One-year Pediatric Bone Marrow Transplant Survival Prediction

Zoey Yu
Brown University
https://github.com/Zoey-Yu98/data1030-midterm.git

1. Introduction

Purpose:

Bone Marrow Transplant (BMT) is a vital medical procedure used to replace unhealthy bone marrow with healthy cells, often serving as an effective treatment for blood-related cancers and diseases such as leukemia, lymphoma, and multiple myeloma. However, BMTs carry significant risks and are associated with a relatively low survival rate. Notably, most patient deaths occur within the first year following the procedure, with survival rates improving over time. The one-year survival rate is approximately 63%, while the 25-year survival rate declines to 41%.

Given this context, predicting post-operative one-year survival status is crucial for facilitating informed decision-making. Accurate predictions can guide treatment planning, optimize post-transplant care, allocate resources effectively, and support patients and their families in navigating complex choices. This underscores the importance of developing robust predictive models for one-year survival following BMT.

Dataset Description

The dataset used in this study originates from a research paper published in *Biology of Blood and Marrow Transplantation* in 2010[1]. The research, conducted at a hospital in Poland between 2000 and 2010, aimed to investigate whether higher CD34(+) and CD3(+) cell doses in the graft could improve long-term survival without significantly increasing the incidence of severe acute or chronic graft-versus-host disease (GvHD).

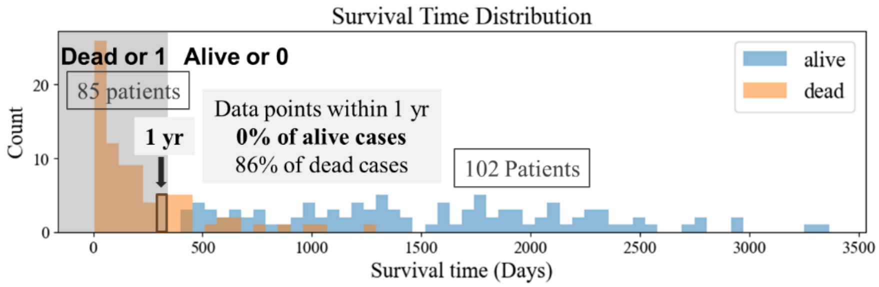
The dataset, also included in the UCI Machine Learning Repository [2] comprises information from 187 pediatric patients who underwent BMT. It contains 37 attributes or features, including pre-operative parameters (e.g., age, gender, blood type, disease type, antigen matching) and post-operative assessments (e.g., the occurrence of GvHD). For this project, the dataset was used to approach the problem as a binary classification task: predicting the one-year survival status (alive or deceased) of patients after surgery.

2. Explanatory Data Analysis

Target Variable Definition

The dataset did not explicitly include the target variable for one-year survival status. Instead, it provided two related attributes: **"Survival Time"**, indicating the time in days from surgery to either the observation point (for living patients) or death, and **"Survival Status"**, reflecting the patient's condition at the time of observation. The observation times varied due to factors such as patients changing hospitals or opting out of follow-ups years after surgery.

Crucially, no cases showed patients marked as "alive" at the time of observation. This allowed me to infer the true target variable, **"1_yr_survival_status"** (alive or deceased one year after surgery). Using the available data, I deduced this variable by considering whether patients survived beyond the one-year mark or not, as illustrated in the shaded section of the figure below. This approach ensured the validity of the classification task.

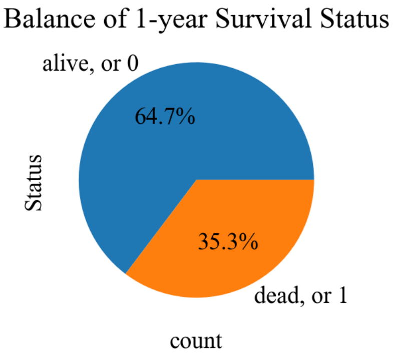


Data Collection Validity

All data points for all features in the dataset were collected within the first year following the bone marrow transplant. This ensures the consistency and validity of using all features and data points for the analysis, as they directly relate to the one-year survival status prediction.

