Classification of Cancer by Gene Expression

Nadia Paquin, Vu Ton, Beatrice Tuda

03/15/2024

Introduction

1. Historical Problem and Solution:

In 1999, Golub et al published a proof-of-concept study in *Science* that demonstrated how new cases of cancer could be classified by gene expression monitoring via DNA microarray. This study provided a novel general approach for identifying new cancer classes and assigning tumors to known classes. Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the paper. Each sample represents an individual, and each individual is tested for over 7000 genes. The data were used to classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The datasets contain measurements of relative concentration for specific genes corresponding to ALL and AML in each patient. Samples were taken from bone marrow and peripheral blood. Intensity values have been re-scaled such that overall intensities for each chip are equivalent. When the paper was published, the classification analysis using genetic expression to differentiate cancer classes would allow for a standardized process to classify tumors by minimally invasive biopsies.

With this report we are attempting to automatically classify acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) using models such as Logistics Regression , Random Forest, Support vector machine, and K-means.

2. About the Data:

There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in this analysis. These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood. Intensity values have been re-scaled such that overall intensities for each chip are equivalent.

Exploratory Data Analysis:

Figure 1: The data set

```
[31]:                     Gene Description Gene Accession Number    1 call    2  \
     0  AFFX-BioB-5_at (endogenous control)        AFFX-BioB-5_at -214    A -139
     1  AFFX-BioB-M_at (endogenous control)        AFFX-BioB-M_at -153    A  -73
     2  AFFX-BioB-3_at (endogenous control)        AFFX-BioB-3_at  -58    A   -1
     3  AFFX-BioC-5_at (endogenous control)        AFFX-BioC-5_at   88    A  283
     4  AFFX-BioC-3_at (endogenous control)        AFFX-BioC-3_at -295    A -264

       call.1    3 call.2    4 call.3  …   29 call.33   30 call.34   31 call.35  \
     0      A  -76      A -135      A  …   15       A -318       A  -32       A
     1      A  -49      A -114      A  … -114       A -192       A  -49       A
     2      A -307      A  265      A  …    2       A  -95       A   49       A
     3      A  309      A   12      A  …  193       A  312       A  230       P
     4      A -376      A -419      A  …  -51       A -139       A -367       A

        32 call.36   33 call.37
     0 -124       A -135       A
     1  -79       A -186       A
     2  -37       A  -70       A
     3  330       A  337       A
     4 -188       A -407       A
```

To data cleaning we had to make the gene number as column number, remove some columns, and merge the cancer data.

