical risk factors for consideration in future scoring models. By leveraging these insights, we aim to contribute to the development of a more efficient and equitable organ allocation system, ultimately enhancing patient outcomes and potentially saving more lives through better patient prioritization.**

Keywords— Liver Transplant; MELD scores; Machine Learning; Organ Allocation; Waitlist Mortality.

I. Introduction

Liver transplantation (LT) is a life-saving treatment for end-stage liver disease. The growing disparity between organ supply and demand, coupled with the complexity of transplant outcomes, has driven the development of predictive models to improve risk assessment and allocation systems. The Model for End-Stage Liver Disease (MELD) [1], introduced in 2002, has been widely used to prioritize patients based on 3-month mortality risk. Later, MELD-Na [2] and the more recent MELD 3.0 [3] were proposed, revising the MELD score by adding serum sodium, albumin and gender information.

Despite these improvements, MELD-based scores remain limited in predicting certain life-threatening conditions, known as MELD exceptions [4-5], such as hepatocellular carcinoma and recurrent cholangitis, which may not be fully captured by these models. This highlights the need for more generalizable and reliable modeling approaches to mortality prediction.

Machine learning (ML) has emerged as a powerful tool in medical outcome modeling, with decision tree-based models commonly used and neural networks explored in recent studies [6-10]. However, these studies have notable limitations affecting their generalizability and applicability. In [8], patient data appeared in both training and test sets, which introduces bias and risks inflating performance metrics. In [10], only 3% of patients were women, limiting the applicability of the findings to a broader population. Additionally, many studies [7-9] applied ML algorithms to numerous features without assessing their relevance. This can lead models to rely on highly correlated features, masking important relationships and complicating interpretability. Also, including more features increases the likelihood of missing data in real-world clinical practice. Finally, some works [9-10] have studied mortality risk in LT by comparing different ML-based models; nevertheless, without data understanding and under the methodological limitations mentioned above.

In this study, we propose to conduct a comparative analysis of traditional MELD-based scores and various ML models—including linear model, neural networks, and tree-based algorithms—to predict 3-month waitlist mortality among LT candidates. Our aim is to identify the most suitable modeling approach for this problematic. By addressing the limitations highlighted in previous studies, we have developed a novel mortality risk score that enhances predictive accuracy and clinical utility. Furthermore, we identified and introduced new risk factors that influence patient outcomes. By overcoming existing system shortcomings and directly addressing prior literature limitations, our work seeks to improve organ allocation strategies for LT candidates.

II. Database and Methods

A. UNOS dataset description

This study uses data from the Organ Procurement and Transplantation Network (OPTN) and the United Network for Organ Sharing (UNOS), as found in the Standard Transplant

979-8-3315-3214-7/24/$31.00 ©2024 IEEE

Analysis and Research (STAR) file. The dataset includes clinical and laboratory information on 259,081 patients listed in the U.S. from February 27, 2002 to September 30, 2023 with multiple observations per patient due to updates in their records.

We focused on patients aged 18 or older at listing, including those who: (i) died or were removed for being too sick within three months, or (ii) survived beyond three months. Patients listed with MELD exception scores or for multi-organ transplants were excluded, aiming to curate a dataset reflective of a broad spectrum of LT candidates. Additionally, patients who received a transplant before the studied time period were excluded as their outcomes cannot be known in the absence of transplantation. Thus, the resultant dataset for the study comprises data on 94,891 patients (83425 survivors and 11466 non-survivors). The cohort demographics including the number of observations *N* (visits) per class, are given in Table I.

TABLE I. COHORT DEMOGRAPHICS.