Data Characteristics

The dataset assumes that all data points are independent and identically distributed (i.i.d.). The target variable, **"1_yr_survival_status"**, is slightly imbalanced, with approximately 65% of patients surviving (labeled as '0') and 35% not surviving (labeled as '1') one year after surgery. This mild imbalance in the target distribution was taken into account during the analysis to ensure the robustness of the predictive model.

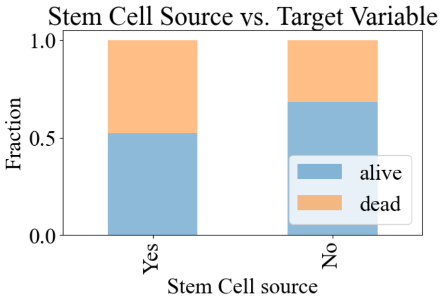


Missing Data Overview

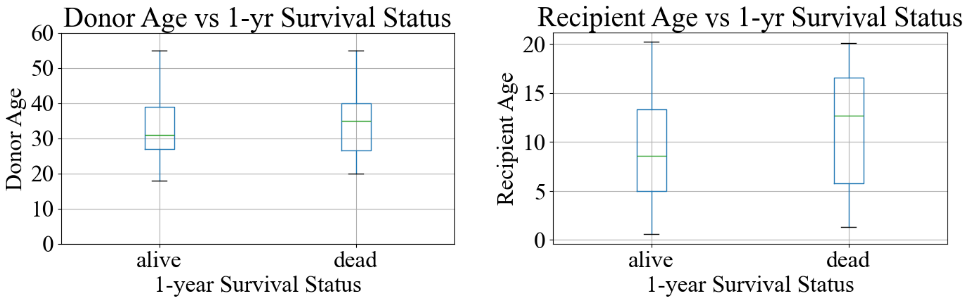
The dataset contains a significant amount of missing data, which required careful handling during analysis. Approximately 40% of the features (15 columns) include missing values, and 85% of the data points have at least one missing value. This high proportion of missing data presented challenges and necessitated the use of appropriate imputation techniques or feature engineering strategies to ensure the reliability of the predictive model.

Key Findings from Exploratory Data Analysis

The exploratory data analysis revealed several noteworthy patterns in the dataset:

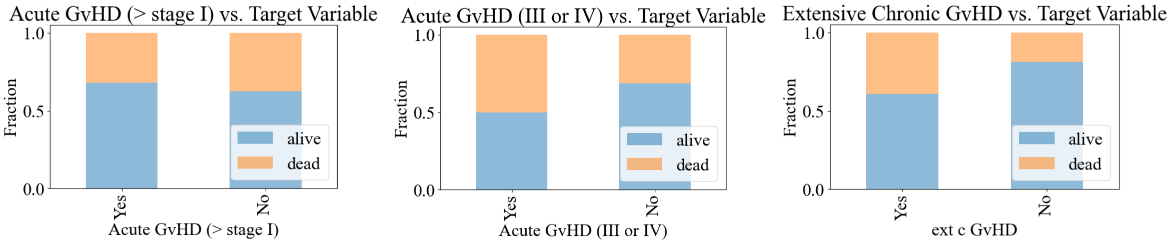


Source of Transplant (above): Bone marrow transplants sourced from peripheral blood were associated with better outcomes compared to those sourced from bone marrow.


Donor Age vs 1-yr Survival Status

Recipient Age vs 1-yr Survival Status

Age Factor (above): Lower ages of both donors and recipients correlated with improved survival rates.


Acute GvHD (> stage I) vs. Target Variable

Acute GvHD (III or IV) vs. Target Variable

Extensive Chronic GvHD vs. Target Variable

Impact of GvHD (above): While acute or extensive graft-versus-host disease (GvHD) increased the risk of death, it was not always fatal, highlighting the complex relationship between GvHD severity and survival outcomes.

These insights provided valuable context for understanding factors influencing post-operative survival and informed feature selection for model development.

