Analyzing Microarray data 200501P

Abstract

This report details the analysis of gene expression data to differentiate between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Using a publicly available dataset containing expression levels for over 7000 genes across training and testing samples, interpretable classification models were developed and evaluated. Logistic Regression achieved the highest test accuracy (97.1%), outperforming Decision Tree, Random Forest, and XGBoost models (all at 91.2%). Feature importance analysis across models identified key genes that potentially distinguish AML from ALL, although the top genes varied between linear and tree-based methods (e.g., M25079_s_at, Y00787_s_at for Logistic Regression; X95735_at for Decision Tree/Random Forest). Unsupervised clustering techniques were applied to the training data without using class labels. Spectral Clustering demonstrated high purity (97.4%), effectively grouping the samples according to their true leukemia type, thus highlighting the potential of unsupervised methods for discovering underlying structures in gene expression data.

1. Introduction

Leukemia, a cancer of blood-forming tissues, encompasses various types, including Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Differentiating between these types is crucial for appropriate treatment strategies. Gene expression microarrays provide a powerful tool for studying cancer by simultaneously measuring the activity levels of thousands of genes. Differences in gene expression patterns can reveal molecular distinctions between cancer subtypes.

This study utilizes a gene expression dataset (Crawford/gene-expression) containing data for over 7000 genes from patients diagnosed with either AML or ALL. The dataset is pre-divided into training (38 samples) and testing (34 samples) sets. The primary goals of this analysis are:

1. To build and evaluate interpretable classification models capable of accurately predicting leukemia type (AML vs. ALL) based on gene expression profiles.
2. To identify specific genes or gene expression patterns that are most influential in distinguishing between AML and ALL using the developed models.
3. To explore the data's inherent structure using unsupervised clustering techniques and assess their ability to group samples according to leukemia type without prior knowledge of the labels.

2. Data Preparation

The analysis utilized three provided CSV files: data_set_ALL_AML_train.csv, data_set_ALL_AML_independent.csv (test set), and actual.csv (containing class labels).

The following preprocessing steps were performed:

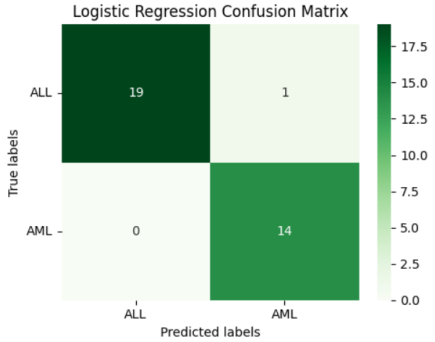
1. **Data Loading:** The training, testing, and label datasets were loaded into pandas.

2. **Label Encoding:** The categorical labels ('ALL', 'AML') in actual.csv were converted to a numerical format (ALL: 0, AML: 1).
3. **Feature Selection:** Columns containing the term "call" (likely related to data quality flags from the microarray analysis) were removed from both training and testing expression datasets.
4. **Data Transposition:** The expression data was transposed so that rows represent individual patient samples and columns represent gene expression features. This aligns the data with standard machine learning input formats. The original training data had 7129 rows (genes) and 78 columns (samples + info), while the test data had 7129 rows and 70 columns.
5. **Header Assignment & Cleaning:** Gene Accession Numbers were set as column headers, and non-numeric metadata rows ('Gene Description', 'Gene Accession Number') were dropped. The data was converted to numeric types.
6. **Train/Test Split:** The data was explicitly split based on the provided files and patient numbering conventions (Patients 1-38 for training, 39-72 for testing), resulting in:
    - X_train: (38 samples, 7129 genes)
    - y_train: (38 samples, 1 label column)
    - X_test: (34 samples, 7129 genes)
    - y_test: (34 samples, 1 label column)

3. Q1: Classification Models for Leukemia Prediction

The goal was to build interpretable models to classify samples as either AML or ALL. Several models were trained using X_train and y_train, and their performance was evaluated on X_test and y_test.