Variable	Survivors (*N*=1299215)	Non-survivors (*N*=58254)
Age at registration	52.97 ± 10.47	55.09 ± 10.48
Sex, male	781207 (60.13%)	33183 (56.96%)
MELD at listing	16.40 ± 6.52	30.76 ± 10.30
Bilirubin (mg/dL)	4.37 ± 5.59	15.44 ± 12.57
Creatinine (mg/dL)	1.07 ± 0.57	1.83 ± 1.32
Sodium (mEq/L)	136.41 ± 4.44	135.04 ± 6.34
Albumin (g/dL)	3.10 ± 0.66	2.96 ± 0.78
INR	1.57 ± 0.59	2.50 ± 1.38
BMI	28.87 ± 5.81	28.91 ± 6.59
Ascites		
Absent	222411 (17.12%)	4386 (7.53%)
Slight	641251 (49.36%)	21256 (36.49%)
Moderate/large	202694 (15.60%)	25245 (43.34%)
N/A	232859 (17.92%)	7367 (12.65%)

B. Data preprocessing

The study focuses on variables recorded prior to either transplantation or removal from the waitlist. A total of 27 variables are considered, encompassing clinical, laboratory, and disease-specific factors. To capture patient health dynamics, we compute differences (DIFF variables) between consecutive measurements for key lab values: SERUM_SODIUM, SERUM_CREAT, ALBUMIN, BILIRUBIN, and INR. Missing values in numerical variables (less than 7%) were imputed with class means, while observations with missing categorical data (under 0.02%) were removed. One-hot encoding, with the first category omitted to prevent multicollinearity, was used to transform categorical variables for model training.

C. Methodology

This study targets predicting mortality within three months on the waiting list, using on one hand the three MELD-based scores (MELD, MELD-Na and MELD 3.0), and on the other hand four ML-based classifiers: (i) Linear Discriminant Analysis (LDA) [11], (ii) Neural Network architecture for tabular data including sequential attention mechanisms (TabNet) [12], (iii) Random Forest (RF) [13] and (iv) Light Gradient Boosting Machine (LightGBM) [14]. We use different ML-based models to investigate the predictive power of a linear classifier, as well as of more complex and non-linear models relying on neural networks (TabNet) and tree-based models (RF and LightGBM).

We propose to evaluate such ML-based models in two steps: first considering only the variables of the MELD-based scores, and second on the 27 variables. MELD, MELD-Na and MELD 3.0 were calculated per observation as detailed in [3].

Due to the unbalanced dataset, we down-sample the majority class into 23 balanced partitions, each containing the same number of observations as the minority class and different patients from the majority class. For each partition, we perform 3-fold cross-validation to train and evaluate classifiers, ensuring patient observations are kept in either the training or test set. For both LDA and TabNet, we normalize the variables using Z-score transformation. This is not necessary for the tree-based models.

The optimal configuration for the classifiers is determined using a grid search for each hyperparameter described in Table II. The other hyperparameters are kept at their default settings in scikit-learn across all models. The models are evaluated using AUROC as the scoring metric and validated through 3-fold cross-validation to ensure robustness.

TABLE II. HYPERPARAMETER SPACE FOR ML CLASSIFIERS.

Methods	Hyperparameters	Grid
LDA	Solver algorithm	'svd', 'lsqr', 'eigen'
	Covariance Shrinkage	None, 'auto', 0 to 1 (step of 0.1)
TabNet	Decision Dim (n_d)	8, 16, 24
	Attention Dim (n_a)	8, 16, 24
	Num. of decision steps	3, 5
RF	Number of trees	25 to 150 (step of 25), 200, 250
	Maximum tree depth	None, 2,3,4,5,7,10
LightGBM	Number of estimators	40 to 100 (step of 5), 125, 150

Models' performance is assessed using AUROC, Accuracy, Sensitivity (correctly classified non-survivors) and Specificity (correctly classified survivors). We compute these metrics for each of the three folds for all subsets and then average them. The optimal decision threshold is selected to maximize both sensitivity and specificity. After evaluating the ML classifiers with all 27 features, we apply Gini importance criterion [13] for feature selection to assess best model's performance using only the most relevant features.

