Classification of Cancer by Gene Expression

Nadia Paquin, Vu Ton, Beatrice Tuda

IST718, 03/15/2024

Introduction

Background: 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, Naive Bayes, and K-means.

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.

Data Manipulation:

	Gene Description	Gene Accession Number	1	call	2	call.1	3	call.2	4	call.3	...	29	call.33	30	call.34	31	call.35	32	call.36
0	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	A	-139	A	-76	A	-135	A	...	15	A	-318	A	-32	A	-124	A
1	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	A	-73	A	-49	A	-114	A	...	-114	A	-192	A	-49	A	-79	A
2	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	A	-1	A	-307	A	265	A	...	2	A	-95	A	49	A	-37	A
3	AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	88	A	283	A	309	A	12	A	...	193	A	312	A	230	P	330	A
4	AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-295	A	-264	A	-376	A	-419	A	...	-51	A	-139	A	-367	A	-188	A

Figure 1: Raw Training Data. Numbered columns correspond to individual patients, rows correspond to individual genes, ‘call’ columns contain no data and will be removed. Raw data set is 7129 rows by 78 columns. Testing data has an identical layout.

	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	...	U58516_at	U73738_at	X06956_at	X16699_at
0	1	-214	-153	-58	88	-295	-558	199	-176	252	...	511	-125	389	-37
1	2	-139	-73	-1	283	-264	-400	-330	-168	101	...	837	-36	442	-17
2	3	-76	-49	-307	309	-376	-650	33	-367	206	...	1199	33	168	52
3	4	-135	-114	265	12	-419	-585	158	-253	49	...	835	218	174	-110
4	5	-106	-125	-76	168	-230	-284	4	-122	70	...	649	57	504	-26

Figure 2: Cleaned Training Data. Rows and columns are flipped so that each row represents one patient, with genes in the columns. The cleaned data has 38 rows by 7131 columns. Test data was cleaned in the same manner, resulting in a 34 row by 7131 column dataset.

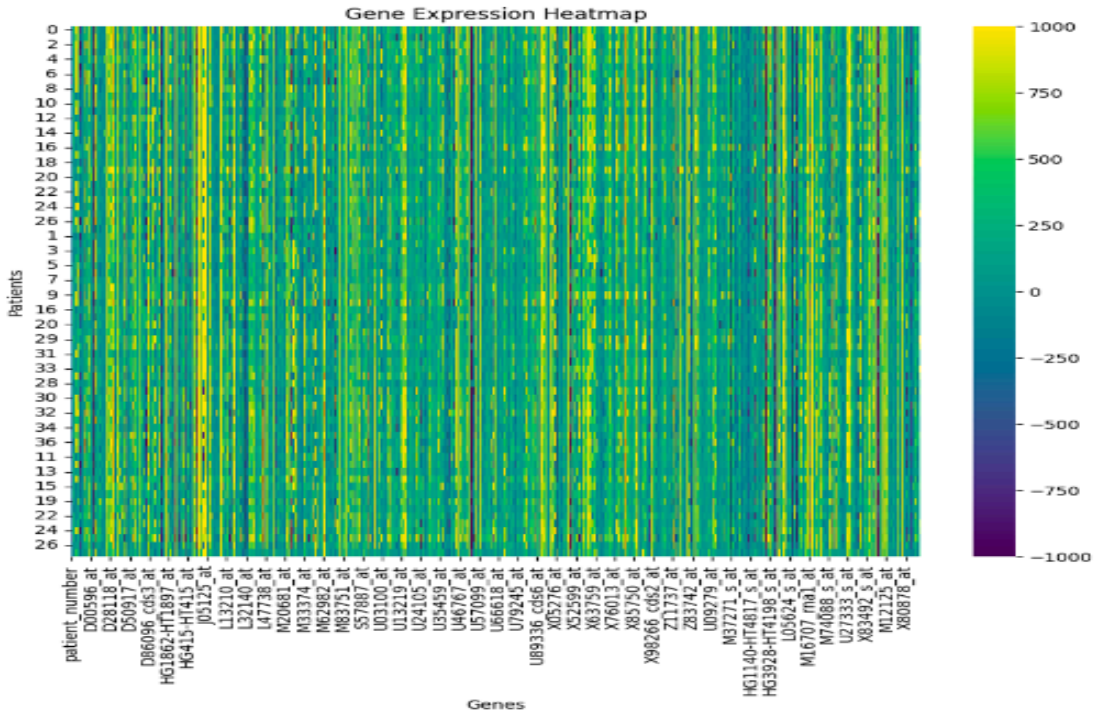

Figure 3: Heatmap of raw gene expression. The first several rows correspond to patients with ALL and the bottom half of the data corresponds to patients with AML. Brighter yellow colors correspond to overexpression whereas deeper blue colors represent underexpression. If the data can be used to differentiate the two leukemias, we would expect to see a visual representation of genetic variance between the two groups, visualized by a horizontal split. However in this heatmap, there are no clear distinctions between the top and bottom half, suggesting the genes might not be significantly different.

