UNIVERSITY OF MORATUWA

Faculty of Engineering

Registered Module No: CS4743

Leukemia Gene Expression Analysis Report

Project Report

Date of Submission:

April 6, 2025

Isuru Gunarathne - 200189M

Supervisors:

Dr. Charith Chitraranjan (University of Moratuwa)

Department of:

Computer Science and Engineering

Abstract

This report details an analysis of gene expression data from patients with Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). We developed an interpretable logistic regression model to predict leukemia type with high accuracy, identified key gene expression patterns that differentiate AML and ALL, and applied unsupervised clustering to further explore the data structure. Our supervised model achieved a test accuracy of approximately 97%, while unsupervised clustering resulted in a clustering purity of about 73.6%. The analysis reveals distinct gene markers, such as those related to myeloid and lymphoid differentiation, which are consistent with known biological characteristics of these leukemias.

Contents

1 Introduction

Gene expression profiling offers a powerful means of distinguishing between different leukemia subtypes by capturing the intricate patterns of gene activity across the genome [1]. In this project, we analyze microarray data comprising expression levels for over 7000 genes measured in patients diagnosed with either Acute Myeloid Leukemia (AML) or Acute Lymphoblastic Leukemia (ALL) [2, 3]. The high dimensionality of the data provides both opportunities and challenges, as it contains rich biological information that can be harnessed to understand the underlying mechanisms of these diseases.

The primary objectives of this analysis are as follows:

1. Develop an interpretable classification model that accurately differentiates AML from ALL. We employ logistic regression to leverage its simplicity and interpretability, allowing us to directly relate model coefficients to gene expression patterns.

2. Identify key gene expression patterns that characterize each leukemia type. By examining the most influential genes and their biological functions, we aim to uncover the molecular signatures that define AML and ALL.

3. Explore the natural grouping of samples through unsupervised clustering techniques. Utilizing methods such as k-means clustering and principal component analysis (PCA), we investigate whether the inherent structure in the gene expression data can reveal distinct clusters corresponding to the leukemia subtypes.

This comprehensive approach not only facilitates accurate disease classification but also enhances our understanding of the biological differences between AML and ALL, potentially guiding future research and therapeutic strategies.

2 Data Description

The analysis utilizes several key datasets, each serving a distinct role in our investigation of gene expression in leukemia patients. These datasets are:

- **Training Data (`data_set_ALL_AML_train.csv`)**: This dataset contains gene expression measurements for patients diagnosed with either AML or ALL. It includes expression values for over 7000 genes along with associated gene information such as gene descriptions and accession numbers. The training data is used to develop and tune the classification model [4].

- **Test Data (`data_set_ALL_AML_independent.csv`)**: An independent dataset reserved exclusively for evaluating the performance of the trained model. By using a separate test set, we can assess the generalizability and robustness of the model on unseen data [4].

- **Actual Labels (`actual.csv`)**: This file provides the true leukemia type (AML or ALL) for each patient. It is essential for assigning ground truth labels to the samples and for validating the predictions made by our classification model [4].

Note: The data has been pre-normalized to ensure consistency across samples. Preprocessing steps included removing non-numeric "call" columns used for quality control and transposing the data matrix so that each row represents an individual patient and each column corresponds to a specific gene feature. This format is optimal for applying machine learning techniques in the subsequent analysis.

3 Methodology

3.1 Data Preprocessing

The raw training data was loaded and non-numeric columns (e.g., "call" columns) were removed. The dataset was then transposed to form an expression matrix where rows correspond to patient samples and columns correspond to genes. Additionally, a mapping from each gene's row index to its "Gene Description" was created for later interpretation [4].

3.2 Supervised Classification Model

We employed logistic regression due to its interpretability [5]. The model was trained on the training dataset, and regularization was applied to manage the high dimensionality of the data (over 7000 features) relative to the number of samples. The trained model was then evaluated on an independent test set to avoid overfitting [1].

3.3 Model Interpretation

To gain biological insights, we analyzed the logistic regression coefficients:

- **Positive coefficients:** Indicate genes more highly expressed in AML.
- **Negative coefficients:** Indicate genes more highly expressed in ALL.