Figure 2: Merged data looks like

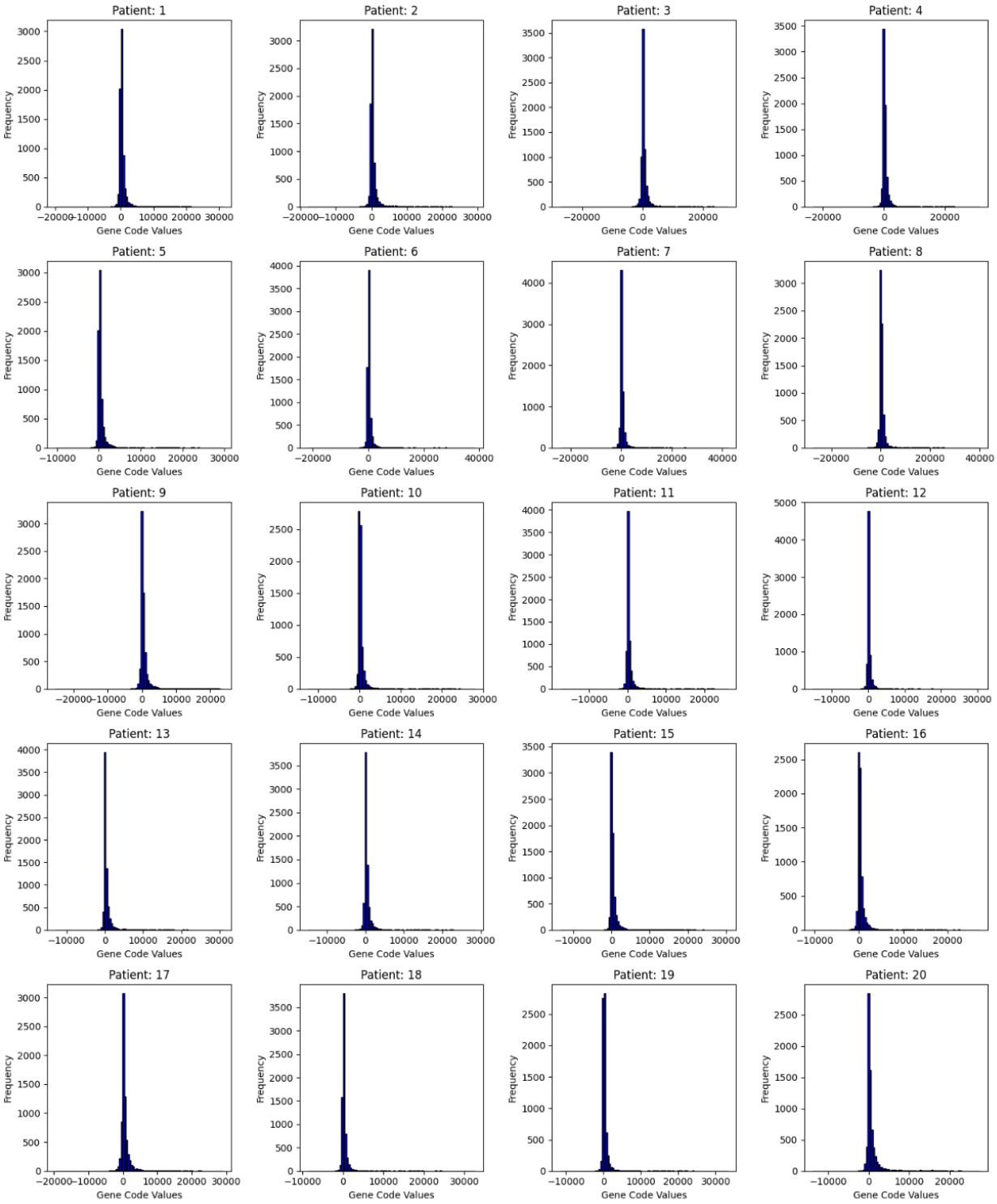
```
train_set.head(5)
```

	patient_number	AFFX-BioB-5_at	AFFX-BioB-M_at	AFFX-BioB-3_at	AFFX-BioC-5_at	AFFX-BioC-3_at	AFFX-BioDn-5_at	AFFX-BioDn-3_at	AFFX-CreX-5_at	AFFX-CreX-3_at
0	1	-214	-153	-58	88	-295	-558	199	-176	252
1	2	-139	-73	-1	283	-264	-400	-330	-168	101
2	3	-76	-49	-307	309	-376	-650	33	-367	206
3	4	-135	-114	265	12	-419	-585	158	-253	49
4	5	-106	-125	-76	168	-230	-284	4	-122	70

5 rows × 7131 columns

Then more data exploring was done. From the gene value we learned that patient gene code value was mostly around 0 with outliers between -10000 to 20000.

