BMJ Open

Identifying the most influential factors in lung transplant patients using a multivariate prediction model

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-089796
Article Type:	Original research
Date Submitted by the Author:	09-Jun-2024
Complete List of Authors:	Gholamzadeh, Marsa; Tehran University of Medical Sciences, Safdari, Reza; Tehran University of Medical Sciences, Asadi Gharabaghi, Mehrnaz; Tehran University of Medical Sciences, Abtahi, Hamidreza; Tehran University of Medical Sciences, Department of Pulmonary Medicine; Tehran University of Medical Sciences, Pulmonary and Critical Care Department
Keywords:	Machine Learning, Pulmonary Disease < Lung Diseases, Transplant medicine < INTERNAL MEDICINE

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Identifying the most influential factors in lung transplant patients using a multivariate prediction model

Marsa Gholamzadeh[1], Reza Safdari[2], Mehrnaz Asadi Gharabaghi[3], Hamidreza Abtahi[3,4*]

1. Ph.D. in Medical Informatics, Health Information Management Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.
2. Professor of Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran
3. Department of Pulmonary Medicine, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran.
4. Associate Professor, Pulmonary and Critical Care Department, Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Tehran, Iran.

***Corresponding author: Hamidreza Abtahi**

E-mail address: hrabtahi2020research@gmail.com

Tel: +9821-66192646

Postal address: Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Qarib Ave, Keshavarz Blv, Tehran, Iran.

ORCID ID:

Marsa Gholamzadeh, https://orcid.org/0000-0001-6781-9342;

Hamidreza Abtahi, https://orcid.org/0000-0002-1111-0497;

Reza Safdari, https://orcid.org/0000-0002-4982-337X

Mehrnaz Asadi Gharabaghi: https://orcid.org/0000-0003-0852-1532

Word count: 3728 words

Abstract

Objectives: In lung transplantation, a priority is assigned to each candidate on the waiting list. Our primary objective was to identify the key factors that influence the allocation of priorities in lung transplantation using machine learning (ML) techniques to enhance the process of prioritizing patients.

Design: Developing a prediction model.

Setting and participants: Our data was retrieved from the UNOS open-source database of transplant patients between 2005 and 2023.

Interventions: After the preprocessing process, a feature engineering technique was employed to select the most relevant features. Then, six ML models with an optimized hyper-parameter including Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model, and a deep learning model (DL) were developed under trained data.

Primary and secondary outcome measures: The performance of each model was evaluated using R-squared (R^2) and other error rate indexes. Next, the Shapley Additive Explanations (SHAP) technique was utilized to identify the most important features in the prediction.

Results: The raw dataset contains 196,270 records with 545 features. After preprocessing, 32,966 records with 15 features remain. Among various models, the RF model achieved a high R2 score. Additionally, the RF model exhibited the lowest error values indicating its superior precision compared to other regression models SHAP technique in conjunction with the RF model revealed the 11 most important features for priority allocation. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model.

Conclusion: The deployment of the ML model has the potential to act as an automated tool to aid physicians in assessing the priority of lung transplants and identifying significant factors that play a role in patient survival.

Keywords: Lung transplantation, allocation score, Machine learning, Prediction.

Strengths and limitations of this study:

- Despite the potential benefits of using ML algorithms in medical sciences, there is a scarcity of studies examining the use of such algorithms in lung transplantation and organ allocation.
- The use of various preprocessing and data cleaning techniques in our survey increased the robustness and performance of the model.
- Understanding the factors influencing the determination of lung transplant priority could support clinicians in designing treatment plans and thus improving the quality of life of patients.
- Deploying the developed ML model in the form of a decision support system increases its applicability in clinical practice.

1-Introduction

Lung transplantation (LTx) is an advanced treatment option for patients suffering from end-stage lung disease. When no other treatment options are available and the patient is likely to die, lung transplant surgery is suggested as a well-established treatment option [1]. When a patient meets the inclusion criteria for transplantation, they are placed on a waiting list and assigned a priority. Various conditions may affect eligibility for lung transplantation and the patient's priority [2]. In some countries, a score is assigned to each patient on the waiting list to enhance the recipient selection process [3, 4]. Understanding the most influential factors in priority allocation for lung transplantation is beneficial for researchers worldwide, as it can improve post-transplant survival. Utilizing data mining methods and developing forecasting models in this field could aid clinicians in uncovering hidden patterns and relationships within patient data and allocation scores.

Machine learning (ML) methods have been developed across various fields of clinical medicine to assist clinicians in predicting and classifying diseases [5]. These methods are used to predict the length of stay in the Intensive Care Unit (ICU), diagnose septic infection[6], and extract disease patterns from big data [7, 8]. Nevertheless, there is a lack of studies on the development of predictive models and identification of important features using ML methods to predict lung transplantation priority [9, 10].

Thus, the primary objective of this study was to utilize ML techniques to identify the most influential factors that strongly impacted outcomes based on various developed ML methods to predict the priority using clinical and demographic data.

2- Methods

Throughout this section, the process of developing, comparing, and evaluating ML models is shown schematically in Fig 1. Python programming language version 10 was used in this study for developing and validating ML algorithms. For data preprocessing, Numpy and Pandas modules

were employed, while the sci-kit learn library was utilized for developing supervised classifier algorithms.

2-1-Dataset description and data retrieval

The data for this study were obtained from the United Network for Organ Sharing (UNOS) online database [11]. Upon receiving written permission from UNOS, we accessed the recorded data pertaining to lung transplantation for our research. Our study included patients over 18 years old with end-stage lung disease who underwent lung transplants between 2005 and 2022. We performed a waiting list analysis using all available data entries from the United Network for Organ Sharing (UNOS) database for our study.

The priority of candidates on the waiting list was considered as the outcome, while the clinical and demographic characteristics of patients were considered as features or predictors.

2-2-Pre-processing process

Data pre-processing is a crucial step in ML techniques, especially when dealing with raw data from clinical databases or medical records that often contain missing or unclear information. To ensure the development of more accurate models based on appropriate data, we followed a series of data pre-processing steps. The following steps were employed in this phase as pre-processing techniques.

1- Checking the duplicated values and records to remove the duplicates

2- De-identify records and remove irreverent features

3- Convert nominal and categorical features to numerical values

4- Identify missing data and missing values imputation

5- Outlier detection

6- Feature engineering and feature selection

7- Data transformation and normalization

Duplicate checking and removal of irreverent features: After duplicate checking, we consulted UNOS guidelines and experts to review all features and their definitions. Under their supervision, we removed identification variables and irrelevant features, such as ID columns, hospital center identification codes, and country of residence, to de-identify patients.

Following this, we converted the post-transplant survival days variable to years and excluded patients with a survival rate of less than two years. Next, we filtered out patients below 18 years of age and excluded any data before 2005. Additionally, records related to heart transplantation were removed from the dataset.

After removing irrelevant features in the first data-cleaning phase, we utilized the discretized operator to convert nominal values to numerical data. Categorical data were encoded using the LabelEncoder class too.

Missing data management: To address missing data in our dataset, we conducted missing data imputation across the entire dataset. Initially, we assessed the specified columns or attributes to determine the extent of missing and unique data in each column. During this analysis, we discovered that the ICU column was empty and decided to delete it due to its lack of meaningful information. To impute the missing data, a threshold of 60% was set for feature removal. As a result, any column with more than 60% missing data was removed. For the missing data imputation, we adopted a strategy of replacing missing data in numerical features with the mean value of each respective feature. This approach allows us to retain the integrity of the dataset while minimizing the impact of missing data on our analysis. By performing these comprehensive steps of missing data imputation, we ensure the dataset is optimized for further analysis and modeling, enabling us to draw more accurate conclusions and insights.

Outliers handling: To address outliers in the dataset, we first create distribution plots to visualize the data. Next, we apply the IQR method and use Box plots to identify outliers. Finally, we remove these outliers to prepare the data for further processing.

Feature engineering and feature selection: Since a high-dimensional dataset was utilized in this study, feature engineering should be employed to reduce the dimensionality of the features and enhance model performance. Feature selection is the process of identifying relevant features while eliminating irrelevant and redundant ones, aiming to derive a subset of features that effectively describe the problem with minimal loss of efficiency[12].

As the first step in this phase, correlation analysis was conducted to assess the relationships between features and target variables. This analysis helps identify highly correlated features that can aid in feature selection and model development. Subsequently, a combination of filtering and embedded techniques, including variance threshold and XGBoost methods were utilized to select the most pertinent features for the model and enhance its performance. After carefully selecting the features, the most effective and appropriate features remained as predictors for modeling.

Data transformation and normalization: In the end, data normalization was carried out to optimize the features for modeling purposes.

2-4- Model development and tuning

The objective of this study was to develop a prediction model for a continuous numerical variable (priority score) using regression techniques to identify the most effective factors in selecting the most appropriate candidate for LTx. In this study, regression models were selected to examine the connection between input variables and output numerical values, as the target variable (outcome) is a continuous numerical value.

During the model development process, the dataset was divided into training and testing data sets in an 80:20 ratio. To determine the most influential factors and identify the best model, the performance of six regression-based models was evaluated: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model (MLP—a class of feedforward artificial neural network), and a deep learning model (DL). The selection of these models was done based on the type of target variable and the study objectives.

A hyperparameter tuning optimization technique was employed in this phase to improve model performance by optimizing the training process by determining the best hyperparameters for each model. This technique was used to prevent models that underfit or overfit the data [13]. After tuning parameters in each model, the models were trained with updated best hyperparameters, and all metrics were calculated again to achieve the best performance. We employed the random search method, a hyperparameter tuning technique where hyperparameters are randomly chosen from a predefined set to train a model.

2-4-1- Multiple linear regression (MLR)

Multiple linear regression (MLR) is a statistical technique used to estimate the relationship between a dependent variable and one or more independent variables. It is an extension of linear regression, which requires more than one predictor variable to forecast the response variable[14]. MLR is a significant regression algorithm that models the linear association between a dependent continuous variable and multiple independent variables[15]. Hence, we have chosen this model to predict the continuous variable (priority score) based on several independent variables. The equation for multiple linear regression is demonstrated below[15]:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7 \quad (1)$$

where y represents the priority; x_i is the considered variables; β_0 is the intercept; and β^i is the regression coefficients.

2-4-2- Random Forest Regressor (RF)

The random forest (RF) regression algorithm is a kind of ML approach that employs a group of decision trees, which are trained on a subset of the data, to make predictions. This technique is designed to stabilize the algorithm and decrease variance by using multiple trees. The RF regression algorithm is widely recognized as a popular model in developing regression models because of its strong performance with large datasets and diverse data types [16, 17].

2-4-3- Support Vector Machines Regressor (SVM)

Support vector machine regression (SVM) is a versatile regression function that can be used to solve both classification and regression problems. SVM is a supervised learning algorithm that fits a regression to the training data by reducing the distance between the sampled points and the fitted hyperplane[18, 19]. One advantage of SVM is that it is a sparse algorithm, meaning that it only needs information from a limited number of data points[20].

2-4-4- XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another ML library that is available for free and offers a powerful and efficient implementation of the gradient boosting algorithm[21]. Gradient boosting is a technique that involves creating an ensemble of tree-based models and then combining them to create a more accurate overall model than any of the individual models in the sequence[22]. XGBoost is a popular choice for those who require an effective and optimized implementation of gradient boosting[23].

2-4-5- Multilayer Perceptron Model (MLP)

The Multilayer Perceptron (MLP) is considered one of the top regression models in the field of artificial neural networks. It is equipped with the capability to learn from training data using a

variety of training algorithms and rules. This feature allows the MLP to acquire numerous advantages, including increased capacity. As a result, the MLP operates as a self-regulating model that utilizes specific learning algorithms to enhance its performance when encountering new inputs [24, 25].

2-4-6- Deep Learning Model (DL)

A deep learning model can be used for regression problems by learning a mapping from input features to the target output. Deep learning is an adaptable model proficient at effectively managing intricate data relationships. It proves especially beneficial when working with extensive datasets where traditional regression methods might not uncover intricate patterns. Nonetheless, to prevent overfitting and attain peak performance, these models necessitate meticulous calibration and validation[26-28]. Occasionally, due to the complex nature of implementing these models, simpler regression models may outperform them.

2-5- Performance evaluation

Typically, regression models are evaluated based on a function that measures the difference between the predicted and actual numerical value of the target variable, such as the priority score[29]. In this study, three popular evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and R-squared (R^2) score to assess the performance of the developed models [30].

$$(1)\quad MAE = \frac{\sum_{i=1}^{n} |R^*(i) - R(i)|}{n}$$

$$(2)\quad RMSE = \left\{ \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{n} \right\}^{1/2}$$

$$(3)\quad R^2 = 1 - \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{\sum_{i=1}^{n} (R^*(i) - m(i))^2}$$

In these formulas; variable n refers to the number of samples; $R^*(i)$ denotes the retrieved value predicted by the model; $R(i)$ denotes the analyzed value; and $m(i)$ denotes the average analyzed value.

To validate the developed machine learning (ML) models and reduce bias, we employed k-fold cross-validation. This technique overcomes the limitations of a simple train/test split by dividing the available data into multiple folds or subsets. By averaging the results across these folds, we achieve a more robust estimate of the model's performance compared to a simple train/test split.

2-6- Feature Importance

In the realm of machine learning models, a technique employed to elucidate the impact of each feature on the model is the SHAP (Shapley Additive Explanations) method. The SHAP method aims to enhance the transparency and interpretability of machine learning models by drawing on cooperative game theory [31]. For instance, linear models utilize their coefficients to gauge the significance of each feature. However, these coefficients are influenced by the scale of the variable itself, potentially resulting in misinterpretations [32]. The same can be found in tree-based models for feature ranking. This is precisely why SHAP (Shapley Additive Explanations) becomes valuable for model interpretation [33]. The absolute value of SHAP provides insight into how significantly an individual feature influences the prediction [34]. Once we identify the optimal model for priority prediction, we'll leverage the SHAP technique to assign weights to the most critical features. These features will then be ranked based on their importance and impact on the final priority score.

Patient and public involvement

None

3-Results

3-1- Dataset description

The raw dataset comprises information on 196,270 patients who underwent lung transplantation, as well as data related to lung donors. This comprehensive dataset includes 545 features, encompassing demographic and clinical details about the organ recipients, biomarkers, laboratory test results, and characteristics of the donated organs. Additionally, it provides insights into various patient outcomes, such as post-transplant survival rates, occurrences of acute organ rejection, priority levels, duration of intensive care unit stay, post-transplant infections, and instances of re-transplantation.

To preprocess the dataset, we converted the transplantation date data type to a string format and extracted the year column by parsing the month, hour, and year components. We removed all data prior to 2005 due to the absence of a prioritization system during that period. Furthermore, to focus exclusively on adult transplants, we excluded information related to pediatric transplants for children and adolescents under 18 years of age. As a result, our initial dataset comprised 183,086 records.

Subsequently, we filtered the dataset to include only post-transplant survival records exceeding one year. After that, we eliminated any records with missing priority scores. Ultimately, our final dataset consisted of 45,966 records for subsequent analysis.

3-2- Exploratory data analysis

Following imputation of missing independent variable data and preprocessing steps, the overall patient population consisted of 66.88% men and 33.20% women, with a median age of 54.27±14.24 years. Our target variable is the priority (or allocation) score, which represents a continuous numerical value. In Table A-1 in Appendix, we present descriptive analyses and the frequency distribution of various demographic and clinical variables within the dataset.

Furthermore, we employed a data visualization method to enhance our comprehension of the data and dataset. This approach aids in verifying the integrity of the data and detecting any apparent inaccuracies. Incorporating data visualization is essential for all data science projects across various fields [35].

3-3- Data cleaning and preprocessing

In the method section, we present detailed information regarding preprocessing procedures applied to the whole dataset. After de-identifying the dataset, it was limited to 40,024 records and 445 features. During the initial data cleaning phase, we removed irrelevant features, reducing the total to 322. Subsequently, we performed missing values imputation, resulting in 215 features available for further analysis. Next, in the correlation analysis phase, our dataset contains over 165 features post-pre-processing.

Due to the dataset's high dimensionality, we applied pre-processing techniques to select only the most important features based on their importance scores. As a result, we narrowed down the dataset to 65 features in the first phase of feature engineering. After further applying feature engineering and selection techniques, our final dataset consisted of 32,966 records, containing 15 features.

3-4- Development and evaluation of regression models

The prediction models were developed by training several selected features obtained during the feature engineering phase. We used 80% of the dataset to train the algorithms and the rest 20% to test and validate their efficacy (80:20) and all six regression algorithms were trained based on trained data.

To address the bias of training using simple data splitting the average score, K-fold cross-validation was done and K was considered as 10 folds. The results showed that average scores of

10-fold cross-validation in six ML algorithms are the same as the simple splitting data process. Subsequently, the hyperparameters were fine-tuned using a hyperparameter tuning technique to enhance the performance of the developed models.

The optimized and selected hyperparameters are documented in Table 1. Next, the MAE, RMSE, and R^2 values for each optimized model were calculated. Finally, all the optimized ML models were compared based on their R^2 scores and other relevant metrics. The evaluation results for the regression models in terms of error rates are represented in Table 2.

In our quest to identify the best regression model, we focused on minimizing the error in terms of the R^2 score. As a result, the RF regression model emerged as the top performer among the developed prediction models. We made this determination based on a comprehensive evaluation of various metrics, utilizing the best features.

3-5- Most important features to select the most appropriate candidate

Upon selecting the best model, we proceeded to identify and weigh the most influential features using the SHAP library within the final model. Initially, a prediction model based on the chosen regression model was created. Subsequently, the importance of each feature was determined by analyzing the set of trees generated by the model using the SHAP technique. The SHAP library assigns a score to each feature based on its impact on the prediction model. The ranking of the variables used in the ultimate model is visually represented in Fig 2, which is a widely recognized and popular chart produced by SHAP.

Ultimately, the researchers pinpointed the 11 most effective features, each receiving the highest score in candidate prioritization. These features, along with their explanations, are detailed in Table 3. Notably, it showed that factors such as a patient's oxygen consumption and diagnosis played a significant role in prioritizing the waiting list. Additionally, the patient's waiting time on

the transplant list emerged as another influential factor. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model.

4- Discussion

The study aimed to explore the feasibility of utilizing machine learning (ML) methods to predict priority levels for patients on the waiting list for lung transplants and to pinpoint the critical factors influencing priority allocation. Despite the potential advantages of employing ML algorithms in organ allocation [36], there is a lack of research on their application specifically in lung transplantation. This investigation led to the development of a decision support tool for estimating transplantation priorities.

Currently, the decision-making process for prioritizing individuals on organ transplant waiting lists is predominantly reliant on physicians' subjective judgments, often following "first-come, first-served" or "longer waiting time" principles rather than utilizing sophisticated mathematical models [37, 38]. Researchers recommend that authorities explore more equitable and innovative solutions for allocating donor organs to patients on waiting lists. As a result, researchers in the field of transplantation have concentrated on developing advanced models to forecast priority rankings and outcomes for recipients based on pre-transplantation factors [39, 40]. Similarly, we employed ML models to investigate more appropriate factors in assigning organs to recipients.

Prior studies on organ allocation have focused only on classification models to predict the risk of mortality following transplantation [39, 41]. However, these approaches have not been highly effective in improving the prioritization of patients on lung transplant waiting lists[42, 43]. In contrast, our developed model takes into account various factors such as disease type, oxygen saturation, demographics, clinical tests, and functional status.

In the context of machine learning (ML), the effectiveness of methods depends not only on their design and techniques but also on the quality and suitability of the data they operate on. To overcome the limitations of prior research, which often relied on a single ML technique and small sample sizes, our study takes a different approach. We incorporate multiple ML techniques to enhance the accuracy of our results, leveraging a large dataset sourced from the United Network for Organ Sharing (UNOS) database.

Our algorithm yields slightly superior results. To enhance the robustness of our model, we employed various data preprocessing techniques and feature engineering methods. These approaches allowed us to identify the most relevant and informative features in the data while discarding redundant or noisy ones [44, 45]. Data preprocessing plays a crucial role in improving data quality and enhancing the accuracy of knowledge extraction [46]. Additionally, by reducing data complexity and dimensions, our models became better equipped to capture underlying patterns and relationships, resulting in improved predictive performance [10, 45].

Our analysis reveals that employing the RF regressor model, which incorporates 15 features from the most significant donor and recipient variables available prior to transplantation, represents an effective approach for assigning an allocation score to each candidate on the waiting list. This outperforms other regression models. RF was specifically chosen due to its favorable prediction performance in previous research [47]. A deployment model in the form of an AI-based decision support tool could assist clinicians in utilizing the survey results within the context of their decision-making process and point-of-care scenarios

ML-based models rely on intricate mathematical structures and multi-dimensional datasets, often yielding complex patterns and relationships that can be challenging for humans to grasp. To address this complexity and limitation, researchers have turned to SHAP (Shapley Additive

explanations) summary analysis. This technique identifies the top 11 influential features within the final model. By doing so, it sheds light on which parameters should take precedence when selecting the most suitable recipient with the highest priority—a factor that has not received extensive exploration in prior studies.