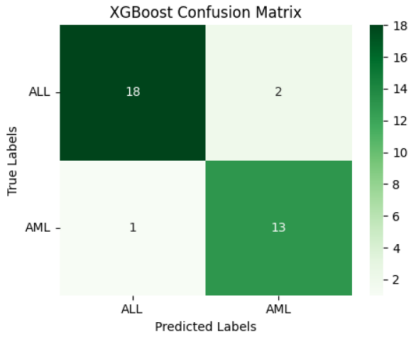
- **Logistic Regression:**
    - A Logistic Regression model was trained (max_iter=1000, random_state=0).
    - **Performance:** Achieved a test accuracy of **97.1%**.
    - **Confusion Matrix:**



(Figure 1) - This indicates only one misclassification (an ALL sample predicted as AML).

- **Interpretability:** Coefficients directly indicate feature importance.

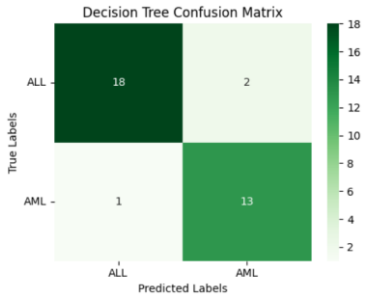
- **XGBoost Classifier:**
  - An XGBoost model was trained (random_state=0).
  - **Performance:** Achieved a test accuracy of **91.2%**.
  - **Confusion Matrix:**



(Figure 2) - This shows 3 misclassifications.

  - **Interpretability:** While powerful, XGBoost is less directly interpretable than LR or DT. SHAP analysis was performed.

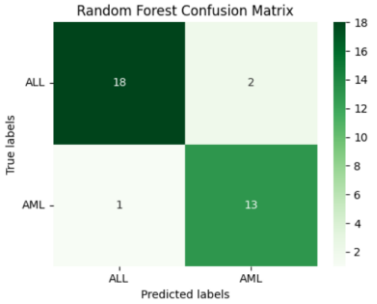
- **Decision Tree Classifier:**
  - A Decision Tree model was trained (random_state=0).
  - **Performance:** Achieved a test accuracy of **91.2%**.
  - **Confusion Matrix:**



(Figure 3) - This shows 3 misclassifications (all ALL samples predicted as AML).

  - **Interpretability:** The tree structure and feature importances provide direct interpretability.

- **Random Forest Classifier:**
    - A Random Forest model was trained using specified parameters (bootstrap=False, max_features=0.6, min_samples_leaf=8, min_samples_split=3, n_estimators=60, random_state=0). *Note: These parameters appear pre-selected; further tuning on a validation set could potentially improve results.*
    - **Performance:** Achieved a test accuracy of **91.2%**.
    - **Confusion Matrix:**



(Figure 4) - Performance identical to the Decision Tree in this case.

- **Interpretability:** Feature importances and SHAP analysis provide insights.

Summary: Based on the test set accuracy, Logistic Regression was the best-performing model (97.1%), followed by the tree-based models (Decision Tree, Random Forest, XGBoost), all achieving 91.2% accuracy. Logistic Regression also provided strong interpretability through its coefficients.

4. Q2: Gene Expression Patterns Distinguishing AML and ALL

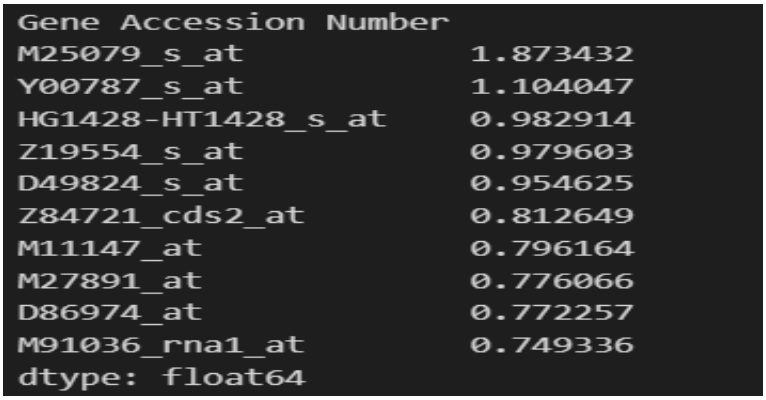
The trained interpretable models were analyzed to identify genes crucial for differentiating between AML and ALL.

