Highlights

Interpretable prediction of mortality in liver transplant recipients based on machine learning

Xiao Zhang, Ricard Gavaldà, Jaume Baixeries

- A framework combining the advanced machine learning model with the SHapley Additive exPlanations (SHAP) was developed to interpret the association between a large number of factors and all-cause mortality in liver transplantation.
- The optimal feature set for the prediction of mortality in liver transplantation was identified by a BPSO-based wrapper model.
- New discoveries have been made in terms of the variation of the effect of features in different age groups and follow-up periods.
- It fills the deficiency in machine learning studies for predicting the risk of death after liver transplantation, especially the mortality risk in the long term.

Interpretable prediction of mortality in liver transplant recipients based on machine learning

Xiao Zhang[a,*], Ricard Gavaldà[b], Jaume Baixeries[a]

[a] *Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain*
[b] *Amalfi Analytics. On Leave from Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain*

Abstract

Background: Accurate prediction of the mortality of post-liver transplantation is an important but challenging task. It relates to optimizing organ allocation and estimating the risk of possible dysfunction. Existing risk scoring models, such as the Balance of Risk (BAR) score and the Survival Outcomes Following Liver Transplantation (SOFT) score, do not predict the mortality of post-liver transplantation with sufficient accuracy. In this study, we evaluate the performance of machine learning models and establish an explainable machine learning model for predicting mortality in liver transplant recipients.
Method: The optimal feature set for the prediction of the mortality was selected by a wrapper method based on binary particle swarm optimization (BPSO). With the selected optimal feature set, seven machine learning models were applied to predict mortality over different time windows. The best-performing model was used to predict mortality through a comprehensive comparison and evaluation. An interpretable approach based on machine learning and SHapley Additive exPlanations (SHAP) is used to explicitly explain the model's decision and make new discoveries.
Results: With regard to predictive power, our results demonstrate that the feature set selected by BPSO outperformed both the feature set in the existing risk score model (BAR score, SOFT score) and the feature set determined

[*] Corresponding author. Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain
Email address: xiao.zhang@upc.edu (Xiao Zhang)

by principal component analysis (PCA). The best performing machine learning model, extreme gradient boosting (XGBoost), was found to improve the Area Under a Curve (AUC) values for mortality prediction by 6.7%, 11.6%, and 17.4% at 3 months, 3 years, and 10 years, respectively, compared to SOFT score. The main predictors of mortality and their impact were discussed for different age groups and different follow-up periods.

Conclusions: Our analysis demonstrates that XGBoost can be an ideal method to assess the mortality risk in liver transplantation. In combination with the SHAP approach, the proposed framework provides a more intuitive and comprehensive interpretation of the predictive model, thereby allowing the clinician to better understand the decision-making process of the model and the impact of factors associated with mortality risk in liver transplantation.

Keywords: Liver transplant, Machine learning, Mortality prediction, Feature selection, Model interpretability

1. Introduction

Liver transplantation is a life-saving therapy for patients suffering from end-stage liver disease. It is estimated that in 2017, about 8,000 liver transplants were performed in the U.S. Meanwhile, approximately 11,500 people were registered on the liver transplant waiting list [1]. A large number of patients die while waiting for liver transplants since the demand for donated livers far exceeds the supply. Despite this reality, the number of transplants does not significantly increase over time. A lack of available organ donors and inefficient allocation of organs are two of the most pressing challenges. A more accurate prediction of mortality would strengthen confidence in performance after liver transplantation and facilitate the efficient allocation of donated livers. Besides, practitioners would be able to better assessment of the risk of early and late graft dysfunction.

The outcome of liver transplantation is influenced by a complex interaction between donor, recipient, and process factors. Several risk score models, devised to predict the post-transplant mortality risk have emerged. The three notable models are End-Stage Liver Disease (MELD) score [2], the Balance of Risk (BAR) score [3], and the Survival Outcomes Following Liver Transplantation (SOFT) score [4]. MELD score is widely used to prioritize organ allocation in practice, but it fails to predict post-transplant mortality risk well [5]. Numerous studies show MELD's poor prediction performance requires the development of a more accurate prediction model [6, 7]. A study by Rana et al. [4] found that the SOFT score offered a more accurate prediction of 3-month post-transplant survival for liver transplant recipients. The BAR score was later proposed to fulfill the short-term prediction for post-transplant survival with fewer features [8]. In contrast with the MELD score, which only takes into account a few factors of the recipient, the SOFT score and the BAR score consider both factors of the recipient as well as the donor.

Several studies have used these two risk score models to predict mortality risk after liver transplantation and have achieved better results than the MELD score [3, 9, 10]. A study by de Campos Junior et al. [11] used the dataset from a Brazilian transplant center, the area under the curve (AUC) value for the BAR score is 0.65 for 3-month mortality prediction, which is clearly unsatisfactory. A study by de Boer et al. [12] used the Scientific Registry of Transplant Recipients (SRTR) database to predict the mortality of liver transplant recipients and the AUC for the SOFT score and the BAR score for 1-year mortality prediction were 0.63 and 0.61, respectively, and 0.59 and 0.56 for mortality prediction at 5 years. These findings suggest that the predictive performance of the SOFT score and the BAR score is not sufficiently satisfactory, and therefore, there remains a need for a method that can provide a more accurate prediction of the mortality risk of liver transplantation, particularly for the long-term mortality risk. The SOFT score and the BAR score only consider a few clinical features. Factors that may have a relevant impact on the results, such as lifestyle, medical history, ethnicity, and socioeconomic factors, are not considered. It is therefore important to develop a more comprehensive mortality prediction model that takes into account a wider range of factors and to examine whether the inclusion of these factors improves the performance of mortality prediction in liver transplantation.

There has been an increase in the use of machine learning algorithms for the prediction of liver transplant mortality risk in recent years [13–16]. Clinical risk score models, such as the SOFT score and the BAR score, are primarily based on linear regression. As a result, they are limited in modeling nonlinear interactions between the predictors. Complex machine learning models, such as neural networks can perform more sophisticated modeling on data, usually resulting in better prediction performance. However, the use of these complex models also poses some challenges concerning the interpretability of the models. An interpretable machine learning model is critical to the medical domain because it can help physicians better understand the decision-making process and the impact of factors.

To solve these limitations, this study identified the optimal feature set for liver transplant mortality prediction from the United Network for Organ Sharing (UNOS) database by using a wrapper method that integrated logistic regression with binary particle swarm optimization (BPSO) algorithm. A combination of the SHapley Additive exPlanations (SHAP) framework and advanced machine learning models was used to interpret the model's decision-making process not only with regard to the importance of attributes but also concerning the individual predictions. The model was trained using different follow-up periods (3-month, 1-year, 3-year, 5-year, and 10-year mortality) and age groups (under 30 years, 30-49, 50-54, 55-65, and over 65). New findings have been derived by comparing the model applied to different recipient groups to identify the variation in the impact of attributes. To our knowledge, this is the first study that uses interpretable complex machine learning models for a comprehensive study of the association between a large number of attributes and all-cause mortality in liver transplantation.

In summary, the contributions of this paper include:

- To address the fact that current research has focused on short-term prediction, we examined the performance of different machine learning models in predicting long-term and short-term mortality risks, for which we established five distinct time frames.

- The useful features for the prediction of mortality risk were identified using the BPSO approach.

- The impact of the important features and individual prediction were explicitly explained by a framework combining the advanced machine learning model with the SHAP approach.

- New findings regarding the impact of features in different follow-up periods and age groups were presented to assist clinicians in better understanding the factors associated with mortality risk in liver transplantation.

The rest of the paper is organized as follows. Section 2 presents the comparative analysis of existing literature and our work. Section 3 provides details about the dataset preparation, the proposed framework for feature selection, the development of risk score models, and machine learning algorithms. Section 4 presents the results of the experiments in detail. Section 5 discusses the results, limitations, and future directions. Finally, the conclusion is presented in Section 6.

2. Comparative Analysis

Machine learning has proven to be a powerful tool in predicting patient mortality and assisting with medical decision-making [17–21]. There have been recent attempts to apply machine learning to electronic health records (EHR) to predict mortality for liver transplantation. Liu et al. [22] used 538 patients' blood test data before surgery to construct the model to predict the patients' survival within 30 days. They used different day ranges of blood for prediction and found that the highest AUC values were obtained when using blood data from patients on days 1-9 (AUC 0.7869). The random forest model was used to select significant features from the dataset and then these features were used to construct predictive models. Experimental results demonstrated that random forest achieves the best results compared to other machine learning algorithms. The study by Lau et al. [23] used two machine learning models, random forest and neural network, to predict graft failure within 30 days after liver transplantation. The study showed that the neural networks had the best prediction performance using the top 15 features selected by the random forest model (AUC, 0.835). Ershoff et al. [13] used a Deep neural network (DNN) to predict the 90-day post-transplant mortality. Though DNN achieves the highest AUC score (0.708), the improvement is limited in comparison with the SOFT score and the sensitivity value is even lower than the SOFT score.

Table 1 summarizes the main characteristics of the related literature. The existing research has focused primarily on the prediction of short-term mortality for post-liver transplantation. The prediction of long-term mortality risk is still not adequately investigated. The UNOS dataset was used in the studies of Ershoff et al. [13], Raji and Chandra [15] and Guijo-Rubio et al. [24]. Based on our study, the highest AUC values for the prediction of 3-month mortality and 1-year mortality are 0.717 and 0.681, respectively. Thus, in terms of prediction results, our prediction for 3-month mortality is higher than those of the Ershoff et al. [13] and Guijo-Rubio et al. [24], and for 1-year mortality is higher than those of the Guijo-Rubio et al. [24]. Raji and Chandra [15] obtained good prediction outcomes with an AUC of 0.9975, but

Table 1: comparative analysis of related work on liver transplant survival prediction using machine learning method

Related Work	Dataset	The best performing model	The time frame of prediction	Number of features used	Number of Patient	Main result
Liu et al. [22]	Chang Gung Memorial Hospital	Random Forest	30-day	13	538	AUC: 0.787; Sensitivity: 0.955; Sensitivity: 0.653
Lau et al. [23]	Austin Hospital Melbourne, Australia	Neural network	30-day	15	180	AUC: 0.835
Raji and Chandra [15]	UNOS	Multilayer Perceptron	3-month	27	383	AUC: 0.9975; Accuracy: 0.9974
Ershoff et al. [13]	UNOS	Neural network	3-month	93	57544	AUC: 0.708; F1-score: 0.212; Sensitivity: 0.348
Guijo-Rubio et al. [24]	UNOS	Logistic regression	3 months, 1-year, 2-year, 5-year	28	11570-34718	AUC: 0.633 (3 months), 0.631 (1-year), 0.629 (2-year), 0.654 (5-year)
Ayllón et al. [14]	King's College Hospital, UK	Neural network	3-month, 1 year	55	858	AUC: 0.94 (3-month), 0.82 (1-year)
Cruz-Ramirez et al. [25]	Spanish liver transplantation units	Neural network	3-month	64	1003	AUC: 0.566; Kappa: 0.0647; RMSE: 0.3207
Byrd et al. [26]	STAR	Gradient boosting	Same-day, 3-month	50	> 100K	AUC: 0.935 (Same-day) 0.834 (3-month)

only 300 samples were included in his study, whereas Ershoff et al. [13] and Guijo-Rubio et al. [24] have used much larger data samples. Additionally, AUC has been used as an evaluation metric across all studies. The obtained AUC value varied considerably across different datasets, which may be due to the different features or the differences in granularity of each dataset.

