Development of an effective classification model for Small B-Cell Lymphoid Neoplasm

Blessy Rajan and Shubhangi Kaushik

Abstract

GOAL : Develop a machine learning-based diagnostic model for accurate classification of the subclass of small B-cell lymphoid neoplasms, improving the performance compared to the previously built model.

RESULTS : Differently expressed genes were obtained from the GEO data, which were then used to train and test the machine learning algorithm used, i.e. gradient boosting. Important features were obtained and compared to the important feature obtained from the paper.

Learned to use GEOquery package in R in order to access and analyze the gene expression data used in the analysis.

Goal

The goal of this project is to develop an effective diagnostic method for the classification of small B-cell lymphoid neoplasms (SBCLNs). Highly and differently expressed genes were discovered by the analysis of gene expression matrices from 8 data sets containing data related to SBCLN. Using the filtered genes, a classifier was developed using Gradient Boosting algorithm[3] for the data set, and the most significant genes were found. Zhang et al.[6] conducted a study where the candidate genes or the most significant genes found by them using the same data set were employed to refine the classification model. To evaluate the performance of our model, we will assess the overlap of significant genes found in both the analysis and the comparison of other scoring values.

Data

The gene expression profiles (GEP) of 718 samples, analyzed using Affymetrix U133 plus 2.0 microarrays, were obtained from 8 Gene Expression Omnibus (GEO) data sets. It is a subset of the 1039 samples from 27 GEO data sets used by Zhang's study[6].

The study included two main groups of samples: 645 samples from individuals with SBCLNs and 73 samples from individuals without malignant conditions, i.e. the control samples. Within the SBCLN group, there are 6 main subgroups, chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL), conventional mantle cell lymphoma(MCL), follicular lymphoma (FL), leukemic non-nodal mantle cell lymphoma (MCL), marginal zone lymphoma (MZL), lymphoplasmacytic lymphoma/Waldenstr¨om's macroglobulinemia (LPL/WM), and other undetermined.

The grouping within the 8 GEO datasets is given in Table 1.

GEO Accession	Subtype 718	CLL/SLL 259	FL 165	MCL 76	Other 42	LPL/WM 41	MZL 57	Control 73
50006	220	41/147	0	0	0	0	0	32
79196	189	54	12	54	42	4	23	0
53820	81	0	81	0	0	0	0	0
55267	69	0	63	0	0	0	0	6
16455	55	17	7	22	5	0	4	0
27928	35	0	0	0	0	0	0	35
39577	32	0	2	0	0	0	30	0
9656	37	0	0	0	0	37	0	0

Table 1: Distribution of the data in the investigated GEO datasets.

In Zhang's study[6], 154 candidate genes were first identified using the 27 GEO data sets. The training cohort consisted of 159 cases enrolled retrospectively, while the validation cohort included 197 SBCLN cases and was obtained by Nanostring profiling for those particular 154 genes, along with 13 other housekeeping genes. The classifier was then trained using 57 SBCLN and 102 nonmalignant control cases, followed by validation using an independent cohort.

Data Preprocessing

The gene expression matrices for the 8 GEO files were retrieved using GEOquery[2] library in R. The matrices were then merged on the same gene ids and a merged dataset with 54675 genes for the 718 samples was retrieved. The dataset was subjected to gene expression analysis using the DESeq2[5] method.

DESeq2, an R library, is used to identify differentially expressed genes across conditions in this project. Initially, a matrix was created with integer values from decimal values, and sample names were stored as metadata. However, duplication issues between the matrix and metadata were encountered. After resolving the duplications, DESeq2 identified around 26,000 differentially expressed genes. This dataset was then stored and utilized for subsequent analysis in Python.

The labels were assigned to each sample based on the description of the data for each GSE file. There were a total of 7 classes; 0 was label for class Normal,1 was used for CLL/SLL label 2 for FL, 3 for MZL, 4 for Other, 5 for LPL/WM, 6 for MCL.

The data set was divided into X and Y sets, followed up by the scaling of X data set. After this the data was split into test and train sets.