- **Logistic Regression Feature Importance:**
    - **Method 1 (Absolute Coefficients):** The simplest approach ranks genes by the absolute magnitude of their learned coefficients. Top genes identified this way include:

```
Gene Accession Number
Y00787_s_at                 0.000186
M19507_at                   0.000175
Z19554_s_at                 0.000173
M27891_at                   0.000170
M17733_at                   0.000167
M25079_s_at                 0.000150
M96326_rna1_at              0.000149
M11147_at                   0.000136
AFFX-HUMRGE/M10098_3_at     0.000134
M14483_rna1_s_at            0.000132
dtype: float64
```

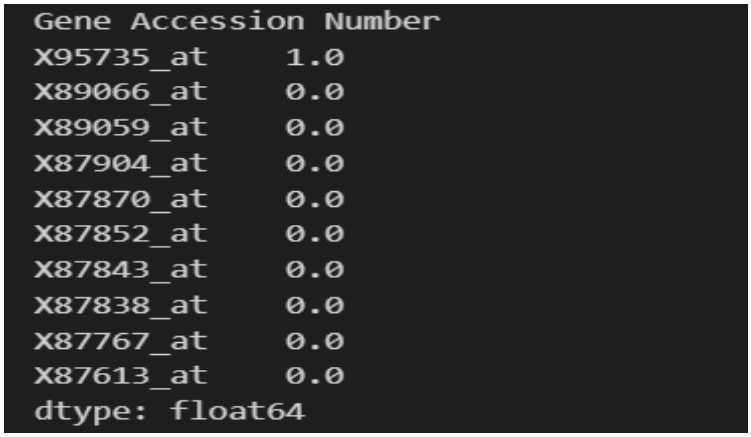
(Figure 5)

- **Method 2 (Coefficient * Standard Deviation):** This method scales the coefficient by the feature's standard deviation in the training data, providing a measure of importance that accounts for feature variability. The top genes by this metric were:



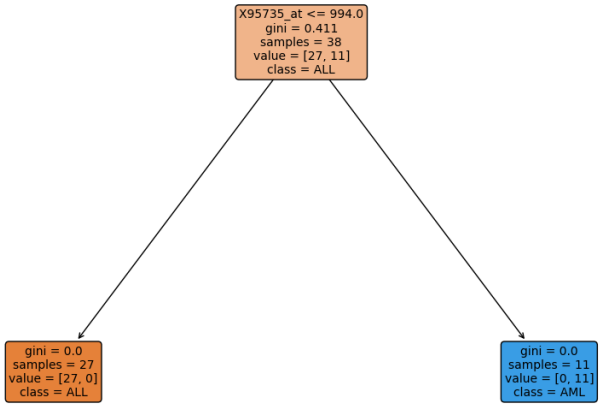
(Figure 6) - This method often gives a more robust indication of importance. Genes like M25079_s_at and HG1428-HT1428_s_at rise significantly in rank here.

- **Decision Tree Feature Importance:**
  - Decision Trees calculate feature importance based on how much a feature contributes to reducing impurity (e.g., Gini impurity) across all splits in the tree.
  - The most important gene identified was **X95735_at**, with an importance score of 1.0, indicating it was the sole feature used for splitting in this particular tree (likely due to the high dimensionality and potential for a single gene to provide good separation, though this can sometimes indicate overfitting or sensitivity to data). All other genes had an importance of 0.0.



(Figure 7)

  - A visualization of the top levels of the tree confirms that gene X95735_at is used for the primary split.