Many studies applied machine learning algorithms to improve the prediction performance for mortality prediction, however, they have not explored which factors in the model increase or reduce mortality risk [13–15, 25]. The lack of intuitive interpretation of machine learning models is one of the major barriers to the application of machine learning methods in liver transplantation decision-making [27, 28]. In our study, the association between a large number of attributes in liver transplantation and all-cause mortality is studied to offer clinicians a more intuitive understanding of the model's decision-making in the prediction of liver transplant mortality.

3. Methods

3.1. Dataset preparation

For this study, the data was sourced from the UNOS database. UNOS database collects all the data related to patient waiting lists, organ donation and matching, and transplantation under the administration of OPTN (The Organ Procurement and Transplantation Network). It is considered to be one of the most comprehensive and well-known data sources for organ transplantation. The UNOS database contains more than 290,000 liver transplant patient records in the US from October 1, 1987, to December 31, 2018.

In order to predict the mortality of post-liver transplantation across different time windows, we used samples from January 1, 2003, to December 31, 2012, except for the 10-year mortality prediction to ensure that we have adequate time for the acquisition of the follow-up outcomes. For the 10-year mortality risk prediction, the sample from January 1, 2003, to December 31, 2007, was used for prediction since more time to obtain follow-up information for each recipient is necessary. The characteristics of the samples that we used for 1-year mortality prediction are shown in Table 2.

Table 2: Characteristics of the study cohort for 1-year mortality (n=47401)

Variable	Survived recipients (n=41455)	Died recipients (n=5946)	Chi-square	P-value
Sex			11.64	< 0.001
Male	28106 (67.80%)	3899 (65.57%)		
Female	13349 (32.20%)	2047 (34.43%)		
Age	53.56 ± 9.99	54.81 ± 10.08		
Ethnicity			39.44	< 0.001
White	30120 (72.66%)	4224 (71.04%)		
Hispanic	5263 (12.70%)	742 (12.48%)		
Black	3675 (8.87%)	683 (11.49%)		
Asian	1918 (4.63%)	233 (3.92%)		
Other	479 (1.16%)	64 (1.08%)		
BMI (kg/m^2)	28.26 ± 5.62	28.20 ± 5.97		
Serum albumin (g/dL)	2.97 ± 0.72	2.90 ± 0.75		
Serum creatinine (mg/dl)	1.38 ± 1.05	1.68 ± 1.23		
Recipient medical condition			198.47	< 0.001
Not hospitalized	29929 (72.20%)	3243 (54.54%)		
Hospitalized not in ICU	6987 (16.85%)	1467 (24.67%)		
ICU	4539 (10.95%)	1236 (20.79%)		
Cold ischemia time (hours)	7.02 ± 3.24	7.37 ± 3.35		
Hepatitis C positive	17018 (41.47%)	2609 (44.64%)	25.62	< 0.001
Type 2 diabetes	6121 (14.77%)	953 (16.03%)	17.01	< 0.001
Previous malignancy	6152 (14.84%)	896 (15.07%)	1.70	0.19
Portal vein thrombosis	2974 (7.17%)	609 (10.24%)	14.37	< 0.001
Donor age	41.03 ± 17.05	43.59 ± 17.19		
Donor ethnicity			0.15	0.700
White	28097 (67.78%)	3910 (65.76%)		
Hispanic	5114 (12.34%)	826 (13.89%)		
Black	6785 (16.37%)	959 (16.13%)		
Asian	920 (2.22%)	168 (2.83%)		
Other	539 (1.30%)	83 (1.40%)		
Donor sex			4.15	0.04
Male	24792 (59.80%)	3473 (58.40%)		
Female	16663 (40.20%)	2473 (41.60%)		

[a] The numerical variable is denoted by mean ± standard deviation and the categorical variable is denoted by number (%).

[b] P value is determined by Chi square test for categorical variable.

We took a series of steps to clean and process the raw dataset so that the predictive model can have better performance (Fig. 1). For processing our study sample, these steps included the exclusion of samples with recipients under the age of 18, living donor transplants, multi-organ transplants, and recipients that were subsequently retransplanted. In addition, we also removed the samples that were lost to follow-up within the corresponding prediction window. As the number of recipients lost to follow-up varies across time windows, the number of study samples used for each time window also varies. The processing flow chart of the study samples is shown in Appendix A. For the processing of features, undesirable features were removed from the dataset, including identifier codes, dates, etc. Additionally, features collected after liver transplantation, as well as features that were missing in more than 90% of the observations, were also removed from the feature set.

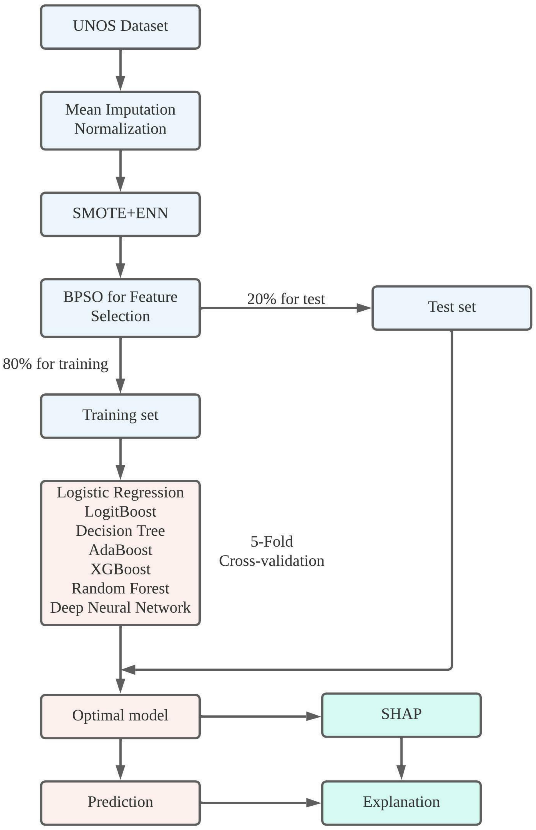

Figure 1: The overall flow for mortality prediction for liver transplant

As a binary event, the death of the recipient within each prediction window was extracted. Take the calculation of labels for 1-year mortality as an example. In the calculation of labels for 1-year mortality, the label is set to 1 if the patient dies within 365 days, and 0 if the patient lives beyond 365 days. Labels were calculated by this method for each prediction window. The dataset was then further processed by imputation, encoding, and min-max normalization. The synthesizing minority oversampling technology combined with edited nearest neighbors (SMOTE+ENN) was applied to deal with the imbalance between positive and negative categories. This sampling method was developed by Batista et al. [29]. It combines the capability of the synthesizing minority oversampling technology (SMOTE) to generate synthetic examples for minority categories with the capability of Wilson's Edited Nearest Neighbor Rule (ENN) to remove from both categories some observations that were identified as having different categories between the observed category and its K-nearest-neighbor majority category[30]. Numerical features with missing values were imputed by the mean. Next, the numerical features were normalized as large values might obscure the impact of other features with relatively smaller values. Concerning categorical features with more than 30 levels, we retained the levels that accounted for more than 95% of all the values for each feature and binned the rest into an 'other' category. These categorical features were then converted into dummy variables. After pre-processing, our data file contained 217 features.

3.2. Feature selection

In this study, a wrapper method was used to select the optimal feature set for predicting mortality, as it has better coverage of the search space and can detect interactions between features. The process of feature selection by wrapper method is shown in Fig. 2. In wrapper methods, a subset of features is evaluated using a machine learning classifier. The approach applies a search strategy to examine all potential subsets of features and evaluates them in accordance with the performance of the machine learning classifier.

3.2.1. Particle swarm optimization

A wrapper method based on BPSO was developed for feature selection for the prediction of mortality risk after liver transplantation. The particle swarm optimization (PSO) is an intelligent algorithm with biological inspirations in swarming behavior aimed at solving the optimization problem [31–33]. In our study, PSO was used to minimize both the classification error and the number of selected features, which contributes to the best model performance.

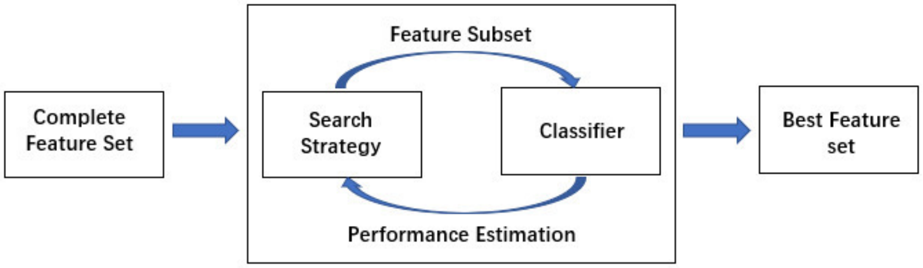

Figure 2: The process of wrapper feature selection

In the PSO technique, the population of particles is initially distributed randomly in the search space. Considering a swarm with K particles, each having its own velocity and position. The position of each particle is denoted by a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where D is the dimension of the search space. As each particle is searching for the best value in the search space, its velocity is represented as $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The best value so far for each particle (local best - p_{best}) and the best value so far for the whole group (global best - g_{best}) are used to update the position and velocity of each particle in the next step, so that the particle continues to search in the search space until the stopping condition is satisfied. The velocity and position of each particle are updated according to Eq. (1) and Eq. (2) [34].

$$\begin{aligned} v_{ij}(t+1) =& w * v_{ij}(t) + c_1 * r_1 * \left(p_{best_{ij}} - x_{ij}(t)\right) \\ &+ c_2 * r_2 * \left(g_{best_j} - x_{ij}(t)\right) \end{aligned} \tag{1}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \tag{2}$$

where $i = 1, 2, \cdots, K$ and $j = 1, 2, \cdots, D$. t represents the iteration number. Variable w represents the inertia weight, which determines how previous velocities affect current velocity. v_{ij} is the velocity of i_{th} particle in the swarm at the j_{th} position index of the particle, and is subject to a predetermined minimum velocity V_{min} and maximum velocity V_{max}. r_1 and r_2 are two random parameters in the range of $[0, 1]$. The variables c_1 and c_2 are the positive constants. $p_{best_{ij}}$ is the local optimum of the particle and g_{best_j} is the global optimum of the whole particle group in j^{th} dimensions. The stopping criterion can be a pre-defined fitness value or reaching the maximum number of iterations [35].

3.2.2. Feature selection based on BPSO

PSO methods were originally designed to solve optimization problems in search spaces with continuous values. However, feature selection is essentially an optimization problem in a discrete search space. The BPSO algorithm, resulting from some improvements to PSO, can solve optimization problems in discrete domain [36, 37].

