Donor-Recipient Matching and Allocation in Pediatric Bone-Marrow Transplantation

Danilo Silva
Department of Artificial Intelligence
Polytechnic School of Porto
Porto, Portugal
1250424@isep.ipp.pt

Luís Magalhães
Department of Artificial Intelligence
Polytechnic School of Porto
Porto, Portugal
1100628@isep.ipp.pt

José Domingues
Department of Artificial Intelligence
Polytechnic School of Porto
Porto, Portugal
1000984@isep.ipp.pt

Ricardo Sousa
Department of Artificial Intelligence
Polytechnic School of Porto
Porto, Portugal
1201856@isep.ipp.pt

Tomás Pereira
Department of Artificial Intelligence
Polytechnic School of Porto
Porto, Portugal
1210830@isep.ipp.pt

***Abstract*—Pediatric allogeneic hematopoietic stem cell transplantation (HSCT) is a potentially curative therapy for malignant and non-malignant diseases, but outcomes depend strongly on donor–recipient compatibility and graft characteristics. Donor selection is therefore a high-stakes, time-sensitive decision in which clinicians must balance immunogenetic risk against urgency and donor availability. This work proposes a decision-support system integrated with predictive artificial intelligence models. Multi-criteria decision-making methods, namely the Analytic Hierarchy Process (AHP) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), are employed to rank and select the most suitable donor for a given recipient. One of the decision criteria is the predicted post-transplant survival time of the recipient, which is estimated using a machine learning model trained on a published pediatric unrelated-donor cohort. The overall goal is to support transparent, reproducible, and clinically meaningful donor–recipient matching and allocation decisions in pediatric HSCT.**

***Index Terms*—Pediatric hematopoietic stem cell transplantation, donor-recipient matching, machine learning, data-driven decision support, multi-criteria decision making**

I. Introduction

Allogeneic hematopoietic stem cell transplantation (HSCT) is an established therapeutic option for a wide range of malignant and non-malignant hematologic conditions in pediatric populations and often represents the standard of care in high-risk cases; however, it carries substantial clinical risks, such as graft-versus-host disease and transplant-related mortality, which necessitate careful donor selection [1]. In clinical practice, transplant success is tightly linked to donorrecipient compatibility and to graft characteristics (the material transplanted from the donor to the recipient, such as $CD34^+$ and $CD3^+$ cells), because immunologic disparity can increase complications such as graft-versus-host disease (GVHD), graft failure, and transplant-related mortality [2]. Donor selection is typically guided by high-resolution Human Leukocyte Antigen (HLA) matching and additional donor/recipient factors (e.g., Cytomegalovirus (CMV) status, age, stem cell source - bone marrow or peripheral blood stem cells), while acknowledging that not all patients have access to an ideal matched sibling and that alternative donor options are frequently required [2]. Beyond "match counts," functional and locus-specific approaches, referring to specific genomic positions associated with immunogenetic variability, can refine immunological risk; for example, classification of HLA-DPB1 mismatches into permissive vs non-permissive groups has been associated with clinically relevant differences in outcomes [3]. In parallel, the availability of large transplantation registries and richer clinical data has motivated machine learning (ML) approaches to outcome prediction (e.g., early mortality, GVHD risk). Prior work has shown that ML can produce clinically meaningful risk stratification, but performance and generalizability depend heavily on data preparation choices and robust validation designs [4], [5], [6], [7], [8].

Despite established donor selection principles, there is still a lack of integrated, transparent frameworks that (i) formalize donor-recipient allocation under multiple clinical criteria and (ii) incorporate data-driven outcome predictions using only pre-decision variables. Proposed solution: This project proposes a decision-support system integrated with predictive artificial intelligence models. Donor allocation is formulated as a multi-criteria decision-making problem and addressed using the Analytic Hierarchy Process (AHP) to derive criterion weights and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to rank feasible donor–recipient pairings. Recipient post-transplant survival time is estimated by a machine learning model trained on a dataset [9] derived from a published pediatric unrelated-donor cohort [10]. The predicted survival outcome is incorporated as one of the decision criteria, yielding an ordered list of candidate matches for clinical review.

