Toward Precision Donor Selection in Acute Leukemia: Machine Learning Models for Predicting Allogeneic Hematopoietic Cell Transplant Outcomes

Dorian Benhamou Goldfajn
dbenhamougol@umass.edu

August 27, 2025

Abstract

Acute leukemia represents an aggressive hematologic malignancy where allogeneic hematopoietic cell transplantation (allo-HCT) provides curative potential. Building on prior work investigating ABO mismatch in transplant outcomes, this project shifts focus to predictive modeling with the ultimate goal of identifying the best donor for a given recipient as part of patient-specific precision medicine. Using the publicly available ds1302 dataset (4946 patients, reduced to 4653 analyzable cases and 20 key variables), we developed classification-based machine learning models to predict overall survival (OS) as a first step. Eight models were trained and evaluated; Support Vector Machine (SVM) achieved the strongest performance (accuracy 0.6015, precision 0.5945, recall 0.6015, F1-score 0.5732). Feature importance analyses consistently identified age, disease status at HCT, GVHD prophylaxis regimens, and donor-recipient ABO matching as influential predictors. These findings suggest that achieving remission prior to allo-HCT and refining prophylaxis regimens remain critical to outcomes, while ABO incompatibility exerts a recurrent, though more modest, effect. This project is currently in Step 1: modeling OS as proof-of-concept for an ML approach. The main next step is to expand prediction beyond OS to additional endpoints (e.g., relapse, GVHD, graft failure). The final step is to flip the prediction paradigm and use outcome-informed models to directly predict the optimal donor for each patient—a substantial stride toward precision donor selection.

Keywords: Acute Leukemia, Allogeneic Hematopoietic Cell Transplantation, Machine Learning, Donor Selection, Predictive Modeling, Overall Survival.

1 Introduction

Acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) are life-threatening hematologic malignancies often requiring allogeneic hematopoietic cell transplantation (allo-HCT) in high-risk or relapsed settings. Although curative, allo-HCT outcomes vary widely depending on patient fitness, disease status, donor compatibility, and post-transplant immune modulation. Expanding evidence suggests that ABO mismatch may modestly influence post-transplant outcomes, yet a broad, integrated model for donor optimization does not yet exist.

This project builds directly on prior work, such as the study by Guru Murthy et al. (2023) , which investigated the association of ABO mismatch with allo-HCT outcomes using traditional survival analyses. Unlike previous work, we specifically pursue a machine learning strategy with the long-term goal of predicting who represents the best donor for a given recipient.

The current step focuses on overall survival (OS) prediction to benchmark performance and establish foundational insights. Future expansion involves modeling other transplant outcomes beyond death, and the final step will reverse the modeling task to explicitly identify the best donor using outcome data as features.

2 Methods

2.1 Data Source and Cohort

We utilized the publicly available `ds1302` dataset, which includes detailed clinical, donor, and transplant outcome information for patients undergoing allo-HCT for acute leukemia. The initial dataset contained 4946 patients and 49 variables. After data cleaning, including the removal of patients with significant missingness in key fields, the final analyzable cohort consisted of 4653 patients and 20 predictor variables. The primary endpoint for this classification task was overall survival, defined as a binary target variable, `dead` (1 = death event, 0 = censored/alive).

Features were removed due to redundancy, high rates of missing data, or potential for outcome contamination. Patient identifiers (e.g., `pseudoid`, `pseudoccn`) and time-to-event variables were also excluded from this initial classification analysis.

2.2 Predictor Variables

The 20 retained predictor variables captured a range of patient, disease, and treatment characteristics:

- **Recipient Demographics & Comorbidities:** `age`, `sex`, `agegp` (age group), `hctcigp` (HCT Comorbidity Index group), `karnofcat` (Karnofsky Performance Status category).
- **Disease-Related Variables:** `dis` (disease type), `alstatprgp` (disease status pre-transplant).
- **Donor/Recipient Compatibility:** `drabomatch` (ABO match status), `drrh` (Rh factor match), `drcmvpr` (CMV serostatus), `drsex` (sex match).
- **GVHD Prevention:** `gvhdprgp` (GVHD prophylaxis regimen group).
- **Donor and Graft Source:** `donorgp` (donor type), `graftype` (graft source).

