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. The multi-criteria decision-making method TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) is 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 (benefit criterion), 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. Experimental results demonstrate that an XGBoost (eXtreme Gradient Boosting) model with the help of Generative Adversarial Networks (GANs) achieved the best predictive performance for post-transplant probability of survival, reaching an F1-Score of 0.65. This probability was then used to predict the expected post-transplant survival time in case of death. For this regression task, the best performing model was an Extra Trees Regressor model, which achieved a Root Mean Square Error (RMSE) of 560, significantly outperforming baseline level (913 RMSE).**

***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 donor-recipient 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].

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. This work proposes a decision-support system integrated with predictive artificial intelligence models. Donor allocation is formulated as a multi-criteria decision-making problem, in which alternative ranking is performed using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), with criterion weights calculated by the Analytic Hierarchy Process (AHP). Recipient post-transplant survival time is estimated by a machine learning model trained on a dataset [7] derived from a published pediatric unrelated-donor cohort [8]. The expected survival time (serves as post-graft prognosis) is incorporated as one of the decision criteria, yielding an ordered list of candidate matches for clinical review.

When a donor is chosen, another model tries to predict the patient's relapse based on previous features and new transplant related ones, providing additional support to the clinician.

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 [4].

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 [9], [10].

This section 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. [8], 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. [6] 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. [9] 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) [4].

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 [9].

1) **Predictive algorithms:** recent literature favors algorithms that balance accuracy with clinical interpretability:
   - **Decision trees and ensemble methods:** Shouval et al. [4] and Arai et al. [5] 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. [6] 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. [5]. 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. [9] and Gupta et al. [10] 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. [9] 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 [9]. 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

The main system aims to use the transformations necessary to join, process and retain the data necessary for each stage of the alocation and estimation process. While the original dataset was used to train the models, exploratory analysis and context of the specialization, synthetic datasets were need to be created in order to be able to test and use the system in the prespective of a professional.

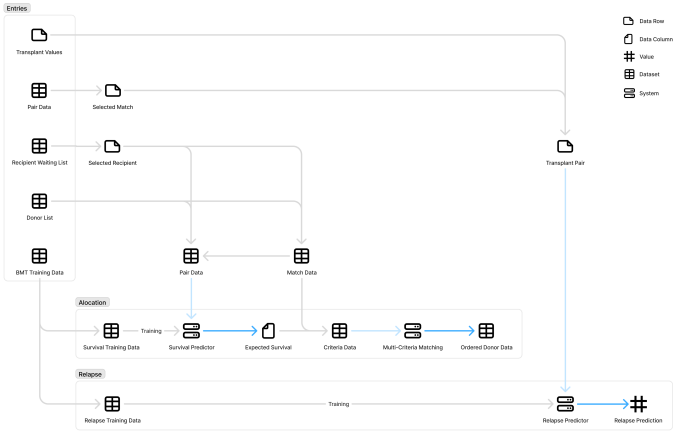

Fig. 1. Data Flow Diagram

The proposed methodology formulates pediatric bone marrow donor–recipient allocation as a multi-criteria decision-making problem. The objective is to generate a ranked list of feasible donor–recipient pairs and provide transparent decision support to the responsible clinician, who retains full responsibility for the final donor selection. Alternative ranking is performed using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Criterion weights are obtained using the Analytic Hierarchy Process (AHP), which is applied exclusively to compute the preference vector (weights) of the decision criteria and validated through consistency analysis (consistency ratio). AHP is not applied directly to rank alternatives due to scalability limitations when the number of potential donors is large, which would require a quadratic number of pairwise comparisons. The proposed framework explicitly separates criteria weighting (AHP), predictive modeling (machine learning), and final ranking (TOPSIS), thereby supporting transparency, clinical interpretability, and auditability. The decision criteria considered in the ranking process are: HLA Match, CMV Serostatus, Donor Age Group, Gender Match, ABO Match, and Expected Survival Time. The relative importance for these criteria was defined through literature review, as shown in Table I. The Predicted Survival Time has the lowest value of importance, as it is a supplementary criterion to support decision-making rather than the primary driver of donor selection. Because hematopoietic stem cell can have two sources (bone marrow or peripheral blood), if the source is bone marrow, the ABO compatibility criterion will have an higher importance than donor-recipient sex match, and if the source is peripheral blood, the ABO compatibility criterion will have a lower importance than donor-recipient sex match, because in this process the blood is not manipulated as in bone marrow donation, where red blood cells are removed to avoid hemolysis in case of ABO incompatibility [13].