II. State-of-the-Art

Allogeneic hematopoietic stem cell transplantation (HSCT) stands as the unique curative option for various pediatric patients with malignant and non-malignant hematologic diseases

[1]. The success of the transplant critically depends on Human Leukocyte Antigen (HLA) compatibility between the donor and recipient. Although the gold standard is a 10/10 genotypic match (HLA-A, -B, -C, -DRB1, -DQB1), approximately 70% of patients do not have a compatible family donor, necessitating the use of unrelated or haploidentical donors [6].

The complexity of immunogenetics, with thousands of known HLA alleles, combined with the heterogeneity of clinical data, has rendered traditional statistical approaches (such as Cox regression) insufficient for capturing complex non-linear interactions. Consequently, the application of Machine Learning (ML) algorithms has emerged as a vital tool for predicting outcomes (survival, Graft-versus-Host Disease (GVHD)) and supporting donor allocation [4], [5].

This document reviews the literature, focusing on data preparation and model validation methodologies.

A. Data preparation and feature engineering

The quality of ML models depends intrinsically on data preparation. The literature identifies three critical vectors in this phase:

1) **HLA resolution and complexity:** high-resolution typing (allelic level) is fundamental. Lee et al. [11] demonstrated in a landmark study that high-resolution matching at HLA-A, -B, -C, and -DRB1 is associated with higher survival rates. Simple binary categorization (matched/mismatched) is insufficient; recent models incorporate the distinction between "permissive" and "non-permissive" mismatches, particularly at the HLA-DPB1 locus, based on T-cell epitopes. Fleischhauer et al. [3] validated that non-permissive mismatches significantly increase mortality, making this a crucial feature for predictive algorithms.

2) **Specific clinical and pediatric variables:** beyond HLA, feature engineering must include donor and graft-specific factors. Kałwak et al. [10], in a study focused on pediatrics, highlighted that higher doses of $CD34^+$ and $CD3^+$ cells in the graft promote better long-term survival without increasing the risk of severe GVHD. The inclusion of these quantitative biological variables enriches ML models. Additionally, Tang et al. [8]. innovated by using longitudinal vital sign data (e.g., temperature, blood pressure) extracted from Electronic Health Records (EHR), demonstrating that temporal trends (slopes) are stronger predictors of acute GVHD than static measurements.

3) **Missing data treatment and feature selection:** because real-world databases may contain noise and missing data, Shouval et al. [4] discuss the need for robust preprocessing, including imputation and discretization. Feature selection is critical to avoid hyper-dimensionality. For instance, in a data mining study involving 28,236 patients, the Alternating Decision Tree (ADTree) algorithm automatically selected 10 out of 20 possible variables, eliminating redundancies (e.g., combining donor/recipient Cytomegalovirus (CMV) serostatus into a single interaction variable) [6].

B. ML models and validation strategies

The transition from classical statistical models to ML requires rigorous validation to avoid overfitting, where the model memorizes training data but fails to generalize [4].

1) **Predictive algorithms** - recent literature favors algorithms that balance accuracy with clinical interpretability:
   - **Decision trees and ensemble methods:** Shouval et al. [6] and Arai et al. [7] successfully used the ADTree algorithm to predict mortality and GVHD, respectively. ADTree was preferred over Artificial Neural Networks (ANN) or Random Forests because it allows for the visualization of decision rules and interactions (e.g., the impact of disease stage varies by age), whereas "black box" models hide this logic.
   - **Penalized logistic regression:** Tang et al. [8] used L2 regularization to handle collinearity in longitudinal vital sign data, outperforming baseline models that used only static characteristics.

