Gene Expression Predictive Signatures for Acute Myeloid Leukemia Using Machine Learning

Neel Jay

Montgomery Blair High School

Under the Guidance of:

Dr. Konstantinos Karagiannis, PhD

Mr. Kamil Kural, MSc

Food and Drug Administration

Abstract

In the past decade, transcriptomics data has been used as a primary method for evaluating the diagnosis and prognosis of various cancers in the past. These data allow for the use of high-dimensional machine learning methods, which further improve our understanding of these diseases. Machine learning has proven in various studies to be successful in identifying acute myeloid leukemia (AML). However, differences in representation between AML subclasses are often disregarded. As not all AML subclasses affect the population at the same rates, these biased data reduce the applicability of these machine learning models in a practical setting. In this study, AML classes were stratified based on French-American-British subclass. Samples from each subclass were compared to control cases, and various machine learning models were used to select important features. Pathway analysis was then performed on selected features, and differentially expressed genes from the top pathways were evaluated. These genes were evaluated for predictive performance and used in a signature search to find chemical therapeutic agents. We showed that genes selected through feature selection on individual subclasses resulted in better clustering after unsupervised learning with autoencoders. We also were able to identify potential novel chemical therapeutic agents for AML. The efficacy of some identified agents has already been discussed in previous studies. Our research demonstrates the applications of machine learning in improving AML diagnosis and therapy development.

Introduction

Acute myeloid leukemia (AML) is a type of blood cancer that commonly starts in myeloid stem cells or myeloid blasts. These are precursors to mature blood cells, such as neutrophils and monocytes. AML is usually characterized by the growth of cancer cells in the bone marrow and in the blood. The progression of AML is usually very fast, and can be fatal if left untreated for a few weeks or months. Thus, it is important to be able to identify AML as soon as possible so a patient can undergo therapy early on. Regarding AML diagnosis, the use of transcriptomics data has been explored extensively in the past. More recently, machine learning models are being implemented, as they complement the high-dimensional nature of these data [1] by finding patterns that alternative methods may not be able to. Machine learning is also able to adapt to different types of data, such as microarray or RNA-seq, and has been demonstrated to have high accuracy. High performance can be seen with not only distinguishing between AML and healthy cells, but also AML and other types of leukemia.

However, many transcriptomic analyses of AML tend to disregard the type of AML when comparing cancer cells to healthy cells [2]. This is a major problem, as not all types of AML affect the general population at equal rates. Data that is biased towards specific classes can be especially harmful when training machine learning models. Without examining AML subtypes, results of these studies could be skewed due to differences in representation between subtypes. Accuracy of these models in practical settings may be significantly lower than tested accuracies.

In this paper, we choose to stratify AML samples based on French-American-British (FAB) classifications. The FAB system was initially proposed in 1976, and classes are divided based on when AML forms in the blood cell maturation process (Figure 1). We will only consider subtypes M0-M5. Since these cancers all start in immature forms of monocytes, they are the most difficult to classify.

Table 1: French-American-British classifications of AML [3]

M0	Undifferentiated acute myeloblastic leukemia
M1	Acute myeloblastic leukemia with minimal maturation
M2	Acute myeloblastic leukemia with maturation
M3	Acute promyelocytic leukemia
M4	Acute myelomonocytic leukemia
M5	Acute monocytic leukemia
M6	Acute erythroid leukemia
M7	Acute megakaryoblastic leukemia

High-dimensional microarray data of AML and healthy bone marrow cells, divided based on FAB classification, was used in this paper. Machine learning models were trained on this data to identify significant genes for each subtype. These genes were then evaluated for predictive performance, and they were also used to aid the identification of novel chemical therapies. The results illustrate how machine learning can be used to improve AML subtype classification and develop therapeutic strategies. They also highlight important limitations of machine learning and transcriptomics that must be addressed in the future.

Materials and Methods

A variety of machine learning models were utilized to identify predictive gene signatures among AML subclasses. Public microarray data was taken from the Gene Expression Omnibus (GEO), maintained by the NCBI.

Data Source

Microarray data was obtained from a published dataset GSE147515 [2], uploaded to the GEO database. The dataset contained 1721 samples of bone marrow cells from both AML patients and healthy donors taken during 28 different studies. All samples had passed a quality control and had undergone Robust Multichip Average (RMA) normalization, which reduces technical variation between arrays used for each sample. Sources of variation added to samples between the different studies, known as the batch effect, were also accounted for using the ComBat algorithm.

