# Predicting Cell Types and Disease Status in Acute Myeloid Leukemia
Life Sciences

Vardhaan Ambati (vambati) and Raymond Yin (raymond7)

**Introduction**: Acute myeloid leukemia is an aggressive form of hematologic cancer characterized by the over proliferation of myeloid blasts in the bone m arrow. It is the most common type of leukemia amongst adults[1] and is responsible for the highest leukemia-related deaths annually[1]. Treatment consists of cytotoxic chemotherapy followed by an allogeneic bone marrow transplant, implemented with limited success[2]. Late detection often reduces the efficacy of treatment, as disease burden tends to rise over time. Patients with higher disease burden at time of diagnosis tend to be much harder to treat clinically. We believe that the way forward in leukemia treatment is with early diagnosis and early treatment to reduce overall disease burden.

One challenge to early diagnosis is the difficulty in obtaining testable samples. AML is thought to originate within the bone marrow, and thus, clinical diagnosis requires a bone marrow aspiration, which is a minor surgical procedure done to extract soft tissue from within the bone. This aspirate is then tested by pathology, who will confirm a diagnosis based on the sample. Bone marrow aspirates are the gold standard in diagnosing leukemia patients, and have extremely high accuracy in determining a patient's disease status.

Bone marrow aspirates, however, are not a routine screening procedure and are only ordered when the provider has adequate suspicion of a hematological malignancy. These tests are expensive and extremely painful for the patient, who is usually only under local anesthetic during the procedure. It is simply impractical and inefficient to make these a routine screening procedure--instead, we want to focus on the screening potential of routine blood tests (which extract peripheral blood).

We believe that peripheral blood samples can also be effectively tested for malignancies. Since all Americans are recommended to obtain routine blood exams at least once per year, malignancy screening could be easily added onto the existing panel of blood exams without necessitating any additional procedures for the patient. Thus, there is no need to scale infrastructure or place additional burden on healthcare providers. In addition, malignancy screening generally consists of basic genomic panels that can be easily integrated into the existing phlebotomy workflow.

One main barrier to implementing peripheral blood malignancy testing is the low presence of disease within the samples. Most of the disease will be located in the bone marrow, as opposed to in the peripheral blood, so a concern arises over the panels' ability to pick up disease. To address this issue, we can explore the integration of newer molecular assays which can cheaply and efficiently perform highly sensitive expression profiling, sometimes on the single cell level. For these to be effective, however, clinicians must be able to: 1. Identify relevant subsets of cells from the overall raw data 2. Generate predictions about disease status based off of the expression patterns of these specific subsets of cells. Our project aims to lay groundwork for the potential of machine learning to solve these problems.

**Datasets:** We will use a total of 4 datasets pulled from the public Gene Expression Omnibus, which is run by the NCBI to help us achieve our aims. First, we have GSE42519 from Rapin et al., 2014, which contains microarray expression data from 10+ blood cell types. This dataset will be used to create profiles for different blood cell types in order to validate the hypothesis that blood cell types can be effectively separated by expression data. Our second and third datasets will be used to train a diagnosis classifier for disease vs healthy individuals. These include GSE12417 from Metzeler et al., 2008, which has gene expression data for 405 AML patients derived from blood sources, and GSE32719 from Pang et al., 2011, which has gene expression data for hematopoietic stem cells (HSCs) for 27 healthy patients. We will use the former dataset as a disease profile, and the latter as a profile for healthy controls since HSCs are often thought to be sound hematological control for disease states. Before working with the data, we will perform PCA to reduce dimensionality to two principal components, since the amount of features (54765) greatly outweighs the amount of samples. After combining these two datasets, we will also split off 15% from this combined dataset to use for validation purposes. Our fourth dataset GSE17054, from Majeti et al., 2009, will serve as our test set, as it includes both leukemic and HSC expression profiles. After testing on GSE17054, from Majeti et al., 2009, we will then combine datasets 2-4 together and perform 4-fold cross validation with the combined datasets (3 repeats). It is worth noting that all four of our datasets were generated using the Affymetrix Human Genome U133 Plus 2.0 Assay, which ensures overlap between the features (genes) of each dataset. The Affymetrix assay measures expression data on the Human Genome U133 Set of genes, which consists of comprehensive whole genome coverage on the most well-substantiated human genes (54675 separate genes).