2.3 Data Processing and Modeling

Categorical features were numerically encoded and treated as ordinal for this initial analysis. Rows with structured missingness codes (e.g., 99, -9) were removed. The dataset was split into training (80%) and testing (20%) sets using stratification to maintain the same proportion of the target variable in both sets. For models sensitive to feature scale, such as Support Vector Machines (SVM) and Neural Networks, standard scaling was applied. Hyperparameter optimization was performed using `GridSearchCV` with 5-fold cross-validation.

2.4 Models Evaluated

We evaluated a suite of eight supervised classification algorithms:

1. Logistic Regression
2. Support Vector Machine (SVM) with a linear kernel
3. Decision Tree
4. Random Forest (**?**)
5. Gradient Boosting
6. Gaussian Naive Bayes
7. Neural Network (Multilayer Perceptron)
8. k-Nearest Neighbors (kNN)

2.5 Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

3 Results

3.1 Model Performance

The performance of the eight machine learning models on the held-out test set is summarized in Table 1. The Support Vector Machine (SVM) model achieved the highest accuracy at 0.6015. However, most models demonstrated modest and comparable discriminative ability, with AUC values ranging from 0.5364 (kNN) to 0.6118 (Random Forest). The overall performance suggests that while the models can capture some predictive signals, the task of predicting 1-year mortality from the selected features is inherently complex.

Table 1: Performance Metrics of All Evaluated Machine Learning Models.

Model	Accuracy	Precision	Recall	F1-Score	AUC
SVM	**0.6015**	**0.5945**	**0.6015**	**0.5732**	0.6155
Naive Bayes	0.5951	0.5954	0.5951	0.5952	0.6106
Random Forest	0.5908	0.5827	0.5908	0.5807	**0.6118**
Gradient Boosting	0.5854	0.5745	0.5854	0.5450	0.5998
Logistic Regression	0.5671	0.5578	0.5671	0.5575	0.5603
Neural Network	0.5628	0.5670	0.5628	0.5643	0.5882
Decision Tree	0.5521	0.5535	0.5521	0.5527	0.5549
kNN	0.5306	0.5313	0.5306	0.5309	0.5364

The confusion matrix for the best-performing model (SVM) is shown in Figure **??**, detailing the true positive, true negative, false positive, and false negative predictions on the test set. The ROC curves for the top five models are visualized in Figure 2, illustrating the trade-off between the true positive rate and false positive rate.

