Predicting transplant success based on gene expression values in kidney tissue and peripheral blood lymphocytes

Mathew Golf
mathewggolf@lewisu.edu
DATA-51000-001, Summer 1
Data Mining and Analytics
Lewis University

I. Introduction

Transplants are an essential component of the healthcare industry especially as the average population age continues to grow in many countries around the world. Understanding the likelihood of rejection or acceptance and the degree of either is important in determining which patient may be best suited for transplant priority. Current intervention methods are based around the use of immunosuppressants to ensure the host immune system does not reject the transplanted organ. Additional intervention may be possible through targeted therapeutic administration that changes the gene expression landscape from all states (acute rejection, dysfunction, etc.) to the well-functioning transplant gene expression landscape, thus ensuring transplant viability and patient health.

Tissue samples were obtained from kidneys and peripheral blood lymphocytes (PBL) in transplant patients[1]. DNA microarrays were then used to analyze gene expression in these samples indicating the stage of transplant related to the gene expression profile[1]. Unfortunately the dataset is quite small as there are only 62 patients sampled for predicting 8 class labels, however there are 9470 features for gene expression values[1]. Nevertheless, prediction of transplant status using these values would also help for monitoring the immune status and the degree of immunosuppression required to ensure transplant success[1].

In section II, we provide an overview of the data describing the features used for predicting the classes associated with transplant success in PBL and kidney tissues. Then in section III, the methodology for obtaining, processing, predicting, and evaluating the prediction is explained. Section IV describes the results from the aforementioned analysis and discusses the relevance of said results. Lastly, in section V conclusions from the results and the discussion of the analysis is provided indicating future directions for further investigation.

II. Data Description

For this analysis one dataset was used and divided into training (90%) and test sets (10%) for prediction of transplant status based on gene expression values. Division of the dataset was required in order to train prediction accuracy in comparing multiple different methods, using 9470 gene expression values for 62 patient samples (Table I). All genes were used in the analysis to ensure best prediction accuracy based on the low total sample number. Differential gene expression profiles of transplant status are clearly apparent in the data as displayed in Figure 1.

TABLE I. Kidney and PBL Gene Expression Values for Transplant PAtients

Attribute	Type	Example Value	Description
CLASS	Nominal (string)	Kidney | Acute Rejection	Transplant status for kidney and PBL samples
1949_at	Numeric (real)	97.400	Gene expression value of 1949_at
⋮	⋮	⋮	⋮
ZZZ3 26009	Numeric (real)	133.200	Gene expression value of ZZZ3 26009