3. Methods

Machine Learning Methods:

1) xgboost.XGBClassifier( )
2) RandomForestClassifier( )
3) SVC( )
4) LogisticRegression(penalty='elasticnet')

Handling Missing Data

For machine learning models like XGBClassifier, missing data can be naturally addressed without explicit preprocessing. However, for other models, preprocessing the missing data was essential. The following strategies were applied:

1) **Categorical and Ordinal Data**: Missing values were treated as an additional category to ensure no information was lost.

2) **Continuous Data**: Depending on the characteristics of the feature, missing values were handled in the following ways:

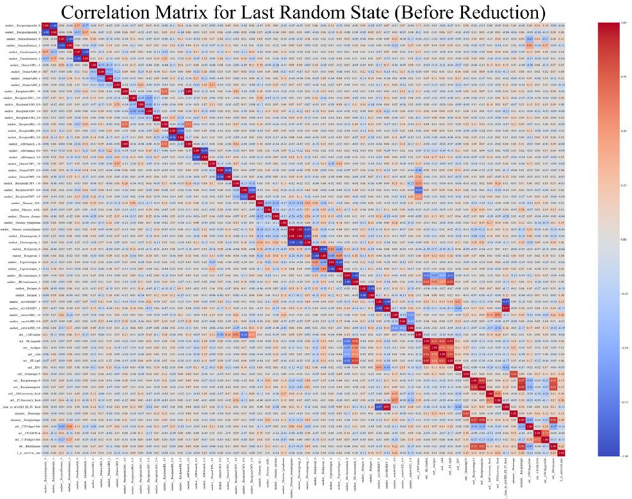
Type	*Examples*	*Characteristics*	*Methods of Preprocessing*
1	Medicine Dose	Missing data means no medicine is added	Filled all missing data to be 0
2	Time to certain side effects	Missing data means no side effects	Transformed all missing data to be bool (meaning 'had' or 'not had' such side effects)
3	Recipient HSC before surgery	Reasons unknown, only 2 rows	Dropped the 2 rows

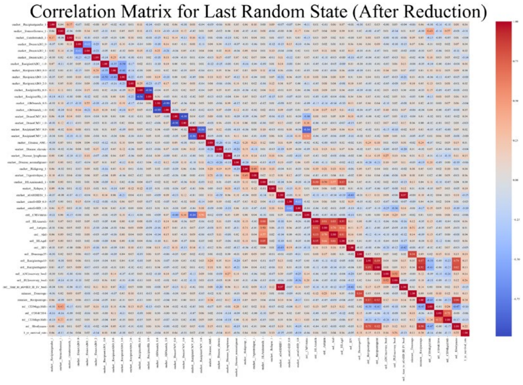
Feature Preprocessing and Correlation Analysis

The training data underwent preprocessing to handle missing values and prepare it for model development. After preprocessing, the target variable was added to the dataset, and the correlations between all features were visualized. We can see that none of the data was very correlated with the target variable (the last line).

Features with a perfect correlation (correlation = 1) were removed to avoid redundancy. However, some features with very high correlations were retained due to their structural or domain-specific importance in the dataset. These features, despite their high correlation, were considered valuable for preserving the predictive power and interpretability of the model.

After the preprocessing, a final preprocessed feature matrix has a shape (185, 47). All preprocessed data then went through another StandardScaler() to be standardized for model training.





Data Splitting for Model Training

The data was split into (train+validation):test = 8:2 using scikitlearn train_test_split, while stratifying the target variable. The (train+validation) set was further validated using **Stratified K-Fold Cross-Validation** with n_splits = 4 and shuffling for cross validation.

Cross-Validation Pipeline

The XGB does not work well with GridsearchCV, so I used three loops:

- Loop 1: random state * 10 for data splitting
- Loop 2: set Hyperparameters for XGB with early_stopping_rounds
- Loop 3: Cross Validation with KFold

In loops, I calculated the mean and standard deviation of the cross-validation scores across all folds. The best model was determined by the highest CV score mean. The final model's performance was evaluated on the test set. I finally got 10 test scores with 10 best parameters due to 10 random states.