2) **Validation methodologies** - robust validation is consistent across high-quality studies:
   - **Train/test split:** Arai et al. [7]. randomly divided a cohort of 26,695 patients into training (70%) and validation (30%) sets. The trained model was tested on the validation cohort, demonstrating clear risk stratification (hazard ratio 2.57 for high risk vs. low risk).
   - **Cross-validation:** Both Shouval et al. [4] and Gupta et al. [5] advocate for the use of 10-fold cross-validation on the training set to optimize hyperparameters prior to final testing.
   - **Calibration:** accuracy (AUC) is not the only metric. Shouval et al. [4] emphasize the importance of calibration (agreement between predicted and observed probability), demonstrating excellent consistency in their 100-day mortality model .

C. From prediction to allocation

While ML models provide a risk score (prediction), clinical decision-making requires selecting the best donor among several available options (allocation). The literature suggests a "prediction-to-decision" gap [4]. Multiple Attribute Decision Making (MADM) methodologies, such as TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), described by Tzeng & Huang [12], allow for the integration of ML predictions (as criteria) into an ordered ranking of alternatives.

This hybrid approach (ML to predict outputs, TOPSIS to rank candidates) represents an improvement for clinical decision support systems, transforming raw probabilities into actionable recommendations.

III. Methodology

A. Data Cleaning

The used dataset called "Bone Marrow Transplant: children" contained 187 children and adolescents between the 2000 and 2008 [9], derived from a published pediatric unrelated-donor cohort [10]. Besides the fact that there are few rows for generalization, the dataset contains a lot of attributes that are very helpful in understanding the factors that influence matching and survivability of the recipient.

There are some problems with this dataset that will need to be addressed. First handle "?", missing and other strange values in donor_CMV, CD3_x1e8_per_kg, CD3_to_CD34_ratio, ANC_recovery, PLT_recovery, time_to_acute_GvHD_III_IV, extensive_chronic_GvHD, HLA_match with caution. Analize the usefulness in outcome of some abstracted attributes like donor_age_belo_35, recipient_age_below_10, HLA_mismatch. Encode various categorical values with onehot-encoding since there is a small range of categorical values. Creation of synthetic data attributes for tissue type and donor gender to make possible HLA and gender match calculation during multi-criteria decision making. Check and handle bias and minority classes in outcome features using stratification and oversampling. Handle highly correlated features by removing the less performing ones, simplifying the generalization with few data rows.

The dataset provides a lot of data to better understand the match and transplant variables influence in survivability. It will be useful for finding correlations and importance of each attribute during the allocation fase, but will need to be simplified in order to help the model predict and generalize better with the small sample size.

B. Data Exploration

One of the first steps in data exploration is to analyze if the target data is biased in any way. This can be done by checking the proportions of the `survival_status` attribute. In this dataset, 54% of people are alive, while 45% are deceased, which means that the target is well balanced. Looking at the data, a possible important factor is the correlation between the ages of the recipient and the donor. Fig.1 shows the correlation between the pairs:

This graph shows that a lot of transplants are made with donor-recipient pairs where the donor is between the ages of 25–45 years old and the recipient between 5–15 years old. Another thing to explore could be the correlation between the age gap of the pair. However, analysis suggests that the donor-recipient age gap doesn't have a strong enough correlation with either survival time or survival status of the recipient. Another attribute worth exploring is how blood type compatibility affects survival. Exploratory analysis showed that, while in