While the suggested model demonstrated satisfactory performance, it does possess evident limitations. Despite the dataset under consideration being of a substantial size, it was obtained from a freely accessible dataset, not the Iranian transplantation data. in future studies, aligning with the structure of the UNOS database will allow for the collection of patient information tailored to researchers' requirements. Leveraging a local dataset can enhance its practical utility in point-of-care.

5-Conclusion

During this study, we succeeded in developing a priority prediction model based on the huge data of the UNOS database using ML models with the least error. Our research is among the pioneering studies that employ the SHAP method to enhance the comprehensibility of the proposed model intended for clinicians. Additionally, the automated auxiliary model that we created can assist clinicians in acquiring a better understanding of the transplant priority estimation and the crucial factors that influence patient survival.

6-Declaration

Ethics approval and consent to participate

The research was approved by the Tehran University of Medical Sciences Ethics Committee (IR.TUMS.IKHC.REC.1401.143). All methods were performed based on the relevant guidelines and regulations. Consent for participation was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

Consent for publication

Consent for publication was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Availability of data and materials

The data used in this article can be obtained from the United Network for Organ Sharing (UNOS) database by visiting www.unos.org/data. However, there are limitations on accessing this data, as it was used under a license for the current study and is not accessible to the general public. The interpretation and reporting of this 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 United States government.

Funding

This research was funded by the Thoracic Research Center through, Tehran University Medical Sciences by Grant No (59042). The funding body played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Authors' contributions

Conception and design of the study: Hamidreza Abtahi, Marsa Gholamzadeh, Reza Safdari, Mehrnaz Asadi Gharabaghi;

Data acquisition: Marsa Gholamzadeh, Hamidreza Abtahi;

Interpretation and/or analysis of data: Marsa Gholamzadeh, Hamidreza Abtahi, Reza Safdari, Mehrnaz Asadi Gharabaghi;

Drafting the manuscript: Marsa Gholamzadeh, Hamidreza Abtahi;

Revising the manuscript critically for important intellectual content: All authors;

Approval of the version of the manuscript to be published: All authors;

Acknowledgments

The data reported here have been supplied by the United Network for Organ Sharing (UNOS/OPTN) as the contractor for the Organ Procurement and Transplantation Network. We express our gratitude to the UNOS organization for allowing access to the data. We would like to extend our sincere thanks to the Thoracic Research Center of the Tehran University of Medical Sciences (TUMS) for their support and cooperation during this research.

References

1. van der Mark SC, Hoek RAS, Hellemons ME: **Developments in lung transplantation over the past decade**. *European Respiratory Review* 2020, **29**(157):190132.
2. Verleden GM, Dupont L, Yserbyt J, Schaevers V, Raemdonck DV, Neyrinck A, Vos R: **Recipient selection process and listing for lung transplantation**. *Journal of Thoracic Disease* 2017, **9**(9):3372-3384.
3. Smits JM, Nossent G, Evrard P, Lang G, Knoop C, Kwakkel-van Erp JM, Langer F, Schramm R, van de Graaf E, Vos R *et al*: **Lung allocation score: the Eurotransplant model versus the revised US model – a cross-sectional study**. *Transplant International* 2018, **31**(8):930-937.
4. Lancaster TS, Miller JR, Epstein DJ, DuPont NC, Sweet SC, Eghtesady P: **Improved waitlist and transplant outcomes for pediatric lung transplantation after implementation of the lung allocation score**. *J Heart Lung Transplant* 2017, **36**(5):520-528.
5. Safdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M: **Using data mining techniques to fight and control epidemics: A scoping review**. *Health and Technology* 2021, **11**(4):759-771.
6. Gholamzadeh M, Abtahi H, Safdari R: **Comparison of different machine learning algorithms to classify patients suspected of having sepsis infection in the intensive care unit**. *Informatics in Medicine Unlocked* 2023, **38**:101236.
7. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, McEvoy D, Stylianopoulos T, Munn LL, Dutta S *et al*: **Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19**. *NPJ Digit Med* 2021, **4**(1):87.
8. Safdari R, Deghatipour A, Gholamzadeh M, Maghooli K: **Applying data mining techniques to classify patients with suspected hepatitis C virus infection**. *Intelligent Medicine* 2022, **2**(4):193-198.
9. Gholamzadeh M, Abtahi H, Safdari R: **Machine learning-based techniques to improve lung transplantation outcomes and complications: a systematic review**. *BMC Medical Research Methodology* 2022, **22**(1):331.
10. Miller PE, Pawar S, Vaccaro B, McCullough M, Rao P, Ghosh R, Warier P, Desai NR, Ahmad T: **Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database**. *Journal of Cardiac Failure* 2019, **25**(6):479-483.
11. LeClaire JM, Smith NJ, Chandratre S, Rein L, Kamalia MA, Kohmoto T, Joyce LD, Joyce DL: **Solid organ donor-recipient race-matching: analysis of the United Network for Organ Sharing database**. *Transpl Int* 2021, **34**(4):640-647.

12. Theng D, Bhoyar KK: **Feature selection techniques for machine learning: a survey of more than two decades of research**. *Knowledge and Information Systems* 2024, **66**(3):1575-1637.
13. Li D, Liu Z, Armaghani DJ, Xiao P, Zhou J: **Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments**. *Sci Rep* 2022, **12**(1):1844.
14. Uyanık GK, Güler N: **A Study on Multiple Linear Regression Analysis**. *Procedia - Social and Behavioral Sciences* 2013, **106**:234-240.
15. Kayri M, Kayri I, Gencoglu MT: **The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data**. In: *2017 14th International Conference on Engineering of Modern Electric Systems (EMES): 1-2 June 2017 2017*; 2017: 1-4.
16. Dai B, Chen RC, Zhu SZ, Zhang WW: **Using Random Forest Algorithm for Breast Cancer Diagnosis**. In: *2018 International Symposium on Computer, Consumer and Control (IS3C): 6-8 Dec. 2018 2018*; 2018: 449-452.
17. Smith PF, Ganesh S, Liu P: **A comparison of random forest regression and multiple linear regression for prediction in neuroscience**. *Journal of Neuroscience Methods* 2013, **220**(1):85-91.
18. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ: **Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes**. *BMC Medical Informatics and Decision Making* 2010, **10**(1):16.
19. Sarker IH: **Machine Learning: Algorithms, Real-World Applications and Research Directions**. *SN Computer Science* 2021, **2**(3):160.
20. Huang H, Wei X, Zhou Y: **An overview on twin support vector regression**. *Neurocomputing* 2022, **490**:80-92.
21. Bentéjac C, Csörgő A, Martínez-Muñoz G: **A comparative analysis of gradient boosting algorithms**. *Artificial Intelligence Review* 2021, **54**(3):1937-1967.
22. Li S, Zhang X: **Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm**. *Neural Computing and Applications* 2020, **32**(7):1971-1979.
23. Liu J, Wu J, Liu S, Li M, Hu K, Li K: **Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model**. *PLOS ONE* 2021, **16**(2):e0246306.
24. Sananmuang T, Mankong K, Chokeshaiusaha K: **Multilayer perceptron and support vector regression models for feline parturition date prediction**. *Heliyon* 2024, **10**(6):e27992.
25. Abiodun OI, Jantan A, Omolara AE, Dada KV, Umar AM, Linus OU, Arshad H, Kazaure AA, Gana U, Kiru MU: **Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition**. *IEEE Access* 2019, **7**:158820-158846.
26. Sarker IH: **Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions**. *SN Computer Science* 2021, **2**(6):420.
27. Ahmed SF, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, Mofijur M, Shawkat Ali ABM, Gandomi AH: **Deep learning modelling techniques: current progress, applications, advantages, and challenges**. *Artificial Intelligence Review* 2023, **56**(11):13521-13617.
28. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**(7553):436-444.
29. Ramesh A, Ramamoorthy S, Buhari SM: **Forecasting Spread of COVID-19 Using Regression Algorithm**. In: *Soft Computing for Problem Solving: 2021// 2021; Singapore*: Springer Singapore; 2021: 461-467.
30. Karunasingha DSK: **Root mean square error or mean absolute error? Use their ratio as well**. *Information Sciences* 2022, **585**:609-629.
31. Ekanayake IU, Meddage DPP, Rathnayake U: **A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)**. *Case Studies in Construction Materials* 2022, **16**:e01059.

32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI: **From Local Explanations to Global Understanding with Explainable AI for Trees**. *Nat Mach Intell* 2020, **2**(1):56-67.
33. Rodríguez-Pérez R, Bajorath J: **Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions**. *Journal of Computer-Aided Molecular Design* 2020, **34**(10):1013-1026.
34. Kim Y, Kim Y: **Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models**. *Sustainable Cities and Society* 2022, **79**:103677.
35. Patel Darshan R, Reddy PVB: **The Importance of Data Visualization in Exploratory Data Analysis**. *Journal of Advanced Zoology* 2023, **44**(S6):923-929.
36. Peloso A, Moeckli B, Delaune V, Oldani G, Andres A, Compagnon P: **Artificial Intelligence: Present and Future Potential for Solid Organ Transplantation**. *Transpl Int* 2022, **35**:10640.
37. Bunnik EM: **Ethics of allocation of donor organs**. *Curr Opin Organ Transplant* 2023, **28**(3):192-196.
38. Madwar S: **United States officials propose further retreat from first-come, first-served organ donation**. *Cmaj* 2011, **183**(10):E639-640.
39. Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, Bailey J: **Machine-Learning Algorithms Predict Graft Failure After Liver Transplantation**. *Transplantation* 2017, **101**(4):e125-e132.
40. Gotlieb N, Azhie A, Sharma D, Spann A, Suo N-J, Tran J, Orchanian-Cheff A, Wang B, Goldenberg A, Chassé M *et al*: **The promise of machine learning applications in solid organ transplantation**. *npj Digital Medicine* 2022, **5**(1):89.
41. Jawitz OK, Raman V, Becerra D, Klapper J, Hartwig MG: **Factors associated with short- versus long-term survival after lung transplant**. *J Thorac Cardiovasc Surg* 2022, **163**(3):853-860.e852.
42. Brahmbhatt JM, Hee Wai T, Goss CH, Lease ED, Merlo CA, Kapnadak SG, Ramos KJ: **The lung allocation score and other available models lack predictive accuracy for post-lung transplant survival**. *J Heart Lung Transplant* 2022, **41**(8):1063-1074.
43. Dalton JE, Lehr CJ, Gunsalus PR, Mourany L, Valapour M: **Refining the Lung Allocation Score Models Fails to Improve Discrimination Performance**. *Chest* 2023, **163**(1):152-163.
44. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM: **A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction**. *Frontiers in Bioinformatics* 2022, **2**.
45. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
46. García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F: **Big data preprocessing: methods and prospects**. *Big Data Analytics* 2016, **1**(1):9.
47. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z: **Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan**. *BMJ Nutrition, Prevention & Health* 2021, **4**(1):140-148.

Tables

Table 1- The best selected hyperparameters

	Algorithm	Hyperparameters
1	Multiple linear regression	positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
2	Random Forest Regressor	n_estimators= 90, min_samples_split= 2, min_samples_leaf= 1, max_samples 10000, max_features: sqrt, max_depth=10
3	SVM Regressor	C =9.11158, loss='epsilon_insensitive', max_iter=5000
4	XGBoost Regressor	subsample=1, min_child_weight= 5, max_depth= 6, learning_rate=0.1, colsample_bytree=0.75
5	MLP	solver= 'sgd', Learning_rate= 'adaptive', hidden_layer_sizes: (20,), alpha: 0.001, activation: logistic
6	DL	Optimizer= 'sgd', batch_size= 16, activation= 'relu'

544 Table 2- The evaluation metrics of developed models and comparison of the model performance

	Model	*R^2*	*MSE*	*MAE*	*RMSE*
1	Random Forest Regressor	95.168	12.548	2.056	3.542
2	XGBoost Regressor	83.012	58.326	4.487	7.637
3	Deep Learning algorithm	68. 736	80.096	42.096	45.05
4	MLP Regressor	66.003	88.97	5.681	9.432
5	Linear Regression	52.259	123.989	6.984	11.131
6	Support Vector Machines	48.590	133.591	6.570	11.555

545

546

547

548 Table 3- The top 11 features identified by the SHAP method based on the prediction model

#	Feature	Description
1	INIT_O2	The amount of oxygen needed when the transplant candidate is on the waiting list
2	GROUPING	Lung transplant candidate diagnosis group
3	DAYSWAIT_CHRON	The amount of waiting time of patients on the waiting list - up-to-date waiting time
4	MED_COND_TRR	The status of the patient's lungs at the time of the last clinical evaluation
5	HEMO_SYS_TRR	The latest status of Hemodynamics Pcw (Sys) MM/Hg
6	END_O2	O2 Requirement at rest
7	VENTILATOR_TCR	The patient's status in terms of the need for a ventilator
8	LIFE_SUP_TCR	The amount of social and financial support
9	CIG_Use	History of cigarette use
10	Vent_Support_TRR	Episode of ventilatory support
11	Transfusion	Events occurring between listing and transplant

549

550

Figure legends

Fig 1. Schematic diagram of the proposed method

Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

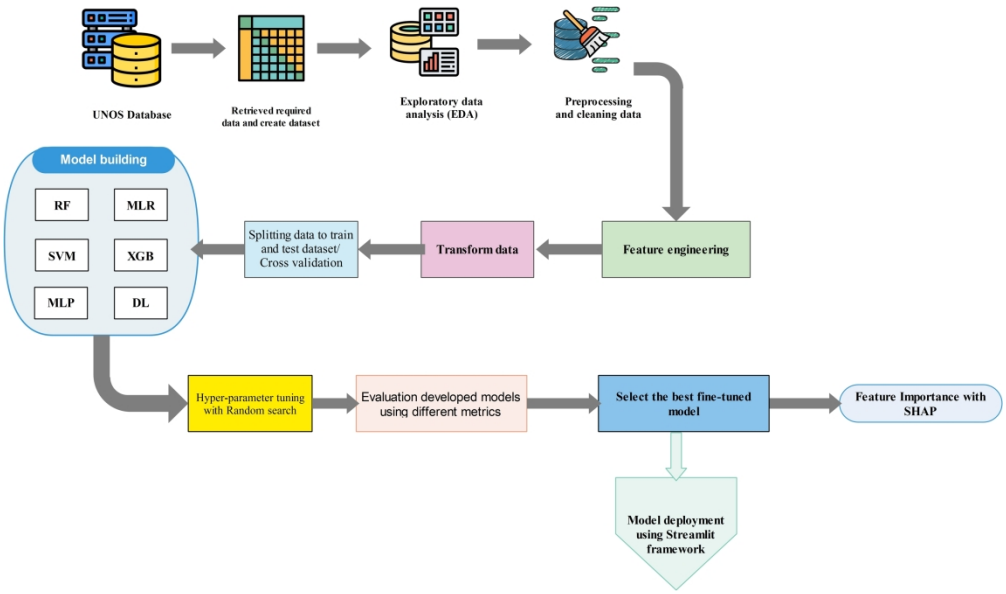
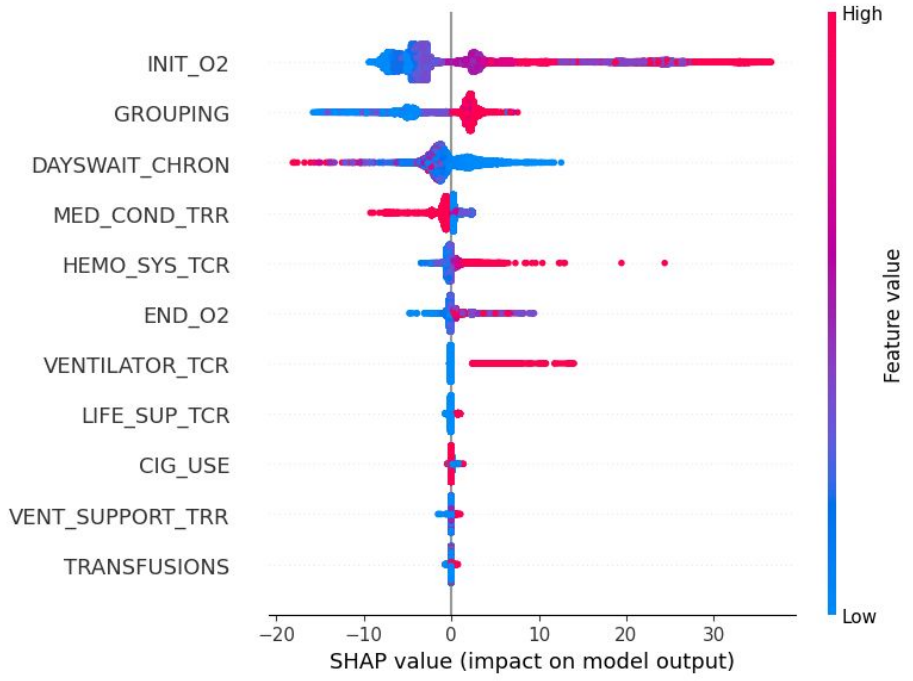
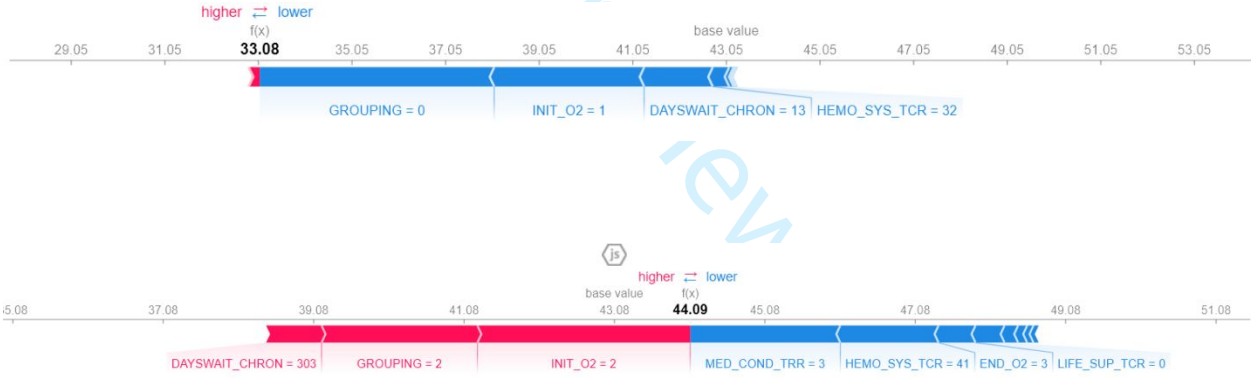


Fig 1. Schematic diagram of the proposed method

282x165mm (300 x 300 DPI)



(a)



(b)

Table A-1- Dataset description

Summary statistics of the selected continuous predictors (N=)					
	Variable	**Range**	**Mean (SD)**	**SE**	**95% Conf.**
Lung recipients	Age	18-58	54.27 (17.30)	0.095	39.2977
	BMI	14.997- 44.77	25.3 (3.83)	0.021	25.2586
	FEV1 value	5-120	39.484 (17.30)	0.095	39.2977
	Initial creatinine	0.1-24	0.841 (0.407)	0.002	0.8369
	Total Albumin serum	0.5-24	3.8787 (0.406)	0.002	3.8743

Summary statistics of selected categorical predictors (N=)				
	Variable		**n**	**Percentage (%)**
Lung recipients	**Gender**	Male	18085	54.86
		Female	14881	45.14
	ABO	A	1513	38.31
		B	4529	11.32
		AB	1513	3.78
		O	18648	46.59
	History of Malignancy	Positive	363	1.10
		Negative	30061	91.19
		Unknown	2542	7.71
	History of previous transplantation	Having	808	2.45
		Not Having	32158	97.55
Donor	**GENDER**	Male	14154	42.94
		Female	14154	42.94
		Unknown	9704	29.44
	ABO	A	279	0.85
		B	2544	7.72
		AB	11832	35.89
		O	13038	39.55
		Unknown	279	0.85
	History of Malignancy	Positive	9756	29.59
		Negative	21300	64.61
		Unknown	1910	5.79

BMJ Open

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-089796.R1
Article Type:	Original research
Date Submitted by the Author:	25-Mar-2025
Complete List of Authors:	Gholamzadeh, Marsa; Tehran University of Medical Sciences, Safdari, Reza; Tehran University of Medical Sciences, Asadi Gharabaghi, Mehrnaz; Tehran University of Medical Sciences, Abtahi, Hamidreza; Tehran University of Medical Sciences, Pulmonary and Critical Care Department; Tehran University of Medical Sciences, Thoracic Research Center
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Health informatics, Respiratory medicine, Health services research, Intensive care
Keywords:	Machine Learning, Pulmonary Disease < Lung Diseases, Transplant medicine < INTERNAL MEDICINE

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets

Marsa Gholamzadeh[1], Reza Safdari[1], Mehrnaz AsadiGharabaghi[2], Hamidreza Abtahi[3,*]

1. Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.
2. Department of Pulmonary Medicine, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran.
3. Pulmonary and Critical Care Department, Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Tehran, Iran.

***Corresponding author: Hamidreza Abtahi**

E-mail address: hrabtahi2020research@gmail.com

Tel: +9821-66192646

Postal address: Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Qarib Ave, Keshavarz Blv, Tehran, Iran.