To prevent the imbalance in the number of labels for each class, Synthetic Minority Over-sampling Technique or SMOTE technique[1] was used to balance the data set which increased the number of samples from 718 to 1444. However, the SMOTE was used after splitting the dataset into test and train sets, in order to prevent bias in the data sets.

Methods

The analysis started with finding the optimized hyper-parameters for the chosen machine learning algorithm, Gradient Boosting[3], for the pre-processed data set. Based on the hyperparameter tuning test, the model worked the best for 500 trees having AUC score of 0.95, and thus, 500 was used as the number of trees when training the actual model.

0.1 Gradient Boosting

Gradient Boosting[3] trains weaker algorithms like Decision Trees sequentially, such that every later model corrects the mistakes made by the former model. The optimization in Gradient Boosting is performed in the function space, where the function estimate $\hat{f}_i(x)$ is parameterized in an additive functional form:

$$\hat{f}(x) = \hat{f}_M(x) = \sum_{i=0}^{M} \hat{f}_i(x)$$

where M is the number of iterations, $\hat{f}_i(0)$ is the initial guess and $\{f_i\}_{i=1}^{M}$ are the function increments or “boosts.”

0.2 Evaluation

The machine learning model was evaluated using the prediction probabilities calculated using K-Fold cross-validation technique where the number of folds was set to 10. Based on the prediction probabilities, the class was predicted for both train and test datasets. These predictions were evaluated using the following:

0.2.1 Confusion Matrix

A confusion matrix is a table that summarizes the performance of a classification model by giving the counts of true positive, true negative, false positive, and false negative predictions (Table 2).

Predicted/Actual	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Table 2: Confusion Matrix

Since the model was a multi-label model, Macro-average values for the sensitivity and specificity were calculated using the confusion matrix.

0.2.2 ROC Curve

(Receiver Operating Characteristic) Curve[4] for each class was also generated using the probabilities. It plots the true positive rate against the false positive rate at different classification thresholds.

Libraries in R like GEOquery was used to get the Gene Expression data, DESeq was used to do the gene expression analysis and hgu133plus2.db and AnnotationDbi were used to annotate the gene IDs.

Results and Discussion

A Gradient Boosting classifier was employed to construct a model using the data collected from the 8 GEO data sets(Table 1), which was subsequently preprocessed.

Sensitivity, Specificity and Accuracy(Table 3) were calculated for the model. The model gave an accuracy of 99% for the training set, whereas the accuracy was 88% for the testing data set.

Model	Accuracy	Sensitivity	Specificity
Gradient Boosting(Train)	99%	99%	99%
Gradient Boosting(Test)	87.5%	77.4%	80.4%
Nested Random Forest	95%	95%	95%

Table 3: Accuracy, Sensitivity, and Specificity of the Gradient Boosting model and also for the Nested Random Forest model used by Zhang et al. The scores for the Gradient Boosting model are for both the training and the test set.

Later the evaluation was conducted for the models using ROC curve. ROC curve for each of the classes were built for both training and testing sets as mentioned in Figure 1 and Figure 2. Based on Figure 1 and Figure 2, the model performed well in classifying the classes from the training and test sets.

The important features were predicted from the data set using the model developed, a subset of these features is mentioned in the Table 4.

Feature Names	Importance
238189 at	0.023621994
233259 at	0.018269168
202931 x at	0.01688599
222099 s at	0.012913862
49327 at	0.012466466
214446 at	0.011775034
227414 at	0.010471852
227199 at	0.010271526
219461 at	0.009676438

Table 4: A subset of the top 500 genes obtained from the dataset used in Zhang et al paper using the gradient boosting method. Along with their importance score.

For the features mentioned in Table 5, the gene symbols were annotated using R's libraries and compared with the important genes obtained from the previous study done by Zhang et al.[6]. During the comparison, from the total of the top 500 genes obtained from this study, only 11 genes were found to be common.

For the important features obtained which were also present in the Zhang et al[6] study, biological significance is noted down in Table 6.