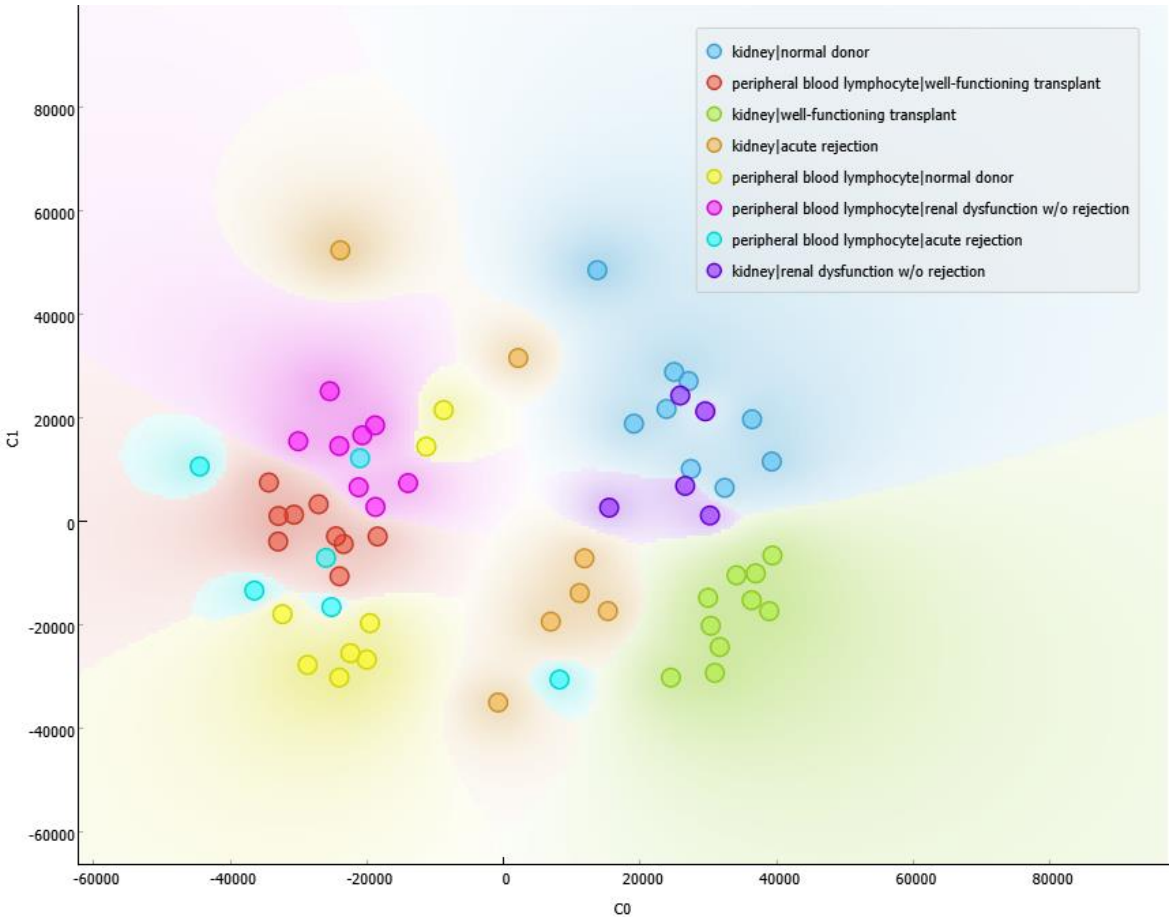

Fig. 1. Manifold learning transformation of all gene expression values for kidney and PBL samples, indicating differential expression for transplant status.

TABLE II. DESCRIPTIVE STATISTICS OF THE MANIFOLD LEARNING TRANSFORMED KIDNEY AND PBL GENE EXPRESSION DATA

Attribute	Mean	Median	Range
C0	-6.45448e-13	-4679.19	-44367.9 to 39454.4
C1	1.46693e-12	1068.42	-35042.3 to 52380.8

III. METHODOLOGY

Data was obtained through the in-built GEO dataset feature of Orange using their bioinformatics add-on package, and the data was provided by Flechner *et al.*[1]. The data was visualized using manifold learning to reduce the dimensionality of the data and to visualize the inherent transplant status subgroups within, while also preserving any non-linear relationships (Fig. 1). Splitting the data into training and testing sets was done by randomly sampling 90% of the data for the training set and 10% of the data for the testing set. Next, we ran the training data through multiple different techniques such as: tree, random forest, gradient boosting, kNN, and adaBoost. Since we used many techniques, we had to adjust each methods parameters to best fit the data and generate the highest F1 score. Afterwards we fed the results into the test and score function for evaluation, and we also examined the resultant confusion matrix and ROC analysis. Once complete and after the techniques were optimized from the training data, we then fed the test data into the test and score function in order to evaluate the prediction accuracy generated from the model learned through the training set.