All samples in the dataset were filtered down to samples that had readily available FAB classifications. After filtering, 464 samples across all subclasses and control cases were identified.

The microarray platform used was Affymetrix U133 plus 2.0. Originally, the data contained over 40,000 different probe IDs. These IDs were matched to gene symbols using the BioMart databases in R, and duplicates were removed.

Feature Selection

Individual comparisons were made between each AML subclass and the control cases. A variety of feature selection methods were employed to improve accuracy: wrapper, filter, and embedded methods were all used. Wrapping methods evaluate the importance of each feature, and iterate until they have reached a desired subset. Recursive feature selection with Logistic Regression was used as a wrapper method. Filter methods rank features based on statistical measures. The chi-squared test was used as a filter method. Embedded methods only pick important features during model creation. Random Forest, Light Gradient Boosting (LGBM), and LASSO regression were used as embedded methods. Implementations of all models were taken from the *sci-kit learn* library in Python. While training models, important hyperparameters were tuned and samples were weighted in order to counter biases regarding unbalanced data. Ultimately, selected genes were chosen using a threshold of at least three of five models that deemed them as important.

Evaluation

Accuracy and precision were used to compare Random Forest and LGBM performance with all features versus selected features. Genes that were selected were also compared to the differentially expressed genes for each pair, in order to identify notable differences and evaluate variations in predictive performance.

Finally, the selected genes were tested for effects on unsupervised deep learning. Samples from all AML subtypes were used to train autoencoders. Autoencoders are a type of neural network that is trained on encoded versions of unlabeled data, which makes them ideal for unsupervised learning. Deep learning was done using Keras, an interface of Tensorflow, in Python. t-Distributed Stochastic Neighbor Embedding (TSNE) was used as a dimensionality reduction technique to graph classification results.

The quality of clusters graphed using TSNE was compared between models that used all features and models that used selective features. Accuracy and the ability for the model to discriminate between AML subtypes were taken into consideration.

Signature Analysis

Gene Ontology (GO) enrichment and pathway analysis was performed on selected genes in order to determine most affected biological processes and pathways for each subtype of AML. Prominent up- and down-regulated genes from each of these pathways were identified through our analysis. The differentially expressed gene signatures were then inputted into the L1000 characteristic direction signatures search (L1000CDS^2) engine. L1000CDS^2 searches through libraries of over a million gene expression profiles of chemically treated cells [4]. The results of this search gave us possible small molecules that may be used as therapeutic agents for AML subtypes.

Results

The accuracy and precision of the Random Forest and LGBM models were 100% for most subtypes, regardless of whether all genes were used as features or just selected genes were used. However, for subtype M2, one false negative case was corrected by the LGBM model when only using selected genes.

Overall, 60-70 genes were selected for each subtype. TSNE mappings of autoencoder results show a stark difference in cluster shapes after selection (Figure 1).