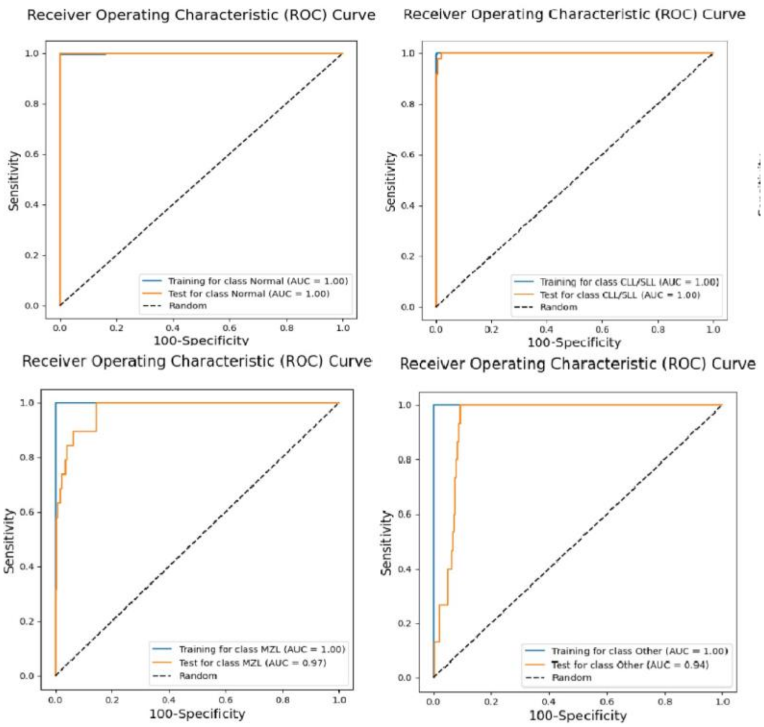

Figure 1: The figure describes the ROC curve of ’Normal’(top-left), ’CLL/SLL’(top-right), ’MZL’(bottom-left), and ’Other’(bottom-right) classes. The curve in blue depicts the training dataset, whereas the curve in orange is for the testing dataset. For the training dataset, the model seems to have done a very good job in classifying the classes ’Normal’, ’CLL/SLL’, ’MZL’, and ’Other’, with the AUC score of 100%. When the testing dataset was evaluated in the model, for the class ’Normal’ and ’CLL/SLL’ the model performed extremely well with AUC 100%. For the ’MZL’ class, the AUC was 97%, and for ’Other’ the AUC was 94%

Feature Names	Gene Names
ELL2	214446 _at
CNR1	208243 s_at
CNR1	213436 _at
NEB	215368 _at
DUSP4	204014 _at
ARHGAP44	205414 s_at
TCF4	212386 _at
BHLHE41	223185 s_at
BASP1	202391 _at
EBF1	232204 _at
TGFBR3	226625 at

Table 5: The overlapping genes between the candidate genes identified in Zhang’s study and the top 500 features obtained from the Gradient Boosting classifier were extracted.

Initially, our plan was to find common genes and build a model based on their dataset. Targeting a larger number of common genes was tried, however, it was discovered that the number of common genes that were found was 11 which was insufficient.