We extracted the top 5 genes for AML and the top 5 genes for ALL. Each gene is labeled with its row number and the first word of its description (e.g., "6200 MPO").

3.4 Unsupervised Clustering

In addition to supervised classification, unsupervised clustering (using k-means with $k = 2$) was applied. The combined dataset was standardized prior to clustering. The clustering results were then compared to the true labels to compute clustering purity. Furthermore, Principal Component Analysis (PCA) was used to visualize the clustering and assess the natural grouping of the samples [3].

Code Repository: `https://github.com/IsuruGunarathne/AML-ALL-Classifier`

4 Results

4.1 Classification Performance

The logistic regression model achieved a perfect fit on the training data, with 100% accuracy, demonstrating that the model can fully separate the training samples. When evaluated on the independent test set, the model maintained a high level of performance with an accuracy of approximately 97%. The confusion matrix further illustrates this performance:

- **ALL:** Out of 20 samples, 19 were correctly classified and 1 was misclassified as AML.

- **AML:** All 14 samples were correctly classified.

This high performance on unseen data indicates that the model generalizes well despite the high dimensionality of the feature space.

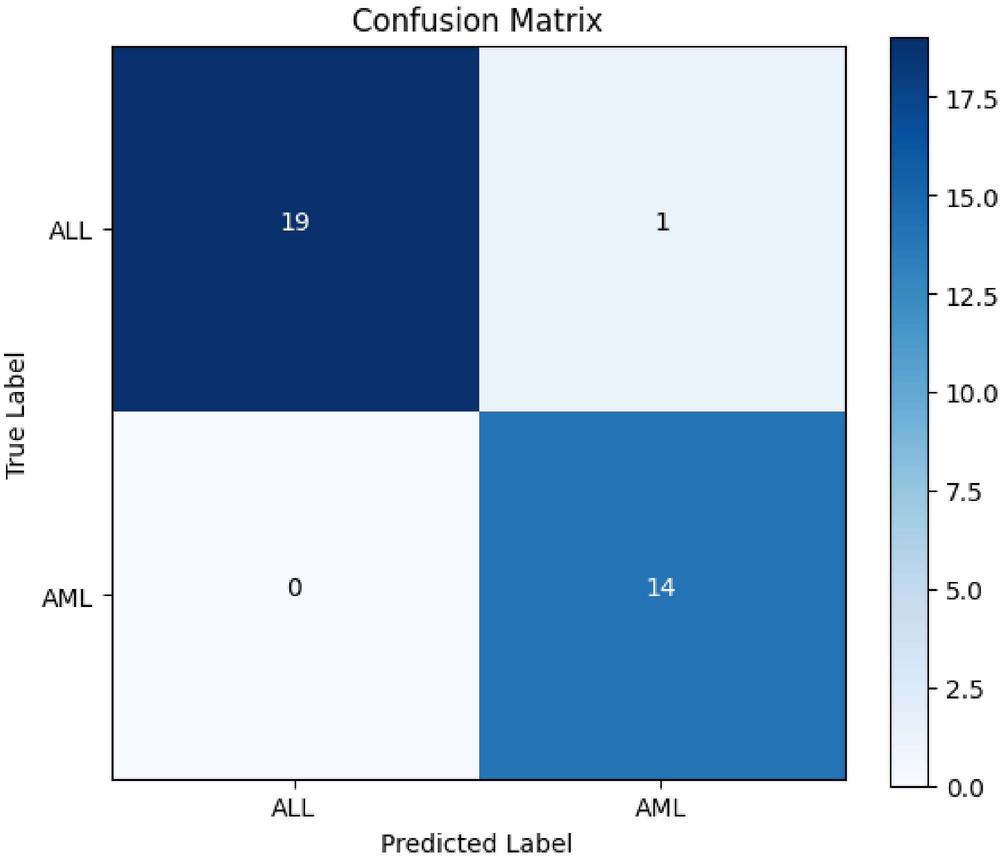

Figure 1: Confusion Matrix for supervised classification

4.2 Interpretation of Gene Expression Patterns

The analysis of logistic regression coefficients provides key insights into the gene expression patterns that distinguish AML from ALL. Specifically, a positive coefficient suggests that the corresponding gene is more highly expressed in AML, while a negative coefficient indicates higher expression in ALL. We extracted the top 5 genes with the highest positive coefficients and the top 5 with the most negative coefficients. Each gene is labeled with its row number and the first word of its description (e.g., “6200 MPO”) for brevity and clarity.