In BPSO, two terms (1 and 0) are used to represent the position of the particles. Assuming that we have a particle with D dimensions, the position of the particle can be expressed as: $x = [x_1, x_2, x_3, \ldots, x_D]$ where $x_j \in 0, 1$. Suppose that we have a dataset with D features and each feature is expressed as a dimension of a particle. We can therefore interpret the binary array as simply turning a feature on or off and eventually finding the best position by implementing the BPSO. The probability distribution of the particle positions can be specified as 0 or 1 depending on the calculated value using the logistic function for the velocity values. The velocity is still updated in the same way as it does in PSO [38] and the position of the particle is determined using Eq. (3) and Eq. (4). rand() is a random number uniformly distributed in the range of $[0, 1]$.

$$X_{ij}(t+1) = \begin{cases} 1 & \text{if } \mathrm{rand}() < S(v_{ij}(t+1)) \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

$$S\left(v_{ij}(t+1)\right) = \frac{1}{1 + e^{-v_{ij}(t+1)}} \tag{4}$$

When considering the feature selection problem as an optimization problem, two issues must be taken into account, one is the number of features selected and the other is the classification accuracy. There is a trade-off between maximizing the prediction accuracy and minimizing the number of selected features. In designing the fitness function for optimization, it is necessary to consider both the number of selected features and the classification accuracy, as we want to select as few features as possible and achieve the best classification accuracy. The fitness function is designed using Eq. (5).

$$\text{Fitness} = \alpha \gamma_R(D) + \beta \frac{|R|}{|F|} \tag{5}$$

where $\gamma_R(D)$ is the error rate of the classifier. In our study, it denotes $1 - AUC$. $|R|$ is the number of features selected, $|F|$ is the total number of features, and $\alpha \in [1, 0]$, $\beta = (1 - \alpha)$ indicate the importance of the accuracy of the classifier and the number of features selected, respectively.

Due to the high computational cost of the wrapper method, a wrapper model that combined logistic regression and BPSO was used to select the important features from the UNOS database. The logistic regression model was chosen because it is easy to implement and computationally efficient compared to other types of models. The wrapper model was developed using the PySwarms and Scikit-learn library. Table 3 summarizes the BPSO parameters utilized in this study. The values of the BPSO parameters were determined by changing one parameter at a time while keeping the other parameters constant and then evaluating the fitness function. The parameters are fine-tuned based on the best fitness values and a reasonable computation time.

Table 3: The parameters for BPSO to use in this study

Parameter	Value
inertia weight w	0.8
acceleration constant c_1	1.8
acceleration constant c_2	1.8
Population size K	200
Dimension D	217
Iterations	150
fitness parameter α	0.92

3.3. Risk score models

The MELD, SOFT, and BAR scores are three notable risk score models that can be used to estimate the risk of mortality for liver transplant recipients as previously mentioned. The MELD score [2] is defined in Eq. (6) and it uses only creatinine, bilirubin, and the international normalized ratio (INR). The features used in the BAR score [3] and the SOFT score [4] are shown respectively in Table 4 and Table 5. Compared with the MELD score, the SOFT and BAR scores use more features, and also the prediction accuracy is higher. The prediction accuracy of these risk score models was then used to compare with different machine learning models in this study.

$$
\begin{aligned}
\text{MELD} =&9.57 * \log_e(\text{creatinine}) + 3.78 * \log_e(\text{bilirubin}) \\
&+ 11.20 \times \log_e(\text{INR}) + 6.43
\end{aligned} \tag{6}
$$

Table 4: The calculation of BAR Score

BAR Score		Assigned Points
MELD score at transplantation	$6 - 15$	0
	$16 - 25$	5
	$26 - 35$	10
	> 35	14
Retransplantation		4
Life support pretransplant		3
Recipient age (years)	≤ 40	0
	$> 40 - 60$	1
	> 60	3
Cold ischemia time (hours)	$0 - 6$	0
	$> 6 - 12$	1
	> 12	2
Donor age (years)	≤ 40	0
	> 40	1

3.4. Model development and performance metrics

Seven different machine learning models were employed in this study. Classifiers include one linear statistical method (linear regression), two tree-based methods (decision tree, random forests), three boosting methods (LogitBoost, AdaBoost, XGBoost), and DNN. These machine learning models have been used extensively in a variety of medical applications [39, 40]. Among them, random forest, LogitBoost, AdaBoost, and XGBoost are ensemble methods. Ensemble methods are supervised machine learning techniques that combine several base models in order to generate a stronger predictive model. The random forest constructs many trees on a subset of data and then combines the output of all the trees to make predictions. As a result, it reduces the problem of overfitting in decision trees, thus increasing its accuracy and making it suitable for large datasets. XGBoost is an optimized implementation of gradient boosting that provides a high level of

Table 5: The calculation of SOFT score

SOFT Score		Assigned Points
Age (years)	> 60	4
BMI	> 35	2
One previous transplant		9
Two previous transplants		14
Previous abdominal surgery		2
Albumin (g/dl)	< 2.0	2
Dialysis prior to transplantation		3
Intensive care unit pre-transplant		6
Admitted to hospital pre-transplant		3
MELD score	> 30	4
Life support pretransplant		9
Encephalopathy		2
Portal vein thrombosis		5
Ascites pretransplant		3
Portal bleed 48h pretransplant		6
Donor age (years)	$10 - 20$	-2
Donor age (years)	> 60	3
Donor cause of death from CVA		2
Donor creatinine (mg/dl)	> 1.5	2
National allocation		2
Cold ischemia time (hours)	$0 - 6$	-3

The feature "Portal bleed within 48h pretransplant" is not available in the UNOS database. The calculation is based on the remaining features.

computational efficiency and is suitable for processing large datasets. Several studies have shown that ensemble learning methods outperform single models in terms of prediction performance [41–43]. With powerful approximation capabilities for non-linear models, DNN models are more efficient in learning complex features and performing more intensive computational tasks. In this study, we will compare the performance of these machine learning models for predicting mortality after liver transplantation over a variety of time windows.

All the models were constructed in Python. Logistic regression, random forest, decision tree, and AdaBoost were implemented using the Scikit-learn library. XGBoost and LogitBoost were built using the XGBoost library and LogitBoost library in Python, respectively. The DNN used in this study is a feedforward network consisting of multiple fully connected layers and a sigmoid output function, which was developed using Pytorch with NVIDIA Quadro P1000 GPU acceleration. Based on the average cross-validation results, the optimal hyperparameters and architecture for each model were determined by performing the 5-fold cross-validation using the training set (80%). The hyperparameters of each model were optimized by maximizing the average AUC through a grid search method. Each model was then trained using the best hyperparameters on the training set (80%) before its performance was tested on a separate test set (20%).

4. Results

4.1. Feature selection

As discussed in Section 3.2, the wrapper method based on BPSO and logistic regression was used to select the significant features from the UNOS database. Comparatively, we applied the machine learning models to the complete feature set, the principal component analysis (PCA)-selected feature set, the feature set of the SOFT score, and the feature set of the BAR score. PCA is an unsupervised dimensionality reduction method that is used to identify critical original features of the principal components. In our study, PCA was applied to cover 95% of the data variance, and the new vector created after processing by PCA was applied to different classification algorithms. We compared the predictive power of different feature sets using five different metrics, including AUC, sensitivity, specificity, accuracy, and F1-score. Table 6 illustrates the mean results for all experiments using the given feature set.

The results indicate that the highest values for all metrics are found either in the complete feature set or in the BPSO feature set. Among them, the best AUC, specificity, accuracy, and F1-score were achieved by the complete set, whereas the BPSO feature set achieved the highest sensitivity. In terms of the classifier, none of the algorithms is optimal in all five metrics. XGBoost achieved the highest AUC and F1-score values on the complete set of features, while logistic regression achieved the optimal sensitivity on the BPSO feature set. Comparing XGBoost's performance with different feature sets, it can be seen that using the BPSO feature set is superior to using the feature set of PCA, SOFT score, and BAR score on almost all metrics, indicating the efficiency of the feature set selected by BPSO.

Table 6: Results for experiments with different feature sets using SMOTE-ENN and machine learning models for 1-year mortality prediction

Feature set	Algorithm	AUC	Sensitivity	Specificity	Accuracy	F1-score
Complete	SE-LR	0.678	0.548	0.706	0.688	0.286
	SE-LB	0.646	0.365	**0.800**	**0.750**	0.250
	SE-DT	0.645	0.434	0.760	0.723	0.263
	SE-AB	0.677	0.572	0.689	0.675	0.287
	SE-XGB	**0.685**	0.613	0.669	0.670	**0.289**
	SE-RF	0.677	0.662	0.595	0.602	0.275
	SE-DNN	0.666	0.565	0.676	0.663	0.277
BPSO	SE-LR	0.672	**0.683**	0.562	0.575	0.268
	SE-LB	0.656	0.523	0.659	0.643	0.251
	SE-DT	0.643	0.507	0.684	0.663	0.253
	SE-AB	0.671	0.533	0.717	0.696	0.286
	SE-XGB	0.681	0.602	0.680	0.668	0.287
	SE-RF	0.676	0.564	0.676	0.663	0.276
	SE-DNN	0.667	0.512	0.723	0.700	0.279
PCA	SE-LR	0.656	0.603	0.629	0.626	0.269
	SE-LB	0.543	0.473	0.605	0.590	0.208
	SE-DT	0.568	0.454	0.653	0.630	0.219
	SE-AB	0.638	0.613	0.589	0.592	0.255
	SE-XGB	0.641	0.523	0.665	0.649	0.254
	SE-RF	0.622	0.558	0.617	0.610	0.246
	SE-DNN	0.656	0.532	0.698	0.680	0.275
SOFT score	SE-LR	0.635	0.569	0.626	0.620	0.255
	SE-LB	0.624	0.506	0.676	0.656	0.251
	SE-DT	0.614	0.386	0.772	0.728	0.245
	SE-AB	0.639	0.509	0.700	0.678	0.265
	SE-XGB	0.651	0.532	0.675	0.658	0.262
	SE-RF	0.647	0.422	0.776	0.753	0.267
	SE-DNN	0.645	0.529	0.670	0.654	0.259
BAR score	SE-LR	0.630	0.470	0.723	0.694	0.260
	SE-LB	0.598	0.371	0.768	0.722	0.234
	SE-DT	0.614	0.482	0.674	0.652	0.240
	SE-AB	0.636	0.479	0.708	0.681	0.255
	SE-XGB	0.643	0.509	0.686	0.666	0.258
	SE-RF	0.640	0.518	0.664	0.648	0.251
	SE-DNN	0.637	0.613	0.578	0.583	0.251

Abbreviations: SE, SMOTE+ENN; LB, LogitBoost; LR, Logistic regression; DT, Decision tree; AB, AdaBoost; XGB, XGBoost; RF, Random forest; DNN, deep neural network.

42 features were selected out of 217 total features by BPSO. Unlike existing risk score models, such as the BAR score and the SOFT score, which contain relatively limited features, the BPSO feature set contains more extensive features including the patient's past lifestyle, the patient's history of chronic disease as well as some socioeconomic features. The importance of some features is also presented in the literature, but these features are not

considered by the existing risk score model. For instance, a study by Qiu et al. [44] illustrated that smoking exerts deleterious effects on transplant survival. a study by Correia et al. [45] presented that obesity and diabetes negatively affect long-term liver transplant survival and these factors should be monitored and assessed for the recipient. As shown in Fig. 3, the three feature sets have unique selections of features, suggesting there is no clinical consensus on what features are important for assessing post-transplantation mortality risk. In the meantime, it is worth noting that the BPSO selected feature set includes most of the features contained in the SOFT score and the BAR score. Four features are contained in both three feature sets, namely recipient's age, donor's age, cold ischemic time, and life support at transplantation.

