HIGH SCHOOL EDITION
Journal of Student Research

Prediction of Chronic Graft vs. Host Disease Using Machine Learning

Sanay Bordia[1] and Ramin Ramezani[#]

[1]Archbishop Mitty High School, San Jose, CA, USA
#Advisor

ABSTRACT

This paper attempts to predict the onset of chronic Graft vs. Host Disease (GVHD) in children with blood cancers who have received a bone marrow or stem cell transplant using machine learning models. It analyzes and compares the results of three different models in terms of how accurate they each are in predicting chronic GVHD. These models are Logistic Regression, J48 algorithm using decision trees, and Multilayer Perceptron. The models are formed using a dataset containing 36 attributes, excluding chronic GVHD itself. Through data preprocessing and analysis in Weka, these 36 attributes are narrowed down for each model to figure out which combination of attributes leads to the best predictive accuracy. The study uses 10-fold cross validation for each model and uses the Receiver Operating Characteristic (ROC) Area as a measure of the accuracy for each model. The study found that Multilayer Perceptron is the best predictor of chronic GVHD. In comparison, Logistic Regression was the worst predictor of chronic GVHD. The J48 algorithm used the least number of attributes to make its prediction.

Introduction

Cancers of all forms are prevalent across the globe. Specifically, in children, cancers such as leukemia and lymphoma are unfortunately a common occurrence. In such cases, bone marrow or stem cell transplants are becoming more and more common, as they provide a potential curative therapy for some of these advanced blood cancers. Critically, stem cells are unique in that they adapt to serve different functions based on what the body needs, living in bone marrow.

This study is based on allogeneic transplants, in which stem cells come from a "donor" – another person. They are administered after a patient has received chemotherapy or radiation therapy. According to Fred Hutch News Service, 70% of such patients get acute GVHD, and 40% get chronic GVHD [1]. Experts concur that initial GVHD is not necessarily a negative sign but instead could be a positive one, as it indicates that the newer cells are actively getting rid of the cancer cells. This initial GVHD is known as acute GVHD.

For the purposes of this research, the study is focusing on chronic GVHD, in which the new donor cells continue to attack the patient's original cells for a prolonged period or at times for life. This occurs when the donor cells respond to being in a foreign body by attacking the host. In response, the host's immune system gets activated, creating a vicious cycle inside the body that can have serious side effects, including organ failures or even death. Currently, there is no cure for chronic GVHD. However, the way this disease is managed is by using immunosuppressants. One problem with immunosuppressants is that they weaken the host's immune system, so they do not fight the donor cells, keeping the body at equilibrium. Unfortunately, this makes the patient prone to even the smallest of infections which itself could lead to serious consequences. This study aims to find the attributes from a donor that could best predict the onset of chronic GVHD in a recipient. If a recipient has an option of using donor cells from more than one potential donor, this information can be used along with the most important marker – survivability – to pick the most viable donor for the recipient. This might not only cure a patient of their original disease but also leave them with a lower chance of having chronic GVHD.

Methodology

Table 1. All Attributes in Dataset

Type	Attributes
Nominal	RecipientGender, StemCellSource, DonorAge35, IIIV, GenderMatch, DonorABO, RecipientABO, RecipientRh, ABOMatch, CMVStatus, DonorCMV, RecipientCMV, Disease, RiskGroup, Txpostrelapse, DiseaseGroup, HLAMatch, HLAMismatch, Antigen, Alel, HLAgrl, RecipientAge10, RecipientAgeint, Relapse, aGVHD, cGVHD, CD3DCD34
Numeric	DonorAge, RecipientAge, Cd34kgx 10d6, Cd34kgx 10d8, Rbodymass, ANCrecovery, PLTRecovery, time_to_aGVHD, survival_time, survival_status

The data set used for this research is obtained from the University of Irvine [2]. It contains data for 187 children who underwent bone marrow or stem cell transplant. The data contains 37 attributes for each child, as represented in Table 1 above.

Weka

This study uses Waikato Environment for Knowledge Analysis (Weka), developed by the University of Waikato, which is a free open-source software that is used for data mining and provides various machine learning algorithms that can be used to build models and test datasets [3].