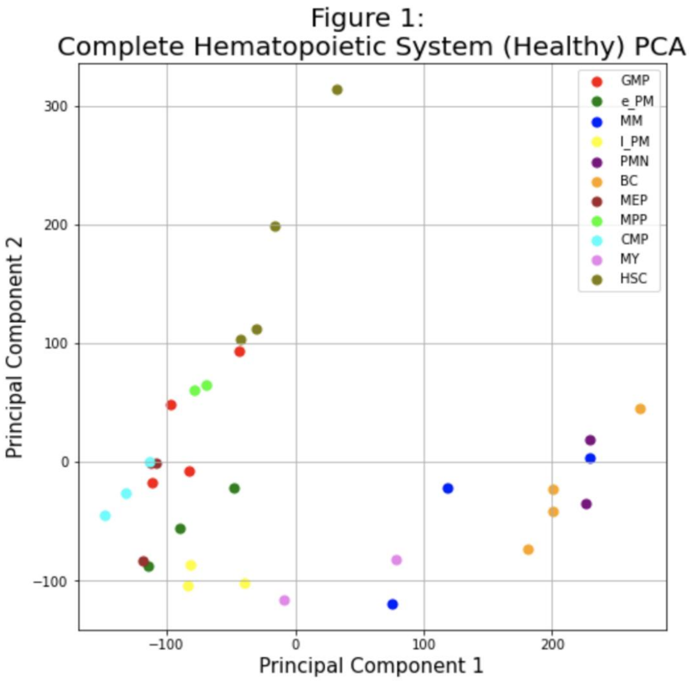
**Methods**: We propose using machine learning to aid in the diagnosis and identification of acute myeloid leukemia. We will first use unsupervised machine learning methods and blood cell expression data from Rapin et al., 2014 to group and classify cells by type in order to create an overall profile of blood cells types based on expression patterns. To do this, we will first parse in the data, which is provided by the Gene Expression Omnibus in a series matrix file. We will aim to create a dataframe with samples and features on either axis. Then, we will run a PCA analysis, which we believe to be perfect for this aim, in which we observe high dimensionality data (in the form of gene expression values) that benefits from reduction. We will also overlay our plot with true labels to assess the validity of such a dimensionality reduction.

Subsequently, we aim to use expression data from Metzeler et al., 2008 and Pang et al., 2011 on hematopoietic and leukemic stem cells to train a logistic regression model. To do this, we must first parse the series matrix files from the Gene Expression Omnibus and standardize the format between the two datasets. We will then need to label each sample as either a ‘disease’ or ‘control’ sample. Since there exist many more disease samples than control samples, we will pick 27 disease samples randomly from our dataset to merge with our 27 control samples from Pang et al., 2011. We will merge the two datasets together by each gene. If there are any genes that do not overlap between the two datasets, we will omit these rows. After merging the two datasets, we will preliminary perform data exploration by conducting PCA and reducing the features to 2 dimensions; we then subsequently perform t-SNE and plot for visual analysis.

We will then use the python scikit-learn package to train a logistic regression classifier with the PCA data. We will apply this classifier on a validation set consisting of 15% of our original dataset and assess performance before moving on. We will test our validated model on Majeti et al., 2009. To do this, we will first parse the dataset from Gene Expression Omnibus, filter out any features that do not match our training features, and conduct PCA dimensionality reduction on the Majeti dataset. After this, we will run our classifier on the 13 samples in the dataset from Majeti et al., 2009, and report subsequent performance metrics.

In addition to the procedure described above, where we train and validate on the Metzeler et al., 2008 and Pang et al., 2011 datasets and test on the Majeti et al., 2009 dataset, we will also aim to combine all three datasets into one large dataset. We will then train and test another logistic model on this combined dataset using 4-fold cross validation with 3 repeats.

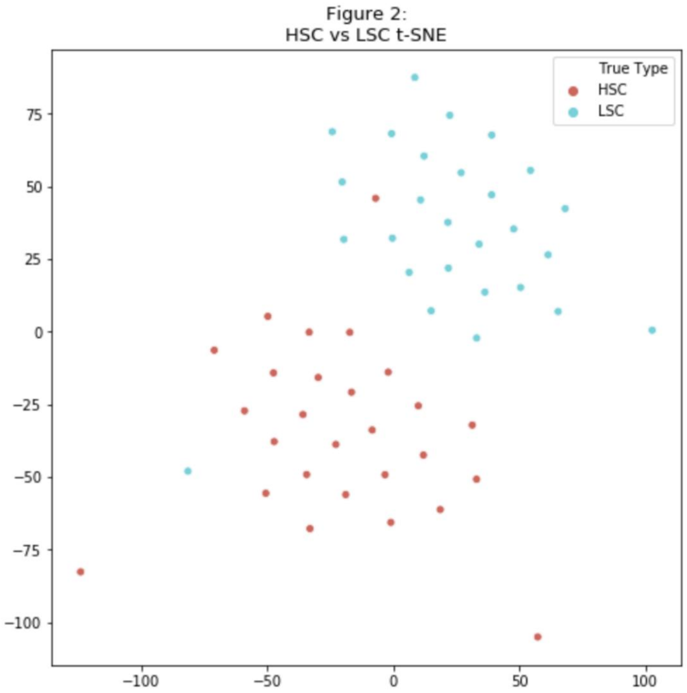
We think that logistic regression is an appropriate model for the binary classification problem of disease status. Since we have adequate training data, we believe that our logistic regression should perform well provided that gene expression profiles of diseased vs healthy patients are significantly different. Since hematology research provides feature rich datasets, there have been previous attempts to utilize gene expression data to produce useful clinical insights. In fact, Metzeler et al., 2008, from which we obtained our disease expression dataset, actually used their dataset for prognostic applications (such as predicting overall survival), suggesting that gene expression data can be a valuable resource for outcome prediction.