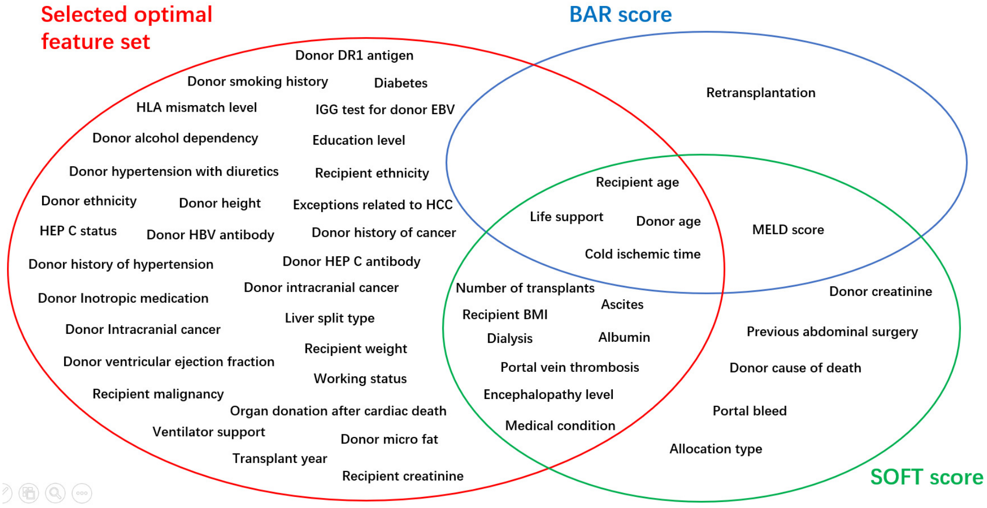

Figure 3: Venn diagram of selected feature set and clinical features used for BAR score and SOFT score.

4.2. Prediction of mortality risk

Machine learning methods were used to predict the mortality risk of post-transplant recipients and the results are shown in Table 7. AUC was calculated for different time frames, including 3 months, 1 year, 3 years, 5 years, and 10 years, which covers both short- and long-term mortality risks comprehensively. A comparison was made between three existing risk scores models and seven machine learning models. Bolded values in Table 7 indicate that the model outperformed the compared model for the particular prediction window.

Our results show that XGBoost outperforms all the other models, by obtaining the highest AUC in all of 3-month (0.717 ± 0.008), 1-year (0.681 ± 0.004), 3-year (0.662 ± 0.006), 5-year (0.660 ± 0.004), 10-year (0.674 ± 0.005) horizons. A very narrow set of features is used by the MELD score, resulting in a poor prediction performance for almost all time windows. Although the SOFT score is the best risk score model among all three models, it performs poorly in predicting long-term liver transplantation mortality, such as 3-year (0.593 ± 0.008), 5-year (0.580 ± 0.006) and 10-year (0.574 ± 0.005), with AUC values not exceeding 0.60. XGBoost improves the AUC for mortality prediction by 6.7%, 11.6%, and 17.4% for 3 months, 3 years, and 10 years, respectively, compared with the SOFT score. As a result, XGBoost is shown to provide consistent improvement in the prediction of post-transplantation mortality as compared to the other machine learning and risk score models. The NRI is a popular indicator that measures the improvement in the percentage of correctly classified cases. Table 8 shows the overall NRI improvement for XGBoost compared with the BAR score which was used as the baseline model. In terms of results, XGBoost achieved the highest overall NRI improvement (15.281%) in its prediction of 1-year mortality. For the 10-year mortality prediction, the overall NRI improvement is 6.860%.

Table 7: Performance of the predictive models for different time frames

Methods	AUC (Mean ± Std)				
	3-month	1-year	3-year	5-year	10-year
MELD	0.610 ± 0.006	0.579 ± 0.005	0.539 ± 0.005	0.523 ± 0.007	0.517 ± 0.005
SOFT	0.672 ± 0.009	0.632 ± 0.007	0.593 ± 0.008	0.580 ± 0.006	0.574 ± 0.005
BAR	0.657 ± 0.008	0.621 ± 0.006	0.578 ± 0.005	0.560 ± 0.007	0.555 ± 0.004
LR	0.703 ± 0.009	0.672 ± 0.005	0.652 ± 0.005	0.649 ± 0.009	0.660 ± 0.003
LB	0.693 ± 0.008	0.656 ± 0.006	0.644 ± 0.005	0.642 ± 0.007	0.661 ± 0.006
DT	0.687 ± 0.007	0.643 ± 0.006	0.643 ± 0.005	$0.623 \pm 0, 006$	0.637 ± 0.005
AB	0.715 ± 0.006	0.671 ± 0.004	$0.655 \pm 0, 003$	0.646 ± 0.004	0.667 ± 0.005
XGB	$\mathbf{0.717 \pm 0.008}$	$\mathbf{0.681 \pm 0.004}$	$\mathbf{0.662 \pm 0.006}$	$\mathbf{0.660 \pm 0.004}$	$\mathbf{0.674 \pm 0.005}$
RF	0.711 ± 0.008	0.676 ± 0.005	0.659 ± 0.008	0.656 ± 0.004	0.663 ± 0.005
DNN	0.705 ± 0.006	0.667 ± 0.005	0.652 ± 0.006	0.649 ± 0.005	0.665 ± 0.004
P-value	<0.001	<0.001	<0.001	<0.001	<0.001

[a] P-value is determined by applying one-way variance analysis to the AUC values of the seven models.

[b] XGBoost is significantly different from other models on the basis of comparisons of Least Significant Difference (LSD).

[c] Abbreviations: LB, LogitBoost; LR, Logistic regression; DT, Decision tree; AB, AdaBoost; XGB, XGBoost; RF, Random forest; DNN, deep neural network.

Table 8: The overall net reclassification improvement (NRI) for risk scoring models and XGBoost over the different follow-up periods

Methods	3-month	1-year	3-year	5-year	10-year
BAR Score			Baseline Model		
SOFT Score	5.514 (1.480)	7.011 (1.302)	0.041 (0.045)	0.152 (0.084)	3.976 (0.779)
XGBoost	13.212 (1.521)	15.281 (1.031)	8.994 (0.255)	7.000 (0.441)	6.860 (0.423)

Values represents the mean percentage improvement and values in brackets represent the standard deviations.

4.3. Feature importance

Assessing the importance of features is of great importance for clinical decision-making. Due to the superior prediction performance of XGBoost over other machine learning algorithms shown in Table 7, a method called SHAP [46] was used to assess which features are essential for XGBoost to generate predictions. The SHAP value measures the significance of the output resulting from the inclusion of a specific feature in all possible combinations other than that particular feature. In this study, the SHAP method was implemented in Python using the SHAP library to interpret the XGBoost model for the mortality prediction of post-liver transplantation from the following three aspects: 1. To identify the important features and explain how they influence the outcome. 2. To provide explanations of the individual samples and indicate the significant factors that contribute to making that prediction. 3. To identify important factors in short- and long-term liver transplant mortality and discover the risk factors associated with different age groups of recipients.

4.3.1. Discovering the feature importance in the model

Fig. 4 illustrates the top 15 risk factors for the 1-year mortality prediction of post-liver transplantation based on the SHAP value. The SHAP value provides a unified index that measures the impact of a certain feature in the model. The five most significant factors that are strongly associated with mortality after liver transplantation include the recipient's age, donor's age, serum creatinine, recipient's medical condition, and cold ischemic time. Note that in Fig. 4B, each dot is generated by the attribution value of a specific feature in the model of each patient, so that the SHAP values not only depict the magnitude of an individual feature's importance, but also its direction of effect. For example, younger donor age reduces the mortality risk in recipients, while lower serum albumin level increases the mortality risk. Fig. 5

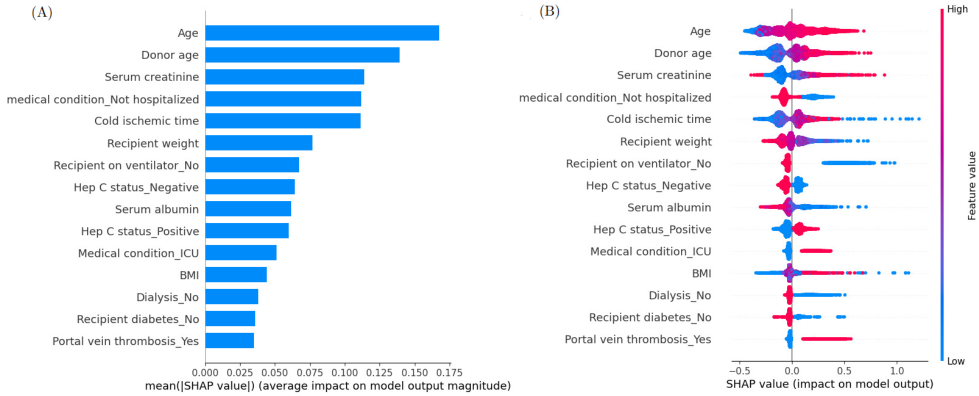

Figure 4: The ranking of feature importance is based on SHAP values for 1-year mortality prediction using the XGBoost model. (A): The mean absolute SHAP values are represented to demonstrate the importance of features. (B): The SHAP summary plot explained the relationship between a feature and mortality outcome. The input variables are presented in descending order of feature importance. In each feature, red represents high feature values, and blue represents low feature values. The positive SHAP value is indicative of increased mortality risk, while the negative SHAP value is indicative of a decreased mortality risk.

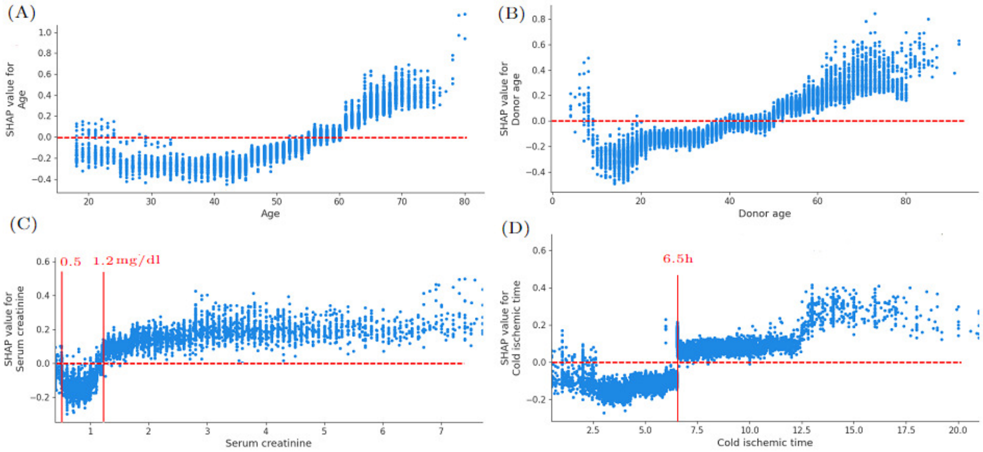

Figure 5: SHAP dependence plot of the XGBoost model for 1-year mortality prediction. (A): The effect of age on 1-year mortality. (B): The effect of donor age on 1-year mortality. (C): The effect of serum creatinine on 1-year mortality. (D): The effect of cold ischemic time on 1-year mortality. The dependence plot illustrates how individual features affect the output of the XGBoost model. The SHAP value above zero indicates an increased risk of mortality for a specific feature, while the value below zero indicates a decreased risk of mortality.

shows the main effects of some top risk factors, where we can roughly determine the thresholds and the extent of effects associated with these variables. In Fig. 5C, the SHAP value is below zero when serum creatinine is approximately between 0.5mg/dl and 1.2mg/dl, which indicates that the mortality risk for post-liver transplantation is decreased within this range. Fig. 5D shows that the mortality risk increases when the recipient's cold ischemic time exceeds 6.5 hours.