III. RESULTS

Table III presents the predictive performance of MELD, MELD-Na, MELD 3.0, and the four ML models. Performance of ML models are evaluated using MELD-based variables and then the 27 variables. For the statistical analysis, we used the Wilcoxon Mann-Whitney test with Bonferroni correction to compare models on each metric. Each model was trained on 23 data partitions and evaluated using 3-fold cross-validation,

yielding 69 values per metric. These values were used to assess the significance of performance differences between models.

TABLE III. PERFORMANCE OF THE DIFFERENT MODELS FOR THE 3-MONTH PERIOD. THE TABLE CELLS CONTAIN MEAN VALUES.

	AUROC	Accuracy (%)	Sensitivity (%)	Specificity (%)
MELD	0.881	80.85	78.94	82.82
MELD-Na	0.888	81.35	81.75	80.95
MELD 3.0	0.884	81.08	81.81	80.30
LDA (MELD)	0.880	80.76	78.79	82.79
LDA (MELD-Na)	0.886	81.08	81.03	81.12
LDA (MELD 3.0)	0.880	80.45	81.10	79.78
TabNet (MELD)	0.884	81.15	79.19	83.17
TabNet (MELD-Na)	0.894	81.92	81.81	82.03
TabNet (MELD 3.0)	0.890	81.47	82.18	80.73
RF (MELD)	0.883	81.04	79.37	82.77
RF (MELD-Na)	0.888	81.35	80.34	82.40
RF (MELD 3.0)	0.895	82.03	81.86	82.20
LightGBM (MELD)	0.883	81.05	79.58	82.57
LightGBM (MELD-Na)	0.909	83.21	82.78	83.64
LightGBM (MELD 3.0)	0.909	83.22	83.43	82.98
LDA (27 vars.)	0.905	82.68	81.44	83.94
TabNet (27 vars.)	0.908	83.08	82.68	83.50
RF (27 vars.)	0.929	85.29	85.32	85.26
LightGBM (27 vars.)	**0.935**	**85.77**	**85.58**	**85.97**

We note that MELD-Na significantly outperforms both the original MELD score and MELD 3.0 in terms of AUROC, achieving 0.888 compared to 0.881 for MELD ($p = 1.23\times10^{-9}$) and 0.884 for MELD 3.0 ($p = 5.69\times10^{-4}$). Additionally, MELD-Na and MELD 3.0 also show a good balance between sensitivity and specificity, both significantly surpassing MELD in sensitivity ($p = 4.35\times10^{-13}$ for MELD-Na and $p = 9.91\times10^{-16}$ for MELD 3.0). These findings highlight that MELD-Na and MELD 3.0 are more effective at identifying high-risk patients than the original MELD score.

Among ML models, TabNet, RF, and LightGBM, when trained on MELD-based variables, outperform MELD-based scores. LightGBM leads to the best performance, especially when using MELD-Na and MELD 3.0 variables, reaching an AUROC of 0.909 and an Accuracy of 83.2%. Notably, LightGBM shows slightly higher sensitivity when trained on the variables of MELD 3.0 compared to those of MELD-Na, at the price of a minor decrease in specificity. By contrast, LDA when trained on MELD variables gives worse performance compared to MELD-based scores.

Finally, using all the 27 variables, the four ML models improve classification performance, with LightGBM achieving significantly the best results (AUROC = 0.935, accuracy = 85.77%, sensitivity = 85.58%, specificity = 85.97%), confirmed by the Wilcoxon Mann-Whitney test ($p < 0.05$ for all comparisons). This underscores the benefit of incorporating additional variables beyond those used in MELD-based scores.

To improve model interpretability, we exploit Gini importance to select the most relevant features for LightGBM, our top-performing model. Figure 1 shows the feature importance results.