For the other three models (RandomForestClassifier, SVC, and LogisticRegression(penalty='elasticnet'), I used GridSearchCV for hyperparameter tuning.

Evaluation Matrix:

The F2 score was chosen as the primary evaluation metric for all models due to its emphasis on minimizing False Negatives (FN) while still accounting for False Positives (FP):

- In this medical context, False Negatives (predicting a dying patient as "healthy") are critical as they might lead to patients missing additional treatment opportunities.
- Although False Positives (predicting a healthy patient as "dying") are less critical, they must still be minimized due to the scarcity and high cost of medical resources.

$$F_2 = \frac{(1 + 2^2) \times Precision \times Recall}{2^2 \times Precision + Recall} = \frac{5\ \mathrm{TP}}{5\ \mathrm{TP} + \mathrm{FP} + 4\ \mathrm{FN}}$$

A prediction of labeling all cases as "positive" (all patients as dying) results in a **baseline F2 score of 0.73**, which serves as a reference point for evaluating model performance.

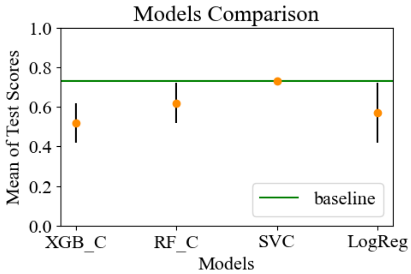
Hyperparameter Tuning

The hyperparameters were tuned as below, the bold number is the best parameter for the model.

Models	Hyperparameters	Hyperparameters Tuning Range
XGBClassifier( )	Learning_rate	[0.01, 0.03, 0.1, 0.3]
	reg_alpha	[0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2]
	max_depth	[1, 3, 10, 30, 100]
SVC( )	C	[0.01, 0.1, 1, 10, 100]
	gamma	[0.01, 0.1, 1, 10, 100]
RandomForestClassifier( )	max_depth	[1, 3, 10, 30, 100]
	max_features	[0.25, 0.5, 0.75, 1.0]
LogisticRegression (penalty = ElasticNet)	C	[0.01, 0.1, 1, 10, 100]
	l1_ratio	[0, 0.2, 0.4, 0.6, 0.8, 1]

4. Results:

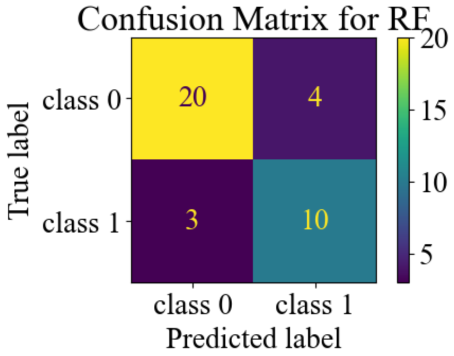
Models	Mean	std
XGBClassifier( )	0.52	0.10
RandomForestClassifier( )	**0.62**	**0.10**
SVC( )	0.73	0.0
LogisticRegression (penalty='elasticnet')	0.57	0.15



We can see that all models perform worse than the baseline, except for SVC. However, examining the SVC confusion matrices reveals that SVC simply predicts all points as positive, resulting in performance equivalent to the baseline. Due to the small size of the data set and the intricate complexity of this prediction task, we were unable to achieve better results than the baseline.

Among the remaining three models, random forest performs the best, showing the lowest standard deviation. After reviewing the best parameters across all ten random states, we observe that max_depth = 10 and max_features = 1.0 yield the best results. Therefore, I selected this configuration as the best model for this task.

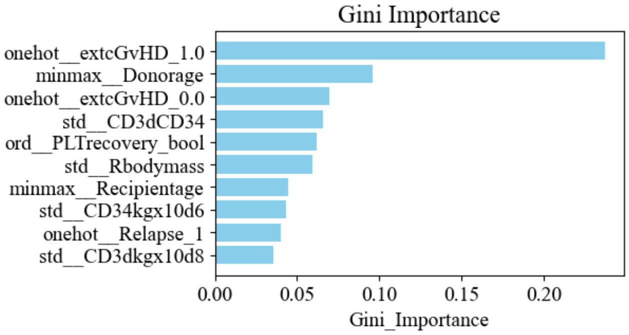
The confusion matrix for the selected model, with the best parameters, is shown below. It aligns with expectations for using the F2 score in the machine learning pipeline, with **FN < FP << TP << TN**.