Figure 4: Gene Densities for each patient in the train_set with ALL cancer type



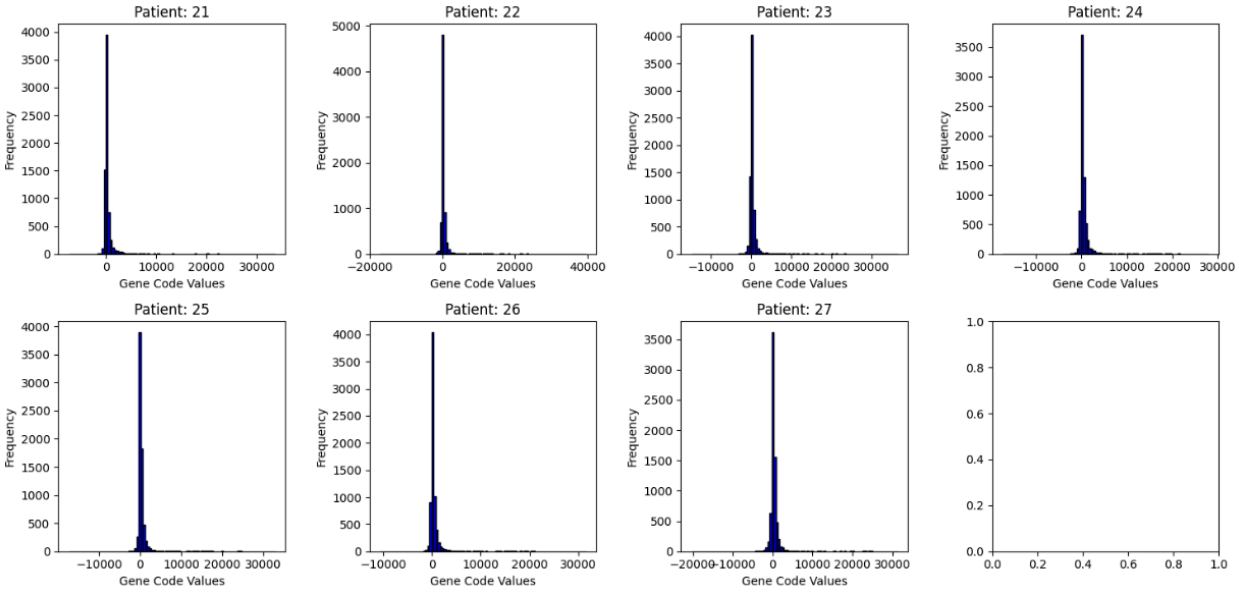


Figure 5: Gene Frequency for each patient in the train_set with AML cancer type

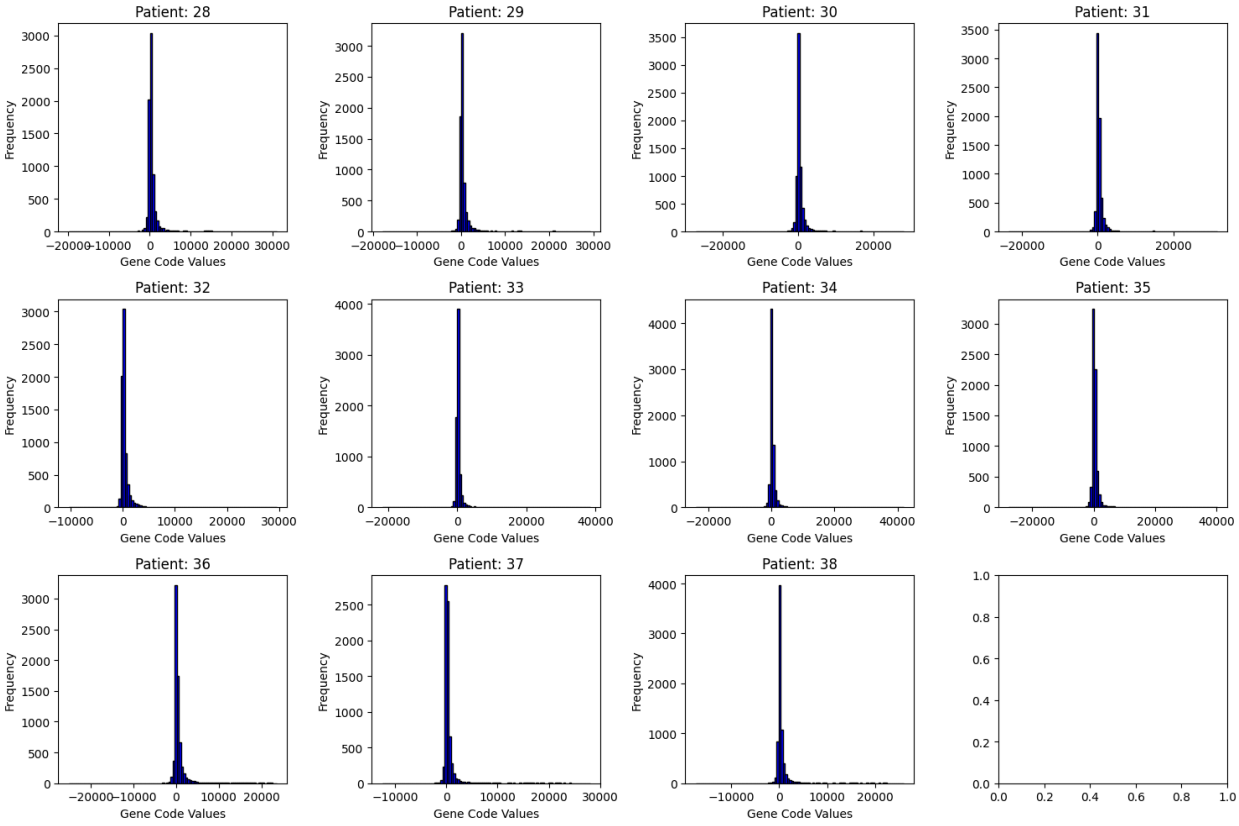
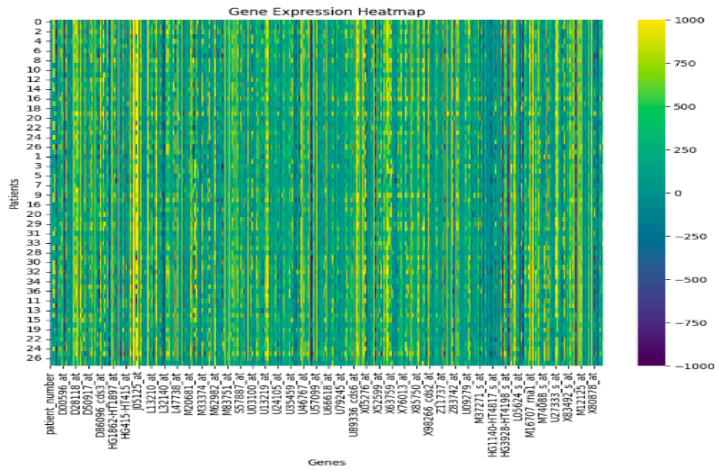
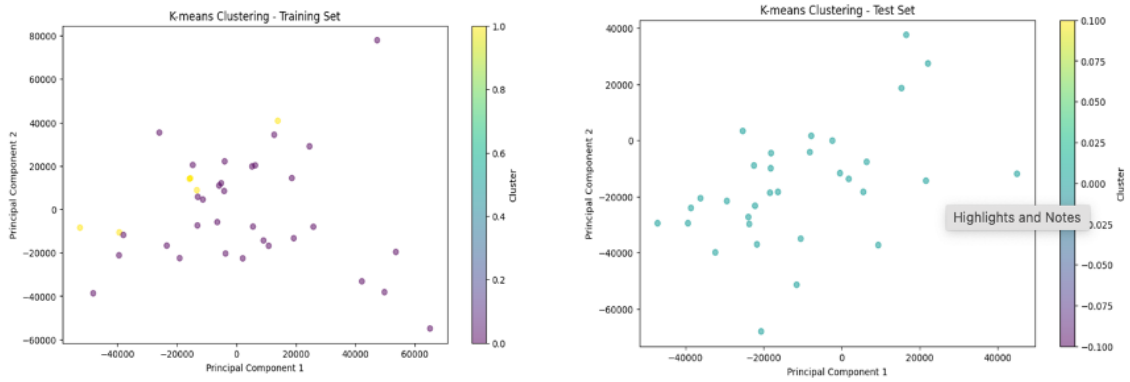