- For AML, genes such as **MPO** and **Vimentin** showed significantly positive coefficients, indicating their strong association with AML.
- For ALL, markers like **PTMA** were identified with highly negative coefficients, underscoring their relevance in ALL.

The balanced bar chart, with red bars representing AML-associated genes and blue bars representing ALL-associated genes, visually summarizes these influential markers.

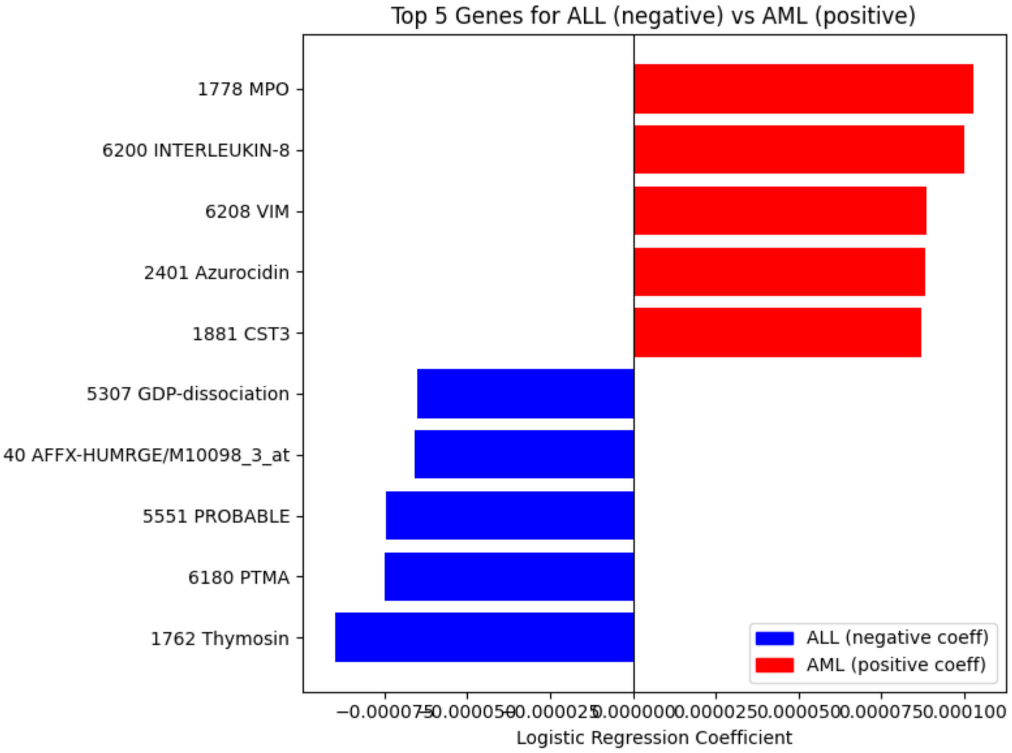

Figure 2: Top 5 genes for ALL and AML

4.3 Clustering Analysis

Unsupervised clustering was performed using the k-means algorithm with $k = 2$ on the standardized gene expression data. The resulting clusters were compared with the true labels to calculate a clustering purity of approximately 73.6%, indicating that the unsupervised approach was able to capture a significant portion of the underlying class structure, albeit with some misclassifications. To visualize these results, Principal Component Analysis (PCA) was applied to reduce the high-dimensional data to two principal components. The PCA scatter plot shows a general separation between AML and ALL samples, though some overlap is evident. This overlap likely reflects underlying biological heterogeneity or the presence of additional subtypes within the leukemia classifications.