ORCID ID:

Marsa Gholamzadeh, https://orcid.org/0000-0001-6781-9342;

Hamidreza Abtahi, https://orcid.org/0000-0002-1111-0497;

Reza Safdari, https://orcid.org/0000-0002-4982-337X

Mehrnaz Asadi Gharabaghi: https://orcid.org/0000-0003-0852-1532

Abstract

Objectives: In lung transplantation, a priority is assigned to each candidate on the waiting list. Our primary objective was to identify the key factors that influence the allocation of priorities in lung transplantation using machine learning (ML) techniques to enhance the process of prioritizing patients.

Design: Developing a prediction model.

Setting and participants: Our data was retrieved from the UNOS open-source database of transplant patients between 2005 and 2023.

Interventions: After the preprocessing process, a feature engineering technique was employed to select the most relevant features. Then, six ML models with an optimized hyper-parameter including Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model, and a deep learning model (DL) were developed based on UNOS dataset.

Primary and secondary outcome measures: The performance of each model was evaluated using R-squared (R^2) and other error rate metrics. Next, the Shapley Additive Explanations (SHAP) technique was utilized to identify the most important features in the prediction.

Results: The raw dataset contains 196,270 records with 545 features in all organs. After preprocessing, 32,966 records with 15 features remain. Among various models, the RF model achieved a high R2 score. Additionally, the RF model exhibited the lowest error values indicating its superior precision compared to other regression models SHAP technique in conjunction with the RF model revealed the 11 most important features for priority allocation. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model.

Conclusion: The deployment of the ML model has the potential to act as an automated tool to aid physicians in assessing the priority of lung transplants and identifying significant factors that play a role in patient survival.

Keywords: Lung transplantation, allocation score, Machine learning, Prediction.

Strengths and limitations of this study:

- To ensure transparency and interpretability in our machine learning models, we employed Explainable Artificial Intelligence (XAI) techniques, specifically the SHAP (Shapley Additive Explanations) method.
- The use of various preprocessing and data cleaning techniques in our survey increased the robustness and performance of the model.
- Understanding the factors influencing the determination of lung transplant priority could support clinicians in designing treatment plans and thus improving the quality of life of patients.
- Deploying the developed ML model in the form of a decision support system increases its applicability in clinical practice.

1-Introduction

Lung transplantation (LTx) is an advanced treatment option for patients suffering from end-stage lung disease. When no other treatment options are available and the patient is likely to die, lung transplant surgery is suggested as a well-established treatment option [1]. When a patient meets the inclusion criteria for transplantation, they are placed on a waiting list and assigned a priority. Various conditions may affect eligibility for lung transplantation and the patient's priority [2]. In some countries, a score is assigned to each patient on the waiting list to enhance the recipient selection process [3, 4]. Understanding the most influential factors in priority allocation for lung transplantation is beneficial for researchers worldwide, as it can improve post-transplant survival. Utilizing data mining methods and developing forecasting models in this field could aid clinicians in uncovering hidden patterns and relationships within patient data and allocation scores.

Machine learning (ML) methods have been developed across various fields of clinical medicine to assist clinicians in predicting and classifying diseases [5]. These methods are used to predict the length of stay in the Intensive Care Unit (ICU), diagnose septic infection[6], and extract disease patterns from big data [7, 8]. Nevertheless, there is a lack of studies on the development of predictive models and identification of important features using ML methods to predict lung transplantation priority [9, 10]. Thus, the primary objective of this study was to utilize ML techniques to identify the most influential factors that strongly impacted outcomes based on various developed ML methods to predict the priority using clinical and demographic data.

2- Methods

Throughout this section, the process of developing, comparing, and evaluating ML models is shown schematically in Fig 1. Python programming language version 10 was used in this study for developing and validating ML algorithms. For data preprocessing, Numpy and Pandas modules

were employed, while the sci-kit learn library was utilized for developing supervised classifier algorithms.

2-1-Dataset description and data retrieval

The data for this study were obtained from the United Network for Organ Sharing (UNOS) online database [11]. Upon receiving written permission from UNOS, we accessed the recorded data pertaining to lung transplantation for our research. Our study included patients over 18 years old with end-stage lung disease who underwent lung transplants between 2005 and 2022. We performed a waiting list analysis using all available data entries from the United Network for Organ Sharing (UNOS) database for our study.

The priority of candidates on the waiting list was considered as the outcome, while the clinical and demographic characteristics of patients were considered as features or predictors.

2-2-Pre-processing process

Data pre-processing is a crucial step in ML techniques, especially when dealing with raw data from clinical databases or medical records that often contain missing or unclear information. To ensure the development of more accurate models based on appropriate data, we followed a series of data pre-processing steps. The following steps were employed in this phase as pre-processing techniques.

1- Checking the duplicated values and records to remove the duplicates

2- De-identify records and remove irreverent features

3- Convert nominal and categorical features to numerical values

4- Identify missing data and missing values imputation

5- Outlier detection

6- Feature engineering and feature selection

7- Data transformation and normalization

2-2-1-Duplicate checking and removal of irreverent features

After duplicate checking, we consulted UNOS guidelines and experts to review all features and their definitions. Under their supervision, we removed identification variables and irrelevant features, such as ID columns, hospital center identification codes, and country of residence, to de-identify patients.

Following this, we converted the post-transplant survival days variable to years and excluded patients with a survival rate of less than two years. Next, we filtered out patients below 18 years of age and excluded any data before 2005. Additionally, records related to heart transplantation were removed from the dataset.

After removing irrelevant features in the first data-cleaning phase, we utilized the discretized operator to convert nominal values to numerical data. Categorical data were encoded using the LabelEncoder class too.

2-2-2- Missing data management:

To address missing data in our dataset, we conducted missing data imputation across the entire dataset. Initially, we assessed the specified columns or attributes to determine the extent of missing and unique data in each column. During this analysis, we discovered that the ICU column was empty and decided to delete it due to its lack of meaningful information. To impute the missing data, a threshold of 80% was set for feature removal with expert consultation. As a result, any column with more than 80% missing data was removed. If the missingness is due to inconsistent reporting rather than clinical irrelevance, dropping the column could exclude critical information about high-risk patients. In this case, domain experts might recommend retaining the column and using advanced imputation techniques or creating a binary indicator for missingness.

Along with feature removal, the sensitivity analysis was conducted on dataset which revealed that the inclusion of omitted features had a detrimental effect on the performance of the Random Forest Regressor model. Specifically, these features led to a decrease in the R^2 score from 0.95 to 0.68 and an increase in RMSE from 3.5 to 4.8.

For the missing data imputation, we adopted a strategy of replacing missing data in numerical features with the mean value of each respective feature. This approach allows us to retain the integrity of the dataset while minimizing the impact of missing data on our analysis. By performing these comprehensive steps of missing data imputation, we ensure the dataset is optimized for further analysis and modeling, enabling us to draw more accurate conclusions and insights.

2-2-3- Outliers handling

Outliers can significantly impact the performance and interpretability of machine learning models. Therefore, it is essential to investigate their causes before deciding whether to exclude, transform, or retain them. This exploration ensures that the preprocessing steps are justified and scientifically sound. To address outliers in the dataset, we first create distribution plots to visualize the data. Next, we apply the IQR method and use Box plots to identify outliers. Finally, we remove these outliers to prepare the data for further processing.

To address outliers in the dataset, a comprehensive approach was employed that included visual and statistical analysis to identify and understand the nature of the outliers. The distribution plots and boxplots were created to visualize outliers and applied the IQR method to quantify their extent. Additionally, statistical analysis was conducted to assess the impact of outliers on the dataset and performed sensitivity analysis to evaluate their influence on model performance. Throughout this process, we ensured transparency and justification by documenting all outliers and providing context-specific reasons for their exclusion, such as data entry errors or clinically irrelevant

extreme values [12, 13]. This rigorous approach ensured that the removal of outliers was methodologically sound and did not compromise the integrity of our analysis.

2-2-4-Feature engineering and feature selection

Given the high-dimensional nature of the dataset used in this study, feature engineering and selection were critical steps to reduce dimensionality, eliminate irrelevant or redundant features, and enhance model performance. Feature selection aims to identify a subset of features that effectively describe the problem with minimal loss of information and computational efficiency [14]. All phases of feature engineering and selection were conducted under the supervision of clinical experts to ensure the relevance and validity of the selected features.

Step one, correlation analysis: As the first step, correlation analysis was performed to assess the relationships between features and the target variable, as well as inter-feature correlations. This analysis helped identify highly correlated features that could introduce multicollinearity and redundancy into the model using heatmap graph. Features with a correlation coefficient above a predefined threshold were flagged for further evaluation.

Step two, variance threshold filtering: To eliminate low-variance features that contribute little to the model's predictive power, a variance threshold was applied. Features with variance below a specified threshold (e.g., 0.01) were removed, as they were deemed to have minimal impact on the target variable.

Step three, embedded feature selection with XGBoost: Following the initial filtering, an embedded feature selection technique was employed using the XGBoost algorithm. XGBoost provides intrinsic feature importance scores based on metrics such as gain, cover, and frequency. Features with very low importance scores (negative scores indicating no correlation with target value) were excluded from the final feature set.

Step four, expert review and validation: All selected features were reviewed and validated by subject matter experts to ensure their clinical, practical, and scientific relevance. This step was critical to avoid eliminating features that, although statistically significant, may not be clinically significant. For example, some features were retained for model customization based on expert consultation, despite having modest statistical significance.

Through a comprehensive and expert-guided feature selection process, we identified a subset of features that were statistically significant, domain-relevant, and impactful for model performance. This rigorous approach ensured the final model was both robust and clinically meaningful, with the selected features deemed critical for predicting the target variable.

2-2-5-Data transformation and normalization: In the end, data normalization was carried out to optimize the features for modeling purposes.

2-3- Splitting data and validation technique

During the model development process, the dataset was divided into training and testing data in an 80:20 ratio where 80% of the data was used for training the models and the remaining 20% was reserved for testing and validation. The training dataset is used to train the model, allowing it to learn patterns and relationships within the data based on the available data. The training dataset typically contains the bulk of the available data. In contrast, the testing dataset is intended solely to evaluate the model's performance on unseen data, ensuring an unbiased assessment of its generalizability. This dataset is kept separate from the training process to provide a true measure of how the model performs in real-world scenarios. Both datasets are often split randomly, with common ratios such as 80:20 or 70:30, depending on the size and nature of the data.

This split ensured that the models were evaluated on unseen data to assess their generalization capability. To mitigate potential bias introduced by simple data splitting, cross-validation

technique was employed. Through this technique, the dataset was divided into k folds, and each model was trained and validated k times, with each fold serving as the validation set once. The average performance metrics across all folds were calculated to ensure robust evaluation. The results of the k-fold cross-validation were consistent with those obtained from the simple 80:20 split, confirming the reliability of the initial approach.

2-4- Model development and tuning

The objective of this study was to develop a prediction model for a continuous numerical variable (priority score) using regression techniques to identify the most effective factors in selecting the most appropriate candidate for LTx. In this study, regression models were selected to examine the connection between input variables and output numerical values, as the target variable (outcome) is a continuous numerical value.

To determine the most influential factors and identify the best model, the performance of six regression-based models was evaluated: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model (MLP—a class of feedforward artificial neural network), and a deep learning model (DL). The selection of these models was done based on the type of target variable and the study objectives.

A hyperparameter tuning optimization technique was employed in this phase to improve model performance by optimizing the training process by determining the best hyperparameters for each model. This technique was used to prevent models that underfit or overfit the data [15]. After tuning parameters in each model, the models were trained with updated best hyperparameters, and all metrics were calculated again to achieve the best performance. We employed the random search

method, a hyperparameter tuning technique where hyperparameters are randomly chosen from a predefined set to train a model.

2-4-1- Multiple linear regression (MLR)

Multiple linear regression (MLR) is a statistical technique used to estimate the relationship between a dependent variable and one or more independent variables. It is an extension of linear regression, which requires more than one predictor variable to forecast the response variable[16]. MLR is a significant regression algorithm that models the linear association between a dependent continuous variable and multiple independent variables[17]. Hence, we have chosen this model to predict the continuous variable (priority score) based on several independent variables. The equation for multiple linear regression is demonstrated below[17]:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7 \quad (1)$$

where y represents the priority; x_i is the considered variables; β_0 is the intercept; and β^i is the regression coefficients.

2-4-2- Random Forest Regressor (RF)

The random forest (RF) Regressor algorithm is a kind of ML approach that employs a group of decision trees, which are trained on a subset of the data, to make predictions. This technique is designed to stabilize the algorithm and decrease variance by using multiple trees. The RF regressor algorithm is widely recognized as a popular model in developing regression models because of its strong performance with large datasets and diverse data types [18, 19].

2-4-3- Support Vector Machines Regressor (SVM)

Support vector machine regression (SVM) is a versatile regression function that can be used to solve both classification and regression problems. SVM is a supervised learning algorithm that fits a regression to the training data by reducing the distance between the sampled points and the fitted

hyperplane[20, 21]. One advantage of SVM is that it is a sparse algorithm, meaning that it only needs information from a limited number of data points[22].

2-4-4- XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another ML library that is available for free and offers a powerful and efficient implementation of the gradient boosting algorithm[23]. Gradient boosting is a technique that involves creating an ensemble of tree-based models and then combining them to create a more accurate overall model than any of the individual models in the sequence[24]. XGBoost is a popular choice for those who require an effective and optimized implementation of gradient boosting[25].

2-4-5- Multilayer Perceptron Model (MLP)

The Multilayer Perceptron (MLP) is considered one of the top regression models in the field of artificial neural networks. It is equipped with the capability to learn from training data using a variety of training algorithms and rules. This feature allows the MLP to acquire numerous advantages, including increased capacity. As a result, the MLP operates as a self-regulating model that utilizes specific learning algorithms to enhance its performance when encountering new inputs [26, 27].

2-4-6- Deep Learning Model (DL)

A deep learning model can be used for regression problems by learning a mapping from input features to the target output. Deep learning is an adaptable model proficient at effectively managing intricate data relationships. It proves especially beneficial when working with extensive datasets where traditional regression methods might not uncover intricate patterns. Nonetheless, to prevent overfitting and attain peak performance, these models necessitate meticulous calibration and validation[28-30]. Occasionally, due to the complex nature of implementing these models, simpler regression models may outperform them.

2-5- Performance evaluation

Typically, regression models are evaluated based on a function that measures the difference between the predicted and actual numerical value of the target variable, such as the priority score[31]. In this study, three popular evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and R-squared (R^2) score to assess the performance of the developed models [32].

$$(1)\quad MAE = \frac{\sum_{i=1}^{n} |R^*(i) - R(i)|}{n}$$

$$(2)\quad RMSE = \left\{ \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{n} \right\}^{1/2}$$

$$(3)\quad R^2 = 1 - \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{\sum_{i=1}^{n} (R^*(i) - m(i))^2}$$

In these formulas; variable n refers to the number of samples; $R^*(i)$ denotes the retrieved value predicted by the model; $R(i)$ denotes the analyzed value; and $m(i)$ denotes the average analyzed value.

To validate the developed machine learning (ML) models and reduce bias, we employed k-fold cross-validation. This technique overcomes the limitations of a simple train/test split by dividing the available data into multiple folds or subsets. By averaging the results across these folds, we achieve a more robust estimate of the model's performance compared to a simple train/test split.

2-6- Feature Importance

To implement Explainable AI (XAI), a technique employed to elucidate the impact of each feature on the model is the SHAP (Shapley Additive Explanations) method. The SHAP method aims to enhance the transparency and interpretability of machine learning models by drawing on cooperative game theory [33]. For instance, linear models utilize their coefficients to gauge the significance of each feature. However, these coefficients are influenced by the scale of the variable itself, potentially resulting in misinterpretations [34]. The same can be found in tree-based models

for feature ranking. This is precisely why SHAP (Shapley Additive Explanations) becomes valuable for model interpretation [35]. The absolute value of SHAP provides insight into how significantly an individual feature influences the prediction [36]. Once we identify the optimal model for priority prediction, we'll leverage the SHAP technique to assign weights to the most critical features. These features will then be ranked based on their importance and impact on the final priority score.

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

3-Results

3-1- Dataset description

The raw dataset including all organs comprises information on 196,270 patients who underwent lung transplantation, as well as data related to lung donors. This comprehensive dataset includes 545 features, encompassing demographic and clinical details about the organ recipients, biomarkers, laboratory test results, and characteristics of the donated organs (Table A-2 in Appendix A). Additionally, it provides insights into various patient outcomes, such as post-transplant survival rates, occurrences of acute organ rejection, priority levels, duration of intensive care unit stay, post-transplant infections, and instances of re-transplantation.

To preprocess the dataset, we converted the transplantation date data type to a string format and extracted the year column by parsing the month, hour, and year components. We removed all data prior to 2005 due to the absence of a prioritization system during that period. Furthermore, to focus exclusively on adult transplants, we excluded information related to pediatric transplants for

children and adolescents under 18 years of age. As a result, our initial dataset comprised 183,086 records.

Subsequently, we filtered the dataset to include only post-transplant survival records exceeding one year. After that, we eliminated any records with missing priority scores. Ultimately, our final dataset consisted of 45,966 records for subsequent analysis.

3-2- Exploratory data analysis

Following imputation of missing independent variable data and preprocessing steps, the overall patient population consisted of 66.88% men and 33.20% women, with a median age of 54.27±14.24 years. Our target variable is the priority (or allocation) score, which represents a continuous numerical value. In Table A-1 in Appendix A, we present descriptive analysis and the frequency distribution of various demographic and clinical variables within the dataset.

Furthermore, we employed a data visualization method to enhance our comprehension of the data and dataset. This approach aids in verifying the integrity of the data and detecting any apparent inaccuracies. Incorporating data visualization is essential for all data science projects across various fields [37].

3-3- Data cleaning and preprocessing

In the method section, we present detailed information regarding preprocessing procedures applied to the whole dataset. After de-identifying the dataset, it was limited to 40,024 records and 445 features. During the initial data cleaning phase, we removed irrelevant features, reducing the total to 322. Subsequently, we performed missing values imputation, resulting in 215 features available for further analysis. Next, in the correlation analysis phase, our dataset contains over 165 features post-pre-processing.

Due to the dataset's high dimensionality, we applied pre-processing techniques to select only the most important features based on their importance scores. As a result, we narrowed down the dataset to 65 features in the first phase of feature engineering. After further applying feature engineering and selection techniques, our final dataset consisted of 32,966 records, containing 15 features.

3-4- Development and evaluation of regression models

The prediction models were developed by training several selected features obtained during the feature engineering phase. We used 80% of the dataset to train the algorithms and the rest 20% to test and validate their efficacy (80:20) and all six regression algorithms were trained based on trained data.

To address the bias of training using simple data splitting the average score, K-fold cross-validation was done and K was considered as 10 folds. The results showed that average scores of 10-fold cross-validation in six ML algorithms are the same as the simple splitting data process. Subsequently, the hyperparameters were fine-tuned using a hyperparameter tuning technique to enhance the performance of the developed models.

The optimized and selected hyperparameters are documented in Table 1. Next, the MAE, RMSE, and R^2 values for each optimized model were calculated and represented in Table 2. Finally, all the optimized ML models were compared based on their R^2 scores in combination with other relevant metrics. In this task, the RF Regressor emerges as the most robust model, demonstrating superior performance with an impressive 95.168% R^2 value, which indicates it explains nearly 96% of the variance in the data, significantly outperforming other techniques. The Adjusted R^2 metric, which penalizes unnecessary model complexity, closely mirrors the standard R^2 here (95.163%). The model's exceptional performance is evidenced by its lowest Mean Squared Error

(12.548), Mean Absolute Error (2.056), and Root Mean Square Error (3.542), suggesting highly accurate and precise predictions. The superior performance of RF model can be attributed to its ability to handle complex, non-linear relationships in medical data through ensemble learning, where multiple decision trees are combined to create a more flexible and generalized predictive model. In contrast, traditional linear methods like Linear Regression and Support Vector Machines struggled, achieving R^2 values below 53%, which suggests the priority prediction requires sophisticated, non-linear modeling approaches that can capture intricate patterns in medical datasets. The progression from linear to ensemble and advanced machine learning techniques clearly demonstrates the importance of selecting appropriate algorithms for complex predictive challenges in healthcare.

As a result, the RF Regressor model emerged as the top performer among the developed prediction models. We made this determination based on a comprehensive evaluation of various metrics, utilizing the best features.

3-5- Most important features to select the most appropriate candidate

Upon selecting the best model, we proceeded to identify and weigh the most influential features using the SHAP library within the final model. Initially, a prediction model based on the chosen regression model was created. Subsequently, the importance of each feature was determined by analyzing the set of trees generated by the model using the SHAP technique. The SHAP library assigns a score to each feature based on its impact on the prediction model. The ranking of the variables used in the ultimate model is visually represented in Fig 2, which is a widely recognized and popular chart produced by SHAP.

Ultimately, the researchers pinpointed the 11 most effective features, each receiving the highest score in candidate prioritization.

These features, along with their explanations, are detailed in Table 3. Notably, it showed that factors such as a patient's oxygen consumption and diagnosis played a significant role in prioritizing the waiting list. Additionally, the patient's waiting time on the transplant list emerged as another influential factor. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model (Fig 3).