Approach

The study tested the attributes in the UC Irvine dataset on each model in Weka, comparing the individual results of each model in terms of the accuracy of predicting the occurrence of chronic GVHD. The models that were used for comparison were Logistic Regression, J48 using decision trees, and Multilayer Perceptron. Each of them is described in more detail below.

Each model started with all 36 attributes (the study does not include chronic GVHD as an attribute because it is the dependent variable; as such, it is separated from the attribute list). Running the models with all 36 attributes provided a baseline ROC Area, accuracy score, precision, recall, and F-Measure. In each subsequent trial, the list of attributes was adjusted to increase the accuracy of each model, in this case the goal being to improve the ROC Area. Attributes that decreased the ROC Area were removed. The final set of attributes resulted in the highest possible ROC Area for each model.

Out of the 187 patients who were analyzed in this study, 156 patients had data available for chronic GVHD. This dataset is imbalanced because only 28 out of these 156 patients had chronic GVHD. Most of the children did not have chronic GVHD. As such, although accuracy is a good measurement, ROC Area is more representative of the quality of the model as it considers the predicted scores. Accuracy can be skewed in that it can account for the majority – in this case being the 128 patients out of 156 that do not have chronic GVHD – but can leave out the minority – the 28 patients that do have chronic GVHD. In contrast, the ROC Area targets the specific value by looking at the false positive fraction versus the true positive fraction, therefore accounting for both. In addition, precision, recall, and F-Measure were all calculated, but they are not the focus of this study. They provide additional context on how accurate the model is, as the study is using an imbalanced dataset. Precision is used to see how accurately the model figures out a true positive, while recall is used to see how accurately the model can predict a positive result overall. The F-Measure balances the precision and recall providing a single number that adds to the understanding of how well the model functions [4]. The closer it is to 1, the better the model is. However, the study only includes these measures in the results and briefly in the discussion, but it does not discuss them in detail in this paper.

Equation 1: The Specificity Equation

$$FP\ /\ (FP\ +\ TN)$$

Equation 1 plots the specificity, or the false positives over the false positives plus true negative instances. It is seen on the x axis of an ROC Area curve graph.

Equation 2: The Sensitivity Equation

$$TP\ /\ (TP\ +\ FN)$$

Equation 2 plots the sensitivity, or the true positives over the true positives plus false negative instances. It is seen on the y axis of the ROC Area curve graph [5].

By default, Weka provides an implementation of stratified cross validation. In all three models, 10-fold cross validation is implemented. This means that 90% of the dataset is used for training, and 10% is used for testing. Cross validation is used to limit overfitting as it divides the data into several test-train splits, in comparison to just one. It is used in this study as the models handle unseen data, meaning that there is limited data attempting to generalize predictions. Chronic GVHD has minimal data available to train and test, and as such 10-fold cross validation is used [6].

Logistic Regression

Multinomial Logistic Regression is one of the machine learning models used in this study to predict the dependent variable chronic GVHD, and it usually assumes that there is a linear pattern between the inputs and output. The model is used to predict the likelihood of whether the person will get chronic GVHD (yes) or not (no). Logistic regression is commonly used with dichotomous variables like chronic GVHD [7].

The output of a logistic regression model ranges from 0 to 1. This study's goal was to find the best Logistic Regression model using the metric of ROC Area. The model uses a sigmoid function to compensate for any outliers in the data [8].

J48

J48 is an algorithm that generates a decision tree. The decision tree consists of nodes that represent tests done on each individual attribute, branches that represent the results of each test, and leaf nodes that contain class labels. J48 generates predictions quickly and works well with smaller datasets. It uses supervised learning, and

Weka added several additional features to the tree, such as whether to prune the tree or not as well as algorithms to reduce overfitting. These parameters are helpful as they make the generalization of data more accurate [9].

Multilayer Perceptron

The Multilayer Perceptron model consists of neural networks continuously stepping forward by feeding each other information from the previous nodes. The output is the weighted sum of all the inputs [10]. Using Weka, the default model consists of an input layer, one hidden layer, and an output layer. One hidden layer – the default set by Weka – was used because the data is nonlinear.