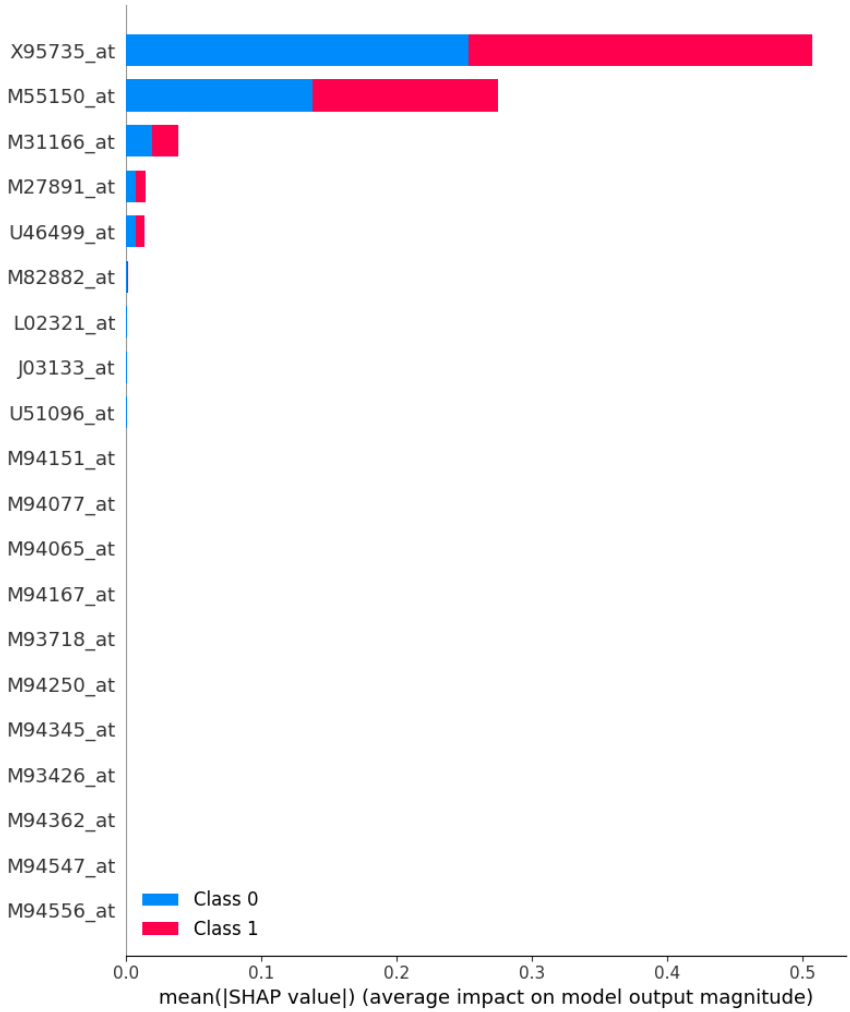
(Figure 8)

- **Random Forest Feature Importance:**
    - Random Forest aggregates feature importances from multiple decision trees. The top genes were:

```
Gene Accession Number
X95735_at     0.533333
M55150_at     0.383333
M31166_at     0.049357
U46499_at     0.016667
M27891_at     0.016452
U51096_at     0.000214
M82882_at     0.000214
L02321_at     0.000214
J03133_at     0.000214
X87767_at     0.000000
dtype: float64
```

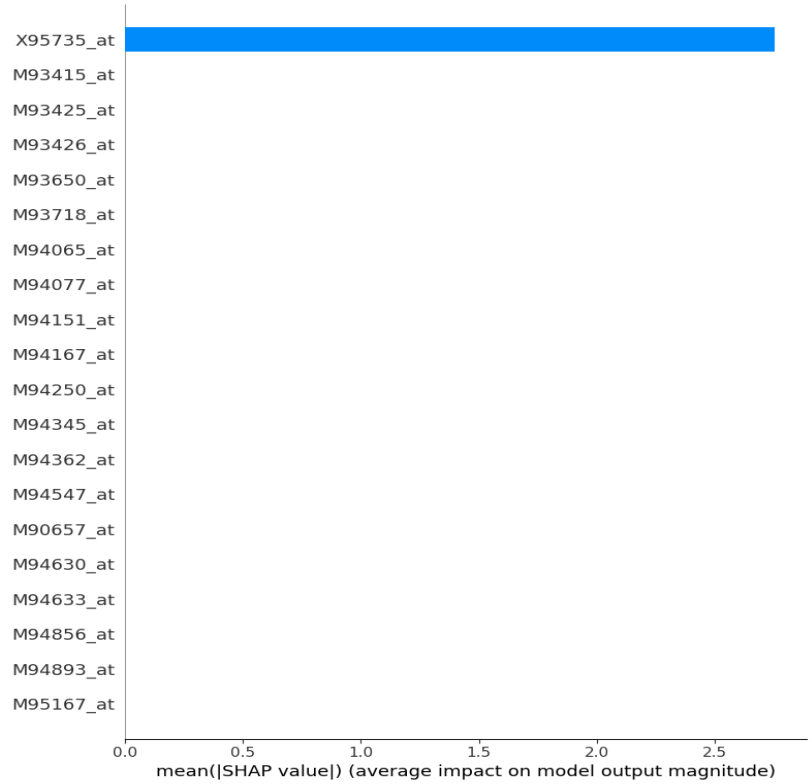
(Figure 9)