4- Discussion

The study aimed to explore the feasibility of utilizing machine learning (ML) methods to predict priority levels for patients on the waiting list for lung transplants and to pinpoint the critical factors influencing priority allocation. Despite the potential advantages of employing ML algorithms in organ allocation [38], there is a lack of research on their application specifically in lung transplantation. This investigation led to the development of a decision support tool for estimating transplantation priorities.

Currently, the decision-making process for prioritizing individuals on organ transplant waiting lists is predominantly reliant on physicians' subjective judgments, often following "first-come, first-served" or "longer waiting time" principles rather than utilizing sophisticated mathematical models [39, 40]. Researchers recommend that authorities explore more equitable and innovative solutions for allocating donor organs to patients on waiting lists. As a result, researchers in the field of transplantation have concentrated on developing advanced models to forecast priority rankings and outcomes for recipients based on pre-transplantation factors [41, 42]. Similarly, we employed ML models to investigate more appropriate factors in assigning organs to recipients.

Prior studies on organ allocation have focused only on classification models to predict the risk of mortality following transplantation [41, 43]. However, these approaches have not been highly effective in improving the prioritization of patients on lung transplant waiting lists[44, 45]. In

contrast, our developed model takes into account various factors such as disease type, oxygen saturation, demographics, clinical tests, and functional status.

In the context of machine learning (ML), the effectiveness of methods depends not only on their design and techniques but also on the quality and suitability of the data they operate on. To overcome the limitations of prior research, which often relied on a single ML technique and small sample sizes, our study takes a different approach. We incorporate multiple ML techniques to enhance the accuracy of our results, leveraging a large dataset sourced from the United Network for Organ Sharing (UNOS) database.

Our algorithm yields slightly superior results. To enhance the robustness of our model, we employed various data preprocessing techniques and feature engineering methods. These approaches allowed us to identify the most relevant and informative features in the data while discarding redundant or noisy ones [46, 47]. Data preprocessing plays a crucial role in improving data quality and enhancing the accuracy of knowledge extraction [48]. Additionally, by reducing data complexity and dimensions, our models became better equipped to capture underlying patterns and relationships, resulting in improved predictive performance [10, 47].

Our analysis reveals that employing the RF regressor model, which incorporates 15 features from the most significant donor and recipient variables available prior to transplantation, represents an effective approach for assigning an allocation score to each candidate on the waiting list. This outperforms other regression models. RF was specifically chosen due to its favorable prediction performance in previous research [49]. Implementing the developed model as an AI-based decision support tool could assist physicians in integrating clinical insights into their decision-making processes and point-of-care scenarios, thereby enhancing the practical utility of the data.

ML-based models rely on intricate mathematical structures and multi-dimensional datasets, often yielding complex patterns and relationships that can be challenging for humans to grasp. To address this complexity and limitation, researchers have turned to SHAP (Shapley Additive explanations) summary analysis. This technique identifies the top influential features within the final model. By doing so, it sheds light on which parameters should take precedence when selecting the most suitable recipient with the highest priority—a factor that has not received extensive exploration in prior studies. On the other hand, as the research community increasingly shifts toward explainable AI (XAI) methods [50, 51], the adoption of this approach represents a significant step forward. By employing XAI techniques, the performance of developed models can be interpreted and explained more transparently, fostering greater trust and understanding in their outcomes.

Our study possesses some limitations. Despite the dataset under consideration being of a substantial size, it was obtained from a freely accessible dataset, aligning with the structure of the UNOS database will allow for the collection of patient information tailored to researchers' requirements. While our study demonstrates the effectiveness of the random forest (RF) model in predicting outcomes for lung transplant patients using the UNOS dataset, it is important to note the lack of external validation as a limitation. The model was developed and validated only on the UNOS dataset, which, although comprehensive, may contain biases related to specific populations and practices in the United States. However, we plan to focus on collaborating with international transplant registries or multicenter studies to validate the performance of the model in different populations and healthcare settings. This will enhance the validity of the model and its potential for widespread clinical adoption. External validation on independent datasets from different geographic regions or healthcare systems is essential to ensure the generalizability and robustness

of our findings. As part of future work, we are developing an intelligent lung transplant patient information system at our center. Building on previous efforts to apply AI-based techniques in solid organ transplantation [52-55], this system aims to integrate the current model with patients' medical records while leveraging additional AI-based models to enhance its performance.

5-Conclusion

During this study, we succeeded in developing a priority prediction model based on the huge data of the UNOS database using ML models with the least error. Our research is among the pioneering studies that employ the SHAP method as an XAI technique to enhance the comprehensibility of the proposed model intended for clinicians. Additionally, the automated auxiliary model that we created can assist clinicians in acquiring a better understanding of the transplant priority estimation and the crucial factors that influence patient survival.

6-Declaration

Ethics approval and consent to participate

The research was approved by the Tehran University of Medical Sciences Ethics Committee (IR.TUMS.IKHC.REC.1401.143). All methods were performed based on the relevant guidelines and regulations. Consent for participation was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

Consent for publication

Consent for publication was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Availability of data and materials

The data used in this article can be obtained from the United Network for Organ Sharing (UNOS) database by visiting www.unos.org/data. However, there are limitations on accessing this data, as it was used under a license for the current study and is not accessible to the general public. The interpretation and reporting of this 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 United States government.

Funding

This research was funded by the Thoracic Research Center through, Tehran University Medical Sciences by Grant No (59042). The funding body played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Authors' contributions

H.A., M.G., R.S., and M.A.G. contributed to the conception and design of the study. M.G. and H.A. acquired the data. M.G., H.A., R.S., and M.A.G. were involved in data interpretation and analysis. M.G. and H.A. drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version to be published. H.A. is the guarantor.;

Acknowledgments

The data reported here have been supplied by the United Network for Organ Sharing (UNOS/OPTN) as the contractor for the Organ Procurement and Transplantation Network. We express our gratitude to the UNOS organization for allowing access to the data. We would like to extend our sincere thanks to the Thoracic Research Center of the Tehran University of Medical Sciences (TUMS) for their support and cooperation during this research.

References

1. Van der Mark SC, Hoek RAS, Hellemons ME: **Developments in lung transplantation over the past decade**. *European Respiratory Review* 2020, **29**(157):190132.

2. Verleden GM, Dupont L, Yserbyt J, Schaevers V, Raemdonck DV, Neyrinck A, Vos R: **Recipient selection process and listing for lung transplantation**. *Journal of Thoracic Disease* 2017, **9**(9):3372-3384.
3. Smits JM, Nossent G, Evrard P, Lang G, Knoop C, Kwakkel-van Erp JM, Langer F, Schramm R, van de Graaf E, Vos R *et al*: **Lung allocation score: the Eurotransplant model versus the revised US model – a cross-sectional study**. *Transplant International* 2018, **31**(8):930-937.
4. Lancaster TS, Miller JR, Epstein DJ, DuPont NC, Sweet SC, Eghtesady P: **Improved waitlist and transplant outcomes for pediatric lung transplantation after implementation of the lung allocation score**. *J Heart Lung Transplant* 2017, **36**(5):520-528.
5. Safdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M: **Using data mining techniques to fight and control epidemics: A scoping review**. *Health and Technology* 2021, **11**(4):759-771.
6. Gholamzadeh M, Abtahi H, Safdari R: **Comparison of different machine learning algorithms to classify patients suspected of having sepsis infection in the intensive care unit**. *Informatics in Medicine Unlocked* 2023, **38**:101236.
7. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, McEvoy D, Stylianopoulos T, Munn LL, Dutta S *et al*: **Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19**. *NPJ Digit Med* 2021, **4**(1):87.
8. Safdari R, Deghatipour A, Gholamzadeh M, Maghooli K: **Applying data mining techniques to classify patients with suspected hepatitis C virus infection**. *Intelligent Medicine* 2022, **2**(4):193-198.
9. Gholamzadeh M, Abtahi H, Safdari R: **Machine learning-based techniques to improve lung transplantation outcomes and complications: a systematic review**. *BMC Medical Research Methodology* 2022, **22**(1):331.
10. Miller PE, Pawar S, Vaccaro B, McCullough M, Rao P, Ghosh R, Warier P, Desai NR, Ahmad T: **Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database**. *Journal of Cardiac Failure* 2019, **25**(6):479-483.
11. LeClaire JM, Smith NJ, Chandratre S, Rein L, Kamalia MA, Kohmoto T, Joyce LD, Joyce DL: **Solid organ donor-recipient race-matching: analysis of the United Network for Organ Sharing database**. *Transpl Int* 2021, **34**(4):640-647.
12. Mazarei A, Sousa R, Mendes-Moreira J, Molchanov S, Ferreira HM: **Online boxplot derived outlier detection**. *International Journal of Data Science and Analytics* 2025, **19**(1):83-97.
13. Kalaivani B, Ranichitra A: **Unveiling the Impact of Outliers: An Improved Feature Engineering Technique for Heart Disease Prediction**. In*: 2024; Singapore*: Springer Nature Singapore; 2024: 469-478.
14. Theng D, Bhoyar KK: **Feature selection techniques for machine learning: a survey of more than two decades of research**. *Knowledge and Information Systems* 2024, **66**(3):1575-1637.
15. Li D, Liu Z, Armaghani DJ, Xiao P, Zhou J: **Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments**. *Sci Rep* 2022, **12**(1):1844.
16. Uyanık GK, Güler N: **A Study on Multiple Linear Regression Analysis**. *Procedia - Social and Behavioral Sciences* 2013, **106**:234-240.
17. Kayri M, Kayri I, Gencoglu MT: **The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data**. In: *2017 14th International Conference on Engineering of Modern Electric Systems (EMES): 1-2 June 2017 2017*; 2017: 1-4.
18. Dai B, Chen RC, Zhu SZ, Zhang WW: **Using Random Forest Algorithm for Breast Cancer Diagnosis**. In: *2018 International Symposium on Computer, Consumer and Control (IS3C): 6-8 Dec. 2018 2018*; 2018: 449-452.
19. Smith PF, Ganesh S, Liu P: **A comparison of random forest regression and multiple linear regression for prediction in neuroscience**. *Journal of Neuroscience Methods* 2013, **220**(1):85-91.

20. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ: **Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes**. *BMC Medical Informatics and Decision Making* 2010, **10**(1):16.
21. Sarker IH: **Machine Learning: Algorithms, Real-World Applications and Research Directions**. *SN Computer Science* 2021, **2**(3):160.
22. Huang H, Wei X, Zhou Y: **An overview on twin support vector regression**. *Neurocomputing* 2022, **490**:80-92.
23. Bentéjac C, Csörgő A, Martínez-Muñoz G: **A comparative analysis of gradient boosting algorithms**. *Artificial Intelligence Review* 2021, **54**(3):1937-1967.
24. Li S, Zhang X: **Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm**. *Neural Computing and Applications* 2020, **32**(7):1971-1979.
25. Liu J, Wu J, Liu S, Li M, Hu K, Li K: **Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model**. *PLOS ONE* 2021, **16**(2):e0246306.
26. Sananmuang T, Mankong K, Chokeshaiusaha K: **Multilayer perceptron and support vector regression models for feline parturition date prediction**. *Heliyon* 2024, **10**(6):e27992.
27. Abiodun OI, Jantan A, Omolara AE, Dada KV, Umar AM, Linus OU, Arshad H, Kazaure AA, Gana U, Kiru MU: **Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition**. *IEEE Access* 2019, **7**:158820-158846.
28. Sarker IH: **Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions**. *SN Computer Science* 2021, **2**(6):420.
29. Ahmed SF, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, Mofijur M, Shawkat Ali ABM, Gandomi AH: **Deep learning modelling techniques: current progress, applications, advantages, and challenges**. *Artificial Intelligence Review* 2023, **56**(11):13521-13617.
30. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**(7553):436-444.
31. Ramesh A, Ramamoorthy S, Buhari SM: **Forecasting Spread of COVID-19 Using Regression Algorithm**. In: *Soft Computing for Problem Solving: 2021// 2021; Singapore*: Springer Singapore; 2021: 461-467.
32. Karunasingha DSK: **Root mean square error or mean absolute error? Use their ratio as well**. *Information Sciences* 2022, **585**:609-629.
33. Ekanayake IU, Meddage DPP, Rathnayake U: **A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)**. *Case Studies in Construction Materials* 2022, **16**:e01059.
34. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI: **From Local Explanations to Global Understanding with Explainable AI for Trees**. *Nat Mach Intell* 2020, **2**(1):56-67.
35. Rodríguez-Pérez R, Bajorath J: **Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions**. *Journal of Computer-Aided Molecular Design* 2020, **34**(10):1013-1026.
36. Kim Y, Kim Y: **Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models**. *Sustainable Cities and Society* 2022, **79**:103677.
37. Patel Darshan R, Reddy PVB: **The Importance of Data Visualization in Exploratory Data Analysis**. *Journal of Advanced Zoology* 2023, **44**(S6):923-929.
38. Peloso A, Moeckli B, Delaune V, Oldani G, Andres A, Compagnon P: **Artificial Intelligence: Present and Future Potential for Solid Organ Transplantation**. *Transpl Int* 2022, **35**:10640.
39. Bunnik EM: **Ethics of allocation of donor organs**. *Curr Opin Organ Transplant* 2023, **28**(3):192-196.
40. Madwar S: **United States officials propose further retreat from first-come, first-served organ donation**. *Cmaj* 2011, **183**(10):E639-640.
41. Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, Bailey J: **Machine-Learning Algorithms Predict Graft Failure After Liver Transplantation**. *Transplantation* 2017, **101**(4):e125-e132.

42. Gotlieb N, Azhie A, Sharma D, Spann A, Suo N-J, Tran J, Orchanian-Cheff A, Wang B, Goldenberg A, Chassé M *et al*: **The promise of machine learning applications in solid organ transplantation**. *npj Digital Medicine* 2022, **5**(1):89.
43. Jawitz OK, Raman V, Becerra D, Klapper J, Hartwig MG: **Factors associated with short- versus long-term survival after lung transplant**. *J Thorac Cardiovasc Surg* 2022, **163**(3):853-860.e852.
44. Brahmbhatt JM, Hee Wai T, Goss CH, Lease ED, Merlo CA, Kapnadak SG, Ramos KJ: **The lung allocation score and other available models lack predictive accuracy for post-lung transplant survival**. *J Heart Lung Transplant* 2022, **41**(8):1063-1074.
45. Dalton JE, Lehr CJ, Gunsalus PR, Mourany L, Valapour M: **Refining the Lung Allocation Score Models Fails to Improve Discrimination Performance**. *Chest* 2023, **163**(1):152-163.
46. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM: **A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction**. *Frontiers in Bioinformatics* 2022, **2**.
47. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
48. García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F: **Big data preprocessing: methods and prospects**. *Big Data Analytics* 2016, **1**(1):9.
49. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z: **Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan**. *BMJ Nutrition, Prevention & Health* 2021, **4**(1):140-148.
50. S Band S, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT, Liang H-W: **Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods**. *Informatics in Medicine Unlocked* 2023, **40**:101286.
51. Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, Bora PK, Almasri A, Alkhawaldeh RS, Hussain S *et al*: **A review of Explainable Artificial Intelligence in healthcare**. *Computers and Electrical Engineering* 2024, **118**:109370.
52. Abtahi H, Shahmoradi L, Amini S, Gholamzadeh M: **Design and evaluation of a Mobile-Based decision support system to enhance lung transplant candidate assessment and management: knowledge translation integrated with clinical workflow**. *BMC Medical Informatics and Decision Making* 2023, **23**(1):145.
53. Gholamzadeh M, Safdari R, Amini S, Abtahi H: **Feasibility study and determination of prerequisites of telecare programme to enhance patient management in lung transplantation: a qualitative study from the perspective of Iranian healthcare providers**. *BMJ Open* 2023, **13**(6):e073370.
54. Abtahi H, Safdari R, Gholamzadeh M: **Pragmatic solutions to enhance self-management skills in solid organ transplant patients: systematic review and thematic analysis**. *BMC Primary Care* 2022, **23**(1):166.
55. Gholamzadeh M, Abtahi H, Safdari R: **Telemedicine in lung transplant to improve patient-centered care: A systematic review**. *International Journal of Medical Informatics* 2022, **167**:104861.

Tables

Table 1- The best selected hyperparameters

	Algorithm	Hyperparameters
1	Multiple linear regression	positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
2	Random Forest Regressor	n_estimators= 90, min_samples_split= 2, min_samples_leaf= 1, max_samples 10000, max_features: sqrt, max_depth=10
3	SVM Regressor	C =9.11158, loss='epsilon_insensitive', max_iter=5000
4	XGBoost Regressor	subsample=1, min_child_weight= 5, max_depth= 6, learning_rate=0.1, colsample_bytree=0.75
5	MLP	solver= 'sgd', Learning_rate= 'adaptive', hidden_layer_sizes: (20,), alpha: 0.001, activation: logistic
6	DL	Optimizer= 'sgd', batch_size= 16, activation= 'relu'

Table 2- The evaluation metrics of developed models and comparison of the model performance

	Model	*R^2(%)*	*Adjusted R^2(%)*	*MSE*	*MAE*	*RMSE*
1	Random Forest Regressor	95.168	95.163	12.548	2.056	3.542
2	XGBoost Regressor	82.88	82.87	58.326	4.487	7.637
3	Deep Learning algorithm	68. 736	68.23	80.096	42.096	45.05
4	MLP Regressor	66.003	65.98	88.97	5.681	9.432
5	Linear Regression	52.259	52.23	123.989	6.984	11.131
6	Support Vector Machines	48.590	48.55	133.591	6.570	11.555

Table 3- The top 11 features identified by the SHAP method based on the prediction model

#	Feature	Description
1	INIT_O2	The amount of oxygen needed when the transplant candidate is on the waiting list
2	GROUPING	Lung transplant candidate diagnosis group
3	DAYSWAIT_CHRON	The amount of waiting time of patients on the waiting list - up-to-date waiting time
4	MED_COND_TRR	The status of the patient's lungs at the time of the last clinical evaluation
5	HEMO_SYS_TRR	The latest status of Hemodynamics Pcw (Sys) MM/Hg
6	END_O2	O2 Requirement at rest
7	VENTILATOR_TCR	The patient's status in terms of the need for a ventilator
8	LIFE_SUP_TCR	The amount of social and financial support
9	CIG_Use	History of cigarette use
10	Vent_Support_TRR	Episode of ventilatory support
11	Transfusion	Events occurring between listing and transplant

Figure legends

Fig 1. Schematic diagram illustrating the proposed methodology for developing machine learning models. The process includes data preprocessing, feature engineering, model training, and evaluation, followed by a systematic comparison of multiple models using performance metrics to identify and select the optimal model for deployment.

Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm

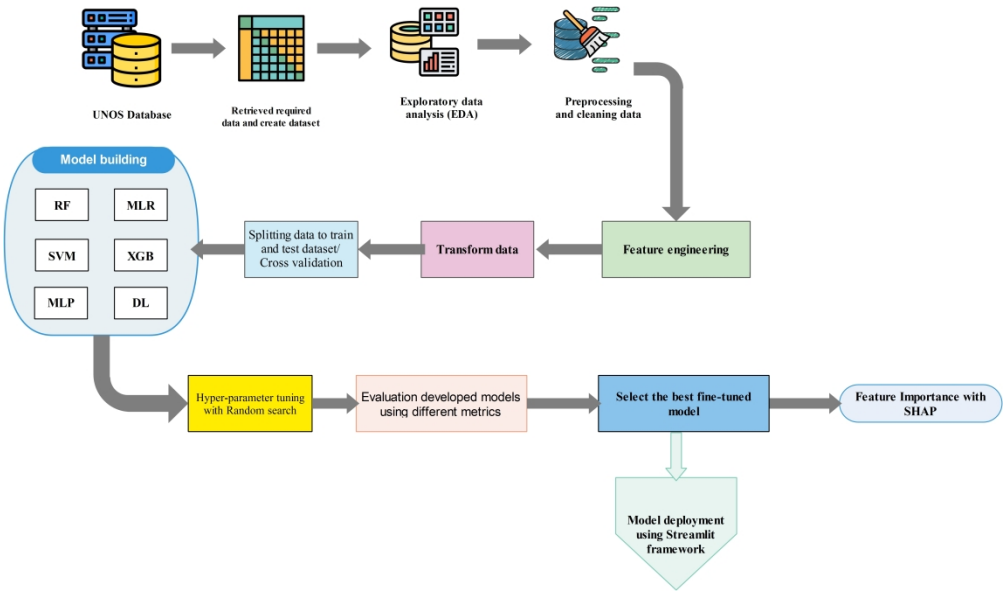


Fig 1. Schematic diagram of the proposed method

282x165mm (600 x 600 DPI)

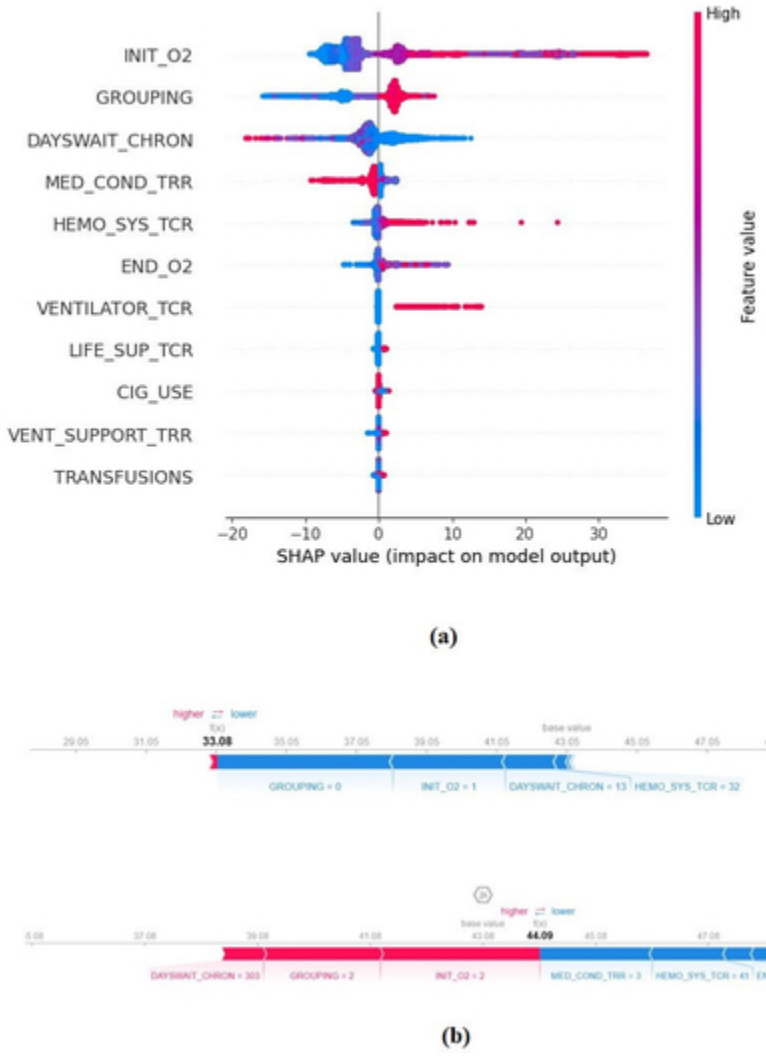


Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

41x45mm (300 x 300 DPI)

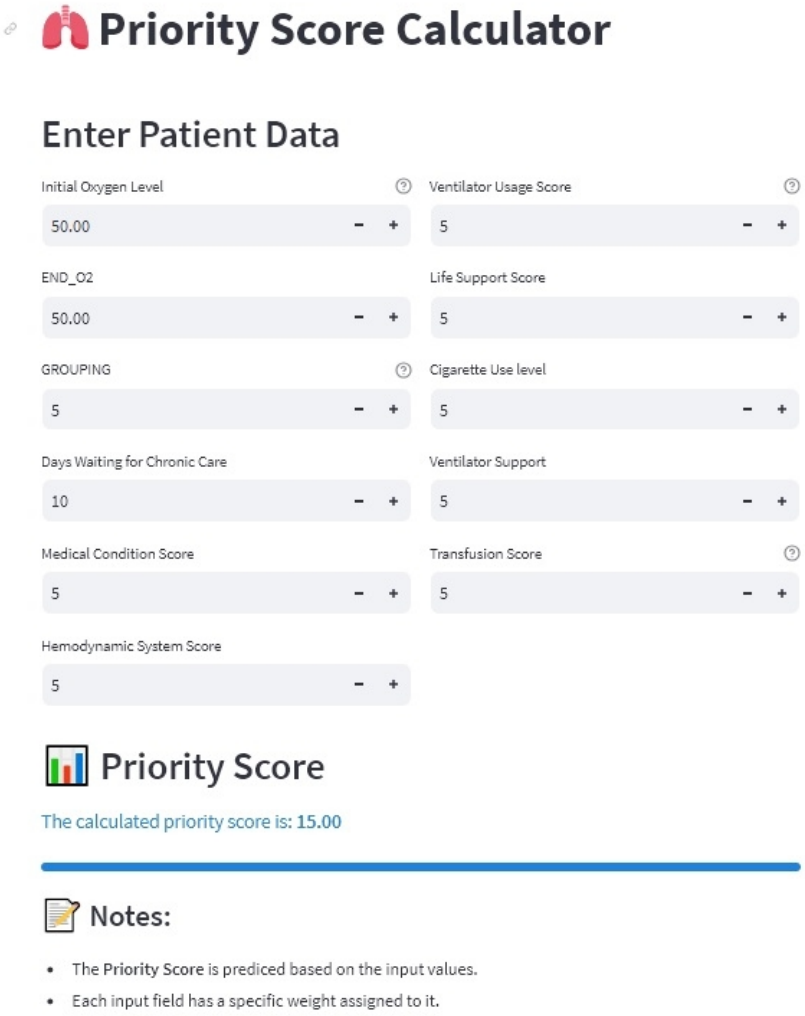


Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm

122x136mm (144 x 144 DPI)

Table A-1- Dataset description

Summary statistics of the selected continuous predictors (N=)					
	Variable	**Range**	**Mean (SD)**	**SE**	**95% Conf.**
Lung recipients	Age	18-58	54.27 (17.30)	0.095	39.2977
	BMI	14.997- 44.77	25.3 (3.83)	0.021	25.2586
	FEV1 value	5-120	39.484 (17.30)	0.095	39.2977
	Initial creatinine	0.1-24	0.841 (0.407)	0.002	0.8369
	Total Albumin serum	0.5-24	3.8787 (0.406)	0.002	3.8743

Summary statistics of selected categorical predictors (N=)				
	Variable		**n**	**Percentage (%)**
Lung recipients	**Gender**	Male	18085	54.86
		Female	14881	45.14
	ABO	A	1513	38.31
		B	4529	11.32
		AB	1513	3.78
		O	18648	46.59
	History of Malignancy	Positive	363	1.10
		Negative	30061	91.19
		Unknown	2542	7.71
	History of previous transplantation	Having	808	2.45
		Not Having	32158	97.55
Donor	**GENDER**	Male	14154	42.94
		Female	14154	42.94
		Unknown	9704	29.44
	ABO	A	279	0.85
		B	2544	7.72
		AB	11832	35.89
		O	13038	39.55
		Unknown	279	0.85
	History of Malignancy	Positive	9756	29.59
		Negative	21300	64.61
		Unknown	1910	5.79

Table A-2- Baseline Dataset description from UNOS database

Variables	count	mean	std	min	25%	50%	75%	max
GROUPING	32966	1.938725	1.316087	0	0	3	3	3
GENDER	32966	0.548596	0.49764	0	0	1	1	1
ABO	32966	1.667658	1.385379	0	0	2	3	3
WGT_KG_TCR	32966	72.41614	17.80507	3.1	59.422	72.575	84.822	212
HGT_CM_TCR	32966	168.2924	12.24525	5	160.02	168.6197	175.5	210.82
FUNC_STAT_TCR	32966	2063.003	278.1843	1	2040	2060	2070	4100
DIAB	32966	6.764454	72.83496	0	1	1	1	998
MALIG_TCR	32966	0.165231	0.540516	0	0	0	0	2
TOT_SERUM_ALBUM	32966	3.878676	0.406892	0.5	3.872302	3.872302	3.872302	9.8
RESIST_INF	32966	0.106049	0.393611	0	0	0	0	2
HEMO_SYS_TCR	32966	42.63604	16.84337	0	32	39	47	180
HEMO_PA_DIA_TCR	32966	17.79379	8.935308	0	12	17	21	110
HEMO_PA_MN_TCR	32966	27.51146	10.9713	0	21	26	31	110
HEMO_PCW_TCR	32966	10.69536	5.374686	0	7	10	14	50
HEMO_CO_TCR	32966	5.261608	1.407182	0.2	4.4	5	5.96	15
CIG_USE	32966	0.565613	0.495684	0	0	1	1	1
TCR_DUR_ABSTAIN	32966	57.32261	166.7262	1	7	7	56	998
LAST_INACT_REASON	32966	5.263635	1.500713	1	5	5	5	16
INIT_STAT	32966	7043.271	178.322	7010	7010	7010	7010	7999
INIT_O2	32966	4.097941	4.138358	0	2	3	4.15641	35
END_O2	32966	5.413994	5.203736	0	2.5	4	6	26.3
INIT_CREAT	32966	0.841288	0.407841	0.1	0.68	0.8	0.98	24
END_CREAT	32966	0.84711	0.42162	0.08	0.66	0.8	0.99	25
CALC_LAS_LISTDATE	32966	43.04133	16.16766	0	33.67689	37.5457	45.01999	96.22491
DAYSWAIT_CHRON	32966	213.5144	367.0553	0	20	73	236	5120

Variables	count	mean	std	min	25%	50%	75%	max
INIT_AGE	32966	54.27398	14.24429	0	48	58	64	81
LIFE_SUP_TCR	32966	0.070527	0.256037	0	0	0	0	1
VENTILATOR_TCR	32966	0.039344	0.194414	0	0	0	0	1
INIT_LLU_FLG	32966	0.356671	0.479024	0	0	0	1	1
INIT_RLU_FLG	32966	0.354638	0.478411	0	0	0	1	1
INIT_BLU_FLG	32966	0.830249	0.375419	0	1	1	1	1
END_LLU_FLG	32966	0.363071	0.480892	0	0	0	1	1
END_RLU_FLG	32966	0.361342	0.480397	0	0	0	1	1
END_BLU_FLG	32966	0.84211	0.364643	0	1	1	1	1
DR51	32966	19.82075	38.82521	0	0	0	0	99
DR51_2	32966	2.935509	16.52555	0	0	0	0	99
DR52	32966	20.40129	39.14027	0	0	0	0	99
DR52_2	32966	2.85661	16.27853	0	0	0	0	99
DR53	32966	20.14518	39.01353	0	0	0	0	99
DR53_2	32966	2.936359	16.51835	0	0	0	0	99
DQ1	32966	2.277498	20.30464	0	0	0	1	609
DQ2	32966	3.00907	29.15105	0	0	0	0	609
MED_COND_TRR	32966	1.877571	1.353962	0	0	3	3	3
CREAT_TRR	32966	0.856159	0.354149	0.1	0.71	0.864311	0.9	25
DIAL_AFTER_LIST	32966	0.311958	0.468832	0	0	0	1	2
FEV1_TRR	32966	39.48449	17.30273	5	26	39.9418	45	120
HEMO_CO_TRR	32966	5.336505	1.149913	1	4.86	5.347303	5.53	15
HEMO_PA_DIA_TRR	32966	17.90618	7.31172	0	14	17.95162	19	110
HEMO_PA_MN_TRR	32966	27.36964	8.662376	0	23	27.43337	28	110
HEMO_PCW_TRR	32966	10.68073	4.350863	0	9	10.68758	12	50
HEMO_SYS_TRR	32966	42.50621	13.37896	0	35	42.6675	42.6675	180

Variables	count	mean	std	min	25%	50%	75%	max
INFECT_IV_DRUG_TRR	32966	0.463447	0.627966	0	0	0	1	2
INOTROP_VASO_CO_TRR	32966	0.670752	0.533664	0	0	1	1	2
INOTROP_VASO_DIA_TRR	32966	0.699539	0.52854	0	0	1	1	2
INOTROP_VASO_MN_TRR	32966	0.682855	0.53295	0	0	1	1	2
INOTROP_VASO_PCW_TRR	32966	0.67694	0.529701	0	0	1	1	2
INOTROP_VASO_SYS_TRR	32966	0.704544	0.527142	0	0	1	1	2
OTH_LIFE_SUP_TRR	32966	0.013802	0.116671	0	0	0	0	1
PCO2_TRR	32966	47.7681	10.66006	10	42.1	47.64814	48	120
PRIOR_LUNG_SURG_TRR	32966	0.392374	0.568063	0	0	0	1	2
STEROID	32966	0.948856	0.825643	0	0	1	2	2
TBILI	32966	0.594413	0.689298	0.1	0.4	0.6	0.611384	36
TRANSFUSIONS	32966	0.3857	0.552885	0	0	0	1	2
VENT_SUPPORT_TRR	32966	0.426985	0.600565	0	0	0	1	2
VENTILATOR_TRR	32966	0.039617	0.19506	0	0	0	0	1
INHALED_NO_TRR	32966	0.00546	0.073692	0	0	0	0	1
PRIOR_CARD_SURG_TYPE_OSTXT_TRR	32966	94.76445	3.836607	0	95	95	95	95
PROSTACYCLIN_TRR	32966	0.004156	0.064332	0	0	0	0	1
TRACHEOSTOMY_TRR	32966	0.35485	0.526356	0	0	0	1	2
ECMO_72HOURS	32966	0.706576	0.492341	0	0	1	1	2
INHALEDNO_72HOURS	32966	0.723928	0.500037	0	0	1	1	2
INTUBATED_72HOURS	32966	0.855942	0.555187	0	1	1	1	2
HBV_CORE	32966	1.042104	1.377525	0	0	0	3	3
HBV_SUR_ANTIGEN	32966	0.961506	1.387078	0	0	0	3	3
HBV_SURF_TOTAL	32966	2.247952	1.226029	0	2	3	3	3
CMV_STATUS	32966	1.73227	1.195281	0	0	2	3	3
HIV_SEROSTATUS	32966	0.983134	1.391581	0	0	0	3	3

Variables	count	mean	std	min	25%	50%	75%	max
HCV_SEROSTATUS	32966	0.98186	1.378863	0	0	0	3	3
EBV_SEROSTATUS	32966	2.148608	0.763592	0	2	2	3	3
CRSMATCH_DONE	32966	1.588151	0.585256	0	1	2	2	2
CPRA	32966	747.2997	431.0568	0	86	999	999	999
PREV_TX	32966	0.02451	0.154629	0	0	0	0	1
PREV_TX_ANY	32966	0.025784	0.158493	0	0	0	0	1
DA1	32966	9.947916	71.84442	1	2	3	10	6802
DA2	32966	45.96509	187.604	0	24	33	48	6802
DB1	32966	43.60744	179.9494	7	13	44	46	5501
DB2	32966	71.46245	234.5561	0	44	60	76	8201
DDR1	32966	10.58812	29.91802	1	4	11	11	1501
DDR2	32966	23.31059	55.5125	0	13	15	24	1602
RA1	32966	20.05406	130.3715	0	2	3	23	6801
RA2	32966	80.44543	330.7992	0	24	68	92	6802
RB1	32966	80.57483	311.7406	0	8	44	92	5703
RB2	32966	142.1197	480.6177	0	44	61	162	8201
RDR1	32966	21.51423	82.27982	0	4	13	24	1601
RDR2	32966	46.76925	139.0017	0	13	17	53	1602
AMIS	32966	35.6986	46.58073	0	1	2	99	99
DRMIS	32966	35.77683	46.56309	0	1	2	99	99
HLAMIS	32966	37.79779	45.08084	0	4	5	99	99
MALIG_TRR	32966	0.625705	0.486636	0	0	1	1	2
CMV_IGG	32966	2.304162	1.077762	0	2	3	3	3
CMV_IGM	32966	2.056331	1.285516	0	1	3	3	3
HIST_COCAINE_DON	32966	0.520142	0.67797	0	0	0	1	2
AGE_DON	32966	34.46505	11.73268	6	26	34	40	76

Variables	count	mean	std	min	25%	50%	75%	max
HBV_CORE_DON	32966	1.917915	1.369888	0	1	1	4	4
HBV_SUR_ANTIGEN_DON	32966	1.885306	1.367688	0	1	1	4	4
ABO_DON	32966	2.821543	1.383585	0	3	3	4	4
ALCOHOL_HEAVY_DON	32966	0.514227	0.67552	0	0	0	1	2
GENDER_DON	32966	1.018079	0.755208	0	0	1	2	2
HEP_C_ANTI_DON	32966	1.904447	1.369103	0	1	1	4	4
ANTIHYPE_DON	32966	0.743766	0.796865	0	0	1	1	2
BUN_DON	32966	19.6837	13.75726	0.4	12	20.21487	20.21487	245
CREAT_DON	32966	1.404116	1.269158	0.07	0.82	1.3	1.42009	37
PT_DIURETICS_DON	32966	1.283868	0.787579	0	1	1	2	2
PT_STEROIDS_DON	32966	1.348025	0.761386	0	1	2	2	2
PT_T3_DON	32966	0.306528	0.469017	0	0	0	1	2
PT_T4_DON	32966	1.229661	0.80519	0	1	1	2	2
PT_OTH2_OSTXT_DON	32966	5972.515	2591.637	0	4203.25	7694	7694	9269
PULM_INF_DON	32966	1.297792	1.174939	0	0	1	3	3
SGOT_DON	32966	96.90449	306.0249	0.3	31	68	97.52713	20000
SGPT_DON	32966	97.76702	365.3933	0.4	25	56	97	44117
TBILI_DON	32966	0.988108	1.057065	0	0.6	0.98776	0.98776	59
URINE_INF_DON	32966	0.963478	1.342444	0	0	0	3	3
VASODIL_DON	32966	0.535764	0.696622	0	0	0	1	2
VDRL_DON	32966	2.198811	1.818314	0	1	1	5	5
CLIN_INFECT_DON	32966	1.238003	0.799413	0	1	1	2	2
HIST_CIG_DON	32966	0.427804	0.607943	0	0	0	1	2
HIST_HYPERTENS_DON	32966	0.623582	0.748324	0	0	0	1	2
HIST_CANCER_DON	32966	0.323697	0.494865	0	0	0	1	2
DIABETES_DON	32966	0.395347	0.580236	0	0	0	1	2

Variables	count	mean	std	min	25%	50%	75%	max
HIST_OTH_DRUG_DON	32966	0.875902	0.823869	0	0	1	2	2
CMV_DON	32966	2.761421	1.145194	0	1	3	4	4
DDAVP_DON	32966	0.530547	0.69316	0	0	0	1	2
HEPARIN_DON	32966	1.663866	0.5103	0	1	2	2	2
ARGININE_DON	32966	1.221865	0.807242	0	1	1	2	2
WGT_KG_DON_CALC	32966	77.36488	14.8769	23.5	70	77.35348	81.6	189
BMI_DON_CALC	32966	26.19033	4.580242	10.58866	23.66783	26.22113	27.1731	66.03564
HBV_NAT_DON	32966	2.028181	1.401618	0	0	3	3	3
ABO_MAT	32966	1.638779	0.905258	1	1	1	3	3
DIAL_PRIOR_TX	32966	0.650064	0.483463	0	0	1	1	2
ISCHTIME	32966	5.292257	1.528211	0.042	4.566406	5.351905	5.632813	25
O2_REQ_CALC	32966	5.375088	4.289068	0	3	5.457995	5.457995	26.3
PRIOR_CARD_SURG_TRR	32966	0.328429	0.490872	0	0	0	1	2
MALIG	32966	0.411818	0.598424	0	0	0	1	2
HGT_CM_CALC	32966	169.9918	8.255988	122	165.1	170.0308	175	210.82
BMI_CALC	32966	25.29995	3.832576	14.99785	23.38251	25.39236	27.33117	44.77787
DISTANCE	32966	211.2596	210.9401	0	67	215	221	4137
VENT_SUPPORT_AFTER_LIST	32966	0.426985	0.600565	0	0	0	1	2
PROTEIN_URINE	32966	0.862131	0.821972	0	0	1	2	2
CARDARREST_NEURO	32966	0.397379	0.568028	0	0	0	1	2
PO2	32966	380.254	121.5752	3.2	368	382.3152	457	754
HIST_MI	32966	0.327246	0.493355	0	0	0	1	2
LV_EJECT	32966	58.08624	9.440397	1	58	58.11868	61	99
CORONARY_ANGIO	32966	2.091974	1.327938	1	1	1	4	4
BIOPSY_DGN	32966	3.081933	1.997858	1	1	5	5	5
HBSAB_DON	32966	3.893739	1.46287	0	3	3	6	6

Variables	count	mean	std	min	25%	50%	75%	max
EBV_IGG_CAD_DON	32966	4.806285	1.547243	0	4	4	7	7
EBV_IGM_CAD_DON	32966	2.707972	2.252266	0	1	1	6	6
CDC_RISK_HIV_DON	32966	0.534581	0.698812	0	0	0	1	2
INOTROP_SUPPORT_DON	32966	0.982254	0.83639	0	0	1	2	2
TRANSFUS_TERM_DON	32966	294.9172	454.811	0	0	1	998	998
PO2_FIO2_DON	32966	86.1261	21.28058	1	86	100	100	100
PCO2_DON	32966	37.02991	5.584925	10	34.7	37	39	110
BRONCHO_LT_DON	32966	321.9415	464.8433	1	2	2	998	998
BRONCHO_RT_DON	32966	331.7915	468.465	1	2	2	998	998
CHEST_XRAY_DON	32966	300.3643	455.5274	1	2	5	999	999
PH_DON	32966	7.415315	0.083511	5	7.4	7.415555	7.44	8
HEMATOCRIT_DON	32966	29.21583	4.461127	2.5	27	29.14996	30.6	71

BMJ Open

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-089796.R2
Article Type:	Original research
Date Submitted by the Author:	06-Apr-2025
Complete List of Authors:	Gholamzadeh, Marsa; Tehran University of Medical Sciences, Safdari, Reza; Tehran University of Medical Sciences, Asadi Gharabaghi, Mehrnaz; Tehran University of Medical Sciences, Abtahi, Hamidreza; Tehran University of Medical Sciences, Pulmonary and Critical Care Department; Tehran University of Medical Sciences, Thoracic Research Center
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Health informatics, Respiratory medicine, Health services research, Intensive care
Keywords:	Machine Learning, Pulmonary Disease < Lung Diseases, Transplant medicine < INTERNAL MEDICINE

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets

Marsa Gholamzadeh[1], Reza Safdari[1], Mehrnaz AsadiGharabaghi[2], Hamidreza Abtahi[3,*]

1. Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.
2. Department of Pulmonary Medicine, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran.
3. Pulmonary and Critical Care Department, Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Tehran, Iran.