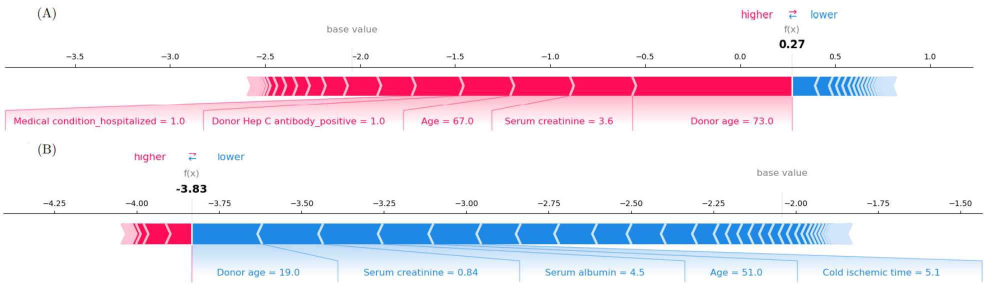

Figure 6: The explanation and analysis of the individual recipient based on the XGBoost model for 1-year mortality prediction. (A): The individualized explanation for the death of a transplant recipient within 1 year. (B): The individualized explanation for a recipient who is alive after 1 year. The values shown in bold indicate the predicted values for that individual recipient, and the base value is the average predicted value. Features shown in red indicate an increased risk of death, while features shown in blue indicate a decreased risk of death.

4.3.2. Interpretation of individual predictions

SHAP can be used to analyze individual predictions and show the impact of each feature on mortality risk for liver transplantation. Fig. 6 provides two examples to illustrate the interpretation of the model at the individual level. The first recipient (Fig. 6A) died within 1 year while the second recipient (Fig. 6B) survived after 1 year. The XGBoost model predicted that the 1-year probability of death for the first recipient was 0.568 and 0.021 for the second recipient. It can be seen from Fig. 6A that the predicted value for the mortality risk is 0.27, which is higher than the base value (average of output values). The red arrows indicate factors associated with an increased risk of mortality. The top 5 significant factors determining the prediction of the mortality risk in this recipient are donor age, serum creatinine, age, positive hepatitis C antibody, and hospitalization before transplantation. For the second recipient, the predicted SHAP value for the mortality risk is -3.83, which is lower than the base value. The five most important factors contributing to making this prediction are donor age, serum creatinine, serum albumin, age, and cold ischemic time.

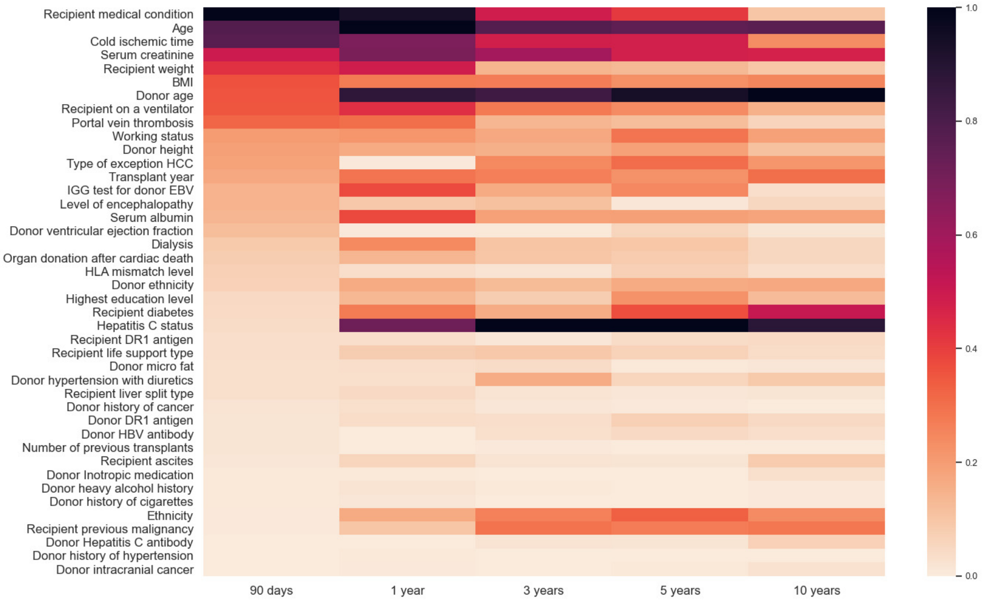

Figure 7: The importance of the selected features for mortality prediction over different time frames.

4.3.3. The interpretation of features over different time frames

To interpret features that have different effects on short-term and long-term mortality risk prediction, we combined different levels of each categorical feature in the selected feature set. The SHAP values were calculated for each time frame, and the relative importance of features was determined by scaling the absolute value of the SHAP value of the feature to a range between 0 and 1. The importance of features over the different time frames was visualized with a heatmap, as shown in Fig. 7. We can see that some factors contribute less to short-term mortality prediction than long-term mortality prediction, such as donor's age, diabetes, recipient's previous malignancy, and hepatitis C status, while recipient's medical condition, cold ischemic time, recipient weight, BMI, ventilator usage, and portal vein thrombosis contribute more to short-term mortality prediction than long-term mortality prediction. The result illustrates the variation in the significance of each feature for a short and long-term mortality prediction and provides a more comprehensive interpretation of mortality risk factors associated with liver transplantation.

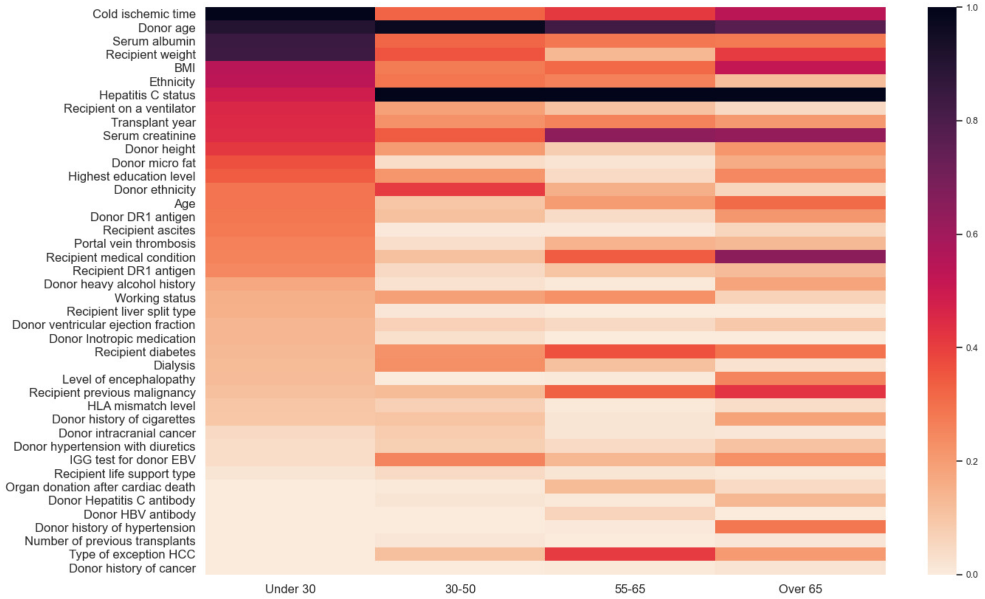

Figure 8: The importance of the selected features for 5-year mortality prediction in different age groups.

4.3.4. The interpretation of features in different age groups

The Fig. 8 shows the relative importance of input features for 5-year mortality prediction in different age groups ($<$ 40, 30-50, 55-65, $>$ 65). As can be seen, some features are more vital for the elder subgroups compared to the younger subgroups such as hepatitis C status, serum creatinine, portal vein thrombosis, recipient medical condition, diabetes, previous malignancy, and donor hypertension history. Results indicate that the mortality risk for the elder subgroups is more influenced by chronic diseases and geriatric conditions. However, most of these diseases are not taken into consideration by the existing risk score models. In addition, some features are more important for the younger subgroups compared to the elder subgroups. The top five most important features for the under-30 subgroup are cold ischemic time, donor age, serum albumin, recipient weight, and BMI. Among these features, cold ischemic time, serum albumin, and recipient weight are significantly more critical in predicting mortality risk in the under-30 subgroup than in the elder subgroups and therefore deserve attention and further exploration.

5. Discussion

Accurate mortality prediction of post-liver transplantation plays an important role in the prediction of the risk of early and late graft dysfunction as well as providing decision support for optimizing organ allocation. The purpose of this study was to assess the risk of death after liver transplantation and provide an extensive interpretation of the model's decision-making and impact of a large number of variables by using machine learning algorithms. More specifically, we will discuss our results in the following sections.

Selection of the optimal feature set. A wrapper model was developed for selecting an optimal feature set to predict mortality risk for liver transplantation. 42 features were selected from the UNOS database, covering various aspects of the recipient, such as the recipient's lifestyle, medical history, and socioeconomic status. We found that the machine learning model trained with the selected optimal feature set has a much higher prediction accuracy than the same model trained with the features of the existing risk scores. Consequently, the results demonstrate the necessity of incorporating a broader range of significant features into the predictive model to enhance prediction performance.

Comparison of machine learning models and risk scores. Our results indicate that XGBoost outperforms other models in all time windows. Existing studies have shown that the XGBoost model has superior predictive performance in a variety of medical applications due to its architecture suitable for training with minimal features and ease of handling missing values [47, 48]. Our results suggested that XGBoost has good interpretability and can be considered as a suitable machine learning method for mortality risk assessment for liver transplantation and aiding clinical decision-making. In addition, we found that DNN models did not outperform XGBoost and random forest, indicating that more sophisticated algorithms are not always superior to other simpler machine learning models. With superior nonlinear modeling capabilities, DNN models may be better suited to analyzing large-scale, heterogeneous, and high-granularity datasets that are not available in the UNOS dataset.

Short- and long-term prediction performance and risk factors determination Most studies on liver transplant mortality prediction have focused only on short-term or long-term risk, but have not shown how the impact of features and performance of the model changes over time. In this study, five different time windows were chosen for comparison. From the experimental result, the predictive performance of all three risk scores decreased significantly as the time windows became longer, with particularly poor predictions of 10-year mortality. However, the prediction accuracy of most machine learning models for the 3- and 5-year mortality risks were found to be slightly lower than that of the 10-year mortality risk. The prediction accuracy of XGBoost is significantly higher than that of the existing three risk score models across all time windows.