Expected Survival Time (in days) is estimated using a machine learning regression model trained and validated on a publicly available dataset of pediatric unrelated-donor bone marrow transplantations, containing 187 children and adolescents between the 2000 and 2008 [7], derived from a published pediatric unrelated-donor cohort [8]. The variables of this dataset are described in table II. One of the predictive features used in this regression model is the Expected Probability of Survival, which represents a prediction of post-transplant survival within a follow-up window of three to eight years after transplantation. This survival probability is predicted using a separate regression machine learning model. At the end of the pipeline, the system outputs an ordered ranking of donor–recipient pairs together with the estimated expected survival time. These results are presented to the clinician as decision support, enabling informed, transparent, and reproducible donor selection.

A. Data Cleaning

Several data quality issues were identified and addressed prior to model development. Missing `recipient_body_mass` values were imputed using recipient age bins, separated by sex, since they are strongly correlated (aprox. 0.89) and can provide more accurate values than simple imputing methods. A simple linear regression model was used to impute missing `CD3_x1e8_per_kg` values using `CD34_x1e6_per_kg` and `recipient_body_mass`, also due to their strong correlation and linear nature. Analize the usefulness in outcome of some abstracted attributes like `donor_age_below_35`, `recipient_age_below_10`, `HLA_mismatch`. Encode various categorical values with one-hot 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.

A new `missing_(feature)` attribute was created for each value equal to "?", particularly

TABLE I
DECISION CRITERIA FOR DONOR-RECIPIENT MATCHING

Rank	Criterion	Priority	Justification	Sources
1	HLA Compatibility	Critical	Remains the primary determinant of transplant outcome. High-resolution matching at HLA-A, -B, -C and -DRB1 is the gold standard. Mismatches are consistently associated with reduced survival and increased complications, making HLA assessment the starting point of donor selection, even among alternative donors.	[13]–[15]
2	Donor Age	Very High	The most consistently validated non-HLA factor associated with overall survival. Younger donors (18–32 years) are associated with superior outcomes. Each 10-year increase in donor age increases mortality risk by approximately 5.5%, and donor age should therefore be prioritized over other secondary factors.	[13], [15]
3	CMV Serostatus	High	Particularly relevant for CMV-negative recipients, who benefit from CMV-negative donors to reduce non-relapse mortality. However, the overall impact on survival has decreased due to improved antiviral prophylaxis, making this criterion secondary to HLA and donor age.	[13], [15], [16]
4	ABO Compatibility	Moderate	The impact on overall survival is modest or not statistically significant in recent cohorts. Its relevance is mainly operational, reducing hemolysis risk and transfusion requirements. It becomes more important when bone marrow is the graft source, where major ABO mismatches may reduce usable cell yield during processing.	[13]
5	Donor–Recipient Sex Match	Low	Female-to-male transplants are associated with increased GVHD risk due to anti-HY alloreactivity and delayed neutrophil recovery. Current guidelines consider this a minor criterion, mainly useful for tie-breaking between otherwise equivalent donors.	[13], [17]
6	Probability of success (Predicted survival time)	Lowest	Sources acknowledge that mortality prediction and performance prognosis are tools used and advocated for the optimal selection of donors.	[18], [19]