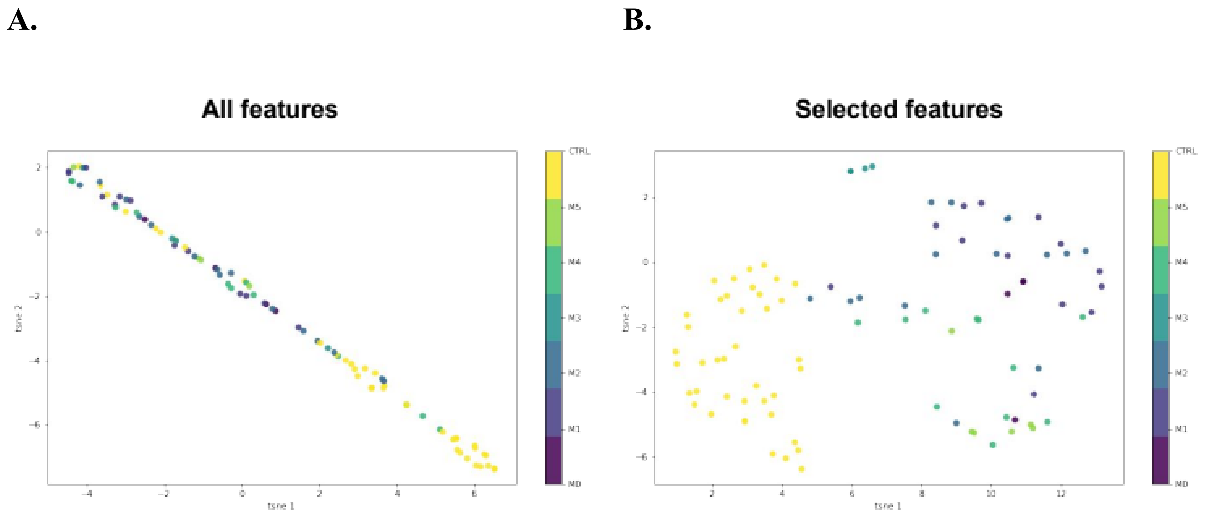

Figure 1. TSNE mapping of latent representations given by autoencoders, using all features (A) and just selected features (B).

Selected genes underwent GO enrichment and pathway analysis. Genes from top processes and pathways were taken and heatmaps were made using the significantly differentially expressed genes (Figure 2a-c). All subtypes show a clear distinction in selected gene expression between the control and AML samples. This distinction seems to be most pronounced in subtypes M0 and M3.

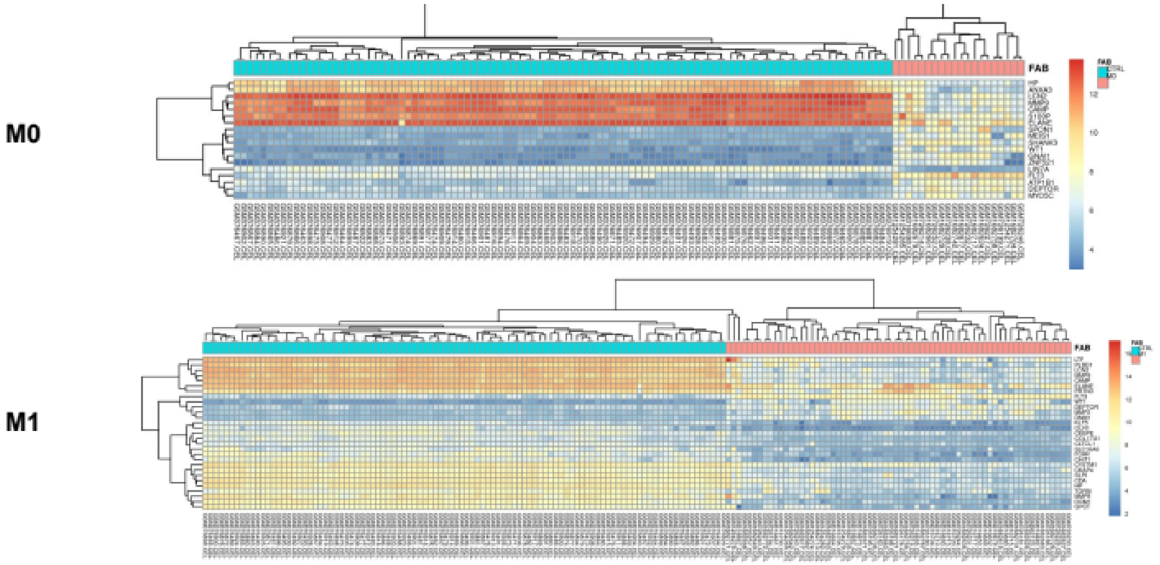

Figure 2a. Heatmaps of differentially expressed genes after pathway analysis, for subtypes M0 and M1.