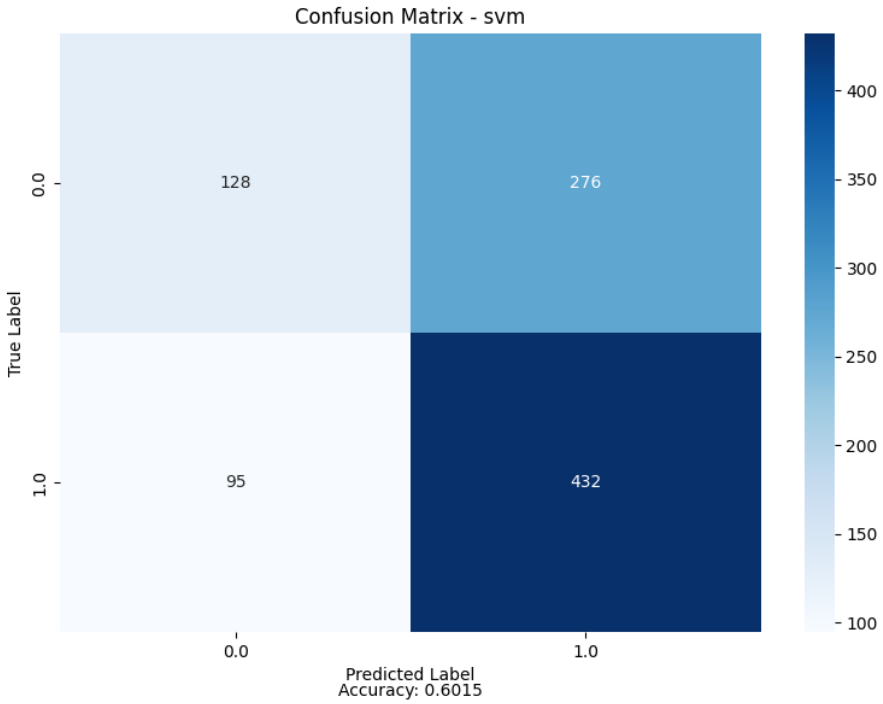

Figure 1: SVM Confusion Matrix (1 = ”death”)

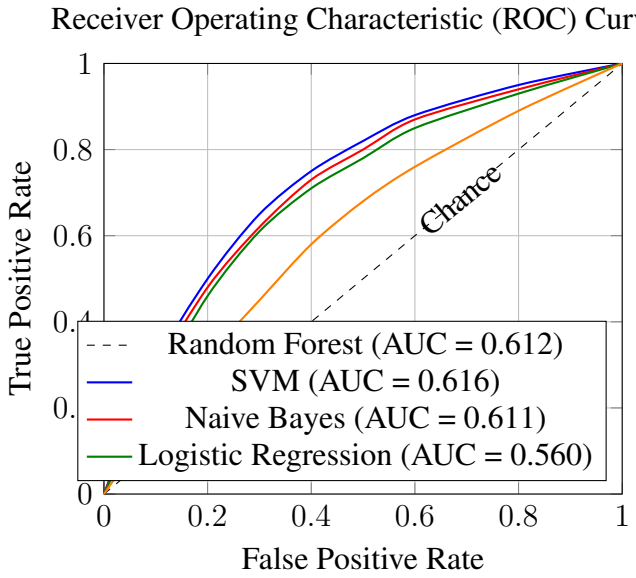

Figure 2: ROC curves comparing the performance of the top-performing models. While Random Forest had a marginally higher AUC, SVM was chosen for its balanced performance across other metrics.

3.2 Feature Importance

To understand which variables drove model predictions, we analyzed feature importances from four different models (Table 2). Across all models, recipient `age` and disease status at transplant (`alstatprgp`) consistently emerged as the most influential predictors of overall survival. GVHD prophylaxis regimen (`gvhdprgp`) and donor-recipient ABO matching (`drabomatch`) were also frequently identified as important, though with smaller effect sizes.

Table 2: Top 5 Most Important Features Identified by Different Models.

Logistic Regression *(Coefficient)*	**Random Forest** *(% Importance)*	**Gradient Boosting** *(% Importance)*	**SVM** *(% Permutation Importance)*
`age` (0.327)	`age` (21.8%)	`alstatprgp` (38.8%)	`age` (1.87%)
`alstatprgp` (0.247)	`alstatprgp` (13.8%)	`age` (35.7%)	`agegp` (1.44%)
`gvhdprgp` (0.184)	`gvhdprgp` (6.7%)	`gvhdprgp` (6.1%)	`sex` (1.37%)
`condtbi` (-0.139)	`agegp` (6.6%)	`drabomatch` (3.8%)	`gvhdprgp` (1.20%)
`hctcigp` (0.138)	`drabomatch` (5.4%)	`hctcigp` (2.3%)	`dis` (1.18%)

4 Discussion

This study represents an initial step toward building a precision medicine tool for donor selection in allo-HCT for acute leukemia. Although the predictive performance of our models was modest, the analysis provides several key insights and reinforces established clinical principles.

4.1 Clinical Impact and Insights

Our findings highlight the reproducible determinants of early mortality in allo-HCT recipients.