Interpretation of the machine learning model In the medical field, it is essential to intuitively interpret the machine learning model, which has always been a challenging task. Compared to previous studies, our study provides a more comprehensive interpretation of the machine learning model based on the SHAP values in the context of liver transplantation. In contrast to existing studies that only showed the importance of features, we also described how these features influence the model's decision-making, including whether the feature increases or decreases the mortality risk of liver transplantation and at what threshold the feature may increase or decrease the risk of death. Clinicians can then better understand the decision-making process of the model and the impact of specific variables. Additionally, Some interesting findings were also presented in this study. For instance, for the elder subgroups, we found that geriatric and chronic diseases have a greater impact on mortality risk for liver transplantation compared to the younger subgroups. This deserves further exploration and research.

Limitations The study has several limitations. First, we performed our study using the UNOS database, which is one of the largest organ transplantation databases in the world. To increase the generalizability of these findings, further external validation of our model using other databases is necessary. In future studies, liver transplantation databases from other regions or countries may serve as an external validation of our model. Besides, the dataset used in this study is structured. Several studies have presented that the combination of clinical text notes with structured data enhances the prediction accuracy in some clinical application areas [49, 50]. Future work may include the integration of clinical text notes with structured EHR data to improve our predictive performance. Third, due to the different recording systems used by different transplant centers, the predictive performance of the model may be affected by distinct application environments. Although it may not be ideal to directly export a well-trained model from one center to another, it is still perfectly practicable to tailor the approach to each transplant center.

6. Conclusion

In contrast to the existing risk score models, which incorporate only a limited number of clinical features, this study considered a broader range of factors, including the patient's lifestyle, medical history, and socioeconomic status. A wrapper model that combines BPSO with logistic regression classifier was used to identify the optimal set of features from the feature-rich UNOS database. Five different time frames were set up to compare the performance of different machine learning algorithms in predicting the mortality of liver transplant recipients over the short, medium, and long term. Our results demonstrate that XGBoost outperformed other machine learning models as well as existing risk score models in all prediction windows and can therefore be considered an ideal method for assessing the mortality risk following liver transplantation. An approach combining machine learning and SHAP was used to explore the association between a large number of attributes in liver transplantation and all-cause mortality. Based on this approach, the individual predictions, the impact of attributes and their corresponding thresholds, and variations in the impact of attributes with respect to follow-up periods and age groups were interpreted explicitly. This comprehensive interpretable paradigm can also be applied to the assessment of the risk of other diseases and provide better interpretations.

Conflict of Interest Statement

We declare that we have no conflict of interest.

Acknowledgement

The authors would like to thank UNOS for providing the dataset as of December 31, 2018. The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government.

Availability of software

Our software code is available at the following web address: `https://github.com/zhangxiaowbl/LT`

References

[1] Liver transplant, `https://www.mayoclinic.org/tests-procedures/liver-transplant/about/pac-20384842`, 2021. [Online; accessed 02-December-2021].

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

[3] P. Dutkowski, C. E. Oberkofler, K. Slankamenac, M. A. Puhan, E. Schadde, B. Müllhaupt, A. Geier, P. A. Clavien, Are there better guidelines for allocation in liver transplantation?: A novel score targeting justice and utility in the model for end-stage liver disease era, Annals of surgery 254 (2011) 745–754.

[4] A. Rana, M. Hardy, K. Halazun, D. Woodland, L. Ratner, B. Samstein, J. Guarrera, R. Brown Jr, J. Emond, Survival outcomes following liver transplantation (soft) score: a novel method to predict patient survival following liver transplantation, American Journal of Transplantation 8 (2008) 2537–2546.

[5] K. B. Klein, T. D. Stafinski, D. Menon, Predicting survival after liver transplantation based on pre-transplant meld score: a systematic review of the literature, PloS one 8 (2013) e80661.

[6] P. H. Hayashi, L. Forman, T. Steinberg, T. Bak, M. Wachs, M. Kugelmas, G. T. Everson, I. Kam, J. F. Trotter, Model for end-stage liver disease score does not predict patient or graft survival in living donor liver transplant recipients, Liver transplantation 9 (2003) 737–740.

[7] S. Habib, B. Berk, C.-C. H. Chang, A. J. Demetris, P. Fontes, I. Dvorchik, B. Eghtesad, A. Marcos, A. O. Shakil, Meld and prediction of post–liver transplantation survival, Liver transplantation 12 (2006) 440–447.

[8] P. Dutkowski, A. Schlegel, K. Slankamenac, C. E. Oberkofler, R. Adam, A. K. Burroughs, E. Schadde, B. Müllhaupt, P.-A. Clavien, The use of fatty liver grafts in modern allocation systems: risk assessment by the balance of risk (bar) score, Annals of surgery 256 (2012) 861–869.

[9] A. Rana, T. Jie, M. Porubsky, S. Habib, H. Rilo, B. Kaplan, A. Gruessner, R. Gruessner, The survival outcomes following liver transplantation (soft) score: validation with contemporaneous data and stratification of high-risk cohorts, Clinical transplantation 27 (2013) 627–632.

[10] H. Schrem, A.-L. Platsakis, A. Kaltenborn, A. Koch, C. Metz, M. Barthold, C. Krauth, V. Amelung, F. Braun, T. Becker, et al., Value and limitations of the bar-score for donor allocation in liver transplantation, Langenbeck's archives of surgery 399 (2014) 1011–1019.

[11] I. D. de Campos Junior, R. S. B. Stucchi, E. Y. Udo, I. d. F. S. F. Boin, Application of the bar score as a predictor of short-and long-term survival in liver transplantation patients, Hepatology international 9 (2015) 113–119.

[12] J. D. de Boer, H. Putter, J. J. Blok, I. P. Alwayn, B. van Hoek, A. E. Braat, Predictive capacity of risk models in liver transplantation, Transplantation direct 5 (2019).

[13] B. D. Ershoff, C. K. Lee, C. L. Wray, V. G. Agopian, G. Urban, P. Baldi, M. Cannesson, Training and validation of deep neural networks for the

prediction of 90-day post-liver transplant mortality using unos registry data, in: Transplantation proceedings, volume 52, Elsevier, 2020, pp. 246–258.

[14] M. D. Ayllón, R. Ciria, M. Cruz-Ramírez, M. Pérez-Ortiz, I. Gómez, R. Valente, J. O'Grady, M. de la Mata, C. Hervás-Martínez, N. D. Heaton, et al., Validation of artificial neural networks as a methodology for donor-recipient matching for liver transplantation, Liver Transplantation 24 (2018) 192–203.

[15] C. Raji, S. V. Chandra, Graft survival prediction in liver transplantation using artificial neural network models, Journal of computational science 16 (2016) 72–78.

[16] D. Guijo-Rubio, P. A. Gutiérrez, C. Hervás-Martínez, Machine learning methods in organ transplantation, Current Opinion in Organ Transplantation 25 (2020) 399–405.

[17] M. Nazari, I. Shiri, H. Zaidi, Radiomics-based machine learning model to predict risk of death within 5-years in clear cell renal cell carcinoma patients, Computers in Biology and Medicine 129 (2021) 104135.

[18] H.-C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strøm, P. J. Chmura, M. Heimann, L. Dybdahl, et al., Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records, The Lancet Digital Health 2 (2020) e179–e191.

[19] K. Wang, J. Tian, C. Zheng, H. Yang, J. Ren, Y. Liu, Q. Han, Y. Zhang, Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and shap, Computers in Biology and Medicine 137 (2021) 104813.

[20] A. Awad, M. Bader-El-Den, J. McNicholas, J. Briggs, Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach, International journal of medical informatics 108 (2017) 185–195.

[21] Y. Gao, G.-Y. Cai, W. Fang, H.-Y. Li, S.-Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P.-F. Cui, et al., Machine learning based early warning

system enables accurate mortality risk prediction for covid-19, Nature communications 11 (2020) 1–10.

[22] C.-L. Liu, R.-S. Soong, W.-C. Lee, G.-W. Jiang, Y.-C. Lin, predicting short-term survival after liver transplantation using machine learning, Scientific reports 10 (2020) 1–10.

[23] L. Lau, Y. Kankanige, B. Rubinstein, R. Jones, C. Christophi, V. Muralidharan, J. Bailey, Machine-learning algorithms predict graft failure after liver transplantation, Transplantation 101 (2017) e125.

[24] D. Guijo-Rubio, J. Briceño, P. A. Gutiérrez, M. D. Ayllón, R. Ciria, C. Hervás-Martínez, Statistical methods versus machine learning techniques for donor-recipient matching in liver transplantation, Plos one 16 (2021) e0252068.

[25] M. Cruz-Ramirez, C. Hervas-Martinez, J. C. Fernandez, J. Briceno, M. De La Mata, Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks, Artificial intelligence in medicine 58 (2013) 37–49.

[26] J. Byrd, S. Balakrishnan, X. Jiang, Z. C. Lipton, Predicting mortality in liver transplant candidates, in: Explainable AI in Healthcare and Medicine, Springer, 2021, pp. 321–333.

[27] L. R. Wingfield, C. Ceresa, S. Thorogood, J. Fleuriot, S. Knight, Using artificial intelligence for predicting survival of individual grafts in liver transplantation: a systematic review, Liver Transplantation 26 (2020) 922–934.

[28] A. Spann, A. Yasodhara, J. Kang, K. Watt, B. Wang, A. Goldenberg, M. Bhat, Applying machine learning in liver disease and transplantation: a comprehensive review, Hepatology 71 (2020) 1093–1105.

[29] G. E. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD explorations newsletter 6 (2004) 20–29.

[30] K. Wang, J. Tian, C. Zheng, H. Yang, J. Ren, C. Li, Q. Han, Y. Zhang, Improving risk identification of adverse outcomes in chronic heart failure using smote+ enn and machine learning, Risk Management and Healthcare Policy 14 (2021) 2453.

[31] R. Poli, J. Kennedy, T. Blackwell, Particle swarm optimization, Swarm intelligence 1 (2007) 33–57.

[32] D. Wang, D. Tan, L. Liu, Particle swarm optimization algorithm: an overview, Soft Computing 22 (2018) 387–408.

[33] S. M. Vieira, L. F. Mendonça, G. J. Farinha, J. M. Sousa, Modified binary pso for feature selection using svm applied to mortality prediction of septic patients, Applied Soft Computing 13 (2013) 3494–3504.

[34] B. S. G. de Almeida, V. C. Leite, Particle swarm optimization: A powerful technique for solving engineering problems, Swarm Intelligence-Recent Advances, New Perspectives and Applications (2019).

[35] M. Mafarja, R. Jarrar, S. Ahmad, A. A. Abusnaina, Feature selection using binary particle swarm optimization with time varying inertia weight strategies, in: Proceedings of the 2nd International Conference on Future Networks and Distributed Systems, 2018, pp. 1–9.

[36] H. Nezamabadi-pour, M. Rostami-Shahrbabaki, M. Maghfoori-Farsangi, Binary particle swarm optimization: challenges and new solutions, CSI J Comput Sci Eng 6 (2008) 21–32.

[37] J. Wei, R. Zhang, Z. Yu, R. Hu, J. Tang, C. Gui, Y. Yuan, A bpso-svm algorithm based on memory renewal and enhanced mutation mechanisms for feature selection, Applied Soft Computing 58 (2017) 176–192.

[38] A. Ismail, A. P. Engelbrecht, Self-adaptive particle swarm optimization, in: Asia-Pacific conference on simulated evolution and learning, Springer, 2012, pp. 228–237.