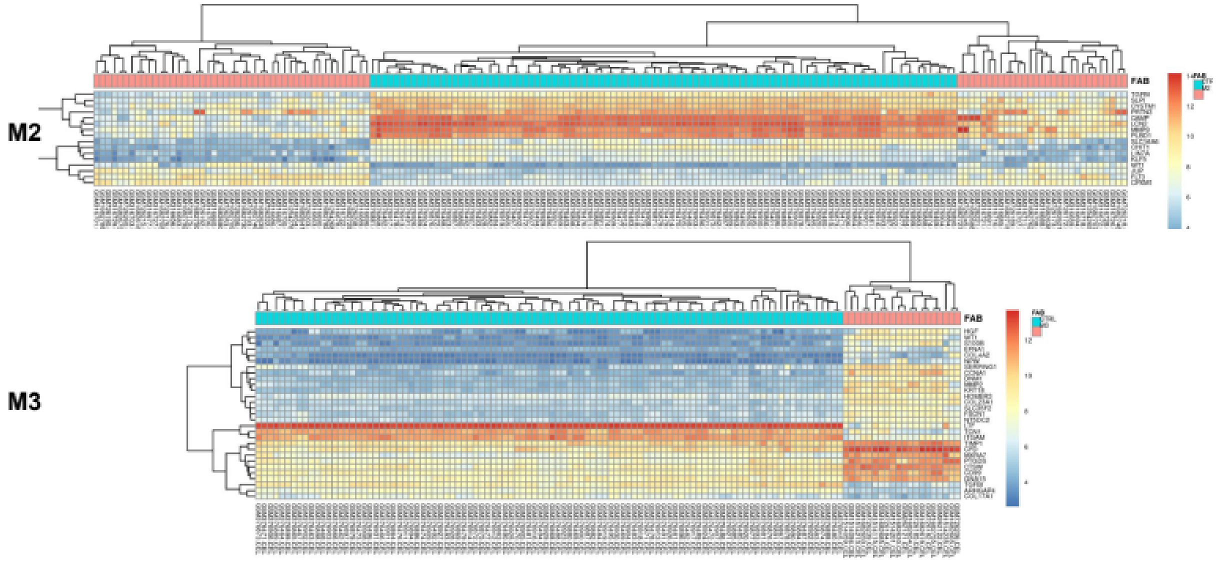

Figure 2b. Heatmaps of differentially expressed genes after pathway analysis, for subtypes M2 and M3.

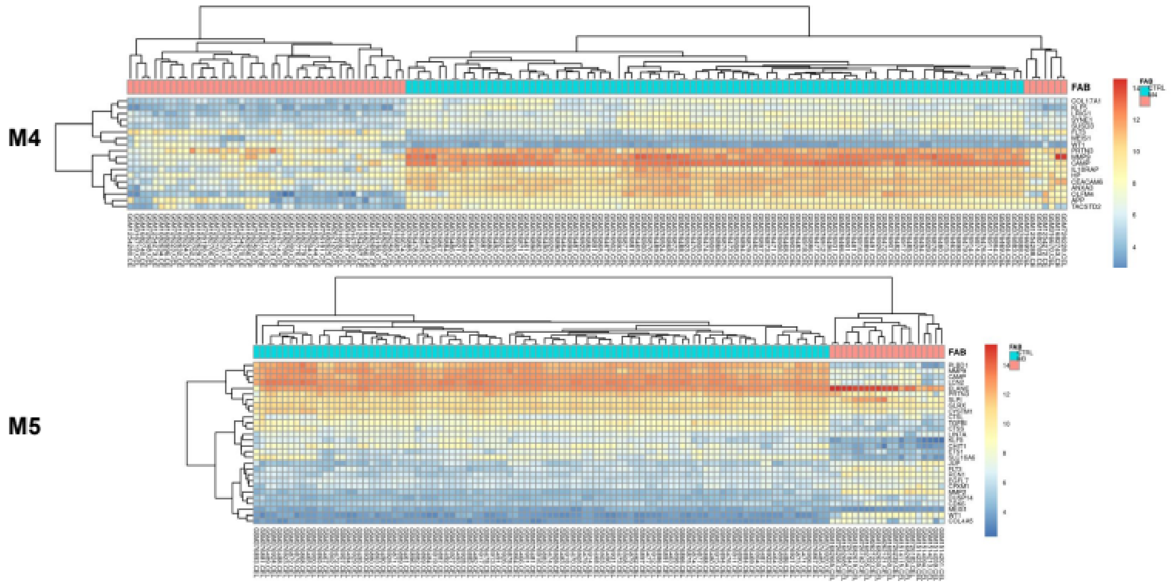

Figure 2c. Heatmaps of differentially expressed genes after pathway analysis, for subtypes M4 and M5.

These genes were inputted into the L1000CDS^2 engine. The database finds small molecules that have the highest chance of reversing the inputted gene signature. Top 3 results are shown in Table 2.

Table 2. Top 3 small molecules and drugs with the highest match to reverse gene signature. Search was done for all subtypes using L1000CDS^2.