There are two requirements when it comes to what function should be used for the nodes:

1) The function should be nonlinear.
2) The function should be continuous and differentiable with relation to the nodes.

As such, all the nodes in Weka use the sigmoid function as their activation function as it fits both these requirements [11].

Equation 3: Sigmoid Function

$$f(x) = 1/(1 + e^{-x})$$

Equation 3 represents a regular sigmoid function.

Equation 4: Derivative of a Sigmoid Function

$$f'(x) \ = \ f(x)(1 \ - \ f(x))$$

To figure out if the sigmoid function is differentiable, the derivative of the sigmoid function is taken and examined, as represented by Equation 4. The derivative is found by applying the quotient rule and the chain rule, and f(x) as seen in the equation represents the regular sigmoid function.

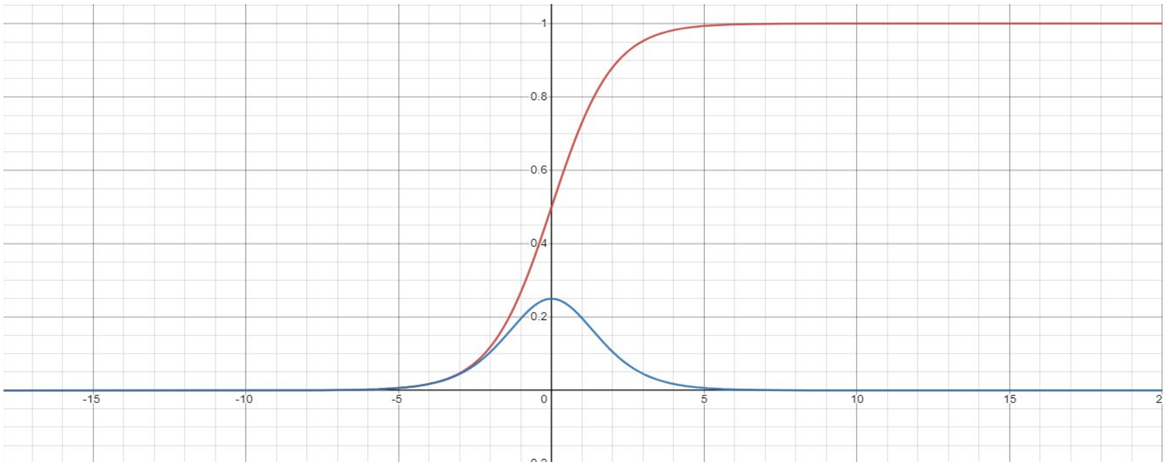

Figure 1. Graph of a Sigmoid Function and its Derivative [12]

In Figure 1, the sigmoid function is both continuous and differentiable. At each point in the function's domain, a derivative at that point exists since it can be evaluated to a real number each time. Since it is established that the function is differentiable throughout, continuity is implied [13].

The model is trained by implementing back-propagation, using a learning rate of 0.3, a momentum of 0.2, and a training time of 500. It is run using 10-fold cross validation as a means of comparison.

Results

Overall, this study found that Multilayer Perceptron had the best ROC Area of 0.736, F-Measure of 0.871, and accuracy of 78.205%, out of all the models. In contrast, Logistic Regression had the worst ROC Area, F-Measure, and accuracy at 0.649, 0.858, and 76.923%, respectively. J48 had results in between these two models, with an ROC Area of 0.669, F-Measure of 0.869, and an accuracy of 77.564%. In terms of the attributes, J48 had the least number of attributes at 19 after testing was finished, while Logistic Regression had the most attributes left at 30. According to the study, all the remaining attributes have some connection in causing chronic GVHD based on how they are labeled.

Logistic Regression

Table 2. Shortlisted Attributes Using Logistic Regression

Type	Attributes
Nominal	RecipientGender, StemCellSource, DonorAge35, IIIV, GenderMatch, DonorABO, RecipientABO, RecipientRh, DonorCMV, RecipientCMV, Disease, RiskGroup, DiseaseGroup, HLAMatch, HLAMismatch, Antigen, Alel, HLAgrl, RecipientAgeint, Relapse, aGVHD, CD3DCD34
Numeric	DonorAge, RecipientAge, Cd34kgx 10d8, ANCrecovery, PLTRecovery, time_to_aGVHD, survival_time, survival_status

Table 2 consists of the 30 attributes that provide the best outcome for determining chronic GVHD with Logistic Regression based on trials.

