Multiclass Classification for GvHD Prognosis Prior to Allogeneic Stem Cell Transplantation

Md. Asif Bin Khaled [1] Md. Junayed Hossain [2] Saifur Rahman[3] Jannatul Ferdaus [4]

Department of Computer Science and Engineering
Independent University, Bangladesh
Dhaka, Bangladesh.
{[1]mdasifbinkhaled,[2]junayed.ndc16,[3]ksaifur13,[4]jannatul0442}@gmail.com

Abstract

Hematopoietic Stem Cell Transplantation (HSCT) is an effective treatment for a variety of blood diseases, including hematologic and lymphoid malignancies, along with numerous other conditions. One of the most prevalent adverse effects of Allogeneic Stem Cell Transplantation (ASCT) in patients is Graft versus Host Disease (GvHD). Inflammation brought on by the donor's stem cells attacking the patient's body can lead to GvHD. Moreover, Acute Graft versus Host Disease (aGvHD) and Chronic Graft versus Host Disease (cGvHD) are the two possible manifestations. The patient has a chance of developing this disease even if the donor and the recipient are a perfect match. Therefore, early diagnosis of the forms of GvHD before a patient receives ASCT treatment is essential. However, it is still necessary to identify the type of GvHD even if the patient has already undergone a transplant to advise any clinical decision. In this research, the types of the GvHD are precisely predicted using a variety of multi-class classification models. The techniques utilized in this study include Random Forest, Decision Tree, K-Nearest Neighbor, Gradient Boosting, XG Boosting, LG Boosting and a feed forward Artificial Neural Network named Multilayer Perceptron. This study revealed that the Random Forest algorithm demonstrated state-of-the-art performance in multi-class classification, **with an accuracy of 98.62% along with 96.38% F1-Score and an area under the ROC curve (AUC) score of 98.02%.** In terms of accuracy and reduced feature dependence for predicting the multi-class target feature, this study offers a useful prognosis tool for medical experts.

Introduction

Bone marrow or stem cell is soft, adipose tissue in the body that is located in the skeletal structures. It is responsible for producing the red blood cells, white blood cells, and platelets of the human body, and it also contains hematopoietic stem cells. Bone marrow or stem cell transplantation is necessary when a cancer patient's bone marrow has been damaged by radiation therapy or intense chemotherapy.There are two different types of stem cell transplants they are: **Autologous Stem Cell Transplant** and **Allogenic Stem Cell Transplant (ASCT)**. But following ASCT, Graft versus Host Disease (GvHD) may occur, which is further classified into acute GvHD and chronic GvHD. Acute GvHD might surface within a day, a week, a month, or within 100 days after transplantation. It can damage several organs, such as the liver, skin, eyes, mucosa, and intestines of a patient. On the other hand acute GvHD is developed after two years.
This study develops several ML and ANN models that can categorize a multi-class feature by estimating the possibility that a patient will have acute GvHD, chronic GvHD, both of these, or none of these diseases. These models only rely on 9 out of the 37 features from the original dataset to make this admiringly accurate prediction.

Problem Statement

The chances of developing GvHD following an ASCT is difficult to predict. Failure to anticipate this disease in its early stages can lead to the development of many other diseases. One condition that is a frequent symptom of chronic GvHD is dry eye disease (DED), which has been identified as a serious side effect of ASCT. In most circumstances, manually foretelling this condition is quite challenging and erroneous. Though donor gene-expression profiling and a new biomarker panel also can predict this GvHD but both of these methods are very time-consuming and imprecise. However, ML methods may be the best option for predicting GvHD using donor and recipient medical data, and most importantly, predicting it before the transplantation.

Proposed Methodology

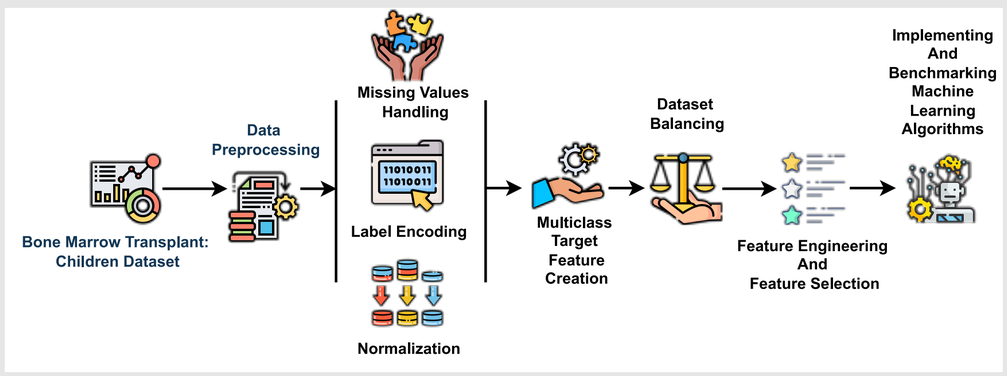


Figure 1. Methodology Employed in this Study.