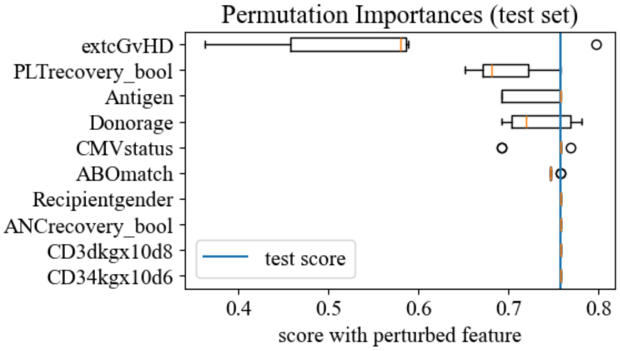


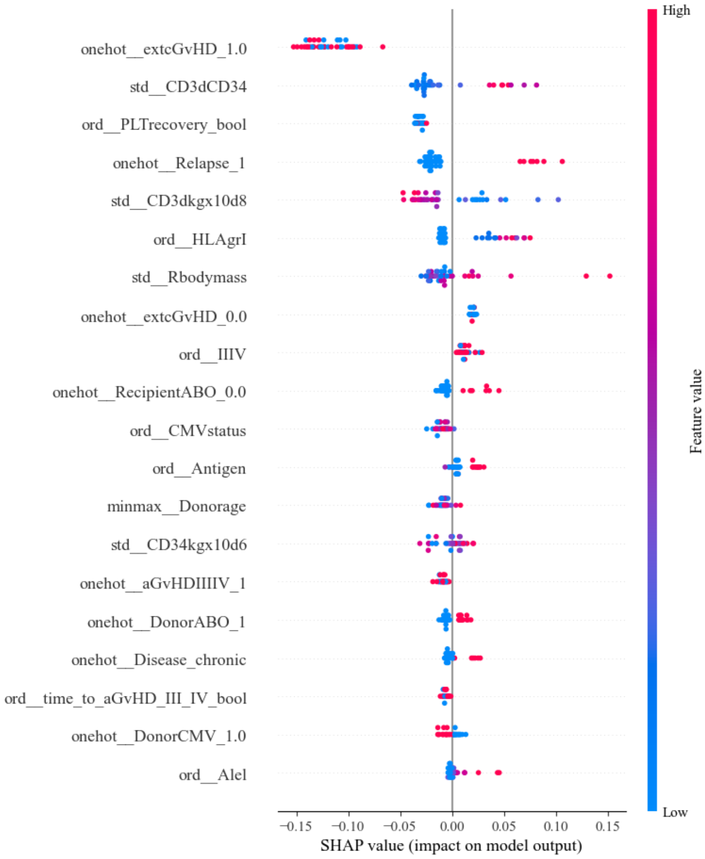
Global Feature Importance

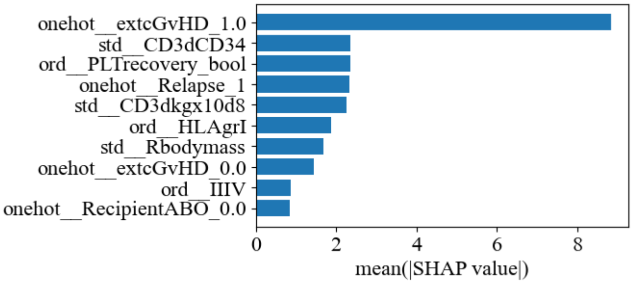
I used the built-in feature importance of the random forest classifier, Permutation importances, and SHAP global importance to understand the global feature importance of the best mode.