Baseline (all attributes)

Table 3. Accuracy and ROC Area before testing with Logistic Regression

Cross Validation	Accuracy	ROC Area	Precision	Recall	F-Measure
10	69.231%	0.558	0.845	0.766	0.803

As mentioned previously, 10-fold cross validation was used, and the dataset was tested in Weka. This original dataset with all 36 attributes rendered an accuracy of 69.231% and an ROC Area of 0.558, as seen in Table 3.

Final (shortlisted attributes)

Table 4. Accuracy and ROC Area after testing with Logistic Regression

Cross Validation	Accuracy	ROC Area	Precision	Recall	F-Measure
10	76.923 %	0.649	0.865	0.852	0.858

10-fold cross validation was used, and 30 attributes remained after testing was complete. They rendered an accuracy of 76.923% and an ROC Area of 0.649, representing an increase in accuracy of 7.692% and an increase in ROC Area of 0.091.

J48

Table 5. Shortlisted Attributes Using J48

Type	Attributes
Nominal	RecipientGender, DonorAge35, IIIV, GenderMatch, DonorABO, RecipientRh, CMVStatus, DonorCMV, Disease, Txpostrelapse, HLAMatch, HLAMismatch, Antigen, Relapse, CD3DCD34
Numeric	DonorAge, RecipientAge, Cd34kgx 10d6, survival_time

Table 5 consists of the 19 attributes that provide the best outcome for determining chronic GVHD with J48 based on trials.

Baseline (all attributes)

Table 6. Accuracy and ROC Area before testing with J48

Cross Validation	Accuracy	ROC Area	Precision	Recall	F-Measure
10	81.410%	0.498	0.856	0.930	0.891

10-fold cross validation was used, and the dataset was tested in Weka. This original dataset with all 36 attributes rendered an accuracy of 81.410% and an ROC Area of 0.498, as seen in Table 6.

Final (shortlisted attributes)

Table 7. Accuracy and ROC Area after testing with J48

Cross Validation	Accuracy	ROC Area	Precision	Recall	F-Measure
10	77.564%	0.669	0.835	0.906	0.869

10-fold cross validation was used, and 19 attributes remained after testing was complete. They rendered an accuracy of 77.564% and an ROC Area of 0.669, representing a decrease in accuracy of 3.846% but an increase in ROC Area of 0.171, the more important measure.

Multilayer Perceptron

Table 8. Shortlisted Attributes Using Multilayer Perceptron

Type	Attributes

Nominal	RecipientGender, StemCellSource, DonorAge35, GenderMatch, DonorABO, RecipientABO, RecipientRh, ABOMatch, CMVStatus, DonorCMV, RecipientCMV, Disease, RiskGroup, Txpostrelapse, DiseaseGroup, HLAMismatch, Antigen, Alel HLAgrl, RecipientAgeint, Relapse
Numeric	RecipientAge, DonorAge, Cd34kgx10d6, survival_time, survival_status

Table 8 consists of the 26 attributes that provide the best outcome for determining chronic GVHD with Multilayer Perceptron based on trials.

Baseline (all attributes)

Table 9: Accuracy and ROC Area before testing with Multilayer Perceptron

Cross Validation	Accuracy	ROC Area	Precision	Recall	F-Measure
10	75.641%	0.629	0.852	0.852	0.852

10-fold cross validation was used, and the dataset was tested in Weka. This original dataset with all 36 attributes rendered an accuracy of 75.641% and an ROC Area of 0.629, as seen in Table 9.

Final (shortlisted attributes)

Table 10: Accuracy and ROC Area after testing with Multilayer Perceptron

Cross Validation	Accuracy	ROC Area	Precision	Recall	F-Measure
10	78.205%	0.736	0.846	0.898	0.871

10-fold cross validation was used, and 26 attributes remained after testing was complete. They rendered an accuracy of 78.205% (increase of 2.564%) and an ROC Area of 0.736 (increase of 0.107).