	M0	**M1**	**M2**	**M3**	**M4**	**M5**
1	BRAZILIN	S1057	656402-250 MG	Parthenolide	S1040	BRD-K5885 3583
2	MLN4924	656402-250 MG	OXIBENDA ZOLE	DC-45-A2	OXIBENDA ZOLE	NCG001826 09-01
3	Cyclosporin A	S1003	Cyclosporin A	NCG001823 82-01	BRD-K8420 3638	S1003

Discussion

About 60 to 70 genes passed the threshold of three methods for each subtype. Among subtypes M0-M2, prominent selected genes included ORM1, WT1, and FLT3. ORM1 encodes acute plasma proteins, which are created during acute inflammation and may be involved in many aspects of immunosuppression. WT1 encodes proteins that play a role in cell growth, cell differentiation, and apoptosis. Finally, FLT3 encodes fmk-like tyrosine kinase, which is part of the large family of receptor tyrosine kinases (RTKs). All of these genes have a strong relationship to cancer development, and dysregulation of these genes would be a major cause of AML

Among subtypes M3-M5, prominent selected genes included ORM1, CAMP, and CEACAM6, HOMER3, and CAMP. CEACAM6 encodes a protein that is part of the carcinoembryonic antigen (CEA) family. CEAs are glycosyl phosphatidylinositol anchored cell surface glycoproteins, and they play a role in cell adhesion and tumor cell sensitivity to outside stressors. HOMER3 encodes a dendritic protein that regulates the metabotropic glutamate receptors. This protein is thought to be involved in cell growth. The prominence of this gene among the selected group for AML cases supports this. Finally, CAMP encodes a member of the antimicrobial peptide family, which has roles in antibacterial, antifungal, and antiviral activities.

The selected genes are similar between all subtypes for the most part, but there are noticeable differences between types M0-M2 and types M3-M6. Subtypes M0-M2 refer to AML where cancer is developed in myeloblasts [Table 1]. These are precursors in the process of WBC development. Subtypes M3-M6 are developed during the transition between myeloid cells and monocytes, so there will definitely be a difference compared to previous types in terms of relevant genes.

The predictive power of these selected genes can be seen in Figure 1. For unsupervised learning using autoencoded neural networks, inputting all 20 thousand features resulted in almost no discernable clusters after TSNE mapping [Figure 1A]. On the other hand, using selected features had much better clusters. There is very noticeable separation between types M0-M2 and types M3-M6 [Figure 1B], which further supports the differences in gene expression profiles between these two groups. In terms of supervised learning using Random Forest or LightGBM, there was very little difference in accuracy and precision between methods with all features versus selected features. Since accuracy was almost always 100%, this indicates overfitting of the models. The most likely cause of this is the limited number of available samples. Many samples that we found during research didn't have FAB classification details available, so there was limited availability of data. In addition, while unbalanced data representation was accounted for during training via sample weights, this may have also played a role in overfitting. Future research needs to explore a larger number of labeled samples, to more accurately see how genes selected through machine learning can be used.

The top pathways determined through pathway analysis of selected genes were relatively similar across all subtypes. Most top pathways related to epidermal growth factor (EGF) and its receptor. EGR proteins stimulate cell growth and differentiation, and dysregulation of these is a common cause of cancer. Other prominent pathways involved fibroblast growth factors, B-cell activation, and transforming growth factors. It is interesting to note that many of the top pathways for subtype M2 involved cellular senescence. Caused by stressors like telomere dysfunction, oncogene activation, DNA damage, or outside factors, cellular senescence causes a complete arrest in cell proliferation. This process was not prominently seen in other subtypes.

Acknowledgements

I would like to thank Dr. Konstantinos Karagiannis and Mr. Kamil Kural from the HIVE lab at the U.S. Food and Drug Administration. I would also like to thank Ms. Angelique Bosse (Montgomery Blair High School) for her support and guidance.

Citations

[1] Warnat-Herresthal, S. et al. (2020). Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics. iScience, 23(1), 100780. doi:10.1016/j.isci.2019.100780

[2] Nehme, A., Dakik *et al.* (2020). Horizontal meta-analysis identifies common deregulated genes across AML subgroups providing a robust prognostic signature. Blood Advances, 4(20), 5322–5335. doi:10.1182/bloodadvances.2020002042

[3] Acute Myeloid Leukemia (AML) Subtypes and Prognostic Factors. American Cancer Society. (n.d.). Retrieved October 19, 2021, from https://www.cancer.org/cancer/acute-myeloid-leukemia/detection-diagnosis-staging/how-classified.html.

[4] Duan, Q., Reid *et al.* (2016). L1000CDS2: LINCS L1000 characteristic direction signatures search engine. Npj Systems Biology and Applications, 2(1). doi:10.1038/npjsba.2016.15