IV. RESULTS AND DISCUSSION

Predicting transplant success based on gene expression values of kidney tissue and PBL's was done using multiple different models. Since multiple models were used for analysis, we will review all models used and suggest the best fit models for practical use. In order to do so, the models learned by training on 90% of the data (55 samples) and were subsequently tested on 10% of the data (7 samples). Unfortunately, we were limited on the size of the dataset as it contained a total of 62 patient samples. Likewise,

we were predicting class labels for 7 different instances regarding kidney samples and PBL samples relating to the rejection or acceptance of transplant. Therefore, as we are predicting 7 classes from 62 samples our dataset for learning is limited, small, and therefore imbalanced. Since our dataset is imbalanced, we will focus more on the F1 score than the accuracy or AUC score to properly reflect the performance of the models on imbalanced data. Unsurprisingly, for the training data the random forest and gradient boosting models performed the best, as indicated by their high F1 scores relative to the other methods, 0.675 and 0.673 respectively (Table III). Furthermore, their AUC values were also high despite the imbalance of entries per class in the data from the given dataset (Table III). Both these methods are ensemble tree methods, therefore it makes sense they performed so well as they take the average of many different model iterations. For comparison, the tree method is a single model iteration and had lower values in all measures as compared to the ensemble tree models of random forest and gradient boosting (Table III). Interestingly, the AdaBoost method did not perform as well as the other ensemble methods and in fact the kNN method performed better as indicated by the higher model statistics in Table III. kNN performed well, which makes sense as the exploration of the data in Fig. 1 indicates there is some underlying cluster structure within the data. Furthermore, confusion matrix data of all models generated from the training data is available and listed in the appendix (Tables V to IX).

TABLE III. MODEL STATISTICS FOR TRAINING DATA OF KIDNEY AND PBL GENE EXPRESSION VALUES

Model	AUC	Accuracy	F1	Precision	Recall
kNN	0.926	0.709	0.636	0.594	0.709
Tree	0.751	0.564	0.554	0.564	0.564
Random Forest	0.895	0.673	0.675	0.688	0.673
Gradient Boosting	0.926	0.709	0.673	0.672	0.709
AdaBoost	0.779	0.618	0.592	0.631	0.618

After learning using the training data, models were tested using the remaining 10% of the randomly sampled data. Since the data available for analysis was limited in size, this 10% accounted for only 7 samples. As sampling of the data into training and testing sets was random, we cannot be sure that the 7 samples selected are evenly distributed throughout the 7 different classes. Even if the remaining data was evenly distributed throughout the classes 1 entry per class is not sufficient to truly test the model performance, especially given the practical impact of changing the transplant patients genetic landscape to ensure transplant acceptance. Nevertheless, model performance on testing data was still evaluated and similar results are seen in Table III and IV. kNN performed the best, classifying all samples correctly and therefore scoring 1.000 in all statistics (Table IV). Similarly, random forest and gradient boosting also performed well only misclassifying one sample in each model (Table IV). All confusion matrices resultant from the testing data are available in the appendix (Table X to XIV). True comparison and judgement of model performance is not possible given the small set of data. Future directions should be focused at increasing the number of patients sampled to construct more robust models and ensure predictions are accurate and not just a byproduct of imbalanced data. Likewise, clinical application of these techniques may be the difference in changing a transplant gene expression landscape from rejection to acceptance based on therapeutic intervention targeted at changing gene expression or modulating the given immunosuppressants.

TABLE IV. MODEL STATISTICS FOR TESTING DATA OF KIDNEY AND PBL GENE EXPRESSION VALUES

Model	AUC	Accuracy	F1	Precision	Recall
kNN	1.000	1.000	1.000	1.000	1.000
Tree	0.838	0.714	0.667	0.643	0.714
Random Forest	0.975	0.857	0.857	0.857	0.857
Gradient Boosting	1.000	0.857	0.810	0.786	0.857
AdaBoost	0.850	0.714	0.714	0.714	0.714

V. CONCLUSIONS

Understanding the gene expression profile in kidney tissue and PBL's of transplant patients is important for modulating immunosuppressants given to patients to ensure viability of the transplant[1]. Similarly, using gene expression from the samples to predict acceptance or rejection of the transplant and the degree of which, i.e. well-functioning transplant or acute rejection may be used in the future for targeted therapeutic intervention to change the gene expression profile from acute rejection to well-functioning transplant in transplant patients[1]. Therefore, we sought to prove transplant success prediction possible based on the gene expression from the given samples. In doing so, we used data from 9470 genes for 62 patient samples in order to train our model for prediction of transplant status. Despite the low sample numbers prediction was still highly successful, indicating that