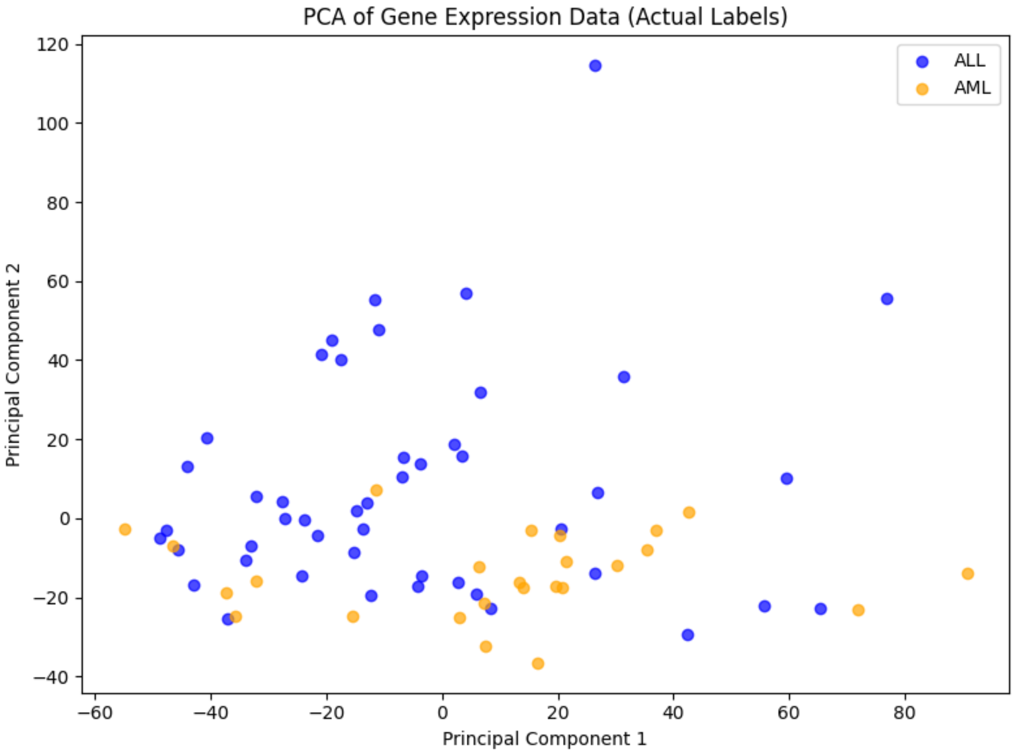

Figure 3: Clustering - colored by actual labels

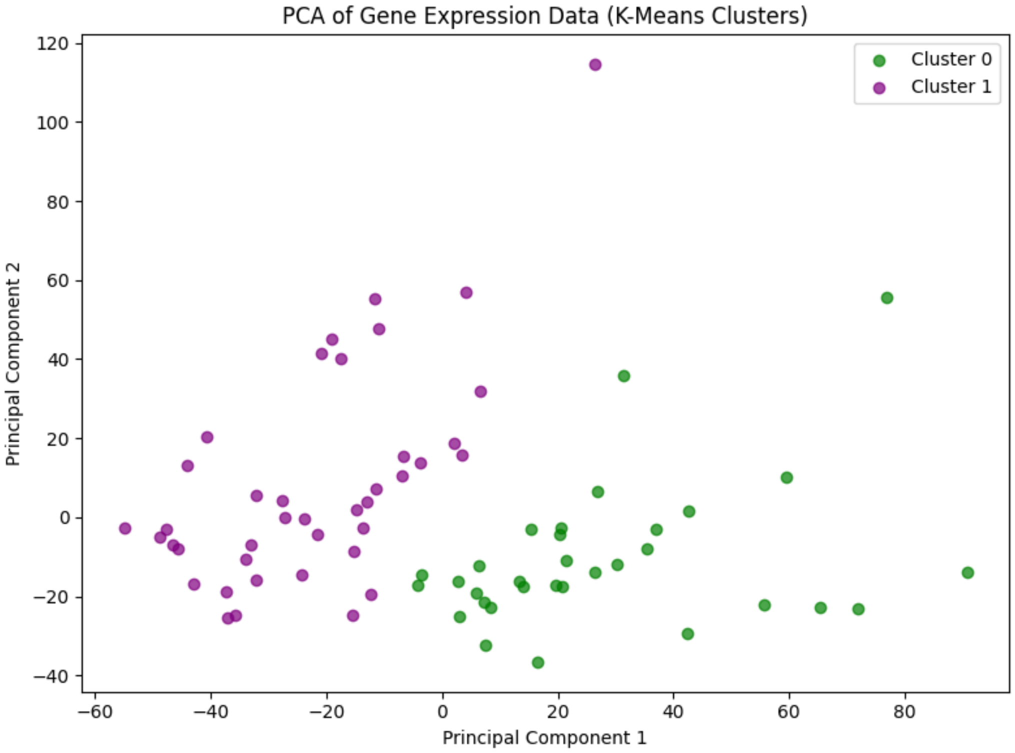

Figure 4: Clustering - colored based on k-means clusters

5 Discussion

Supervised Learning:

The logistic regression model performed with high accuracy on both the training and test datasets [1]. Its interpretability allowed us to directly associate specific genes with either AML or ALL. For instance, the identification of genes such as **MPO** (for AML) corroborates known biological insights into myeloid differentiation [2].

Unsupervised Learning:

The clustering results, with a purity of around 73.6%, suggest that while gene expression data inherently carries the signal differentiating AML and ALL, additional factors (such as underlying subtypes within ALL) may influence the clustering outcome [3]. Unsupervised methods, therefore, offer a valuable approach for discovering novel subgroups within the data that might not be evident from supervised analysis alone.

Limitations:

The high dimensionality of the data relative to the number of samples poses significant challenges, including potential overfitting and reduced model generalizability [5]. Future work could involve incorporating feature selection or dimensionality reduction techniques to enhance both the robustness and interpretability of the models [6, 7].

6 Conclusion

Conclusion

This analysis demonstrates that an interpretable logistic regression model can effectively differentiate between AML and ALL based on high-dimensional gene expression data. The supervised approach yielded near-perfect accuracy on the training set and approximately 97% accuracy on the test set, underscoring the model's robustness.

Key gene markers, identified through the examination of model coefficients, provide biologically meaningful insights into the underlying differences between the leukemia types. For instance, markers such as **MPO** and **Vimentin** in AML, and **PTMA** in ALL, not only validate known biological associations but also suggest potential targets for further research.