***Corresponding author: Hamidreza Abtahi**

E-mail address: hrabtahi2020research@gmail.com

Tel: +9821-66192646

Postal address: Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Qarib Ave, Keshavarz Blv, Tehran, Iran.

ORCID ID:

Marsa Gholamzadeh, https://orcid.org/0000-0001-6781-9342;

Hamidreza Abtahi, https://orcid.org/0000-0002-1111-0497;

Reza Safdari, https://orcid.org/0000-0002-4982-337X

Mehrnaz Asadi Gharabaghi: https://orcid.org/0000-0003-0852-1532

Abstract

Objectives: In lung transplantation, a priority is assigned to each candidate on the waiting list. Our primary objective was to identify the key factors that influence the allocation of priorities in lung transplantation using machine learning (ML) techniques to enhance the process of prioritizing patients.

Design: Developing a prediction model.

Setting and participants: Our data was retrieved from the UNOS open-source database of transplant patients between 2005 and 2023.

Interventions: After the preprocessing process, a feature engineering technique was employed to select the most relevant features. Then, six ML models with an optimized hyper-parameter including Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model, and a deep learning model (DL) were developed based on UNOS dataset.

Primary and secondary outcome measures: The performance of each model was evaluated using R-squared (R^2) and other error rate metrics. Next, the Shapley Additive Explanations (SHAP) technique was utilized to identify the most important features in the prediction.

Results: The raw dataset contains 196,270 records with 545 features in all organs. After preprocessing, 32,966 records with 15 features remain. Among various models, the RF model achieved a high R2 score. Additionally, the RF model exhibited the lowest error values indicating its superior precision compared to other regression models SHAP technique in conjunction with the RF model revealed the 11 most important features for priority allocation. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model.

Conclusion: The deployment of the ML model has the potential to act as an automated tool to aid physicians in assessing the priority of lung transplants and identifying significant factors that play a role in patient survival.

Keywords: Lung transplantation, allocation score, Machine learning, Prediction.

Strengths and limitations of this study:

- The use of various preprocessing and data cleaning techniques in our survey increased the robustness and performance of the model.
- To ensure transparency and interpretability in our machine learning models, we employed Explainable Artificial Intelligence (XAI) techniques, specifically the SHAP (Shapley Additive Explanations) method.
- Understanding the factors influencing the determination of lung transplant priority could support clinicians in designing treatment plans and thus improving the quality of life of patients.
- Deploying the developed ML model in the form of a decision support system increases its applicability in clinical practice.
- The model was validated on the UNOS dataset but requires external validation across diverse populations and healthcare systems to ensure generalizability and clinical applicability.

1-Introduction

Lung transplantation (LTx) is an advanced treatment option for patients suffering from end-stage lung disease. When no other treatment options are available and the patient is likely to die, lung transplant surgery is suggested as a well-established treatment option [1]. When a patient meets the inclusion criteria for transplantation, they are placed on a waiting list and assigned a priority. Various conditions may affect eligibility for lung transplantation and the patient's priority [2]. In some countries, a score is assigned to each patient on the waiting list to enhance the recipient selection process [3, 4]. Understanding the most influential factors in priority allocation for lung transplantation is beneficial for researchers worldwide, as it can improve post-transplant survival. Utilizing data mining methods and developing forecasting models in this field could aid clinicians in uncovering hidden patterns and relationships within patient data and allocation scores.

Machine learning (ML) methods have been developed across various fields of clinical medicine to assist clinicians in predicting and classifying diseases [5]. These methods are used to predict the length of stay in the Intensive Care Unit (ICU), diagnose septic infection[6], and extract disease patterns from big data [7, 8]. Nevertheless, there is a lack of studies on the development of predictive models and identification of important features using ML methods to predict lung transplantation priority [9, 10]. Thus, the primary objective of this study was to utilize ML techniques to identify the most influential factors that strongly impacted outcomes based on various developed ML methods to predict the priority using clinical and demographic data.

2- Methods

Throughout this section, the process of developing, comparing, and evaluating ML models is shown schematically in Fig 1. Python programming language version 10 was used in this study for developing and validating ML algorithms. For data preprocessing, Numpy and Pandas modules

were employed, while the sci-kit learn library was utilized for developing supervised classifier algorithms.

2-1-Dataset description and data retrieval

The data for this study were obtained from the United Network for Organ Sharing (UNOS) online database [11]. Upon receiving written permission from UNOS, we accessed the recorded data pertaining to lung transplantation for our research. Our study included patients over 18 years old with end-stage lung disease who underwent lung transplants between 2005 and 2022. We performed a waiting list analysis using all available data entries from the United Network for Organ Sharing (UNOS) database for our study.

The priority of candidates on the waiting list was considered as the outcome, while the clinical and demographic characteristics of patients were considered as features or predictors.

2-2-Pre-processing process

Data pre-processing is a crucial step in ML techniques, especially when dealing with raw data from clinical databases or medical records that often contain missing or unclear information. To ensure the development of more accurate models based on appropriate data, we followed a series of data pre-processing steps. The following steps were employed in this phase as pre-processing techniques.

1- Checking the duplicated values and records to remove the duplicates

2- De-identify records and remove irreverent features

3- Convert nominal and categorical features to numerical values

4- Identify missing data and missing values imputation

5- Outlier detection

6- Feature engineering and feature selection

7- Data transformation and normalization

2-2-1-Duplicate checking and removal of irreverent features

After duplicate checking, we consulted UNOS guidelines and experts to review all features and their definitions. Under their supervision, we removed identification variables and irrelevant features, such as ID columns, hospital center identification codes, and country of residence, to de-identify patients.

Following this, we converted the post-transplant survival days variable to years and excluded patients with a survival rate of less than two years. Next, we filtered out patients below 18 years of age and excluded any data before 2005. Additionally, records related to heart transplantation were removed from the dataset.

After removing irrelevant features in the first data-cleaning phase, we utilized the discretized operator to convert nominal values to numerical data. Categorical data were encoded using the LabelEncoder class too.

2-2-2- Missing data management:

To address missing data in our dataset, we conducted missing data imputation across the entire dataset. Initially, we assessed the specified columns or attributes to determine the extent of missing and unique data in each column. During this analysis, we discovered that the ICU column was empty and decided to delete it due to its lack of meaningful information. To impute the missing data, a threshold of 80% was set for feature removal with expert consultation. As a result, any column with more than 80% missing data was removed. If the missingness is due to inconsistent reporting rather than clinical irrelevance, dropping the column could exclude critical information about high-risk patients. In this case, domain experts might recommend retaining the column and using advanced imputation techniques or creating a binary indicator for missingness.

Along with feature removal, the sensitivity analysis was conducted on dataset which revealed that the inclusion of omitted features had a detrimental effect on the performance of the Random Forest Regressor model. Specifically, these features led to a decrease in the R^2 score from 0.95 to 0.68 and an increase in RMSE from 3.5 to 4.8. It means that the sensitivity analysis demonstrated that omitting features with missing data significantly degraded model performance (R^2 decreased from 0.95 to 0.68; RMSE increased from 3.5 to 4.8), suggesting these missing values were not missing at random (MNAR). The observed performance drop implies that the missingness may depend on unobserved factors systematically related to the outcome, warranting MNAR-specific methods for robust inference.

For the missing data imputation, we adopted a strategy of replacing missing data in numerical features with the mean value of each respective feature. This approach allows us to retain the integrity of the dataset while minimizing the impact of missing data on our analysis. By performing these comprehensive steps of missing data imputation, we ensure the dataset is optimized for further analysis and modeling, enabling us to draw more accurate conclusions and insights.

2-2-3- Outliers handling

Outliers can significantly impact the performance and interpretability of machine learning models. Therefore, it is essential to investigate their causes before deciding whether to exclude, transform, or retain them. This exploration ensures that the preprocessing steps are justified and scientifically sound. To address outliers in the dataset, we first create distribution plots to visualize the data. Next, we apply the IQR method and use Box plots to identify outliers. Finally, we remove these outliers to prepare the data for further processing.

To address outliers in the dataset, a comprehensive approach was employed that included visual and statistical analysis to identify and understand the nature of the outliers. The distribution plots

and boxplots were created to visualize outliers and applied the IQR method to quantify their extent. Additionally, statistical analysis was conducted to assess the impact of outliers on the dataset and performed sensitivity analysis to evaluate their influence on model performance. Throughout this process, we ensured transparency and justification by documenting all outliers and providing context-specific reasons for their exclusion, such as data entry errors or clinically irrelevant extreme values [12, 13]. This rigorous approach ensured that the removal of outliers was methodologically sound and did not compromise the integrity of our analysis.

2-2-4-Feature engineering and feature selection

Given the high-dimensional nature of the dataset used in this study, feature engineering and selection were critical steps to reduce dimensionality, eliminate irrelevant or redundant features, and enhance model performance. Feature selection aims to identify a subset of features that effectively describe the problem with minimal loss of information and computational efficiency [14]. All phases of feature engineering and selection were conducted under the supervision of clinical experts to ensure the relevance and validity of the selected features.

Step one, correlation analysis: As the first step, correlation analysis was performed to assess the relationships between features and the target variable, as well as inter-feature correlations. This analysis helped identify highly correlated features that could introduce multicollinearity and redundancy into the model using heatmap graph. Features with a correlation coefficient above a predefined threshold were flagged for further evaluation.

Step two, variance threshold filtering: To eliminate low-variance features that contribute little to the model's predictive power, a variance threshold was applied. Features with variance below a specified threshold (e.g., 0.01) were removed, as they were deemed to have minimal impact on the target variable.

Step three, embedded feature selection with XGBoost: Following the initial filtering, an embedded feature selection technique was employed using the XGBoost algorithm. XGBoost provides intrinsic feature importance scores based on metrics such as gain, cover, and frequency. Features with very low importance scores (negative scores indicating no correlation with target value) were excluded from the final feature set.

Step four, expert review and validation: All selected features were reviewed and validated by subject matter experts to ensure their clinical, practical, and scientific relevance. This step was critical to avoid eliminating features that, although statistically significant, may not be clinically significant. For example, some features were retained for model customization based on expert consultation, despite having modest statistical significance.

Through a comprehensive and expert-guided feature selection process, we identified a subset of features that were statistically significant, domain-relevant, and impactful for model performance. This rigorous approach ensured the final model was both robust and clinically meaningful, with the selected features deemed critical for predicting the target variable.

2-2-5-Data transformation and normalization: In the end, data normalization was carried out to optimize the features for modeling purposes.

2-3- Splitting data and validation technique

During the model development process, the dataset was divided into training and testing data in an 80:20 ratio where 80% of the data was used for training the models and the remaining 20% was reserved for testing and validation. The training dataset is used to train the model, allowing it to learn patterns and relationships within the data based on the available data. The training dataset typically contains the bulk of the available data. In contrast, the testing dataset is intended solely to evaluate the model's performance on unseen data, ensuring an unbiased assessment of its

generalizability. This dataset is kept separate from the training process to provide a true measure of how the model performs in real-world scenarios. Both datasets are often split randomly, with common ratios such as 80:20 or 70:30, depending on the size and nature of the data.

This split ensured that the models were evaluated on unseen data to assess their generalization capability. To mitigate potential bias introduced by simple data splitting, cross-validation technique was employed. Through this technique, the dataset was divided into k folds, and each model was trained and validated k times, with each fold serving as the validation set once. The average performance metrics across all folds were calculated to ensure robust evaluation. The results of the k-fold cross-validation were consistent with those obtained from the simple 80:20 split, confirming the reliability of the initial approach.

2-4- Model development and tuning

The objective of this study was to develop a prediction model for a continuous numerical variable (priority score) using regression techniques to identify the most effective factors in selecting the most appropriate candidate for LTx. In this study, regression models were selected to examine the connection between input variables and output numerical values, as the target variable (outcome) is a continuous numerical value.

To determine the most influential factors and identify the best model, the performance of six regression-based models was evaluated: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model (MLP—a class of feedforward artificial neural network), and a deep learning model (DL). The selection of these models was done based on the type of target variable and the study objectives.

A hyperparameter tuning optimization technique was employed in this phase to improve model performance by optimizing the training process by determining the best hyperparameters for each model. This technique was used to prevent models that underfit or overfit the data [15]. After tuning parameters in each model, the models were trained with updated best hyperparameters, and all metrics were calculated again to achieve the best performance. We employed the random search method, a hyperparameter tuning technique where hyperparameters are randomly chosen from a predefined set to train a model.

2-4-1- Multiple linear regression (MLR)

Multiple linear regression (MLR) is a statistical technique used to estimate the relationship between a dependent variable and one or more independent variables. It is an extension of linear regression, which requires more than one predictor variable to forecast the response variable[16]. MLR is a significant regression algorithm that models the linear association between a dependent continuous variable and multiple independent variables[17]. Hence, we have chosen this model to predict the continuous variable (priority score) based on several independent variables. The equation for multiple linear regression is demonstrated below[17]:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7 \quad (1)$$

where y represents the priority; x_i is the considered variables; β_0 is the intercept; and β^i is the regression coefficients.

2-4-2- Random Forest Regressor (RF)

The random forest (RF) Regressor algorithm is a kind of ML approach that employs a group of decision trees, which are trained on a subset of the data, to make predictions. This technique is designed to stabilize the algorithm and decrease variance by using multiple trees. The RF regressor algorithm is widely recognized as a popular model in developing regression models because of its strong performance with large datasets and diverse data types [18, 19].

2-4-3- Support Vector Machines Regressor (SVM)

Support vector machine regression (SVM) is a versatile regression function that can be used to solve both classification and regression problems. SVM is a supervised learning algorithm that fits a regression to the training data by reducing the distance between the sampled points and the fitted hyperplane[20, 21]. One advantage of SVM is that it is a sparse algorithm, meaning that it only needs information from a limited number of data points[22].

2-4-4- XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another ML library that is available for free and offers a powerful and efficient implementation of the gradient boosting algorithm[23]. Gradient boosting is a technique that involves creating an ensemble of tree-based models and then combining them to create a more accurate overall model than any of the individual models in the sequence[24]. XGBoost is a popular choice for those who require an effective and optimized implementation of gradient boosting[25].

2-4-5- Multilayer Perceptron Model (MLP)

The Multilayer Perceptron (MLP) is considered one of the top regression models in the field of artificial neural networks. It is equipped with the capability to learn from training data using a variety of training algorithms and rules. This feature allows the MLP to acquire numerous advantages, including increased capacity. As a result, the MLP operates as a self-regulating model that utilizes specific learning algorithms to enhance its performance when encountering new inputs [26, 27].

2-4-6- Deep Learning Model (DL)

A deep learning model can be used for regression problems by learning a mapping from input features to the target output. Deep learning is an adaptable model proficient at effectively managing intricate data relationships. It proves especially beneficial when working with extensive datasets

where traditional regression methods might not uncover intricate patterns. Nonetheless, to prevent overfitting and attain peak performance, these models necessitate meticulous calibration and validation[28-30]. Occasionally, due to the complex nature of implementing these models, simpler regression models may outperform them.

2-5- Performance evaluation

Typically, regression models are evaluated based on a function that measures the difference between the predicted and actual numerical value of the target variable, such as the priority score[31]. In this study, three popular evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and R-squared (R^2) score to assess the performance of the developed models [32].

$$(1)\quad MAE = \frac{\sum_{i=1}^{n} |R^*(i) - R(i)|}{n}$$

$$(2)\quad RMSE = \left\{ \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{n} \right\}^{1/2}$$

$$(3)\quad R^2 = 1 - \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{\sum_{i=1}^{n} (R^*(i) - m(i))^2}$$

In these formulas; variable n refers to the number of samples; $R^*(i)$ denotes the retrieved value predicted by the model; $R(i)$ denotes the analyzed value; and $m(i)$ denotes the average analyzed value.

To validate the developed machine learning (ML) models and reduce bias, we employed k-fold cross-validation. This technique overcomes the limitations of a simple train/test split by dividing the available data into multiple folds or subsets. By averaging the results across these folds, we achieve a more robust estimate of the model's performance compared to a simple train/test split.

2-6- Feature Importance

To implement Explainable AI (XAI), a technique employed to elucidate the impact of each feature on the model is the SHAP (Shapley Additive Explanations) method. The SHAP method aims to

enhance the transparency and interpretability of machine learning models by drawing on cooperative game theory [33]. For instance, linear models utilize their coefficients to gauge the significance of each feature. However, these coefficients are influenced by the scale of the variable itself, potentially resulting in misinterpretations [34]. The same can be found in tree-based models for feature ranking. This is precisely why SHAP (Shapley Additive Explanations) becomes valuable for model interpretation [35]. The absolute value of SHAP provides insight into how significantly an individual feature influences the prediction [36]. Once we identify the optimal model for priority prediction, we'll leverage the SHAP technique to assign weights to the most critical features. These features will then be ranked based on their importance and impact on the final priority score.

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

3-Results

3-1- Dataset description

The raw dataset including all organs comprises information on 196,270 patients who underwent lung transplantation, as well as data related to lung donors. This comprehensive dataset includes 545 features, encompassing demographic and clinical details about the organ recipients, biomarkers, laboratory test results, and characteristics of the donated organs (Table A-2 in Appendix A). Additionally, it provides insights into various patient outcomes, such as post-transplant survival rates, occurrences of acute organ rejection, priority levels, duration of intensive care unit stay, post-transplant infections, and instances of re-transplantation.

To preprocess the dataset, we converted the transplantation date data type to a string format and extracted the year column by parsing the month, hour, and year components. We removed all data prior to 2005 due to the absence of a prioritization system during that period. Furthermore, to focus exclusively on adult transplants, we excluded information related to pediatric transplants for children and adolescents under 18 years of age. As a result, our initial dataset comprised 183,086 records.

Subsequently, we filtered the dataset to include only post-transplant survival records exceeding one year. After that, we eliminated any records with missing priority scores. Ultimately, our final dataset consisted of 45,966 records for subsequent analysis.

3-2- Exploratory data analysis

Following imputation of missing independent variable data and preprocessing steps, the overall patient population consisted of 66.88% men and 33.20% women, with a median age of 54.27±14.24 years. Our target variable is the priority (or allocation) score, which represents a continuous numerical value. In Table A-1 in Appendix A, we present descriptive analysis and the frequency distribution of various demographic and clinical variables within the dataset.

Furthermore, we employed a data visualization method to enhance our comprehension of the data and dataset. This approach aids in verifying the integrity of the data and detecting any apparent inaccuracies. Incorporating data visualization is essential for all data science projects across various fields [37].

3-3- Data cleaning and preprocessing

In the method section, we present detailed information regarding preprocessing procedures applied to the whole dataset. After de-identifying the dataset, it was limited to 40,024 records and 445 features. During the initial data cleaning phase, we removed irrelevant features, reducing the total

to 322. Subsequently, we performed missing values imputation, resulting in 215 features available for further analysis. Next, in the correlation analysis phase, our dataset contains over 165 features post-pre-processing.

Due to the dataset's high dimensionality, we applied pre-processing techniques to select only the most important features based on their importance scores. As a result, we narrowed down the dataset to 65 features in the first phase of feature engineering. After further applying feature engineering and selection techniques, our final dataset consisted of 32,966 records, containing 15 features.

3-4- Development and evaluation of regression models

The prediction models were developed by training several selected features obtained during the feature engineering phase. We used 80% of the dataset to train the algorithms and the rest 20% to test and validate their efficacy (80:20) and all six regression algorithms were trained based on trained data.

To address the bias of training using simple data splitting the average score, K-fold cross-validation was done and K was considered as 10 folds. The results showed that average scores of 10-fold cross-validation in six ML algorithms are the same as the simple splitting data process. Subsequently, the hyperparameters were fine-tuned using a hyperparameter tuning technique to enhance the performance of the developed models.

The optimized and selected hyperparameters are documented in Table 1. Next, the MAE, RMSE, and R^2 values for each optimized model were calculated and represented in Table 2. Finally, all the optimized ML models were compared based on their R^2 scores in combination with other relevant metrics. In this task, the RF Regressor emerges as the most robust model, demonstrating superior performance with an impressive 95.168% R^2 value, which indicates it explains nearly

96% of the variance in the data, significantly outperforming other techniques. The Adjusted R^2 metric, which penalizes unnecessary model complexity, closely mirrors the standard R^2 here (95.163%). The model's exceptional performance is evidenced by its lowest Mean Squared Error (12.548), Mean Absolute Error (2.056), and Root Mean Square Error (3.542), suggesting highly accurate and precise predictions. The superior performance of RF model can be attributed to its ability to handle complex, non-linear relationships in medical data through ensemble learning, where multiple decision trees are combined to create a more flexible and generalized predictive model. In contrast, traditional linear methods like Linear Regression and Support Vector Machines struggled, achieving R^2 values below 53%, which suggests the priority prediction requires sophisticated, non-linear modeling approaches that can capture intricate patterns in medical datasets. The progression from linear to ensemble and advanced machine learning techniques clearly demonstrates the importance of selecting appropriate algorithms for complex predictive challenges in healthcare.