there are distinct expression profiles which can predict the status and the success of transplants. Reasoning from this conclusion it is apparent that analyzing expression profiles before and after transplant may help for ensuring success by modulating gene expression levels through drug interventions and administering the correct amount of immunosuppressants. Multiple models were tested with the best performing models – as indicated by the F1 score and AUC values for both testing and training data, were the kNN, random forest, and gradient boosting models (Table III & IV). Unsurprisingly, ensemble methods were represented in 2 of the 3 best models (random forest and gradient boosting), performing the best for the training data and the second and third best in the test data (Table III & IV). As well, it is important to note that both the random forest and gradient boosting models are ensemble tree models.

Future directions should be based around gathering more patient data to ensure that the predictions are true and not a byproduct of a small imbalanced dataset. As well, increasing the number of patients may increase the variance of the data or it may further delineate the underlying cluster structures apparent in the data between different class labels (Fig. 1). More research may be needed to understand the why of the differential expression profiles relating to acceptance and rejection of transplant. Thus, investigating the significantly different expression levels between patients may result in finding key genes as targets for intervention to correct rejection to acceptance. Studies that focus on the how of changing the expression profiles may suggest that no prior knowledge of function is required if the problem turns out to be as simple as tweaking a combination of gene levels within the transplant's local environment. Combining immunosuppressants and gene expression modulators in tandem may not only result in a significant increase in transplant acceptance likelihood, but they may also result in the optimization of transplant from good to great to best. With the increasing use of animal hybrids for transplant in some regions globally and the advancement of tissue regeneration research we may soon be at the stage of growing organs in vats or printing them derived from patient stem cells, thus mitigating the requirement of immunosuppressants and increasing the likelihood of transplant success. Until that time however, it is important to understand the differential gene expression profiles between transplant samples. Focusing on the ability to predict the transplant status based on gene expression, and subsequently how to change the expression of patients transplants to ensure acceptance and survival of both the patient and the transplanted organ.

REFERENCES

[1] S. M. Flechner *et al.*, "Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes," *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, vol. 4, no. 9, pp. 1475–1489,Sep.2004,doi:10.1111/j.1600-6143.2004.00526.x.

APPENDIX

TABLE V. KNN CONFUSION MATRIX FOR TRAINING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

kNN	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	8	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	7	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	9	0	0	0	0	0
Acute rejection (K)	1	0	0	4	1	0	0	0
Normal donor (PBL)	0	1	0	0	5	1	0	0
Renal dysfunction w/o rejection (PBL)	0	1	0	0	0	6	0	0
Acute rejection (PBL)	0	4	0	1	1	0	0	0
Renal dysfunction w/o rejection (K)	4	0	0	1	0	0	0	0

TABLE VI. TREE CONFUSION MATRIX FOR TRAINING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

Tree	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)

Tree	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	4	0	1	0	0	0	0	3
Well-functioning transplant (PBL)	0	7	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	7	2	0	0	0	0
Acute rejection (K)	0	1	0	4	0	0	0	1
Normal donor (PBL)	0	0	0	0	3	1	3	0
Renal dysfunction w/o rejection (PBL)	0	1	0	0	1	3	2	0
Acute rejection (PBL)	0	1	0	0	1	1	3	0
Renal dysfunction w/o rejection (K)	3	0	0	2	0	0	0	0

TABLE VII. RANDOM FOREST CONFUSION MATRIX FOR TRAINING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

Random Forest	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	5	0	0	0	1	0	0	2
Well-functioning transplant (PBL)	0	4	0	0	1	0	2	0
Well-functioning transplant (K)	0	0	9	0	0	0	0	0
Acute rejection (K)	0	1	0	4	1	0	0	0
Normal donor (PBL)	0	1	0	0	5	1	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	1	5	1	0
Acute rejection (PBL)	0	0	0	0	2	1	3	0
Renal dysfunction w/o rejection (K)	2	0	0	1	0	0	0	2