[39] A. Ogunleye, Q.-G. Wang, Xgboost model for chronic kidney disease diagnosis, IEEE/ACM transactions on computational biology and bioinformatics 17 (2019) 2131–2140.

[40] X. Su, Y. Xu, Z. Tan, X. Wang, P. Yang, Y. Su, Y. Jiang, S. Qin, L. Shang, Prediction for cardiovascular diseases based on laboratory data: an analysis of random forest model, Journal of Clinical Laboratory Analysis 34 (2020) e23421.

[41] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, N. Khovanova, Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation, Biomedical Signal Processing and Control 52 (2019) 456–462.

[42] P. Chen, C. Pan, Diabetes classification model based on boosting algorithms, BMC bioinformatics 19 (2018) 1–9.

[43] N. I. Khan, T. Mahmud, M. N. Islam, S. N. Mustafina, Prediction of cesarean childbirth using ensemble machine learning methods, in: Proceedings of the 22nd international conference on information integration and web-based applications & services, 2020, pp. 331–339.

[44] F. Qiu, P. Fan, G. D. Nie, H. Liu, C.-L. Liang, W. Yu, Z. Dai, effects of cigarette smoking on transplant survival: extending or shortening it?, Frontiers in immunology 8 (2017) 127.

[45] I. M. Correia, L. O. Rego, A. S. Lima, Post-liver transplant obesity and diabetes, Current Opinion in Clinical Nutrition & Metabolic Care 6 (2003) 457–460.

[46] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.

[47] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, N. M. Luscombe, Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, PloS one 13 (2018) e0202344.

[48] P. C. Austin, D. S. Lee, E. W. Steyerberg, J. V. Tu, Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods?, Biometrical journal 54 (2012) 657–673.

[49] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: a deep learning approach, BMC medical informatics and decision making 20 (2020) 1–11.

[50] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani, Natural language processing of clinical notes on chronic diseases: systematic review, JMIR medical informatics 7 (2019) e12239.

Appendix A. The chart for number of data used

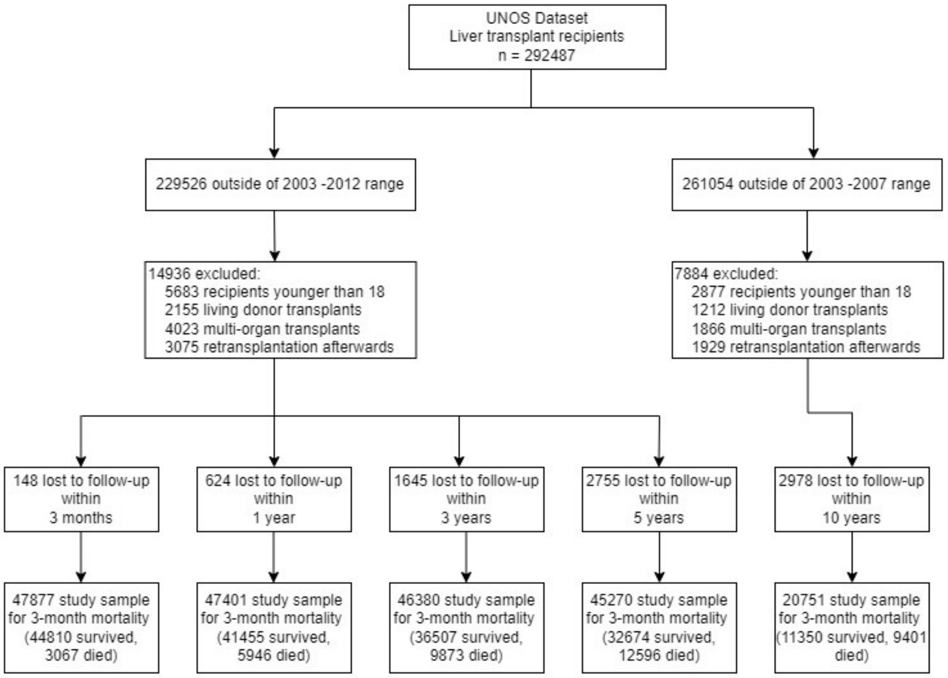
UNOS Dataset
Liver transplant recipients
n = 292487
229526 outside of 2003 -2012 range
261054 outside of 2003 -2007 range
14936 excluded:
5683 recipients younger than 18
2155 living donor transplants
4023 multi-organ transplants
3075 retransplantation afterwards
7884 excluded:
2877 recipients younger than 18
1212 living donor transplants
1866 multi-organ transplants
1929 retransplantation afterwards
148 lost to follow-up within 3 months
624 lost to follow-up within 1 year
1645 lost to follow-up within 3 years
2755 lost to follow-up within 5 years
2978 lost to follow-up within 10 years
47877 study sample for 3-month mortality (44810 survived, 3067 died)
47401 study sample for 3-month mortality (41455 survived, 5946 died)
46380 study sample for 3-month mortality (36507 survived, 9873 died)
45270 study sample for 3-month mortality (32674 survived, 12596 died)
20751 study sample for 3-month mortality (11350 survived, 9401 died)

Figure A.1: The flow chart of the study cohort. The study cohort is not the same for each time window because the number of recipients lost to follow-up at different periods is different. As the 10-year mortality analysis requires a longer period of time to collect follow-up information from the recipients, five years of data from 2003 to 2007 were used.

Appendix B. The feature set selected by BPSO-LR

Appendix C. The number and percentage of missing value

Appendix D. Characteristics of the study cohort

Appendix E. Hyperparameters for the classifiers

Table B.1: Description of the feature set selected by BPSO-LR model

#	Feature name	Feature description
1	AGE	Recipient age
2	AGE_DON	Donor age
3	ALBUMIN_TX	Recipient's serum albumin at transplant
4	ALCOHOL_HEAVY_DON	Donor had history of alcohol dependency
5	AMIS	HLA mismatch level
6	ASCITES_TX	Recipient had ascites at transplant
7	BMI_CALC	BMI of recipient at transplant
8	COLD_ISCH	Total cold ischemic time
9	CONTIN_CIG_DON	The smoking history of the donor
10	CREAT_TX	Recipient's serum creatinine at transplant
11	DDR1	Donor's DR1 antigen
12	DIAB	Recipient had diabetes at registration
13	DIURETICS_DON	Donor treated hypertension with diuretics
14	EBV_IGG_CAD_DON	Result of IGG test for donor's EBV
15	EDUCATION	The highest education level of the recipient
16	ENCEPH_TX	Recipient's level of encephalopathy at transplant
17	ETHCAT	Recipient's ethnicity
18	ETHCAT_DON	Donor's ethnicity
19	EXC_HCC	Type of exception relative to HCC: HCC/HBL/NON-HCL
20	FINAL_DIALYSIS_PRIOR_WEEK	Patient dialysis twice in prior week
21	HBV_CORE_DON	Donor HBV core antibody
22	HCV_SEROSTATUS	Recipient Hepatitis C virus status
23	HEP_C_ANTI_DON	Hepatitis C virus antibody of Donor
24	HGT_CM_DON_CALC	Donor's height
25	HIST_CANCER_DON	Donor's history of cancer
26	HIST_HYPERTENS_DON	Donor's history of hypertension
27	INIT_WGT_KG	Recipient's weight at waiting list
28	INOTROP_SUPPORT_DON	Donor took inotropic medications at procurement
29	INTRACRANIAL_CANCER_DON	Donor has intracranial cancer at procurement
30	LITYP	The type of liver graft was whole or split
31	LV_EJECT_METH_DON	Left ventricular ejection fraction
32	MALIG	Recipient had any malignancy previously
33	MED_COND_TRR	Medical condition of the recipient prior to transplant
34	MICRO_FAT_LI_DON	Donor's micro fat (%)
35	NON_HRT_DON	Donor's organ was donated after cardiac death
36	NUM_PREV_TX	The number of previous transplants
37	ON_VENT_TRR	Recipient was on ventilator at transplant
38	OTH_LIFE_SUP_TRR	Other life support type for recipient at transplant
39	PORTAL_VEIN_TRR	Recipient suffered from portal vein thrombosis
40	RDR1	Recipient's DR1 antigen
41	TX_YEAR	Transplant year
42	WORK_INCOME_TCR	Recipient was working at registration

Table C.1: Statistics of missing values (n=47401)

#	Feature name	Number of missing value (Percentage)
1	AGE	0 (0%)
2	AGE_DON	0 (0%)
3	ALBUMIN_TX	7 (0.01%)
4	ALCOHOL_HEAVY_DON	5882 (12.41%)
5	AMIS	27478 (57.97%)
6	ASCITES_TX	0 (0%)
7	BMI_CALC	32 (0.07%)
8	COLD_ISCH	2516 (5.31%)
9	CONTIN_CIG_DON	33752 (71.21%)
10	CREAT_TX	49 (0.10%)
11	DDR1	1250 (2.64%)
12	DIAB	14 (0.03%)
13	DIURETICS_DON	31489 (0.66%)
14	EBV_IGG_CAD_DON	13770 (29.05%)
15	EDUCATION	16 (0.03%)
16	ENCEPH_TX	0 (0%)
17	ETHCAT	0 (0%)
18	ETHCAT_DON	0 (0%)
19	EXC_HCC	0 (0%)
20	FINAL_DIALYSIS_PRIOR_WEEK	0 (0%)
21	HBV_CORE_DON	0 (0%)
22	HCV_SEROSTATUS	524 (1.11%)
23	HEP_C_ANTI_DON	0 (0%)
24	HGT_CM_DON_CALC	1 (0%)
25	HIST_CANCER_DON	1 (0%)
26	HIST_HYPERTENS_DON	1 (0%)
27	INIT_WGT_KG	2 (0%)
28	INOTROP_SUPPORT_DON	30 (0.06%)
29	INTRACRANIAL_CANCER_DON	1 (0%)
30	LITYP	2 (0%)
31	LV_EJECT_METH_DON	18166 (38.32%)
32	MALIG	0 (0%)
33	MED_COND_TRR	0 (0%)
34	MICRO_FAT_LI_DON	35270 (74.41%)
35	NON_HRT_DON	1 (0%)
36	NUM_PREV_TX	0 (0%)
37	ON_VENT_TRR	0 (0%)
38	OTH_LIFE_SUP_TRR	0 (0%)
39	PORTAL_VEIN_TRR	0 (0%)
40	RDR1	27354 (57.71%)
41	TX_YEAR	0 (0%)
42	WORK_INCOME_TCR	8223 (17.35%)

Table D.1: Characteristics of the study cohort for 3-month mortality (n=47877)

Variable	Survived recipients (n=44810)	Died recipients (n=3067)	Chi-square	P-value
Sex			28.03	< 0.001
Male	30382 (67.80%)	1937 (63.16%)		
Female	14428 (32.20%)	1130 (36.84%)		
Age	53.67 ± 9.98	54.27 ± 10.59		
Ethnicity			10.79	0.001
White	32443 (72.40%)	2184 (71.21%)		
Hispanic	5736 (12.80%)	386 (12.59%)		
Black	4053 (9.04%)	334 (10.89%)		
Asian	2063 (4.60%)	131 (4.27%)		
Other	515 (1.16%)	32 (1.04%)		
BMI (kg/m^2)	28.23 ± 5.62	28.48 ± 6.16		
Serum albumin (g/dL)	2.97 ± 0.72	2.90 ± 0.76		
Serum creatinine (mg/dl)	1.40 ± 1.06	1.75 ± 1.27		
Recipient medical condition			301.20	< 0.001
Not hospitalized	32035 (71.49%)	1466 (47.80%)		
Hospitalized not in ICU	7696 (17.17%)	600 (19.56%)		
ICU	5079 (11.34%)	1001 (32.64%)		
Cold ischemia time (hours)	7.04 ± 3.24	7.49 ± 3.33		
Hepatitis C positive	18735 (42.26%)	1105 (36.57%)	5.07	0.02
Type 2 diabetes	6659 (14.87%)	477 (15.55%)	10.02	0.002
Previous malignancy	6698 (14.95%)	417 (13.60%)	4.15	0.04
Portal vein thrombosis	3248 (7.25%)	367 (11.97%)	37.65	< 0.001
Donor age	41.24 ± 17.07	42.73 ± 17.30		
Donor ethnicity			0.51	0.47
White	30254 (67.52%)	2043 (66.61%)		
Hispanic	7329 (16.36%)	476 (15.52%)		
Black	5617 (12.54%)	424 (13.82%)		
Asian	1027 (2.28%)	77 (2.51%)		
Other	583 (1.30%)	47 (1.53%)		
Donor sex			3.96	0.05
Male	26801 (59.81%)	1778 (57.97%)		
Female	18009 (40.19%)	1289 (42.03%)		

[a] The numerical variable is denoted by mean ± standard deviation and the categorical variable is denoted by number (%).