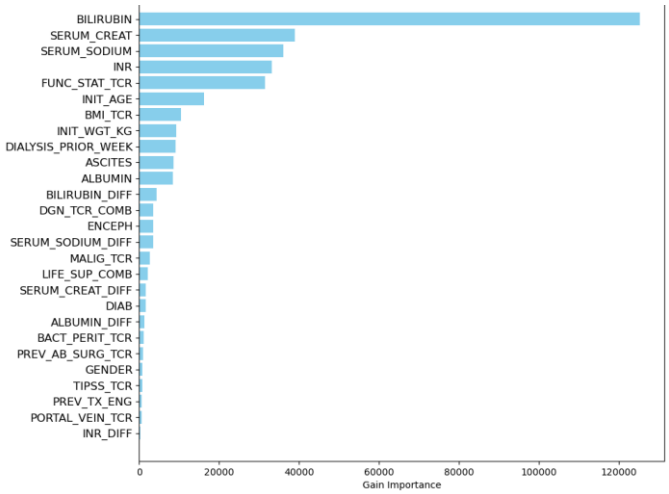

Fig. 1. Feature importance of LightGBM for 3-month mortality prediction.

We note that BILIRUBIN, SERUM CREAT, SERUM SODIUM and INR emerge as the top four features. Interestingly, these variables are included in the MELD-based scores. Additionally, factors such as functional state, age at registration (INIT_AGE), BMI, patient's weight and the degree of ascites (ASCITES) emerge in the top predictors of mortality on the waiting list. It is noteworthy that these features are not taken into consideration by traditional risk score models. This finding highlights a gap in existing models and emphasizes the need to consider these variables in future risk scoring systems. Conversely, variables such as type of diabetes (DIAB) and GENDER (considered in MELD 3.0), are found to be less significant in predicting mortality for the studied time period. This result aligns with the findings in [7].

We continue our study by analyzing the performance of LightGBM against the number of features, ordered by Gini importance. The analysis reveals that the differences in all performance metrics become non-significant when more than 12 features are considered, compared to using the full set of 27 features ($p > 0.05$ for all metrics). This indicates that additional features offer minimal benefits and can potentially introduce noise or overfitting. As shown in Table IV, LightGBM trained with the optimized set of 12 features achieves an AUROC of 0.933, comparable to the AUROC of 0.935 obtained using the full set of 27 variables. This underscores the high predictive power of the selected features and validate the efficiency of the reduced feature set.

TABLE IV. LIGHTGBM PERFORMANCE CONSIDERING THE 12 MOST RELEVANT FEATURES IDENTIFIED WITH GINI METHOD.

	AUROC	Accuracy (%)	Sensitivity (%)	Specificity (%)
LightGBM	0.933	85.59	85.31	85.88

IV. DISCUSSION

This study presents a comparative analysis of several ML models against MELD-based scores for predicting 3-month waitlist mortality in LT candidates. Our results showed that both MELD-Na and MELD 3.0 significantly outperform MELD in predicting 3-month waitlist mortality on our dataset. Our findings are consistent with the state-of-the-art [2-3], demonstrating that enhancements such as the inclusion of serum sodium, albumin, and gender in the MELD score significantly improve its predictive accuracy.

ML models like TabNet, RF, and LightGBM, when trained on MELD-based variables, demonstrate higher accuracy than traditional MELD scores. Furthermore, among the ML approaches, non-linear models outperform the linear one (LDA). This suggests that non-linear interactions between variables are crucial for more accurate predictions, unlike the fixed scoring of MELD-based models.

When considering all the 27 variables, LightGBM achieved the best performance (AUROC = 0.935), emphasizing the importance of considering additional variables beyond MELD scores to capture the complexities of patient conditions. Traditional fixed scoring systems fall short in modeling these nuances, whereas ML models can effectively account for them. Furthermore, decision tree-based models, such as RF and LightGBM, outperformed Neural Network models like TabNet in this study. This indicates that, in the realm of structured medical data, the decision tree paradigm seems more effective in identifying the nuanced patterns necessary for accurate mortality prediction [15].