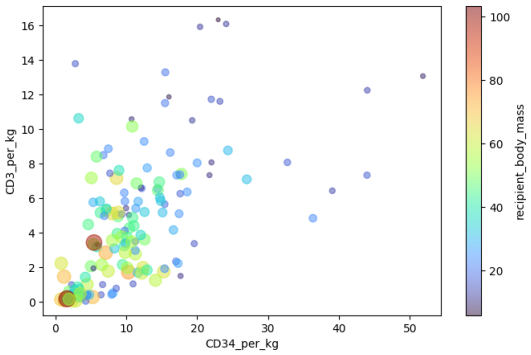

Fig. 2. Correlation between $CD3^+$, $CD34^+$ and recipient body mass

in `(donor-recipient)_CMV`, `ABO_match`, `(donor-recipient)_ABO` and `CMV_status`.

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 slightly unbalanced. Looking at the data, a possible important factor is the correlation between the ages of the recipient and the donor. Fig.3 shows the correlation between the pairs:

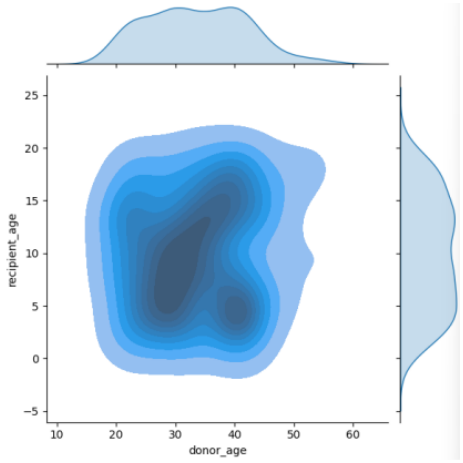

Fig. 3. Age correlation of donor-recipient pairs

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

TABLE II
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 [20]
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 [21]
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
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

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 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

By default, the dataset already comes with various abstracted features. A small number were added or altered. Some of the new engineered features include the addition of the `age_gap` feature, which is calculated using the absolute difference between the recipient-donor pair; the binning of the `donor_age` and `recipient_age` features; and the `total_mismatches` feature, which is calculated using the total number of mismatched allels and antigens between donor and recipient.

Careful selection was important in order to reduce the high dimensionality of the data and help the models generalize better as well as reduce overfitting.

D. Oversampling

1) SMOTE: Regarding oversampling, two types were tested in this project. One of them is Synthetic Minority Over-sampling Technique (SMOTE) [22]. SMOTE is used in unbalanced datasets to create more samples of the minority class using a k-nearest neighbours algorithm. This way, new samples are not equal to existing ones, but slightly different variations of existing minority class samples. In this case, slight variations of SMOTE were used such as SMOTE-NC (for nominal and countinous data) and Adaptive Synthetic Sampling (ADASYN).

2) GANs: Generative Adversarial Networks (GANs) utilize neural networks to learn and generate synthetic data, which makes them a more robust and powerful alternative to other techniques [23].

IV. EXPERIMENTS

Approximately 150 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. [24] [25]

V. SYSTEM ARCHITECTURE

For this experiment, the system architecture is composed woth two modules that work together to achieve the desired functionality. These two modules are:

- A prediction sytem with a api that exposes endpoints for predicting survival classification, survival time regression and relapse with BentoML. The prediction models with the best scores were trainned and imported from MLflow.

- A simple flask api that serves a web application for user interaction and calls the BentoML prediction system api for predictions.

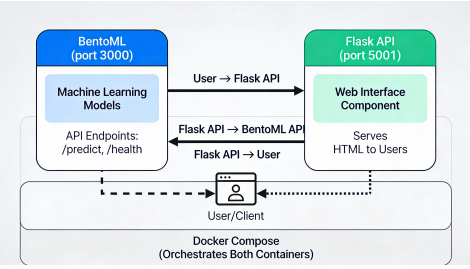

Fig. 4. Architecture Diagram

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. Every tuned model was also tested with different feature engineering steps. 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. [26]

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.