TABLE I
Clinical, Matching, Transplantation, and Outcome Attributes

Attribute	Description
Donor-specific attributes	
donor_age	Refers to the donor age at donation
donor_age_below_35	35 years cutoff age that has significantly lower risk of grade II to IV acute GVHD and lower likelihood of non-relapse mortality with miss-matched recipients [13]
donor_ABO	The blood type of the donor
donor_CMV	Presence of cytomegalovirus infection. A virus that is harmless and asymptomatic to most people but can be life-threatening for people with compromised immune systems
Recipient-specific attributes	
recipient_age	The donor age at transplant
recipient_age_below_10	10 years cutoff
recipient_age_int	Stores an age bin text
recipient_gender	The gender of the recipient
recipient_body_mass	The body mass of the recipient
recipient_ABO	The blood type of the recipient
recipient_rh	The rh of the recipient's blood
recipient_CMV	Presence of cytomegalovirus infection
disease	Type of disease
disease_group	Malignant disease or not
risk_group	The explicit meaning is still to be discovered, but it's assumed to be a value based in disease and disease status to categorize patients into 2 risk groups with significantly different overall survival and progression-free survival on the basis of primarily differences in the relapse risk [14]
Matching-related attributes	
gender_match	Checks if female donor to male recipient or any other case
ABO_match	If blood types are compatible
CMV_status	Level of serological compatibility
HLA_match	Allele level donor-recipient matching
HLA_mismatch	If HLA match if superior 8 alleles
antigen	Difference of antigens between donor and recipient
allel	Difference of alleles between donor and recipient
HLA_group_1	Description of donor-recipient matching/mismatching
Transplantation-related attributes	
stem_cell_source	Where the stem cells were obtained
tx_post_relapse	If it is the second transplant done (after relapse)
CD34_x1e6_per_kg	The $CD34^+$ cell dose (10^6) per kg of recipient body weight
CD3_x1e8_per_kg	The $CD3^+$ cell dose (10^8) per kg of recipient body weight
CD3_to_CD34_ratio	The $CD3^+$ to $CD34^+$ ratio
ANC_recovery	Time in days to achieve an absolute neutrophil count $> 0.5 \times 10^9$/L for 3 consecutive days
PLT_recovery	Time in days to achieve an absolute platelet count $> 0.5 \times 10^9$/L for 3 consecutive days
acute_GvHD_II_III_IV	If the recipient developed acute GVHD stage II, III or IV
acute_GvHD_III_IV	If the recipient developed acute GVHD stage III or IV
time_to_acute_GvHD_III_IV	Time in days that took the recipient to develop acute GVHD stage III or IV
extensive_chronic_GvHD	Time in days that took the recipient to develop extensive chronic GVHD
Survivability attributes	
relapse	If the disease has recurred
survival_time	Time in days the recipient survived from transplant to death (if dead); time in days the recipient is alive from transplantation to time of data collection (if alive)
survival_status	If the recipient is dead or alive

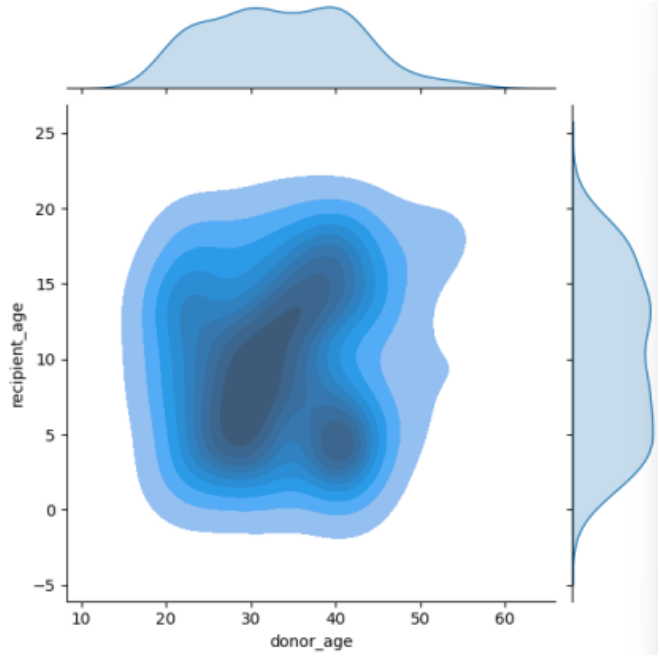

Fig. 1. Age correlation of donor-recipient pairs

case there is a blood type match the rate of survival is very close to 50%, in case of a blood type mismatch the rate of survival drops to close to 40%, showing a possible small correlation between these attributes. Gender matching can also be an important factor to consider in compatibility; however, analysis shows that gender matching alone isn't sufficient to form a correlation.