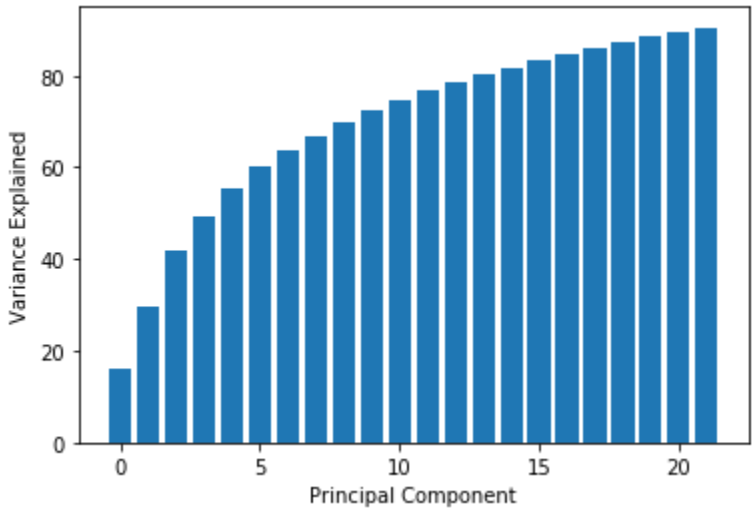

Figure 4: Principal component analysis (PCA). PCA is a technique that uses linear algebra to compress data from the large scale to significantly fewer principal components or dimensions.

Typically this method is used to reduce data to a set number of dimensions. In this analysis, we investigated how many principal components were needed to meet 90% of the variance in the data. Analysis showed that only 22 dimensions were required, which could reduce our training data from 7131 x 36 to 22 x 36.

Models:

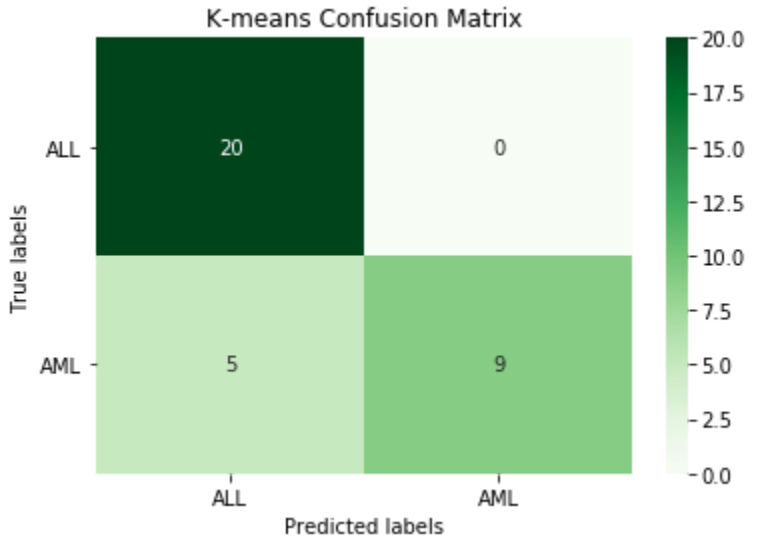

Figure 5: K-means clustering confusion matrix (accuracy of 85.3%). K-means is the only unsupervised model used in this report, which partitions the data into clusters by minimizing the variance within each cluster. The only parameter specified was the number of clusters, which was set to two, representing the two cancer groups.

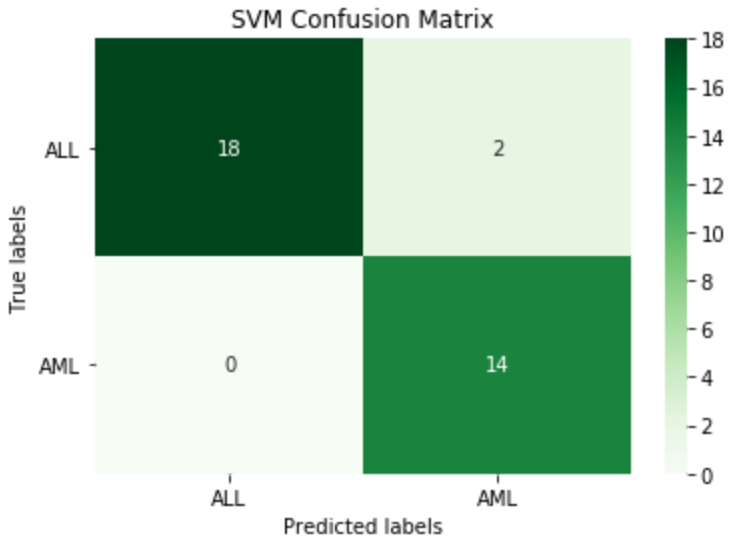

Figure 6: Support vector machine (SVM) confusion matrix (accuracy of 94.1%). SVM is a supervised learning model used for classification tasks, which finds the optimal hyperplane that maximizes the margin between different classes. GridSearch was utilized to optimize the model parameters for best performance.