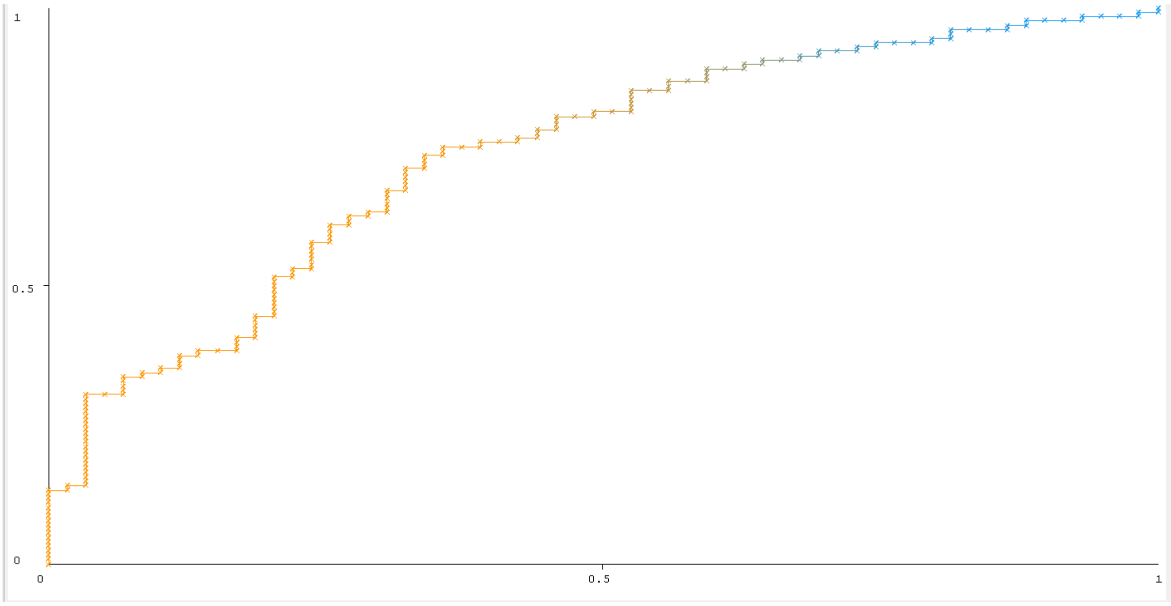

Figure 2. Graph of Area Under ROC Curve for Multilayer Perceptron

In Figure 2, the X Axis represents the false positive rate, and the Y Axis represents the true positive rate. The graph represents the ROC Area of 0.736. The closer the graph is to 1, the more accurate the graph is.

Discussion

Table 11. Comparison of Accuracy and ROC Area of all models used

Algorithm	Cross Validation	Accuracy	ROC Area	Number of Attributes	Precision	Recall	F-Measure
Logistic Regression	10	76.9231 %	0.649	30	0.865	0.852	0.858
J48	10	77.5641%	0.669	19	0.835	0.906	0.869
Multilayer Perceptron	10	78.2051%	0.736	26	0.846	0.898	0.871

As seen in Table 11, from all the models studied above, the Multilayer Perceptron model had the best ROC Area of 0.736 and F-Measure of 0.871, while also maintaining the best overall accuracy at 78.2051%. This can be seen as an acceptable model when looking purely at the ROC Area [14] and a good model when looking at the F-Measure. Since chronic GVHD is a complex problem that medical science is still actively researching, this model works well in relation to how difficult a problem it is attempting to tackle.

While Logistic Regression is accurate for binary classification, it can be hypothesized that it was not as accurate as the Multilayer Perceptron model due to how Logistic Regression works better with data that can be linearly separated – even in Weka. The data in this study is made up of different attributes that may or may not have any causation with chronic GVHD. As such, trying to determine a linear pattern could be erroneous. In comparison, the Multilayer Perceptron model has an additional complexity due to the use of neural networks that can handle nonlinear data, possibly giving more accurate results [15].