C. Feature Engineering

Feature Engineering is being performed to better encode significant relationships and reduce ambiguity in the raw data. Testing will be performed with various models in order to understand which changes help the model create better correlations and thus achieve better results. For example, certain features can be derived from existing ones, like the `age_gap`, calculated using the absolute difference between the recipient-donor pair. Missing values were handled using imputed data relevant to each feature, and an explicit missingness indicator. Categorical data was encoded using one hot encoding or ordinal encoding when the data was binary.

All feature engineering steps were integrated into a preprocessing pipeline

D. Oversampling

IV. Experiments

The section evaluates the performance of different ML models for the prediction of both a classification target and a regression target and analyzes the impact of feature engineering, hyperparameter tuning and data augmentation techniques.

Approximately 100 experiments with different models and model configurations were made. To facilitate this, MLflow and Hydra were used for tracking and result reproducibility across model configurations. [15] [16]

A. Survival Classification

For survival classification, several model families were tested. All models were tested in three different configurations: default parameters, tuned and tuned with SMOTE. Additionally, the best models were also tested using GANs with different combinations of epochs and synthetic sample sizes. Different number of training cross validation folds and train-test splits were also considered. Finally, different methods of hyperparameter optimization (Grid Search, Random Search and Optuna) were explored and compared. [17]

As metrics, the F1-score on the test set was the main metric used for comparison but several other metrics were also used such as the accuracy, precision, recall and cross-validation F1-score of the train set.

Linear classification models, including the Logistic Regression were evaluated as a baseline. Tree ensemble models like the Random Forest (RF) and the Extra Trees (ET), and gradient boosting models like the XGBoost (XGB) and the LightGBM (LGBM) were also tested. Alternating Decision Trees (ADTrees) and Support Vector Machines (SVMs) were also considered.

Out of all the experiments, the tuned Random Forest Classifier with GAN-augmented data obtained the best result, with an F1-score of 0.62 on the test set, a precision of 0.57, a recall of 0.61 and an accuracy of 0.62. However, the model that, on average, achieved the best results was the Logistic Regression, having the base untuned model ranked as the best model not utilizing GANs.

Bayesian hyperparameter tuning using Optuna did not consistently outperform grid search or random search in test performance. However, it reached very similar results significantly faster and sometimes even outperforming other methods.

Further, it was observed that SMOTE oversampling for the minority class didn't significantly help in predicting the survival of the patient, having the models with SMOTE achieved an F1-score on the test set of 0.45, and 0.44 without, with precision and recall being very similar. On the other hand, GAN oversampling improved significantly the results, reaching an average F1-score on the test set of 0.50, against 0.44 without.

On average, the tuned models without oversampling had better overall performance than the untuned models, however, it is important to note that some untuned models performed significantly better than their tuned conter-parts, even when confining their grid search, as can be seen in Fig.2. It was observed that simpler models like Linear Regression and SVM often had better results than more complex models.

GAN-augmentation had a significant impact in model performance for tree models, while negatively affecting linear models. Gradient boosting models also were positively affected. Several GAN configuration were attempted and it was observed that the one that had best performance overall was

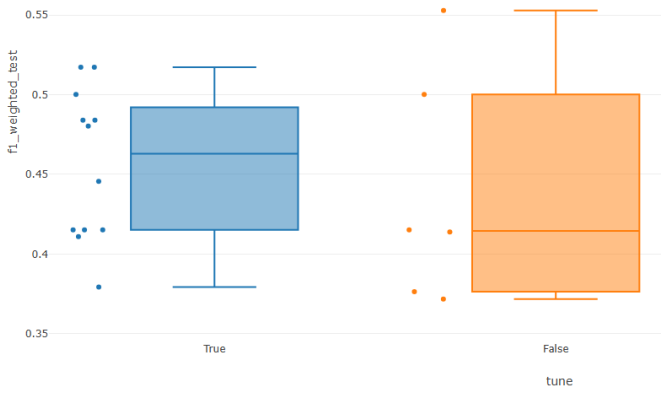

Fig. 2. Performance of tuned (no oversampling) vs. untuned models

a sample size of 10 synthetic entries for class 0, 20 synthetic entries for class 1 and the the number of epochs set to 300.