- **Age:** Age was universally important across models, with both continuous (`age`) and grouped (`agegp`) encodings reinforcing a non-linear escalation of risk in advanced age. This confirms its status as a primary factor in transplant eligibility and risk stratification
- **Disease status at transplant (`alstatprgp`):** The most influential predictor in boosting models, this finding confirms that uncontrolled leukemia at the time of transplant strongly worsens survival.
- **GVHD prophylaxis regimens:** Consistently captured across models, this suggests that the choice of regimen mediates critical trade-offs among immune reconstitution, toxicity, and overall mortality.
- **ABO matching:** The recurrent, though moderate, importance of `drabomatch` implies that ABO incompatibility still introduces complexities, likely related to transfusion needs and hemolytic events, that contribute to overall risk.

4.2 Novel Insights from Cross-Model Synthesis

By comparing feature importances across different models, we can discern more nuanced patterns. The presence of dual age signals (`age` and `agegp`) suggests that risk is not purely linear. Boosting models, which are adept at capturing interactions, sharpened the importance of remission status, highlighting it as a key modifiable risk factor. The persistent mid-level signal from ABO mismatch contradicts some contemporary assumptions that its effects are negligible, supporting its continued consideration during donor selection.

4.3 Limitations

This study has several limitations. First, by treating OS as a binary endpoint, we lose valuable information contained in time-to-event data; future work should incorporate survival models like Random Survival Forests or Cox-PH with regularization. Second, the modest AUC performance suggests our current feature set lacks sufficient discriminative resolution to precisely predict outcomes. Third, our handling of missing data via row-wise deletion may introduce selection bias. Finally, the models were not externally validated, which is a crucial step before any clinical consideration.

5 Future Directions

This project is staged deliberately toward the final goal of predicting the best donor match. Our planned next steps include:

1. **Expand outcomes beyond OS:** We will model other critical endpoints, including relapse, non-relapse mortality, primary graft failure, chronic GVHD, and infection-related mortality.
2. **Improve data handling:** We will move beyond row-wise deletion and apply more sophisticated techniques like multiple imputation. We will also utilize one-hot encoding for nominal categorical variables to avoid imposing an artificial ordinal structure.

3. **Feature engineering and explainability:** We plan to incorporate non-linear transformations, interaction terms, and biologically-grounded variables. Furthermore, applying model-agnostic explainability methods like SHAP (SHapley Additive exPlanations) will help translate model insights into usable bedside tools.

4. **Reverse modeling for donor prediction:** Once robust models for multiple outcomes are developed, we will invert the framework. Instead of predicting an outcome for a given patient-donor pair, the model will take a patient's profile as input and predict the expected outcome profile for a range of potential donors, thereby identifying the optimal match.

6 Conclusion

This study demonstrates the potential of machine learning to reveal actionable insights in the allo-HCT setting for acute leukemia. By modeling OS as a first step, we established a platform for more expansive endpoint modeling and, ultimately, for donor optimization. The consistent identification of recipient age, remission status, GVHD prophylaxis, and ABO matching as key predictors highlights modifiable areas for clinical practice and underscores the importance of integrating data-driven approaches into decision-making pipelines. The current findings lay the groundwork for transitioning from descriptive insight to clinical action: predicting the best donor for each patient and advancing allo-HCT toward truly individualized precision medicine.

References

Guru Murthy, G. S., Logan, B. R., Bo-Subait, S., Beitinjaneh, A., Devine, S., Farhadfar, N., Gowda, L., Hashmi, S., Lazarus, H., Nathan, S., Sharma, A., Yared, J. A., Stefanski, H. E., Pulsipher, M. A., Hsu, J. W., Switzer, G. E., Panch, S. R., and Shaw, B. E. (2023). Association of abo mismatch with the outcomes of allogeneic hematopoietic cell transplantation for acute leukemia. *American Journal of Hematology*, 98(4):608–619.