    - **SHAP Analysis (Random Forest):** A SHAP summary bar plot visually confirms the feature importance ranking derived from the Random Forest model, with X95735_at showing the highest mean absolute SHAP value.



(Figure 10)

- **XGBoost Feature Importance (via SHAP):**
    - SHAP values provide a more detailed explanation for gradient-boosted models like XGBoost.
    - The SHAP summary dot plot shows the magnitude (importance) and direction of effect for top features. Key genes influencing the XGBoost prediction include:



(Figure 11) - The dot plot illustrates, for example, that high values of X95735_at tend to push the prediction towards one class (likely AML, class 1), while low values push towards the other (ALL, class 0).

Summary: Gene X95735_at consistently emerged as highly important in the tree-based models (DT, RF, XGBoost). Logistic Regression highlighted a different set of top genes, including M25079_s_at and Y00787_s_at when feature scale was considered. This difference suggests that linear and non-linear models capture distinct aspects of the gene expression patterns separating AML and ALL. Further biological investigation into these specific genes could provide insights into the molecular differences between the leukemia types.

5. Q3: Unsupervised Clustering

Unsupervised clustering was performed on the *training data* (X_train) without using the y_train labels during the grouping process. The goal was to see if the algorithms could naturally identify the two leukemia types based solely on gene expression patterns. Purity was calculated post-clustering by comparing the assigned cluster labels to the true AML/ALL labels.

- **K-Means Clustering:**
    - Applied with n_clusters=2 and random_state=0.
    - **Purity:** Achieved a purity score of **0.816**.
    - **Confusion Matrix (Cluster vs. True Label):**

```
Confusion Matrix:
 [[ 0  4]
 [27  7]]
Purity of the kmeans clustering: 0.816
```

This indicates that Cluster 0 perfectly captured 4 AML samples, but Cluster 1 mixed most ALL samples (27) with the remaining 7 AML samples.

- **Agglomerative Clustering:**
    - Applied with n_clusters=2.
    - **Purity:** Achieved a purity score of **0.816**.
    - **Confusion Matrix (Cluster vs. True Label):**

```
Confusion Matrix:
 [[27  7]
 [ 0  4]]
Purity of the AgglomerativeClustering clustering: 0.816
```

Performance was identical to K-Means in terms of purity, although the cluster assignments were swapped.

- **Spectral Clustering:**
    - Applied with n_clusters=2, random_state=0, and affinity='nearest_neighbors'.
    - **Purity:** Achieved a high purity score of **0.974**.
    - **Confusion Matrix (Cluster vs. True Label):**

```
Confusion Matrix:
 [[27  1]
 [ 0 10]]
Purity of the SpectralClustering clustering: 0.974
```

Spectral Clustering performed significantly better, almost perfectly separating the two classes with only one AML sample being grouped with the ALL samples.