[b] P value is determined by Chi square test for categorical variable.

Table D.2: Characteristics of the study cohort for 3-year mortality (n=46380)

Variable	Survived recipients (n=36507)	Died recipients (n=9873)	Chi-square	P-value
Sex			0.01	0.93
Male	24643 (67.50%)	6659 (67.45%)		
Female	11864 (32.50%)	3214 (32.55%)		
Age	53.51 ± 9.97	54.78 ± 9.93		
Ethnicity			110.61	< 0.001
White	26684 (73.09%)	6994 (70.84%)		
Hispanic	4593 (12.58%)	1200 (12.15%)		
Black	3103 (8.50%)	1194 (12.29%)		
Asian	1700 (4.66%)	382 (3.87%)		
Other	427 (1.17%)	103 (1.04%)		
BMI (kg/m^2)	28.30 ± 5.62	28.14 ± 5.83		
Serum albumin (g/dL)	2.98 ± 0.72	2.92 ± 0.74		
Serum creatinine (mg/dl)	1.38 ± 1.04	1.58 ± 1.18		
Recipient medical condition			101.00	< 0.001
Not hospitalized	26411 (72.34%)	6081 (54.54%)		
Hospitalized not in ICU	6100 (16.71%)	1932 (19.57%)		
ICU	3996 (10.95%)	1860 (18.84%)		
Cold ischemia time (hours)	7.02 ± 3.24	7.37 ± 3.35		
Hepatitis C positive	17018 (41.05%)	2609 (43.88%)	66.91	< 0.001
Type 2 diabetes	5345 (14.65%)	1580 (16.00%)	23.28	< 0.001
Previous malignancy	5269 (14.43%)	1655 (16.76%)	6.69	0.010
Portal vein thrombosis	2651 (7.17%)	867 (8.78%)	17.12	< 0.001
Donor age	40.73 ± 17.02	43.85 ± 17.12		
Donor ethnicity			0.08	0.78
White	24836 (68.03%)	6498 (65.82%)		
Hispanic	4397 (12.04%)	1383 (14.01%)		
Black	6005 (16.45%)	1586 (16.06%)		
Asian	797 (2.18%)	272 (2.76%)		
Other	472 (1.29%)	134 (1.36%)		
Donor sex			2.63	0.10
Male	21823 (59.78%)	5812 (58.87%)		
Female	14684 (40.22%)	4061 (41.13%)		

[a] The numerical variable is denoted by mean ± standard deviation and the categorical variable is denoted by number (%).
[b] P value is determined by Chi square test for categorical variable.

Table D.3: Characteristics of the study cohort for 5-year mortality (n=45270)

Variable	Survived recipients (n=32674)	Died recipients (n=12596)	Chi-square	P-value
Sex			3.51	0.061
Male	21946 (67.16%)	8577 (68.10%)		
Female	10728 (32.84%)	4019 (31.90%)		
Age	53.51 ± 9.97	54.78 ± 9.93		
Ethnicity			108.46	< 0.001
White	23893 (73.13%)	9003 (71.47%)		
Hispanic	4104 (12.56%)	1505 (11.95%)		
Black	2741 (8.39%)	1479 (11.74%)		
Asian	1555 (4.76%)	477 (3.79%)		
Other	381 (1.16%)	132 (1.05%)		
BMI (kg/m^2)	28.31 ± 5.63	28.16 ± 5.78		
Serum albumin (g/dL)	2.98 ± 0.72	2.94 ± 0.74		
Serum creatinine (mg/dl)	1.38 ± 1.04	1.55 ± 1.16		
Recipient medical condition			77.14	< 0.001
Not hospitalized	23600 (72.23%)	8096 (64.27%)		
Hospitalized not in ICU	5482 (16.78%)	2362 (18.75%)		
ICU	3592 (10.99%)	2138 (16.97%)		
Cold ischemia time (hours)	7.00 ± 3.25	7.26 ± 3.25		
Hepatitis C positive	12710 (39.28%)	5950 (47.93%)	68.31	< 0.001
Type 2 diabetes	4726 (14.47%)	2049 (16.27%)	37.59	< 0.001
Previous malignancy	4640 (14.20%)	2143 (17.01%)	11.23	< 0.001
Portal vein thrombosis	2371 (7.26%)	1074 (8.53%)	14.43	< 0.001
Donor age	40.50 ± 16.98	43.85 ± 17.1		
Donor ethnicity			0.06	0.80
White	22236 (68.05%)	8371 (66.46%)		
Hispanic	3911 (11.97%)	2020 (16.04%)		
Black	5407 (16.55%)	1697 (16.06%)		
Asian	696 (2.14%)	342 (2.72%)		
Other	424 (1.29%)	166 (1.32%)		
Donor sex			2.44	0.12
Male	19530 (59.77%)	7427 (58.96%)		
Female	13144 (40.23%)	5169 (41.04%)		

[a] The numerical variable is denoted by mean ± standard deviation and the categorical variable is denoted by number (%).

[b] P value is determined by Chi square test for categorical variable.

Table D.4: Characteristics of the study cohort for 10-year mortality (n=20751)

Variable	Survived recipients (n=11350)	Died recipients (n=9401)	Chi-square	P-value
Sex			9.93	0.002
Male	7561 (66.62%)	6457 (68.68%)		
Female	3789 (33.38%)	2944 (31.32%)		
Age	52.26 ± 9.86	53.98 ± 9.95		
Ethnicity			23.63	< 0.001
White	8288 (73.90%)	6987 (74.32%)		
Hispanic	1366 (12.03%)	1052 (11.19%)		
Black	902 (7.95%)	955 (10.16%)		
Asian	581 (5.12%)	316 (3.36%)		
Other	213 (1.88%)	91 (0.97%)		
BMI (kg/m^2)	28.05 ± 5.50	28.04 ± 5.74		
Serum albumin (g/dL)	2.94 ± 0.70	2.88 ± 0.71		
Serum creatinine (mg/dl)	1.36 ± 1.05	1.52 ± 1.14		
Recipient medical condition			25.58	< 0.001
Not hospitalized	8373 (73.77%)	6452 (68.63%)		
Hospitalized not in ICU	1752 (15.44%)	1542 (16.40%)		
ICU	1225 (10.79%)	1407 (14.97%)		
Cold ischemia time (hours)	7.42 ± 3.48	7.68 ± 3.61		
Hepatitis C positive	3988 (35.84%)	4016 (43.89%)	42.42	< 0.001
Type 2 diabetes	1006 (8.87%)	1095 (11.65%)	33.54	< 0.001
Previous malignancy	1167 (10.28%)	1258 (13.38%)	15.52	< 0.001
Portal vein thrombosis	526 (4.63%)	548 (5.83%)	2.69	< 0.101
Donor age	39.52 ± 17.21	43.63 ± 17.56		
Donor ethnicity			0.06	0.81
White	7978 (70.29%)	6418 (68.27%)		
Hispanic	1738 (15.31%)	1384 (14.72%)		
Black	1286 (11.33%)	1250 (13.30%)		
Asian	202 (1.78%)	233 (2.48%)		
Other	146 (1.29%)	116 (1.23%)		
Donor sex			1.36	0.24
Male	6796 (59.88%)	7427 (59.07%)		
Female	4554 (40.12%)	5169 (40.93%)		

[a] The numerical variable is denoted by mean ± standard deviation and the categorical variable is denoted by number (%).

[b] P value is determined by Chi square test for categorical variable.

Table E.1: Hyperparameters for the classifiers

Model	Parameter optimization range	Parameter
Logistic Regression	C: [0.001,0.01,0.1,1,10,100], penalty: ['l1','l2','elasticnet','none'], solver: ['newton-cg','lbfgs', 'liblinear', 'sag', 'saga']	penalty='l1', solver='liblinear', C=0.1
LogitBoost	n_estimators: [100,200,500,1000,2000], learning_rate: [0.1,0.2,0.3,0.5,1,2]	n_estimators= 1000, learning_rate = 1

Decision tree	criterion: ['gini','entropy'], max_depth :[4,5,6,7,8,9,10,12,15,20,30,40,50,100],
	min_samples_split: [1,2,3,4,5,10,15,20], min_samples_leaf: [1,2,3,4,5,6,7,8,10,15]
		criterion = 'entropy', max_depth = 6,
		min_samples_split = 3, min_samples_leaf = 5

AdaBoost	n_estimators: [100,200,500,1000,2000], learning_rate: [0.1,0.2,0.3,0.5,1,2]	n_estimators= 500, learning_rate = 0.2

Random Forest	n_estimators: [100,200,500,1000,2000],max_depth:[4,5,6,7,8,9,10,12,15,20,30,40,50,100],
	min_samples_split: [1,2,3,4,5,10,15,20,30], min_samples_leaf: [1,2,3,4,5,6,7,8,10,15]
		n_estimators= 500, max_depth = 20,
		min_samples_split = 20, min_samples_leaf = 4

XGBoost	max_depth: [2,3,4,5,6,7,8,9,10,15,20], learning_rate:[0.01,0.05,0.1,0.15,0.2,0.5,1]
	n_estimators: [100,200,500,1000,2000], reg_lambda:[0.05,0.01,0.15,0.20,0.3,0.4,0.5,1.0]
		max_depth=3, learning_rate=0.15
		n_estimators=500, reg_lambda = 0.15

DNN	Num_hidden_layer:[1,2,3,4,5,6], learning_rate:[0.0001,0.001,0.01,0.1,0.2,0.3,0.4,0.5,1] Num_Neurons in each hidden layer (depend on number of hidden layer) :[1024,512,256,128,64,32,16] , batch size:[256,512,1024,2048,4096]	Number of Hidden layer = 4, learning rate = 0.001, Number of Neurons in each hidden layer = [1024, 128, 16], batch size = 2048