TABLE VIII. GRADIENT BOOSTING CONFUSION MATRIX FOR TRAINING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

Gradient Boosting	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	7	0	0	0	0	0	0	1
Well-functioning transplant (PBL)	0	7	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	9	0	0	0	0	0
Acute rejection (K)	0	0	1	2	1	0	0	2
Normal donor (PBL)	0	0	0	0	6	1	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	0	6	1	0
Acute rejection (PBL)	0	1	1	0	1	1	2	0
Renal dysfunction w/o rejection (K)	2	1	0	1	0	1	0	0

TABLE IX. ADABOOST CONFUSION MATRIX FOR TRAINING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

AdaBoost	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	8	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	6	0	0	0	0	1	0
Well-functioning transplant (K)	0	0	8	0	1	0	0	0
Acute rejection (K)	0	1	1	3	1	0	0	0
Normal donor (PBL)	0	1	0	0	4	1	1	0
Renal dysfunction w/o rejection (PBL)	0	1	0	0	0	4	2	0
Acute rejection (PBL)	0	3	1	1	0	1	0	0
Renal dysfunction w/o rejection (K)	1	0	0	3	0	0	0	1

TABLE X. KNN CONFUSION MATRIX FOR TESTING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

kNN	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	1	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	2	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	1	0	0	0	0	0
Acute rejection (K)	0	0	0	1	0	0	0	0
Normal donor (PBL)	0	0	0	0	1	0	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	0	1	0	0
Acute rejection (PBL)	0	0	0	0	0	0	0	0
Renal dysfunction w/o rejection (K)	0	0	0	0	0	0	0	0

TABLE XI. TREE CONFUSION MATRIX FOR TESTING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

Tree	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	1	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	2	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	0	1	0	0	0	0
Acute rejection (K)	0	0	0	1	0	0	0	0
Normal donor (PBL)	0	0	0	0	1	0	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	0	0	1	0
Acute rejection (PBL)	0	0	0	0	0	0	0	0
Renal dysfunction w/o rejection (K)	0	0	0	0	0	0	0	0

TABLE XII. RANDOM FOREST CONFUSION MATRIX FOR TESTING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

Random Forest	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	1	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	2	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	1	0	0	0	0	0
Acute rejection (K)	0	0	0	0	0	0	0	1
Normal donor (PBL)	0	0	0	0	1	0	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	0	1	0	0
Acute rejection (PBL)	0	0	0	0	0	0	0	0
Renal dysfunction w/o rejection (K)	0	0	0	0	0	0	0	0

TABLE XIII. GRADIENT BOOSTING CONFUSION MATRIX FOR TESTING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

Gradient Boosting	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	1	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	2	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	1	0	0	0	0	0
Acute rejection (K)	0	0	1	0	0	0	0	0
Normal donor (PBL)	0	0	0	0	1	0	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	0	1	0	0
Acute rejection (PBL)	0	0	0	0	0	0	0	0
Renal dysfunction w/o rejection (K)	0	0	0	0	0	0	0	0

TABLE XIV. ADABOOST CONFUSION MATRIX FOR TESTING DATA FROM KIDNEY AND PBL GENE EXPRESSION VALUES

AdaBoost	Normal donor (K)	Well-functioning transplant (PBL)	Well-functioning transplant (K)	Acute rejection (K)	Normal donor (PBL)	Renal dysfunction w/o rejection (PBL)	Acute rejection (PBL)	Renal dysfunction w/o rejection (K)
Normal donor (K)	1	0	0	0	0	0	0	0
Well-functioning transplant (PBL)	0	2	0	0	0	0	0	0
Well-functioning transplant (K)	0	0	0	0	0	0	0	1
Acute rejection (K)	0	0	0	1	0	0	0	0
Normal donor (PBL)	0	0	0	0	1	0	0	0
Renal dysfunction w/o rejection (PBL)	0	0	0	0	0	0	1	0
Acute rejection (PBL)	0	0	0	0	0	0	0	0
Renal dysfunction w/o rejection (K)	0	0	0	0	0	0	0	0