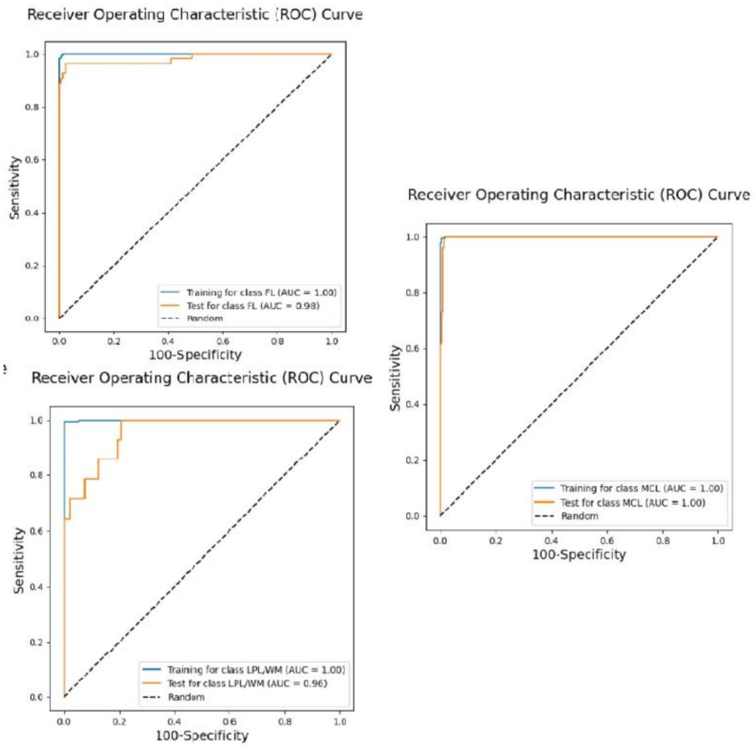

Figure 2: The figure describes the Roc curve of ’FL’(top-left), ’LPL/WM’(bottom-left) and ’MCL’(right) classes. The curve in blue depicts the training dataset, whereas the curve in orange is for the testing dataset. For the training dataset, the model seems to have done a very good job in classifying the classes ’FL’, ’LPL/WM’ and ’MCL’, with AUC score of 100%. When the testing dataset was evaluated for the model, for the class ’MCL’, the model performed extremely well with an AUC score of 100%. For the ’FL’ class, the AUC was 98%, and for ’LPL/WM’, the AUC was at 96%

This could have led to potential issues such as overfitting and biased predictions. Therefore, it was decided to not proceedwith the analysis.

Nonetheless, a classification model was successfully developed using the initial dataset. When comparing the model’s performance with Zhang’s paper, it was found that their nested random forest approach outperformed our regular gradientboosting model(Table 3). In the future, we could explore the nested gradient boosting method as a potential improvement to our results. Other than that, the way the NanoString profiling was conducted to accurately measure the expression levels of the candidate genes in Zhang’s analysis, the top genes found by our analysis could be studied further using similar profiling methods.

Overall, preprocessing step using DESeq2 to obtain the differently expressed genes played an important role in filtering the genes. The model did well while classifying classes like ’Normal’, ’CLL/SLL’, and ’MCL’ with auc score of 100%. For classes like ’FL’, ’LPL/WM’, ’MZL’, and ’Other’ the model gave auc score above 90%. Hence, it could be said that the model built in this project worked decently for the multi-class classification. With a little bit more improvement in the model, this model could benefit the medical community for tasks like multi-class classification for different diseases.

Gene Symbol	Gene ID	Biological Significance
ELL2	214446 at	Found to be downregulated in chronic lymphocytic leukemia (CLL) cancer subtype.
CNR1	208243 s at	Found to influence immune responses, apoptosis, and cell proliferation in various lymphoma subtypes.
CNR1	213436 at	Found to influence immune responses, apoptosis, and cell proliferation in various lymphoma subtypes.
NEB	215368 at	Found to be mutated in the case of LPL/WM subtypes. However further studies are required to understand its precise role in cancer.
DUSP4	204014 at	Found to be mutated, hence influencing the growth and survival of lymphoma cells.
ARHGAP44	205414 s at	Not much is known about its influence in lymphoma cell cancer.
TCF4	212386 at	Found to be mutated and influencing abnormal B-cell development and lymphomagenesis
BHLHE41	223185 s at	Maintains proper cellular differentiation and homeostasis in lymphoid cells, any mutation in these disrupts lymphoid cells stability.
BASP1	202391 at	May contribute to cell signaling pathways and cellular processes involved in lymphoma development.
EBF1	232204 at	Mutation in these disrupt normal B-cell differentiation and contribute to lymphoma pathogenesis
TGFBR	226625 at	Mutation in these contribute to the development and progression of small B-cell lymphoid neoplasms

Table 6: This table lists the biological significance of the common genes found, in small B-cell lymphoid neoplasms

Appendix

References

1. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
2. Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.
3. Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
4. Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.
5. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.
6. Wei Zhang, Qilin Ao, Yuqi Guan, Zhoujie Zhu, Dong Kuang, Monica MQ Li, Kefeng Shen, Meilan Zhang, Jiachen Wang, Li Yang, et al. A novel diagnostic approach for the classification of small b-cell lymphoid neoplasms based on the nanostring platform. *Modern Pathology*, 35(5):632–639, 2022.