**Results:**

We initially conducted PCA analysis of single cell gene expression data of healthy hematopoietic system from Rapin et al., 2014 with the goal of seeing if PCA could categorize different cell types (Figure 1). We noticed that PCA was generally able to group related blood cells, but due to our small dataset size (34 samples, 54675 features), it was difficult to interpret.

We then conducted t-SNE analysis of single cell gene expression data for 27 healthy hematopoietic stem cells (HSCs) and 27 cancerous Leukemic stem cells (LSCs) from our combined Metzeler et al., 2008 and Pang et al.,



2011 dataset to predict if an unsupervised training model would be able to discriminate between healthy and diseased stem cells (Figure 2). Similar to the previous PCA analysis, the clustering looks promising but it is difficult to tell because of the small dataset size (54 samples, 54675 features) .

We then began to train our model. We performed PCA on the combined dataset, which contained both healthy and cancerous samples. From here, we randomly selected 27 disease samples and 27 healthy samples and trained a logistic regression model on 85% of these samples and validated using the remaining 15%. We obtained a validation accuracy of 100%. After validation, we tested on the Majeti et al., 2009 dataset and obtained a testing accuracy of 38%, a sensitivity of 0%, and a specificity of 56%.

Lastly, we performed PCA dimensionality reduction on the 13 samples from Majeti et al., 2009 and added these samples into the training set for our logistic regression model. We used the RepeatedKFold() function from scikit-learn to perform 4 fold validation with 3 repeats on this combined dataset. Our validation obtained an accuracy of 97%.

**Discussion**: From our initial experimentation, it appears that PCA and t-SNE analysis can be performed somewhat successfully on single cell gene expression data. In Figure 1, we show how PCA is able to separate the major cellular groups visually based on expression data. In Figure 2, we note that t-SNE also seems to work well distinguishing between healthy and cancerous stem cells. We even note that the cancerous and healthy stem cells can be easily separated with a line, and thus it may be promising to use SVMs for future classification problems.

Where we start to run into an issue is with our logistic regression model. Despite obtaining a validation accuracy of 100% initially, our model only obtains an accuracy of 38% when testing on the Majeti et al., 2009 dataset. This suggests two things; first, that our model must initially have been overfit to our training dataset. Second, there must be obvious challenges with interpreting the expression data that make it difficult to perform transfer learning on new datasets. One problem that may have occurred is our usage of PCA. Since PCA was done independently (between the testing and training set), it does not guarantee that the same metrics were used. Another challenge is the calibration of scientific assays that are used to measure single cell expression data. In other words, the values from one dataset may not be on the same scale or magnitude as the values from another dataset, even though the assays themselves are physically identical. This is an unavoidable problem when working with medical/biological data, and the only solution is often to work within a dataset, as opposed to between datasets.

This is in part why we decided to add the Majeti et al., 2009 dataset with our original training set, perform PCA on the combined dataset, and retrain our model with the combined dataset. We thought that if we were to integrate the foreign dataset into training, we would be able to account for some of the miscalibration between datasets. We found that when we performed cross validation after retraining with the Majeti data, we obtained an accuracy of 97%, which confirmed some of these suspicions. This suggests that when working with expression data, researchers should either aim to train from diverse datasets, or should try their best to work within a dataset so as to not introduce potential problems with data calibration.

**Conclusion:** Despite our classifier holding poor performance when testing on novel datasets, we have shown that when working within a dataset, machine learning can be incredibly powerful in making predictions with high accuracy. We are led to believe that issues with data calibration between datasets can be overcome by integrating more datasets into model training. Thus, as more and more cellular level data is generated in the field of hematology, the integration of machine learning becomes more and more viable. We believe that future efforts can contribute to a better understanding of a healthy blood cell expression environment and lead to machine learning-based tools that can aid physicians in diagnosing acute myeloid leukemia early, improving patients' prognosis.

**Contributions**: Both Raymond and Vardhaan worked together on all aspects of the project including dataset searching, initial data exploration, model training and testing, and the write up.

## References

1. SEER9 Database 1975-2013.
2. NCCN Clinical Practice Guidelines in Oncology 2016.
3. Rapin N, Bagger FO, Jendholm J, Mora-Jensen H et al. Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. *Blood* 2014 Feb 6;123(6):894-904. PMID: 24363398
4. Majeti R, Becker MW, Tian Q, Lee TL et al. Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc Natl Acad Sci U S A* 2009 Mar 3;106(9):3396-401. PMID: 19218430
5. Metzeler KH, Hummel M, Bloomfield CD, Spiekermann K et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* 2008 Nov 15;112(10):4193-201. PMID: 18716133
6. Pang WW, Price EA, Sahoo D, Beerman I et al. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc Natl Acad Sci U S A* 2011 Dec 13;108(50):20012-7. PMID: 22123971