Unsupervised clustering, while not perfectly aligning with the known classes, revealed additional structure in the data. This indicates that there may be underlying subgroups or further heterogeneity within the leukemia samples, which could be crucial for understanding disease complexity and guiding future studies.

Future work may include exploring advanced explainable AI techniques, such as SHAP and LIME, to better elucidate the contribution of individual genes to model predictions. Furthermore, integrating additional biological data—such as genomic mutations, proteomic profiles, or epigenetic markers—could refine these findings and contribute to a more comprehensive understanding of leukemia pathogenesis.

References

[1] T. R. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class

discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[2] P. J. M. Valk, R. G. W. Verhaak, M. Beijen, C. Erpelinck, S. Barjesteh van Waalwijk van Doorn-Khosrovani, J. Boer, H. B. Beverloo, W. L. J. van Putten, A. Kelder, J. Valk *et al.*, "Prognostically useful gene-expression profiles in acute myeloid leukemia," *New England Journal of Medicine*, vol. 350, no. 16, pp. 1617–1628, 2004.

[3] E. J. Yeoh, M. Ross, S. Shurtleff, J. Gaudet, C. F. Connelly, C.-H. Pui, and S. Raimondi, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Blood*, vol. 99, no. 6, pp. 2089–2099, 2002.

[4] Crawford, "Gene expression data," https://www.kaggle.com/datasets/crawford/gene-expression, 2019, accessed: 2025-04-05.

[5] E. R. Gamazon *et al.*, "Using machine learning methods to predict disease: A review," *Journal of Genetics*, 2019.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2016, pp. 1135–1144.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.