Imputing Missing Values

Missing Values are imputed two different way.

- **Miss Forest :** For numerical values.
- **Random Forest Classifier :** For categorical values.

Normalization

The purpose of normalization is to convert the values of the dataset's numeric columns to a standard scale without losing information or distorting the ranges of values.

Target Feature Creation

The two features termed "extensive chronic GvHD" and "acute GvHD III IV" from the dataset have merged so that the ML models can perform a multi-class classification. A new multi-class attribute called "GvHD Diseases" is created using the mentioned features with binary values.

Balancing Imbalance Dataset

The dataset appears to be slightly unbalanced, so Synthetic Minority Over-sampling Technique (SMOTE) technique and the "Random Over Sampler" approach have been applied to generate instances for the minor category to balance this dataset.

Feature Engineering and Feature Selection

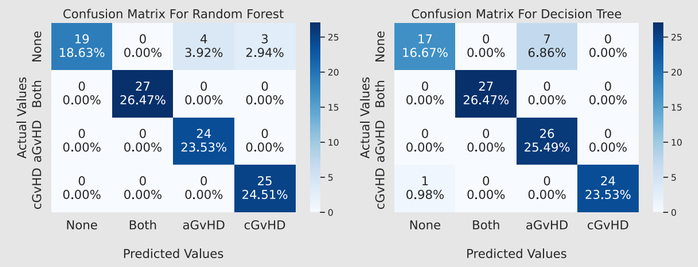
The features with a threshold value of 80% are removed from the dataset before creating the correlation matrix. A feature selection approach is then applied among the selected features following the correlation matrix. This feature selection method makes use of lasso regression by taking α as 0.05.

Results and Discussion

Here the result has been analyzed in different segments. Among them the confusion matrix, validation curves and the accuracy of different matrices are shown bellow.

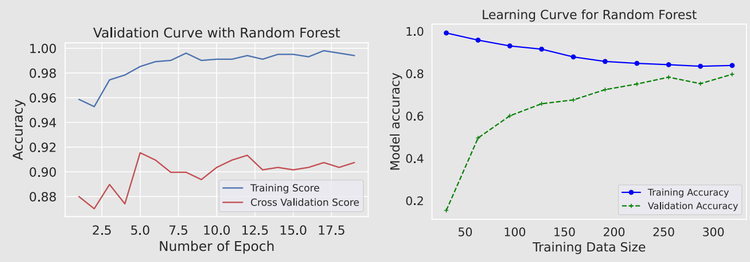
Confusion Matrix

Here is the confusion matrix of Random Forest and Decision Tree algorithms.



Validation and Learning Curve

It is seen that Random Forest maintains its performance as the number of epochs grows. As the number of epochs rises, the validation accuracy increases from 84% to 96%. It is clear from the learning curve that the cross-validation and training curves converge as the size of the training data increases. The cross-validation accuracy increases as more training data are added. Therefore, it is advantageous in this situation to add more training data.



Accuracy, Precision, Recall and F1-Scores

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC Score(%)
Random Forest	98.62	96.38	99.92	98.09	98.02
Decision Tree	92.54	96.53	98.58	97.58	97.58
K-Nearest Neighbor	83.25	80.06	82.13	81.08	93.1
Gradient Boosting	94.7	93.79	93.01	93.38	96.01
XG Boosting	92.12	92.12	92.06	92.09	96.06
LGBM Boosting	94.88	95.21	93.34	94.77	97.24
Multilayer Perceptron	82.12	83.12	81.33	81.34	87.23

Conclusion

To predict this GvHD, this study made use of methods like imputation, normalization, data balance, and feature selection. Since this study is primarily concerned with a multi-class classification, a new multi-class target feature is produced. Subsequently, several models are benchmarked to find the best model to help the medical personnel estimate the risk of transplantation before ASCT and take the necessary precautions to lower that risk.