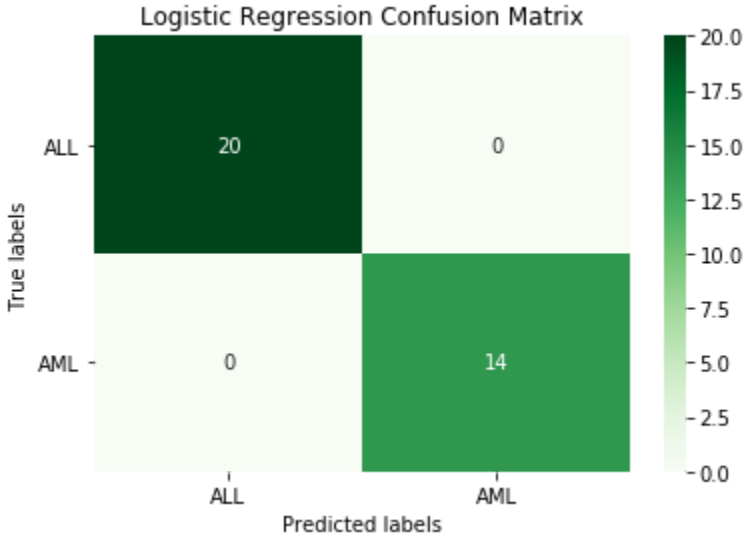

Figure 7: Logistic regression confusion matrix (accuracy of 100%). Logistic Regression is a supervised learning model typically used for binary classification tasks, which estimates the probability that a given input belongs to a particular class using a logistic function. GridSearch was utilized to optimize the model parameters for best performance.

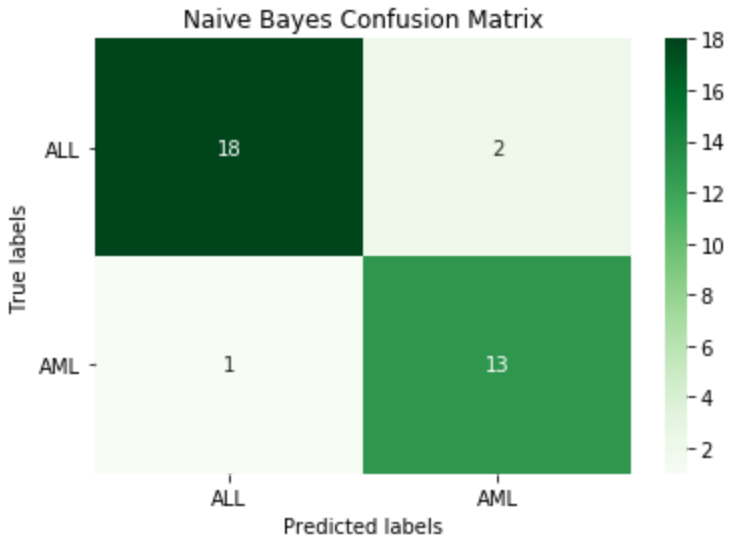

Figure 8: Naïve Bayes confusion matrix (accuracy of 91.2%). Naive Bayes is a probabilistic classifier that calculates the probability that a given input belongs to a particular class by leveraging Bayes' theorem.

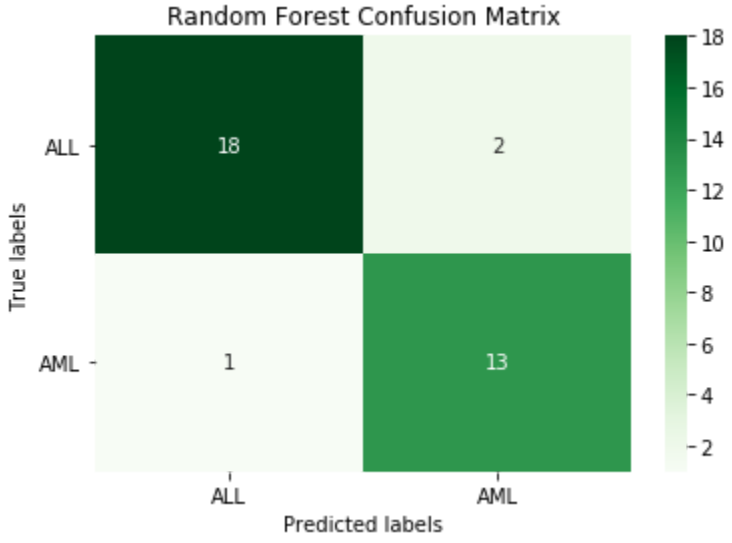

Figure 9: Random forest confusion matrix (accuracy of 91.2%). Random forest constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. Gridsearch was used to optimize parameters.

Model	Accuracy
K-means	0.853
SVM	0.941
Logistic Regression	1.00
Naive Bayes	0.912
Random Forest	0.912

Figure 10: Accuracy of models.

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.