Use Case of Unsupervised Clustering in Gene Expression Analysis:

Unsupervised clustering is highly valuable in analyzing gene expression data for several reasons:

1. **Discovering Novel Subtypes:** It can reveal previously unknown subgroups within known disease categories (e.g., identifying distinct molecular subtypes of AML) based on shared expression patterns, which might correlate with clinical outcomes or treatment responses.
2. **Identifying Sample Groups Without Prior Labels:** In exploratory analysis, where class labels might be uncertain, unavailable, or potentially inaccurate, clustering can group samples based purely on their molecular profiles, suggesting natural biological groupings.
3. **Data Quality Control:** Outlier samples that do not cluster well with others might indicate experimental issues or unique biological characteristics, warranting further investigation.
4. **Hypothesis Generation:** Clusters identified can generate hypotheses about the biological mechanisms driving the observed groupings, guiding further research.
5. **Validation of Known Classes:** As demonstrated here, clustering can assess whether known biological classes (like AML vs. ALL) exhibit distinct enough molecular profiles to be separated algorithmically without supervision. The high purity achieved by Spectral Clustering suggests strong underlying biological differences reflected in the gene expression data.

6. Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) was applied to the training data to explore its dimensionality. The cumulative explained variance plot *(See Figure 12 in Appendix)* shows how much of the data's total variance is captured by increasing numbers of principal components. The plot indicates that a relatively small number of principal components capture a significant portion of the variance, suggesting that the high-dimensional gene expression data might lie on a lower-dimensional manifold and that dimensionality reduction could be effective for visualization or potentially feature engineering (though not used for classification here).

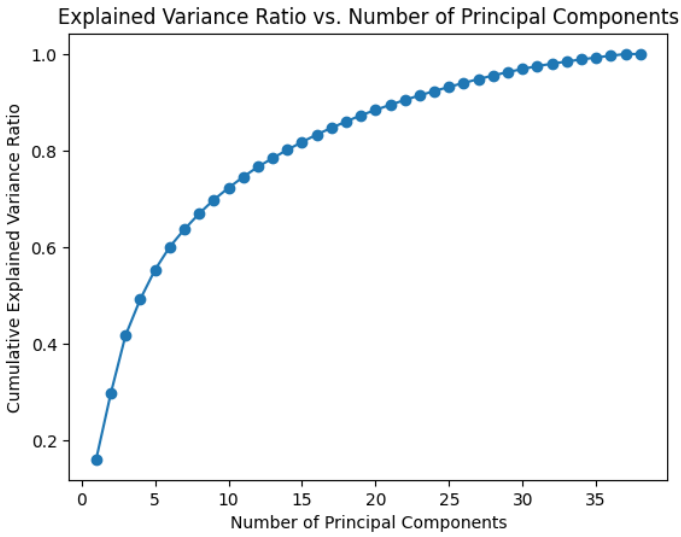


Figure 12

7. Conclusion

This analysis successfully demonstrated the potential of using gene expression data to classify leukemia types and identify distinguishing molecular patterns.

- **Classification:** Logistic Regression provided the highest test accuracy (97.1%) for predicting AML vs. ALL, also offering good interpretability. Tree-based models (DT, RF, XGBoost) achieved lower but still respectable accuracy (91.2%).
- **Gene Patterns:** Different models highlighted different sets of important genes. Tree-based methods strongly favored X95735_at, while Logistic Regression pointed towards genes like M25079_s_at and Y00787_s_at as highly influential when scaled by standard deviation. These genes warrant further biological investigation.
- **Clustering:** Unsupervised clustering, particularly Spectral Clustering, proved highly effective (97.4% purity) at grouping the samples according to their true leukemia types based solely on expression patterns, underscoring the distinct molecular signatures of AML and ALL present in the data and the utility of clustering for biological discovery.

Overall, the results indicate clear molecular distinctions between AML and ALL that can be captured by both supervised and unsupervised machine learning techniques applied to gene expression data.

Appendix: Figures

- Figure 1: Logistic Regression Confusion Matrix
- Figure 2: XG-Boost Confusion Matrix
- Figure 3: Decision Tree Confusion Matrix
- Figure 4: Random Forest Confusion Matrix
- Figure 5: Top 10 LR Features (Absolute Coefficients)
- Figure 6: Top 10 LR Features (Scaled Importance)
- Figure 7: Top 10 DT Features (Importance)
- Figure 8: Decision Tree Visualization (Top Levels)
- Figure 9: Top 10 RF Features (Importance)
- Figure 10: SHAP Summary Plot (Random Forest - Bar)
- Figure 11: SHAP Summary Plot (XG-Boost - Dot)
- Figure 12: PCA Cumulative Explained Variance