The J48 algorithm is based on decision trees, which are great at handling smaller datasets with nonlinear data. Still, the inaccuracy that occurred in J48 in this study could possibly be a result of overfitting, which leads to fluctuations in data and results in many errors. Even though certain Weka algorithms as well as the use of 10-fold cross validation limit this problem, it is hard to completely get rid of. As such, its effects can be hypothesized to still affect the model's accuracy. Another issue is that decision trees are sensitive to changes in data, so the removal of certain attributes after each step following the method done in this study could have disrupted the model [16].

Conclusion

This study's analysis finds that the Multilayer Perceptron model is the best predictor of chronic GVHD, as determined by its ROC Area in comparison to the other two models, Logistic Regression and J48. This research adds to other findings on how to predict chronic GVHD but instead of using biomarkers, it uses machine learning models to predict chronic GVHD. It also adds onto possible biomarkers that could be used in predicting chronic GVHD, as the attribute Cd34kgx10d8 (CD34+) appears as a shortlisted attribute in all three models [17]. For the future, I would like to incorporate more models into the study and expand the dataset to attempt to get a higher ROC Area that can more effectively point to specific attributes more closely related to chronic GVHD.

Limitations

One limitation of this study is not having a larger data set for the children. This is a heavily imbalanced dataset, and the models are only able to use a few data points to make larger generalizations in predicting chronic GVHD.

Acknowledgments

I would like to thank UCLA Professor Ramin Ramezani and Teaching Assistant Joanna Gilberti for giving me an opportunity to conduct this research. This was an enriching experience!

References

[1] Tompa, Rachel. "Life with graft-vs.-host disease: When the transplant is just the beginning." Fred Hutch, 21 April 2015, Life with graft-vs.-host disease: When the transplant is just the beginning (fredhutch.org)
[2] Bone marrow transplant: children. (2020). UCI Machine Learning Repository, UCI Machine Learning Repository
[3] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[4] Brownlee, Jason. "How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification ." Machine Learning Mastery, 3 January 2020, How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification (machinelearningmastery.com)

[5] Ekelund, Suzanne. "ROC curves – what are they and how are they used?" acutecaretesting, January 2011, ROC curves – what are they and how are they used? (acutecaretesting.org)

[6] Brownlee, Jason. “A Gentle Introduction to k-fold Cross-Validation.” Machine Learning Mastery, 3 August 2020, A Gentle Introduction to k-fold Cross-Validation (machinelearningmastery.com)

[7] “What is Logistic Regression?” Statistics Solutions, What is Logistic Regression? - Statistics Solutions

[8] Ketha, Santhosh. “Effect of outliers on Neural Network’s performance.” Medium, 29 October 2019, Effect of outliers on Neural Network’s performance | by Santhosh Ketha | Analytics Vidhya | Medium

[9] “J48 Classifier Parameters.” The Schank Academy, j48_parameters.pdf (schankacademy.com)

[10]”Multilayer Perceptron.” Science Direct, Multilayer Perceptron - an overview | ScienceDirect Topics

[11] “The Backpropagation Algorithm-PART(1): MLP and Sigmoid.” ML-DAWN, The Backpropagation Algorithm-PART(1): MLP and Sigmoid | ML-DAWN (mldawn.com)

[12] Chakraborty, Arunava. “Derivative of the Sigmoid function.” Towards Data Science, 7 July 2018, Derivative of the Sigmoid function | by Arc | Towards Data Science

[13] McMullin, Lin. “Differentiability Implies Continuity.” Teaching Calculus, 17 September 2019, Differentiability Implies Continuity | Teaching Calculus

[14] Zach. “What is Considered a Good AUC Score?” Statology, 9 September 2021, What is Considered a Good AUC Score? - Statology

[15] Raschka, Sebastian. “What is the relation between Logistic Regression and Neural Networks and when to use which?”, What is the relation between Logistic Regression and Neural Networks and when to use which? (sebastianraschka.com)

[16] “Decision Tree Advantages and Disadvantages.” eduCBA, Decision Tree Advantages and Disadvantages | Decision Tree Regressor (educba.com)

[17] Pidala, J., Sarwal, M., Roedder, S. *et al.* Biologic markers of chronic GVHD. *Bone Marrow Transplant* 49, 324–331 (2014). https://doi.org/10.1038/bmt.2013.97