As a result, the RF Regressor model emerged as the top performer among the developed prediction models. We made this determination based on a comprehensive evaluation of various metrics, utilizing the best features.

3-5- Most important features to select the most appropriate candidate

Upon selecting the best model, we proceeded to identify and weigh the most influential features using the SHAP library within the final model. Initially, a prediction model based on the chosen regression model was created. Subsequently, the importance of each feature was determined by analyzing the set of trees generated by the model using the SHAP technique. The SHAP library assigns a score to each feature based on its impact on the prediction model. The ranking of the

variables used in the ultimate model is visually represented in Fig 2, which is a widely recognized and popular chart produced by SHAP.

Ultimately, the researchers pinpointed the 11 most effective features, each receiving the highest score in candidate prioritization.

These features, along with their explanations, are detailed in Table 3. Notably, it showed that factors such as a patient's oxygen consumption and diagnosis played a significant role in prioritizing the waiting list. Additionally, the patient's waiting time on the transplant list emerged as another influential factor. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model (Fig 3).

4- Discussion

The study aimed to explore the feasibility of utilizing machine learning (ML) methods to predict priority levels for patients on the waiting list for lung transplants and to pinpoint the critical factors influencing priority allocation. Despite the potential advantages of employing ML algorithms in organ allocation [38], there is a lack of research on their application specifically in lung transplantation. This investigation led to the development of a decision support tool for estimating transplantation priorities.

Currently, the decision-making process for prioritizing individuals on organ transplant waiting lists is predominantly reliant on physicians' subjective judgments, often following "first-come, first-served" or "longer waiting time" principles rather than utilizing sophisticated mathematical models [39, 40]. Researchers recommend that authorities explore more equitable and innovative solutions for allocating donor organs to patients on waiting lists. As a result, researchers in the field of transplantation have concentrated on developing advanced models to forecast priority rankings

and outcomes for recipients based on pre-transplantation factors [41, 42]. Similarly, we employed ML models to investigate more appropriate factors in assigning organs to recipients.

Prior studies on organ allocation have focused only on classification models to predict the risk of mortality following transplantation [41, 43]. However, these approaches have not been highly effective in improving the prioritization of patients on lung transplant waiting lists[44, 45]. In contrast, our developed model takes into account various factors such as disease type, oxygen saturation, demographics, clinical tests, and functional status.

In the context of machine learning (ML), the effectiveness of methods depends not only on their design and techniques but also on the quality and suitability of the data they operate on. To overcome the limitations of prior research, which often relied on a single ML technique and small sample sizes, our study takes a different approach. We incorporate multiple ML techniques to enhance the accuracy of our results, leveraging a large dataset sourced from the United Network for Organ Sharing (UNOS) database.

Our algorithm yields slightly superior results. To enhance the robustness of our model, we employed various data preprocessing techniques and feature engineering methods. These approaches allowed us to identify the most relevant and informative features in the data while discarding redundant or noisy ones [46, 47]. Data preprocessing plays a crucial role in improving data quality and enhancing the accuracy of knowledge extraction [48]. Additionally, by reducing data complexity and dimensions, our models became better equipped to capture underlying patterns and relationships, resulting in improved predictive performance [10, 47].

Our analysis reveals that employing the RF regressor model, which incorporates 15 features from the most significant donor and recipient variables available prior to transplantation, represents an effective approach for assigning an allocation score to each candidate on the waiting list. This

outperforms other regression models. RF was specifically chosen due to its favorable prediction performance in previous research [49]. Implementing the developed model as an AI-based decision support tool could assist physicians in integrating clinical insights into their decision-making processes and point-of-care scenarios, thereby enhancing the practical utility of the data.

ML-based models rely on intricate mathematical structures and multi-dimensional datasets, often yielding complex patterns and relationships that can be challenging for humans to grasp. To address this complexity and limitation, researchers have turned to SHAP (Shapley Additive explanations) summary analysis. This technique identifies the top influential features within the final model. By doing so, it sheds light on which parameters should take precedence when selecting the most suitable recipient with the highest priority—a factor that has not received extensive exploration in prior studies. On the other hand, as the research community increasingly shifts toward explainable AI (XAI) methods [50, 51], the adoption of this approach represents a significant step forward. By employing XAI techniques, the performance of developed models can be interpreted and explained more transparently, fostering greater trust and understanding in their outcomes.

Our study possesses some limitations. Despite the dataset under consideration being of a substantial size, it was obtained from a freely accessible dataset, aligning with the structure of the UNOS database will allow for the collection of patient information tailored to researchers' requirements. While our study demonstrates the effectiveness of the random forest (RF) model in predicting outcomes for lung transplant patients using the UNOS dataset, it is important to note the lack of external validation as a limitation. The model was developed and validated only on the UNOS dataset, which, although comprehensive, may contain biases related to specific populations and practices in the United States. However, we plan to focus on collaborating with international

transplant registries or multicenter studies to validate the performance of the model in different populations and healthcare settings. This will enhance the validity of the model and its potential for widespread clinical adoption. External validation on independent datasets from different geographic regions or healthcare systems is essential to ensure the generalizability and robustness of our findings. As part of future work, we are developing an intelligent lung transplant patient information system at our center. Building on previous efforts to apply AI-based techniques in solid organ transplantation [52-55], this system aims to integrate the current model with patients' medical records while leveraging additional AI-based models to enhance its performance.

5-Conclusion

During this study, we succeeded in developing a priority prediction model based on the huge data of the UNOS database using ML models with the least error. Our research is among the pioneering studies that employ the SHAP method as an XAI technique to enhance the comprehensibility of the proposed model intended for clinicians. Additionally, the automated auxiliary model that we created can assist clinicians in acquiring a better understanding of the transplant priority estimation and the crucial factors that influence patient survival.

6-Declaration

Ethics approval and consent to participate

The research was approved by the Tehran University of Medical Sciences Ethics Committee (IR.TUMS.IKHC.REC.1401.143). All methods were performed based on the relevant guidelines and regulations. Consent for participation was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

Consent for publication

Consent for publication was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Availability of data and materials

The data used in this article can be obtained from the United Network for Organ Sharing (UNOS) database by visiting www.unos.org/data. However, there are limitations on accessing this data, as it was used under a license for the current study and is not accessible to the general public. The interpretation and reporting of this 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 United States government.

Funding

This research was funded by the Thoracic Research Center through, Tehran University Medical Sciences by Grant No (59042). The funding body played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

Authors' contributions

H.A., M.G., R.S., and M.A.G. contributed to the conception and design of the study. M.G. and H.A. acquired the data. M.G., H.A., R.S., and M.A.G. were involved in data interpretation and analysis. M.G. and H.A. drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version to be published. H.A. is the guarantor.;

Acknowledgments

The data reported here have been supplied by the United Network for Organ Sharing (UNOS/OPTN) as the contractor for the Organ Procurement and Transplantation Network. We

express our gratitude to the UNOS organization for allowing access to the data. We would like to extend our sincere thanks to the Thoracic Research Center of the Tehran University of Medical Sciences (TUMS) for their support and cooperation during this research.

References

1. Van der Mark SC, Hoek RAS, Hellemons ME: **Developments in lung transplantation over the past decade**. *European Respiratory Review* 2020, **29**(157):190132.
2. Verleden GM, Dupont L, Yserbyt J, Schaevers V, Raemdonck DV, Neyrinck A, Vos R: **Recipient selection process and listing for lung transplantation**. *Journal of Thoracic Disease* 2017, **9**(9):3372-3384.
3. Smits JM, Nossent G, Evrard P, Lang G, Knoop C, Kwakkel-van Erp JM, Langer F, Schramm R, van de Graaf E, Vos R *et al*: **Lung allocation score: the Eurotransplant model versus the revised US model – a cross-sectional study**. *Transplant International* 2018, **31**(8):930-937.
4. Lancaster TS, Miller JR, Epstein DJ, DuPont NC, Sweet SC, Eghtesady P: **Improved waitlist and transplant outcomes for pediatric lung transplantation after implementation of the lung allocation score**. *J Heart Lung Transplant* 2017, **36**(5):520-528.
5. Safdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M: **Using data mining techniques to fight and control epidemics: A scoping review**. *Health and Technology* 2021, **11**(4):759-771.
6. Gholamzadeh M, Abtahi H, Safdari R: **Comparison of different machine learning algorithms to classify patients suspected of having sepsis infection in the intensive care unit**. *Informatics in Medicine Unlocked* 2023, **38**:101236.
7. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, McEvoy D, Stylianopoulos T, Munn LL, Dutta S *et al*: **Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19**. *NPJ Digit Med* 2021, **4**(1):87.
8. Safdari R, Deghatipour A, Gholamzadeh M, Maghooli K: **Applying data mining techniques to classify patients with suspected hepatitis C virus infection**. *Intelligent Medicine* 2022, **2**(4):193-198.
9. Gholamzadeh M, Abtahi H, Safdari R: **Machine learning-based techniques to improve lung transplantation outcomes and complications: a systematic review**. *BMC Medical Research Methodology* 2022, **22**(1):331.
10. Miller PE, Pawar S, Vaccaro B, McCullough M, Rao P, Ghosh R, Warier P, Desai NR, Ahmad T: **Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database**. *Journal of Cardiac Failure* 2019, **25**(6):479-483.
11. LeClaire JM, Smith NJ, Chandratre S, Rein L, Kamalia MA, Kohmoto T, Joyce LD, Joyce DL: **Solid organ donor-recipient race-matching: analysis of the United Network for Organ Sharing database**. *Transpl Int* 2021, **34**(4):640-647.
12. Mazarei A, Sousa R, Mendes-Moreira J, Molchanov S, Ferreira HM: **Online boxplot derived outlier detection**. *International Journal of Data Science and Analytics* 2025, **19**(1):83-97.
13. Kalaivani B, Ranichitra A: **Unveiling the Impact of Outliers: An Improved Feature Engineering Technique for Heart Disease Prediction**. In*: 2024; Singapore*: Springer Nature Singapore; 2024: 469-478.
14. Theng D, Bhoyar KK: **Feature selection techniques for machine learning: a survey of more than two decades of research**. *Knowledge and Information Systems* 2024, **66**(3):1575-1637.
15. Li D, Liu Z, Armaghani DJ, Xiao P, Zhou J: **Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments**. *Sci Rep* 2022, **12**(1):1844.

16. Uyanık GK, Güler N: **A Study on Multiple Linear Regression Analysis**. *Procedia - Social and Behavioral Sciences* 2013, **106**:234-240.
17. Kayri M, Kayri I, Gencoglu MT: **The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data**. In: *2017 14th International Conference on Engineering of Modern Electric Systems (EMES): 1-2 June 2017 2017*; 2017: 1-4.
18. Dai B, Chen RC, Zhu SZ, Zhang WW: **Using Random Forest Algorithm for Breast Cancer Diagnosis**. In: *2018 International Symposium on Computer, Consumer and Control (IS3C): 6-8 Dec. 2018 2018*; 2018: 449-452.
19. Smith PF, Ganesh S, Liu P: **A comparison of random forest regression and multiple linear regression for prediction in neuroscience**. *Journal of Neuroscience Methods* 2013, **220**(1):85-91.
20. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ: **Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes**. *BMC Medical Informatics and Decision Making* 2010, **10**(1):16.
21. Sarker IH: **Machine Learning: Algorithms, Real-World Applications and Research Directions**. *SN Computer Science* 2021, **2**(3):160.
22. Huang H, Wei X, Zhou Y: **An overview on twin support vector regression**. *Neurocomputing* 2022, **490**:80-92.
23. Bentéjac C, Csörgő A, Martínez-Muñoz G: **A comparative analysis of gradient boosting algorithms**. *Artificial Intelligence Review* 2021, **54**(3):1937-1967.
24. Li S, Zhang X: **Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm**. *Neural Computing and Applications* 2020, **32**(7):1971-1979.
25. Liu J, Wu J, Liu S, Li M, Hu K, Li K: **Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model**. *PLOS ONE* 2021, **16**(2):e0246306.
26. Sananmuang T, Mankong K, Chokeshaiusaha K: **Multilayer perceptron and support vector regression models for feline parturition date prediction**. *Heliyon* 2024, **10**(6):e27992.
27. Abiodun OI, Jantan A, Omolara AE, Dada KV, Umar AM, Linus OU, Arshad H, Kazaure AA, Gana U, Kiru MU: **Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition**. *IEEE Access* 2019, **7**:158820-158846.
28. Sarker IH: **Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions**. *SN Computer Science* 2021, **2**(6):420.
29. Ahmed SF, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, Mofijur M, Shawkat Ali ABM, Gandomi AH: **Deep learning modelling techniques: current progress, applications, advantages, and challenges**. *Artificial Intelligence Review* 2023, **56**(11):13521-13617.
30. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**(7553):436-444.
31. Ramesh A, Ramamoorthy S, Buhari SM: **Forecasting Spread of COVID-19 Using Regression Algorithm**. In: *Soft Computing for Problem Solving: 2021// 2021; Singapore*: Springer Singapore; 2021: 461-467.
32. Karunasingha DSK: **Root mean square error or mean absolute error? Use their ratio as well**. *Information Sciences* 2022, **585**:609-629.
33. Ekanayake IU, Meddage DPP, Rathnayake U: **A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)**. *Case Studies in Construction Materials* 2022, **16**:e01059.
34. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI: **From Local Explanations to Global Understanding with Explainable AI for Trees**. *Nat Mach Intell* 2020, **2**(1):56-67.
35. Rodríguez-Pérez R, Bajorath J: **Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions**. *Journal of Computer-Aided Molecular Design* 2020, **34**(10):1013-1026.
36. Kim Y, Kim Y: **Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models**. *Sustainable Cities and Society* 2022, **79**:103677.

37. Patel Darshan R, Reddy PVB: **The Importance of Data Visualization in Exploratory Data Analysis**. *Journal of Advanced Zoology* 2023, **44**(S6):923-929.
38. Peloso A, Moeckli B, Delaune V, Oldani G, Andres A, Compagnon P: **Artificial Intelligence: Present and Future Potential for Solid Organ Transplantation**. *Transpl Int* 2022, **35**:10640.
39. Bunnik EM: **Ethics of allocation of donor organs**. *Curr Opin Organ Transplant* 2023, **28**(3):192-196.
40. Madwar S: **United States officials propose further retreat from first-come, first-served organ donation**. *Cmaj* 2011, **183**(10):E639-640.
41. Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, Bailey J: **Machine-Learning Algorithms Predict Graft Failure After Liver Transplantation**. *Transplantation* 2017, **101**(4):e125-e132.
42. Gotlieb N, Azhie A, Sharma D, Spann A, Suo N-J, Tran J, Orchanian-Cheff A, Wang B, Goldenberg A, Chassé M *et al*: **The promise of machine learning applications in solid organ transplantation**. *npj Digital Medicine* 2022, **5**(1):89.
43. Jawitz OK, Raman V, Becerra D, Klapper J, Hartwig MG: **Factors associated with short- versus long-term survival after lung transplant**. *J Thorac Cardiovasc Surg* 2022, **163**(3):853-860.e852.
44. Brahmbhatt JM, Hee Wai T, Goss CH, Lease ED, Merlo CA, Kapnadak SG, Ramos KJ: **The lung allocation score and other available models lack predictive accuracy for post-lung transplant survival**. *J Heart Lung Transplant* 2022, **41**(8):1063-1074.
45. Dalton JE, Lehr CJ, Gunsalus PR, Mourany L, Valapour M: **Refining the Lung Allocation Score Models Fails to Improve Discrimination Performance**. *Chest* 2023, **163**(1):152-163.
46. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM: **A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction**. *Frontiers in Bioinformatics* 2022, **2**.
47. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
48. García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F: **Big data preprocessing: methods and prospects**. *Big Data Analytics* 2016, **1**(1):9.
49. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z: **Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan**. *BMJ Nutrition, Prevention & Health* 2021, **4**(1):140-148.
50. S Band S, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT, Liang H-W: **Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods**. *Informatics in Medicine Unlocked* 2023, **40**:101286.
51. Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, Bora PK, Almasri A, Alkhawaldeh RS, Hussain S *et al*: **A review of Explainable Artificial Intelligence in healthcare**. *Computers and Electrical Engineering* 2024, **118**:109370.
52. Abtahi H, Shahmoradi L, Amini S, Gholamzadeh M: **Design and evaluation of a Mobile-Based decision support system to enhance lung transplant candidate assessment and management: knowledge translation integrated with clinical workflow**. *BMC Medical Informatics and Decision Making* 2023, **23**(1):145.
53. Gholamzadeh M, Safdari R, Amini S, Abtahi H: **Feasibility study and determination of prerequisites of telecare programme to enhance patient management in lung transplantation: a qualitative study from the perspective of Iranian healthcare providers**. *BMJ Open* 2023, **13**(6):e073370.
54. Abtahi H, Safdari R, Gholamzadeh M: **Pragmatic solutions to enhance self-management skills in solid organ transplant patients: systematic review and thematic analysis**. *BMC Primary Care* 2022, **23**(1):166.
55. Gholamzadeh M, Abtahi H, Safdari R: **Telemedicine in lung transplant to improve patient-centered care: A systematic review**. *International Journal of Medical Informatics* 2022, **167**:104861.

Tables

Table 1- The best selected hyperparameters

	Algorithm	Hyperparameters
1	Multiple linear regression	positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
2	Random Forest Regressor	n_estimators= 90, min_samples_split= 2, min_samples_leaf= 1, max_samples 10000, max_features: sqrt, max_depth=10
3	SVM Regressor	C =9.11158, loss='epsilon_insensitive', max_iter=5000
4	XGBoost Regressor	subsample=1, min_child_weight= 5, max_depth= 6, learning_rate=0.1, colsample_bytree=0.75
5	MLP	solver= 'sgd', Learning_rate= 'adaptive', hidden_layer_sizes: (20,), alpha: 0.001, activation: logistic
6	DL	Optimizer= 'sgd', batch_size= 16, activation= 'relu'

Table 2- The evaluation metrics of developed models and comparison of the model performance

	Model	*R^2(%)*	*Adjusted R^2(%)*	*MSE*	*MAE*	*RMSE*
1	Random Forest Regressor	95.168	95.163	12.548	2.056	3.542
2	XGBoost Regressor	82.88	82.87	58.326	4.487	7.637
3	Deep Learning algorithm	68. 736	68.23	80.096	42.096	45.05
4	MLP Regressor	66.003	65.98	88.97	5.681	9.432
5	Linear Regression	52.259	52.23	123.989	6.984	11.131
6	Support Vector Machines	48.590	48.55	133.591	6.570	11.555

Table 3- The top 11 features identified by the SHAP method based on the prediction model

#	Feature	Description
1	INIT_O2	The amount of oxygen needed when the transplant candidate is on the waiting list
2	GROUPING	Lung transplant candidate diagnosis group
3	DAYSWAIT_CHRON	The amount of waiting time of patients on the waiting list - up-to-date waiting time
4	MED_COND_TRR	The status of the patient's lungs at the time of the last clinical evaluation
5	HEMO_SYS_TRR	The latest status of Hemodynamics Pcw (Sys) MM/Hg
6	END_O2	O2 Requirement at rest
7	VENTILATOR_TCR	The patient's status in terms of the need for a ventilator
8	LIFE_SUP_TCR	The amount of social and financial support
9	CIG_Use	History of cigarette use
10	Vent_Support_TRR	Episode of ventilatory support
11	Transfusion	Events occurring between listing and transplant

Figure legends

Fig 1. Schematic diagram illustrating the proposed methodology for developing machine learning models. The process includes data preprocessing, feature engineering, model training, and evaluation, followed by a systematic comparison of multiple models using performance metrics to identify and select the optimal model for deployment.

Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm

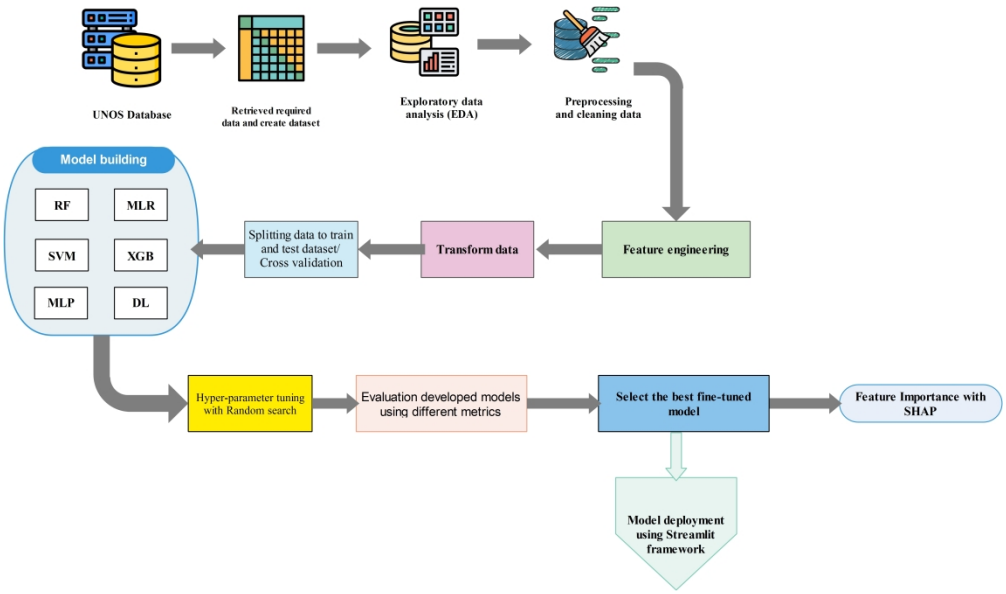
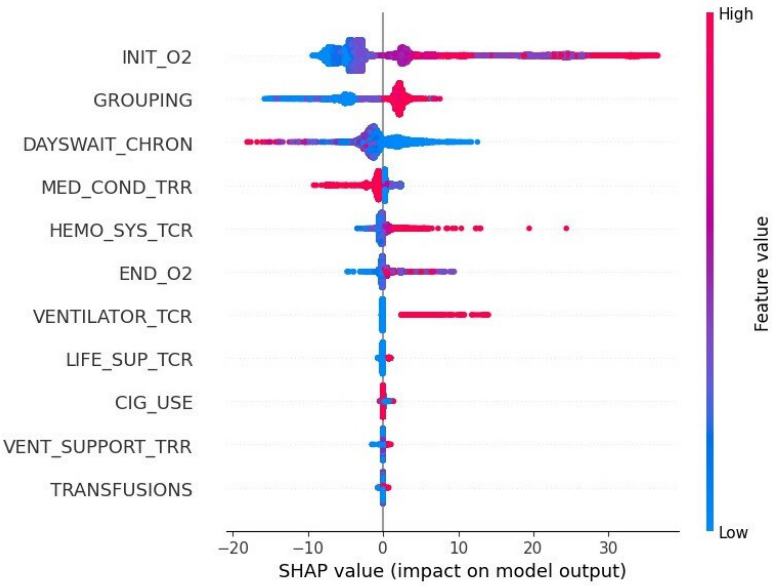


Fig 1. Schematic diagram of the proposed method

282x165mm (600 x 600 DPI)



(a)

(b)

Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

41x45mm (600 x 600 DPI)

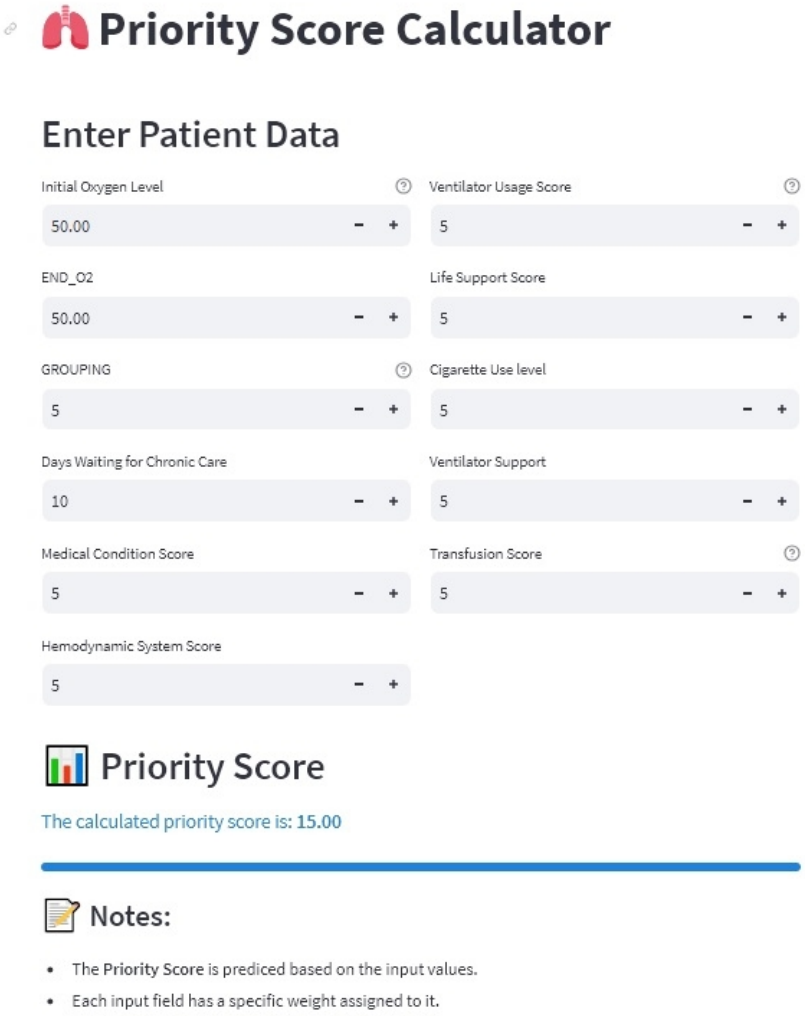


Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm

122x136mm (144 x 144 DPI)

Table A-1- Dataset description

Summary statistics of the selected continuous predictors (N=)					
	Variable	**Range**	**Mean (SD)**	**SE**	**95% Conf.**
Lung recipients	Age	18-58	54.27 (17.30)	0.095	39.2977
	BMI	14.997- 44.77	25.3 (3.83)	0.021	25.2586
	FEV1 value	5-120	39.484 (17.30)	0.095	39.2977
	Initial creatinine	0.1-24	0.841 (0.407)	0.002	0.8369
	Total Albumin serum	0.5-24	3.8787 (0.406)	0.002	3.8743

Summary statistics of selected categorical predictors (N=)				
	Variable		**n**	**Percentage (%)**
Lung recipients	**Gender**	Male	18085	54.86
		Female	14881	45.14
	ABO	A	1513	38.31
		B	4529	11.32
		AB	1513	3.78
		O	18648	46.59
	History of Malignancy	Positive	363	1.10
		Negative	30061	91.19
		Unknown	2542	7.71
	History of previous transplantation	Having	808	2.45
		Not Having	32158	97.55
Donor	**GENDER**	Male	14154	42.94
		Female	14154	42.94
		Unknown	9704	29.44
	ABO	A	279	0.85
		B	2544	7.72
		AB	11832	35.89
		O	13038	39.55
		Unknown	279	0.85
	History of Malignancy	Positive	9756	29.59
		Negative	21300	64.61
		Unknown	1910	5.79

Table A-2- Baseline Dataset description from UNOS database

Variables	count	mean	std	min	25%	50%	75%	max
GROUPING	32966	1.938725	1.316087	0	0	3	3	3
GENDER	32966	0.548596	0.49764	0	0	1	1	1
ABO	32966	1.667658	1.385379	0	0	2	3	3
WGT_KG_TCR	32966	72.41614	17.80507	3.1	59.422	72.575	84.822	212
HGT_CM_TCR	32966	168.2924	12.24525	5	160.02	168.6197	175.5	210.82
FUNC_STAT_TCR	32966	2063.003	278.1843	1	2040	2060	2070	4100
DIAB	32966	6.764454	72.83496	0	1	1	1	998
MALIG_TCR	32966	0.165231	0.540516	0	0	0	0	2
TOT_SERUM_ALBUM	32966	3.878676	0.406892	0.5	3.872302	3.872302	3.872302	9.8
RESIST_INF	32966	0.106049	0.393611	0	0	0	0	2
HEMO_SYS_TCR	32966	42.63604	16.84337	0	32	39	47	180
HEMO_PA_DIA_TCR	32966	17.79379	8.935308	0	12	17	21	110
HEMO_PA_MN_TCR	32966	27.51146	10.9713	0	21	26	31	110
HEMO_PCW_TCR	32966	10.69536	5.374686	0	7	10	14	50
HEMO_CO_TCR	32966	5.261608	1.407182	0.2	4.4	5	5.96	15
CIG_USE	32966	0.565613	0.495684	0	0	1	1	1
TCR_DUR_ABSTAIN	32966	57.32261	166.7262	1	7	7	56	998
LAST_INACT_REASON	32966	5.263635	1.500713	1	5	5	5	16
INIT_STAT	32966	7043.271	178.322	7010	7010	7010	7010	7999
INIT_O2	32966	4.097941	4.138358	0	2	3	4.15641	35
END_O2	32966	5.413994	5.203736	0	2.5	4	6	26.3
INIT_CREAT	32966	0.841288	0.407841	0.1	0.68	0.8	0.98	24
END_CREAT	32966	0.84711	0.42162	0.08	0.66	0.8	0.99	25
CALC_LAS_LISTDATE	32966	43.04133	16.16766	0	33.67689	37.5457	45.01999	96.22491
DAYSWAIT_CHRON	32966	213.5144	367.0553	0	20	73	236	5120

Variables	count	mean	std	min	25%	50%	75%	max
INIT_AGE	32966	54.27398	14.24429	0	48	58	64	81
LIFE_SUP_TCR	32966	0.070527	0.256037	0	0	0	0	1
VENTILATOR_TCR	32966	0.039344	0.194414	0	0	0	0	1
INIT_LLU_FLG	32966	0.356671	0.479024	0	0	0	1	1
INIT_RLU_FLG	32966	0.354638	0.478411	0	0	0	1	1
INIT_BLU_FLG	32966	0.830249	0.375419	0	1	1	1	1
END_LLU_FLG	32966	0.363071	0.480892	0	0	0	1	1
END_RLU_FLG	32966	0.361342	0.480397	0	0	0	1	1
END_BLU_FLG	32966	0.84211	0.364643	0	1	1	1	1
DR51	32966	19.82075	38.82521	0	0	0	0	99
DR51_2	32966	2.935509	16.52555	0	0	0	0	99
DR52	32966	20.40129	39.14027	0	0	0	0	99
DR52_2	32966	2.85661	16.27853	0	0	0	0	99
DR53	32966	20.14518	39.01353	0	0	0	0	99
DR53_2	32966	2.936359	16.51835	0	0	0	0	99
DQ1	32966	2.277498	20.30464	0	0	0	1	609
DQ2	32966	3.00907	29.15105	0	0	0	0	609
MED_COND_TRR	32966	1.877571	1.353962	0	0	3	3	3
CREAT_TRR	32966	0.856159	0.354149	0.1	0.71	0.864311	0.9	25
DIAL_AFTER_LIST	32966	0.311958	0.468832	0	0	0	1	2
FEV1_TRR	32966	39.48449	17.30273	5	26	39.9418	45	120
HEMO_CO_TRR	32966	5.336505	1.149913	1	4.86	5.347303	5.53	15
HEMO_PA_DIA_TRR	32966	17.90618	7.31172	0	14	17.95162	19	110
HEMO_PA_MN_TRR	32966	27.36964	8.662376	0	23	27.43337	28	110
HEMO_PCW_TRR	32966	10.68073	4.350863	0	9	10.68758	12	50
HEMO_SYS_TRR	32966	42.50621	13.37896	0	35	42.6675	42.6675	180

Variables	count	mean	std	min	25%	50%	75%	max
INFECT_IV_DRUG_TRR	32966	0.463447	0.627966	0	0	0	1	2
INOTROP_VASO_CO_TRR	32966	0.670752	0.533664	0	0	1	1	2
INOTROP_VASO_DIA_TRR	32966	0.699539	0.52854	0	0	1	1	2
INOTROP_VASO_MN_TRR	32966	0.682855	0.53295	0	0	1	1	2
INOTROP_VASO_PCW_TRR	32966	0.67694	0.529701	0	0	1	1	2
INOTROP_VASO_SYS_TRR	32966	0.704544	0.527142	0	0	1	1	2
OTH_LIFE_SUP_TRR	32966	0.013802	0.116671	0	0	0	0	1
PCO2_TRR	32966	47.7681	10.66006	10	42.1	47.64814	48	120
PRIOR_LUNG_SURG_TRR	32966	0.392374	0.568063	0	0	0	1	2
STEROID	32966	0.948856	0.825643	0	0	1	2	2
TBILI	32966	0.594413	0.689298	0.1	0.4	0.6	0.611384	36
TRANSFUSIONS	32966	0.3857	0.552885	0	0	0	1	2
VENT_SUPPORT_TRR	32966	0.426985	0.600565	0	0	0	1	2
VENTILATOR_TRR	32966	0.039617	0.19506	0	0	0	0	1
INHALED_NO_TRR	32966	0.00546	0.073692	0	0	0	0	1
PRIOR_CARD_SURG_TYPE_OSTXT_TRR	32966	94.76445	3.836607	0	95	95	95	95
PROSTACYCLIN_TRR	32966	0.004156	0.064332	0	0	0	0	1
TRACHEOSTOMY_TRR	32966	0.35485	0.526356	0	0	0	1	2
ECMO_72HOURS	32966	0.706576	0.492341	0	0	1	1	2
INHALEDNO_72HOURS	32966	0.723928	0.500037	0	0	1	1	2
INTUBATED_72HOURS	32966	0.855942	0.555187	0	1	1	1	2
HBV_CORE	32966	1.042104	1.377525	0	0	0	3	3
HBV_SUR_ANTIGEN	32966	0.961506	1.387078	0	0	0	3	3
HBV_SURF_TOTAL	32966	2.247952	1.226029	0	2	3	3	3
CMV_STATUS	32966	1.73227	1.195281	0	0	2	3	3
HIV_SEROSTATUS	32966	0.983134	1.391581	0	0	0	3	3

Variables	count	mean	std	min	25%	50%	75%	max
HCV_SEROSTATUS	32966	0.98186	1.378863	0	0	0	3	3
EBV_SEROSTATUS	32966	2.148608	0.763592	0	2	2	3	3
CRSMATCH_DONE	32966	1.588151	0.585256	0	1	2	2	2
CPRA	32966	747.2997	431.0568	0	86	999	999	999
PREV_TX	32966	0.02451	0.154629	0	0	0	0	1
PREV_TX_ANY	32966	0.025784	0.158493	0	0	0	0	1
DA1	32966	9.947916	71.84442	1	2	3	10	6802
DA2	32966	45.96509	187.604	0	24	33	48	6802
DB1	32966	43.60744	179.9494	7	13	44	46	5501
DB2	32966	71.46245	234.5561	0	44	60	76	8201
DDR1	32966	10.58812	29.91802	1	4	11	11	1501
DDR2	32966	23.31059	55.5125	0	13	15	24	1602
RA1	32966	20.05406	130.3715	0	2	3	23	6801
RA2	32966	80.44543	330.7992	0	24	68	92	6802
RB1	32966	80.57483	311.7406	0	8	44	92	5703
RB2	32966	142.1197	480.6177	0	44	61	162	8201
RDR1	32966	21.51423	82.27982	0	4	13	24	1601
RDR2	32966	46.76925	139.0017	0	13	17	53	1602
AMIS	32966	35.6986	46.58073	0	1	2	99	99
DRMIS	32966	35.77683	46.56309	0	1	2	99	99
HLAMIS	32966	37.79779	45.08084	0	4	5	99	99
MALIG_TRR	32966	0.625705	0.486636	0	0	1	1	2
CMV_IGG	32966	2.304162	1.077762	0	2	3	3	3
CMV_IGM	32966	2.056331	1.285516	0	1	3	3	3
HIST_COCAINE_DON	32966	0.520142	0.67797	0	0	0	1	2
AGE_DON	32966	34.46505	11.73268	6	26	34	40	76

Variables	count	mean	std	min	25%	50%	75%	max
HBV_CORE_DON	32966	1.917915	1.369888	0	1	1	4	4
HBV_SUR_ANTIGEN_DON	32966	1.885306	1.367688	0	1	1	4	4
ABO_DON	32966	2.821543	1.383585	0	3	3	4	4
ALCOHOL_HEAVY_DON	32966	0.514227	0.67552	0	0	0	1	2
GENDER_DON	32966	1.018079	0.755208	0	0	1	2	2
HEP_C_ANTI_DON	32966	1.904447	1.369103	0	1	1	4	4
ANTIHYPE_DON	32966	0.743766	0.796865	0	0	1	1	2
BUN_DON	32966	19.6837	13.75726	0.4	12	20.21487	20.21487	245
CREAT_DON	32966	1.404116	1.269158	0.07	0.82	1.3	1.42009	37
PT_DIURETICS_DON	32966	1.283868	0.787579	0	1	1	2	2
PT_STEROIDS_DON	32966	1.348025	0.761386	0	1	2	2	2
PT_T3_DON	32966	0.306528	0.469017	0	0	0	1	2
PT_T4_DON	32966	1.229661	0.80519	0	1	1	2	2
PT_OTH2_OSTXT_DON	32966	5972.515	2591.637	0	4203.25	7694	7694	9269
PULM_INF_DON	32966	1.297792	1.174939	0	0	1	3	3
SGOT_DON	32966	96.90449	306.0249	0.3	31	68	97.52713	20000
SGPT_DON	32966	97.76702	365.3933	0.4	25	56	97	44117
TBILI_DON	32966	0.988108	1.057065	0	0.6	0.98776	0.98776	59
URINE_INF_DON	32966	0.963478	1.342444	0	0	0	3	3
VASODIL_DON	32966	0.535764	0.696622	0	0	0	1	2
VDRL_DON	32966	2.198811	1.818314	0	1	1	5	5
CLIN_INFECT_DON	32966	1.238003	0.799413	0	1	1	2	2
HIST_CIG_DON	32966	0.427804	0.607943	0	0	0	1	2
HIST_HYPERTENS_DON	32966	0.623582	0.748324	0	0	0	1	2
HIST_CANCER_DON	32966	0.323697	0.494865	0	0	0	1	2
DIABETES_DON	32966	0.395347	0.580236	0	0	0	1	2

Variables	count	mean	std	min	25%	50%	75%	max
HIST_OTH_DRUG_DON	32966	0.875902	0.823869	0	0	1	2	2
CMV_DON	32966	2.761421	1.145194	0	1	3	4	4
DDAVP_DON	32966	0.530547	0.69316	0	0	0	1	2
HEPARIN_DON	32966	1.663866	0.5103	0	1	2	2	2
ARGININE_DON	32966	1.221865	0.807242	0	1	1	2	2
WGT_KG_DON_CALC	32966	77.36488	14.8769	23.5	70	77.35348	81.6	189
BMI_DON_CALC	32966	26.19033	4.580242	10.58866	23.66783	26.22113	27.1731	66.03564
HBV_NAT_DON	32966	2.028181	1.401618	0	0	3	3	3
ABO_MAT	32966	1.638779	0.905258	1	1	1	3	3
DIAL_PRIOR_TX	32966	0.650064	0.483463	0	0	1	1	2
ISCHTIME	32966	5.292257	1.528211	0.042	4.566406	5.351905	5.632813	25
O2_REQ_CALC	32966	5.375088	4.289068	0	3	5.457995	5.457995	26.3
PRIOR_CARD_SURG_TRR	32966	0.328429	0.490872	0	0	0	1	2
MALIG	32966	0.411818	0.598424	0	0	0	1	2
HGT_CM_CALC	32966	169.9918	8.255988	122	165.1	170.0308	175	210.82
BMI_CALC	32966	25.29995	3.832576	14.99785	23.38251	25.39236	27.33117	44.77787
DISTANCE	32966	211.2596	210.9401	0	67	215	221	4137
VENT_SUPPORT_AFTER_LIST	32966	0.426985	0.600565	0	0	0	1	2
PROTEIN_URINE	32966	0.862131	0.821972	0	0	1	2	2
CARDARREST_NEURO	32966	0.397379	0.568028	0	0	0	1	2
PO2	32966	380.254	121.5752	3.2	368	382.3152	457	754
HIST_MI	32966	0.327246	0.493355	0	0	0	1	2
LV_EJECT	32966	58.08624	9.440397	1	58	58.11868	61	99
CORONARY_ANGIO	32966	2.091974	1.327938	1	1	1	4	4
BIOPSY_DGN	32966	3.081933	1.997858	1	1	5	5	5
HBSAB_DON	32966	3.893739	1.46287	0	3	3	6	6

Variables	count	mean	std	min	25%	50%	75%	max
EBV_IGG_CAD_DON	32966	4.806285	1.547243	0	4	4	7	7
EBV_IGM_CAD_DON	32966	2.707972	2.252266	0	1	1	6	6
CDC_RISK_HIV_DON	32966	0.534581	0.698812	0	0	0	1	2
INOTROP_SUPPORT_DON	32966	0.982254	0.83639	0	0	1	2	2
TRANSFUS_TERM_DON	32966	294.9172	454.811	0	0	1	998	998
PO2_FIO2_DON	32966	86.1261	21.28058	1	86	100	100	100
PCO2_DON	32966	37.02991	5.584925	10	34.7	37	39	110
BRONCHO_LT_DON	32966	321.9415	464.8433	1	2	2	998	998
BRONCHO_RT_DON	32966	331.7915	468.465	1	2	2	998	998
CHEST_XRAY_DON	32966	300.3643	455.5274	1	2	5	999	999
PH_DON	32966	7.415315	0.083511	5	7.4	7.415555	7.44	8
HEMATOCRIT_DON	32966	29.21583	4.461127	2.5	27	29.14996	30.6	71