Another observation made was comparing the cross-validation F1-Score on the train set with the test set. As can be seen in the sample of parallel coordinates plot in Fig.3, 12 random experiments were chosen, and compared. It was observed that experiments that scored a high F1-score on cross-validation with the train set performed consistently worse on the test set than those that scored a lower value.

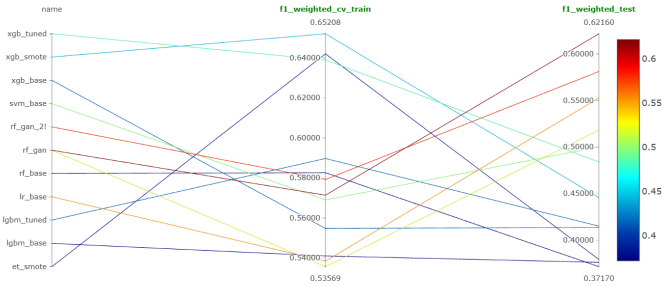

Fig. 3. Comparison between train and test set performance across experiments

Some additional experiments were made on the best-performing models comparing different values for number of cross validation folds and train/test split ratios. It was observed that training the model with 5 folds and a test set size of 15% of the total dataset yielded better results.

B. Survival Time Regression

To predict survival time tests were performed on largely the same model families as the step above. For the linear models family, ridge regression and elastic-net were tested. All models were tested with and without the addition of the classification variable predicted by the best classification model, tuned and untuned. Additionally, the classification variable was tested being predicted as a binary variable or as a probability. Metrics such as the Root Mean Squared Error (RMSE), Mean Average Error (MAE) and the coefficient of determination (R^2), were saved for both train and test sets and used to compare the regression models. A baseline RMSE value was also set (RMSE = 913), which would represent a simple model in which every patient is given the average survival time.

As can be seen in Fig. 4, using classification as a feature to predict regression considerably improved model performance.

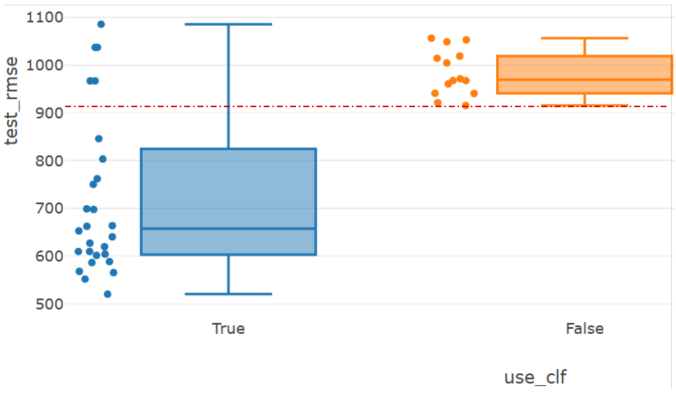

Fig. 4. Use of classification as feature for regression

Using the red line to indicate the baseline level, the RMSE of the models not using classification were all observed to be above that level. However, by adding the survival status as a feature the RMSE was cut almost in half. Moreover, predicting the survivability as a probability instead of a binary feature produced mixed results across models and configurations, with tree and linear models performing particularly bad, however, gradient boosting models like LightGBM and XGBoost had better results than the same model with a binary classification feature.

As with classification, different numbers of cross-validation folds were tested and 5 folds was observed to produce the best results.

In the end, the regression model that preformed best was a tuned LightGBM model using survival status as a probability.

V. Discussion

VI. Conclusion

Acknowledgment

The authors thank the lecturers of the Master in Artificial Intelligence Engineering (MEIA) at the Porto School of Engineering (ISEP) for the guidance and support provided throughout the development of this project.