Figure 5: Heatmap



Heat map is the graphical representation of data where the colors are shown by colors. Since there are a lot of genes the heatmap doesn't give too much information.

Figure 6: K-means



K-means is a clustering algorithm that helps in pattern recognition which is what we are trying in this sample. The best centroids for this project is 2 and it is seen in figure 6 the Training set is separated into 2 clusters and the Test set in one.

Modeling and Predictions

Model 1: Logistics Regression

• Since the data is very well-cleaned, the result is very good with an accuracy around 97%. • The execution time is also very low (under 3 seconds). • True positive is 100% accurate.	Confusion Matrix: Predicted ALL Predicted AML Actual ALL 19 1 Actual AML 0 14 Accuracy: 0.9705882352941176

• There is no modification on the model.	

Model 2: Random Forest

• The execution time for the model is also low (under 3 seconds) with an accuracy around 82% • True positive is about 76.92% and true negative is 100% • This model is modified by using different numbers of trees 50, 100, 200, 500, 1000, and 50000. The best accuracy is when the number of trees equals 100. • The accuracy seems to converge to 74% when increasing the numbers of trees.	Confusion Matrix: Predicted ALL Predicted AML Actual ALL 20 0 Actual AML 6 8 Accuracy: 0.8235294117647058 n = 100

Model 3: Support Vector Machine: Grid search

• Model relies on PCA transformed data • PCA analysis yielded around 22 components to be responsible for 90% of the variance in the data. • GridSearch function used to tune hyperparameters. • Accuracy of 94.1%, true positives at 90% true negatives at 100%	The best combination of paremeters is SVC(C=0.1, decision_function_shape='ovo', gamma=1, kernel='linear') SVM accuracy: 0.941 Confusion Matrix: Predicted ALL Predicted AML Actual ALL 18 2 Actual AML 0 14

Conclusion

Most of the algorithms had high accuracy except the random forest model, which might result from overfitting or needing improvement in pre-processing. Pre-processing is one of the main steps in achieving the most accuracy of each algorithm. The model technique with the best accuracy is the Logistic Regression, which has 97.06% accuracy. Most of the models had few false positives, only the Random Forest predicted false negatives.

Based on the high accuracy of the models, we conclude that we are able to effectively differentiate between the two types of cancer using sequenced DNA with approximately 95%

confidence. Doctors can diagnose using a bone marrow biopsy alone, eliminating the need for several assays in distinct, highly specialized labs. This finding, in the years of its original publication, opened the door to early detection and proper treatment for patients with leukemias.

These results suggest that there is a relationship between gene expression and cancer pathologies. This raises the question that there perhaps might be a pathological relationship between specific genes and leukemia. More research should be done to determine which genes are most critical, the proteins and pathways they are associated with, and if these relationships are causal. Future research can also be done to see if these models work for other cancers.