As we can see from the three figures below, showing the top ten feature importance, the extcGvHD (extensive chronic GvHD) is always the most important feature of the prediction. The other features' importance showing up and down based on the methods that I am using. But generally, donor's and recipient's age, if the platelet count ended up being >50,000/mm³ (PLTrecovery), incompatibility between donor and recipient in terms of HLA, antigen and blood type, always shows high importance in different method of importance calculation.





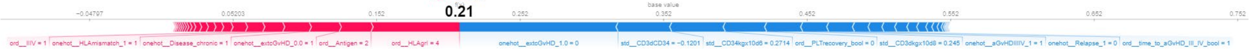




Local Feature Importance

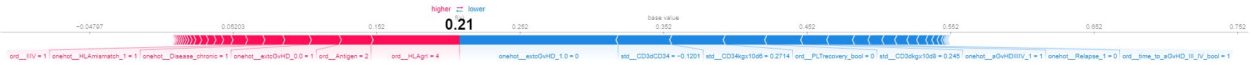
I used the SHAP for local feature importance calculations. The baseline is 0.35 for class 1.

Index = 6

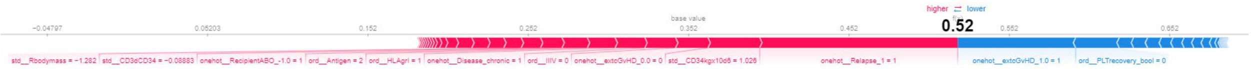


For point 6, the prediction is 0.21, smaller than 0.35, so the model predicted the patient to be alive. Although the patient developed extcGvHD (extensive chronic graft-versus-host disease), which accounted most in reducing the chance of death, their medication doses helped decrease the chance of death. Additionally, PLTrecovery_bool = 0, indicating that the platelet count ended up being >50,000/mm³, also played an important role in reducing the risk of death. Furthermore, the absence of acute GvHD stage III or IV decreased the patient's likelihood of death.

However, a high level of HLA incompatibility between the donor and the recipient (HLAgrl = 4) significantly increased the risk of death. The large difference in antigens between the donor and the recipient (Antigen = 2) also played a very important role in raising the patient's chance of death. This was followed by other factors, including developing extcGvHD, having a chronic disease, HLA mismatch, and the development of acute GvHD stages II, III, or IV.



Index = 27



For point 27, the prediction is 0.52, smaller than 0.35, so the model predicted the patient to be dead. The fact that the patient did not develop extcGvHD (extensive chronic graft-versus-host disease) accounted most in decreasing their chance of death, following by the fact that their platelet count ending up being >50,000/mm³ (**PLTrecovery_bool = 0).** But high reoccurrence of the disease (Relapse=1) greatly increased their chance of death, following by the medicine dose, and so on.

5. Outlook and Future Directions

The performance of the model on this dataset highlights some key challenges:

1. **Complexity of the Problem**: Predicting survival outcomes post-BMT is inherently difficult due to the multifactorial nature of the problem, involving a wide range of medical and physiological variables.
2. **Insufficient Data**: The dataset's size and potential limitations in capturing all relevant clinical features may hinder the model's ability to generalize effectively.

To improve prediction accuracy and model performance, the following strategies are proposed:

1. **Consult Medical Experts**: Collaborate with healthcare professionals to identify additional key features relevant to survival outcomes, such as graft source, donor-recipient compatibility, and pre-transplant health conditions.
2. **Expand the Dataset**: Seek out additional data sources to enhance statistical robustness. This could involve leveraging larger datasets, partnering with hospitals, or accessing multicenter clinical studies.
3. **Experiment with Models and Metrics**:
    - Test a broader range of machine learning models to identify those better suited for this task.
    - Fine-tune the F_β score, adjusting β to explore its impact on the trade-off between recall and precision.

6. Reference:

[1] Kaƚwak K, et al. Biol Blood Marrow Transplant. 2010 Oct;16(10):1388-401. doi: 10.1016/j.bbmt.2010.04.001.

[2] Sikora, M., Wróbel, Ł., & Gudyś, A. (2020). Bone marrow transplant: children [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NP6Z.

The grammar was checked by ChatGPT4.0.