References

[1] M. A. Diaz, "Editorial: Allogeneic transplantation in pediatric patients with hematologic malignancies," *Frontiers in Pediatrics*, vol. 12, p. 1411922, 2024, doi: 10.3389/fped.2024.1411922.

[2] J.-M. Tiercy, "How to select the best available related or unrelated donor of hematopoietic stem cells?" *Haematologica*, vol. 101, no. 6, pp. 680–687, 2016, doi: 10.3324/haematol.2015.141119.

[3] K. Fleischhauer *et al.*, "Effect of t-cell-epitope matching at hla-dpb1 in recipients of unrelated-donor haemopoietic-cell transplantation: a retrospective study," *The Lancet Oncology*, vol. 13, no. 4, pp. 366–374, 2012, doi: 10.1016/S1470-2045(12)70004-9.

[4] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct," *Bone Marrow Transplantation*, vol. 49, no. 3, pp. 332–337, 2014, doi: 10.1038/bmt.2013.146.

[5] V. Gupta, T. M. Braun, M. Chowdhury, M. Tewari, and S. W. Choi, "A systematic review of machine learning techniques in hematopoietic stem cell transplantation (hsct)," *Sensors*, vol. 20, no. 21, p. 6100, 2020, doi: 10.3390/s20216100.

[6] R. Shouval *et al.*, "Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: A european group for blood and marrow transplantation acute leukemia working party retrospective data mining study," *Journal of Clinical Oncology*, vol. 33, no. 28, pp. 3144–3151, 2015, doi: 10.1200/JCO.2014.59.1339.

[7] Y. Arai *et al.*, "Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation," *Blood Advances*, vol. 3, no. 22, pp. 3626–3634, 2019, doi: 10.1182/bloodadvances.2019000934.

[8] S. Tang, G. T. Chappell, A. Mazzoli, M. Tewari, S. W. Choi, and J. Wiens, "Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records," *JCO Clinical Cancer Informatics*, no. 4, pp. 128–135, 2020, doi: 10.1200/CCI.19.00105.

[9] M. Sikora, L. Wróbel, and A. Gudyś, "Bone marrow transplant: Children," https://www.kaggle.com/datasets/adamgudys/bone-marrow-transplant-children, 2010, kaggle dataset.

[10] K. Kałwak *et al.*, "Higher cd34+ and cd3+ cell doses in the graft promote long-term survival, and have no impact on the incidence of severe acute or chronic graft-versus-host disease after in vivo t cell-depleted unrelated donor hematopoietic stem cell transplantation in children," *Biology of Blood and Marrow Transplantation*, vol. 16, no. 10, pp. 1388–1401, 2010, doi: 10.1016/j.bbmt.2010.04.001.

[11] S. J. Lee *et al.*, "High-resolution donor-recipient hla matching contributes to the success of unrelated donor marrow transplantation," *Blood*, vol. 110, no. 13, pp. 4576–4583, 2007, doi: 10.1182/blood-2007-06-097386.

[12] G.-H. Tzeng and J.-J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. CRC Press, 2011.

[13] A. Nagler *et al.*, "Young (< 35 years) haploidentical versus old ($>= 35$ years) mismatched unrelated donors and vice versa for allogeneic stem cell transplantation with post-transplant cyclophosphamide in patients with acute myeloid leukemia in first remission," *Bone Marrow Transplantation*, vol. 59, no. 11, pp. 1552–1562, 2024, doi: 10.1038/s41409-024-02400-5.

[14] P. Armand *et al.*, "A disease risk index for patients undergoing allogeneic stem cell transplantation," *Blood*, vol. 120, no. 4, pp. 905–913, 2012, doi: 10.1182/blood-2012-03-418202.

[15] M. Zaharia *et al.*, "Accelerating the machine learning lifecycle with mlflow," in *Proceedings of the IEEE International Conference on Big Data*, 2018.

[16] O. Yadan, "Hydra - a framework for elegantly configuring complex applications," Github, 2019. [Online]. Available: https://github.com/facebookresearch/hydra

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.