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.

C. Relapse Classification

Various models were tested: Logistic Regression, k-Nearest Neighbors, Stochastic Gradient Descent, Random Forests and more. Model performance was primarily assessed through confusion matrices, ROC curves and precision–recall curves. Resampling techniques were applied to address class imbalance. In addition, the class weight parameter was set to balanced (when resampling was not used), and recall was selected as the main scoring metric during model optimization. These strategies improved significantly positive classification performance across all models. Feature selection was preferred over dimensionality reduction, because in the medical field, understanding the contribution of each feature to the predicted outcome is critical.

Despite the best efforts, achieving favorable results with a such limited dataset proved to be quite challenging. It was required to implement multiple iterative adjustments while carefully avoiding extensive hyperparameter tuning or frequent evaluation on the test data to prevent overfitting.

VI. Results and Discussion

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.

A. Survival Classification

Out of all the experiments, the tuned XGBoost with GAN-augmented data obtained the best result, with an F1-score of 0.65 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 did not significantly help in predicting the patient survival, 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.5. It was observed that simpler models like Logistic Regression and SVM often had better results than more complex models.

This is possibly due to the small dataset size which causes many of the more complex models, even on default parameters, to overfit the training data, causing poor performance on the

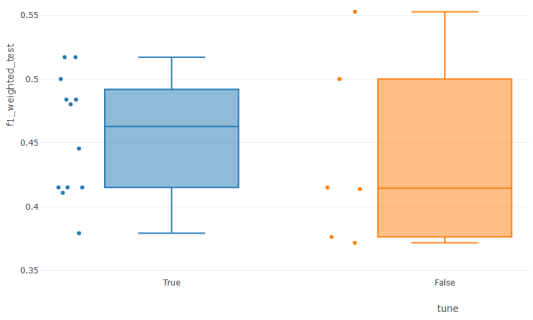

Fig. 5. Performance of tuned (no oversampling) vs. untuned models

test set. Due to this, simpler models are often able to create a better generalization.

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 produced the best results overall was 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.

This observation is consistent with the previous hypothesis. With more data, tree models don't overfit the training set as much, and get better results. However, it is important to note that data quality also plays a big role, because simply generating more data using GANs doesn't always yield better results. With little data, GANs can also overfit the train set and generate synthetic data that is too close to the real set which only adds noise. On the other hand, reducing the number of epochs too much can also cause them to not learn enough about the data.

Another observation made was when 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.6, 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.

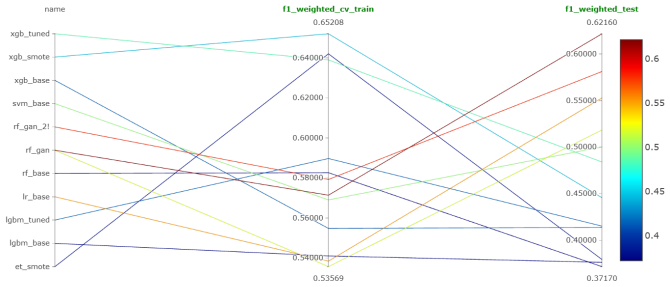

Fig. 6. Comparison between train and test set performance across experiments

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

As can be seen in Fig. 7, using classification as a feature to predict regression considerably improved model performance.

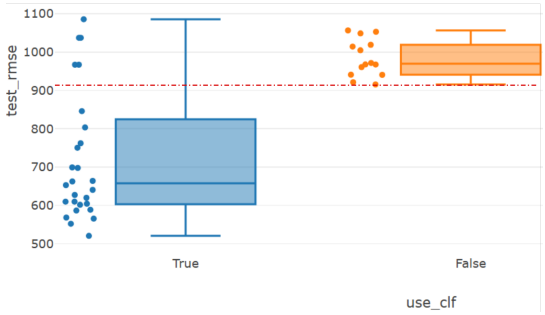

Fig. 7. Use of classification as feature for regression

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 Extra Trees Regressor model using survival status as a probability.