Using the Gini importance method, we identified 12 key variables critical for predicting waitlist mortality. LightGBM, when trained on these variables, achieved similar performance to when all 27 variables were used, demonstrating their strong predictive power. Notably, six of the selected variables are components of MELD-based scores, which demonstrates the effectiveness of our methodology. Our study emphasizes the importance of factors not included in MELD-based scores, such as patient's functional state, age at registration, BMI, weight, degree of ascites, and changes in bilirubin over time. Through feature selection, we highlighted their specific contribution to waitlist mortality prediction.

V. CONCLUSIONS

Our study provides valuable insights into predicting LT waitlist mortality. We demonstrated the importance of non-linear interactions in clinical data for more accurate predictions, in contrast to the fixed scoring of MELD-based models. Additionally, incorporating variables beyond MELD components significantly improved performance. Despite the growing interest in deep learning, decision tree-based models proved more effective in structured medical data. Our research also identified 12 key variables for mortality prediction, with six drawn from MELD-based scores. This reinforces the relevance of MELD components while also highlighting the importance of additional factors. These findings underscore the need to consider these factors more closely for improved organ allocation and waitlist management.

In future work, we aim to validate our model on excluded populations and enhance its interpretability using explainability tools.

ACKNOWLEDGMENT

We thank BOPA (Bloc OPératoire Augmenté) Innovation Chair for funding this research. The data were supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network (OPTN). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the OPTN or the US government.

REFERENCES

[1] P. S. Kamath et al., "A model to predict survival in patients with end-stage liver disease", Hepatology, vol. 33, nº 2, pp. 464-470, 2001.

[2] W. R. Kim et al., "Hyponatremia and mortality among patients on the liver-transplant waiting list", N Engl J Med, vol. 359, nº 10, pp. 1018-1026, 2008.

[3] W. R. Kim et al., "MELD 3.0: The Model for End-Stage Liver Disease Updated for the Modern Era", Gastroenterology, vol. 161, nº 6, pp. 1887-1895.e4, 2021.

[4] C. Francoz et al., "Model for end-stage liver disease exceptions in the context of the French model for end-stage liver disease score-based liver allocation system", Liver Transpl, vol. 17, nº 10, pp. 1137-1151, 2011.

[5] D. S. Goldberg et K. M. Olthoff, "Standardizing MELD Exceptions: Current Challenges and Future Directions", Curr Transplant Rep, vol. 1, nº 4, pp. 232-237, 2014.

[6] M. Bhat et al., "Artificial intelligence, machine learning, and deep learning in liver transplantation", Journal of Hepatology, vol. 78, nº 6, p. 1216-1233, 2023.

[7] S. Nagai et al., "Use of neural network models to predict liver transplantation waitlist mortality", Liver Transpl, vol. 28, nº 7, p. 1133-1143, 2022.

[8] D. Bertsimas et al., "Development and validation of an optimized prediction of mortality for candidates awaiting liver transplantation", American Journal of Transplantation, vol. 19, nº 4, p. 1109-1118, 2019.

[9] A. Guo, N. R. Mazumder, D. P. Ladner, et R. E. Foraker, "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning", PLOS ONE, vol. 16, nº 8, p. e0256428, 2021.

[10] F. Kanwal et al., "Development, Validation, and Evaluation of a Simple Machine Learning Model to Predict Cirrhosis Mortality", JAMA Netw Open, vol. 3, nº 11, p. e2023780, 2020.

[11] R. A. Fisher, "The use of multiple measurements in taxonomic problems", Annals of Eugenics, vol. 7, no. 2, pp. 179-188, 1936.

[12] S. Ö. Arik et T. Pfister, "TabNet: Attentive Interpretable Tabular Learning", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, nº 8, Art. nº 8, 2021.

[13] L. Breiman, "Random Forests", Machine Learning, vol. 45, nº 1, pp. 5-32, 2001.

[14] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," NeurIPS, vol. 30, 2017.

[15] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?", NeurIPS, vol. 35, pp. 507-520, 2022.