C. Relapse Classification

Finally, for relapse prediction, due to the big similarity of the dataset to the previous classification model, only the models that performed best were tested. In this case, the XGBoost with GAN augmentation was also the best model. However, due to the limited dataset size, and significant class inbalance, prediction results were not satisfactory. Even utilizing GAN augmentation, the best model achieved a macro F1-Score of 0.47.

D. SHAP Analysis

To improve model explainability, a general SHapley Additive exPlanation (SHAP) analysis was performed and plotted for the best-performing classification and regression model [27]. Each dot represents a single data point of the respective feature. The x-axis represents the SHAP value, which is a measure of the impact of each feature on the model's prediction (a negative shap means the model is more likely to predict the patient's survival, while a positive value means the opposite). The color of each dot represents the value of the feature for that data

point. A red feature represents a high value and a blue feature represents a low value. For example, in the feature `age_gap`, a red dot represents a high age gap and a blue dot represents a low age gap. The ordering along the y-axis does not affect interpretability and is used only to improve visualization. It should be noted that SHAP analysis was performed strictly post hoc and was not used to guide feature engineering decisions. Fig.8 shows a beeswarm plot of the most important features for the classification model.

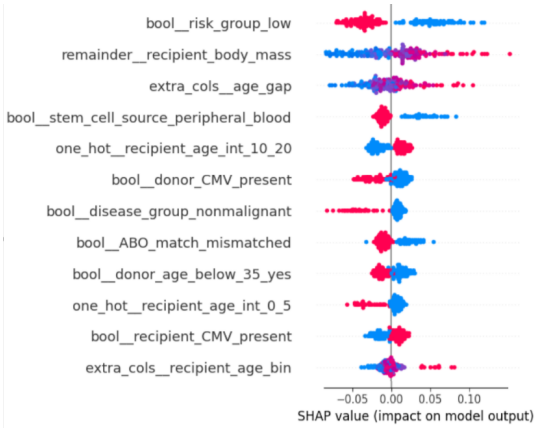

Fig. 8. SHAP Analysis for classification model

Fig.9 shows a beeswarm plot of the top 3 features for the regression model.

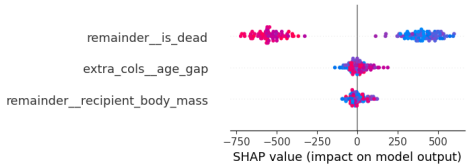

Fig. 9. SHAP Analysis for regression model

Some features are very strong and clean indicators of the model's prediction. For example, the feature `risk_group_low` being positive is a big indicator that the model will predict the patients survival. To add to this, we can observe that most binary features form separated clusters, indicating a consistent directional impact on the model's prediction depending on feature value. However, that's not always the case, as we can see in the feature `ABO_match_mismatched` or `donor_CMV_present`. We can also observe that the engineered features `age_gap` and `recipient_age_bin` had a strong impact on the model's prediction.

Furthermore, we can clearly see the importance of predicting the survival status of the patient as a feature for the regression model.

VII. Conclusion

A. Main findings

The results indicate that simpler models, particularly logistic regression, exhibit strong generalization performance in small clinical datasets, while more complex models are prone to overfitting. Although GAN-based data augmentation improved the performance of tree-based and gradient boosting models, its performance was highly dependent on data quality. Furthermore, incorporating survival classification outputs as features for survival time regression significantly reduced prediction error, demonstrating the benefit of multi-stage modeling strategies.

B. Limitations

1) Small Dataset: The biggest limitation of this project was the limited dataset size. A size of around 180 total samples, means there are not enough samples for the model to create a good generalization and the test is too small to have a clear idea of how the model would preform with new data.

2) Class imbalance: Additionally the dataset shows a small class imbalance for the `survival_status` feature which was able to be remedied utilizing oversampling techniques. However, there is a much more significant unbalance for the feature `relapse`.

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 *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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[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. M. Jimenez Jimenez, S. R. Spellman, I. Politikos, S. R. McCurdy, S. M. Devine, M. M. A. Malki, Y.-T. Bolon, S. J. Lee, J. Dehn, J. Pidala, M. Maiers, M. Askar, C. Malmberg, J. J. Auletta, H. Stefanski, L. Broglie, M. Qayed, M. Horwitz, J. S. Wilder, M. Gooptu, R. S. Mehta, M. Fernandez-Viña, B. E. Shaw, and B. C. Shaffer, "Allogeneic Hematopoietic Cell Donor Selection: Contemporary Guidelines from the NMDP/CIBMTR," *Transplantation and Cellular Therapy*, vol. 31, no. 12, pp. 973–988, Dec. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666636725012904

[14] J. Dehn, S. Spellman, C. K. Hurley, B. E. Shaw, J. N. Barker, L. J. Burns, D. L. Confer, M. Eapen, M. Fernandez-Vina, R. Hartzman, M. Maiers, S. R. Marino, C. Mueller, M.-A. Perales, R. Rajalingam, and J. Pidala, "Selection of unrelated donors and cord blood units for hematopoietic cell transplantation: guidelines from the NMDP/CIBMTR," *Blood*, vol. 134, no. 12, pp. 924–934, Sep. 2019. [Online]. Available: https://ashpublications.org/blood/article/134/12/924/374909/Selection-of-unrelated-donors-and-cord-blood-units

[15] C. Kollman, S. R. Spellman, M.-J. Zhang, A. Hassebroek, C. Anasetti, J. H. Antin, R. E. Champlin, D. L. Confer, J. F. DiPersio, M. Fernandez-Viña, R. J. Hartzman, M. M. Horowitz, C. K. Hurley, C. Karanes, M. Maiers, C. R. Mueller, M.-A. Perales, M. Setterholm, A. E. Woolfrey, N. Yu, and M. Eapen, "The effect of donor characteristics on survival after unrelated donor transplantation for hematologic malignancy," *Blood*, vol. 127, no. 2, pp. 260–267, Jan. 2016. [Online]. Available: https://ashpublications.org/blood/article/127/2/260/34808/The-effect-of-donor-characteristics-on-survival

[16] P. Ljungman, R. De La Camara, C. Robin, R. Crocchiolo, H. Einsele, J. A. Hill, P. Hubacek, D. Navarro, C. Cordonnier, and K. N. Ward, "Guidelines for the management of cytomegalovirus infection in patients with haematological malignancies and after stem cell transplantation from the 2017 European Conference on Infections in Leukaemia (ECIL 7)," *The Lancet Infectious Diseases*, vol. 19, no. 8, pp. e260–e272, Aug. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1473309919301070

[17] A. Gratwohl, C. Ruiz De Elvira, M. Gratwohl, H. T. Greinix, and R. Duarte, "Gender and Graft-versus-Host Disease after Hematopoietic Stem Cell Transplantation," *Biology of Blood and Marrow Transplantation*, vol. 22, no. 6, pp. 1145–1146, Jun. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1083879116001695

[18] I. J. Ratul, U. H. Wani, M. M. Nishat, A. Al-Monsur, A. M. Ar-Rafi, F. Faisal, and M. R. Kabir, "Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-Square Test and Hyperparameter Optimization: A Retrospective Analysis," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–14, Sep. 2022. [Online]. Available: https://www.hindawi.com/journals/cmmm/2022/9391136/

[19] A. Islam Rifat, M. Hossain, N. Nahid, S. Akter, and A. Islam, "Children Hematopoietic Stem Cell Transplant Survival Status Prediction using Machine Learning," in *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*. Kanjirapally, India: IEEE, Nov. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10420035/

[20] 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.

[21] 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.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002. [Online]. Available: http://dx.doi.org/10.1613/jair.953

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

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

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

[26] 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.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.