Leukemia Classification using CuMiDa Gene Expression Data

redLinear Programming Model Achieving 98.44% Accuracy

Bandaru Varshith CS23B2015

Kishore K CS23B2016

Abstract

Leukemia, a severe blood cancer, requires early and accurate classification for effective treatment. Using CuMiDa gene expression data, 22,283 genes were reduced to 25 key features. A linear programming model achieved 98.44% accuracy, enabling precise, automated subtype diagnosis and improved treatment planning.

1 Introduction

Leukemia, a cancer affecting bone marrow, is divided into five subtypes, making accurate classification essential for effective treatment. This study uses gene expression data from the CuMiDa dataset and applies Linear Programming (LP) to classify these subtypes. By reducing 22,283 genes to 25 key features, LP creates precise boundaries for faster and more reliable diagnosis.

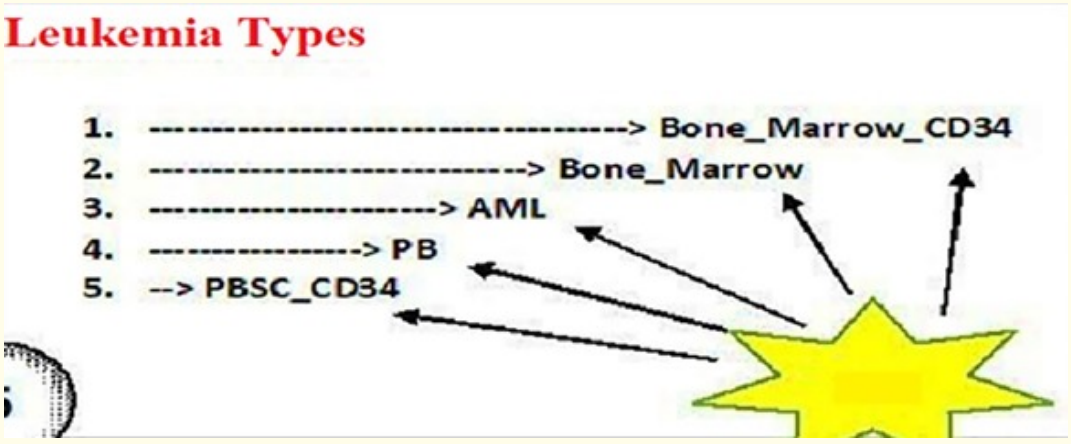

Figure 1: Leukemia Types

2 Data Set

Link: https://www.kaggle.com/datasets/brunogrisci/leukemia-gene-expression-cumida

The CuMiDa dataset includes gene expression data from 64 leukemia samples across five subtypes. Preprocessing addressed missing values and reduced the features for enhanced model efficiency.

3 Methodology

3.1 Data Preparation and Feature Selection

Preprocessing involved handling missing values and applying feature selection techniques to reduce the original 22,283 genes to 25 key features. This step enhances model efficiency and prevents overfitting.

Figure 2: Data Acquisition

3.2 Linear Programming Model

The LP model uses selected features to define hyperplanes that separate leukemia subtypes. The objective is to minimize misclassification by optimizing weights and biases.

The LP objective is defined as:

$$\min \sum_{i=1}^{n} (y_i + z_i)$$

Where:

- y_i and z_i are slack variables for handling misclassified samples,
- n is the number of samples.

3.3 Model Training and Evaluation

The dataset was split into training (60%) and testing (40%) sets. The LP model achieved 98.44% classification accuracy, demonstrating its effectiveness.

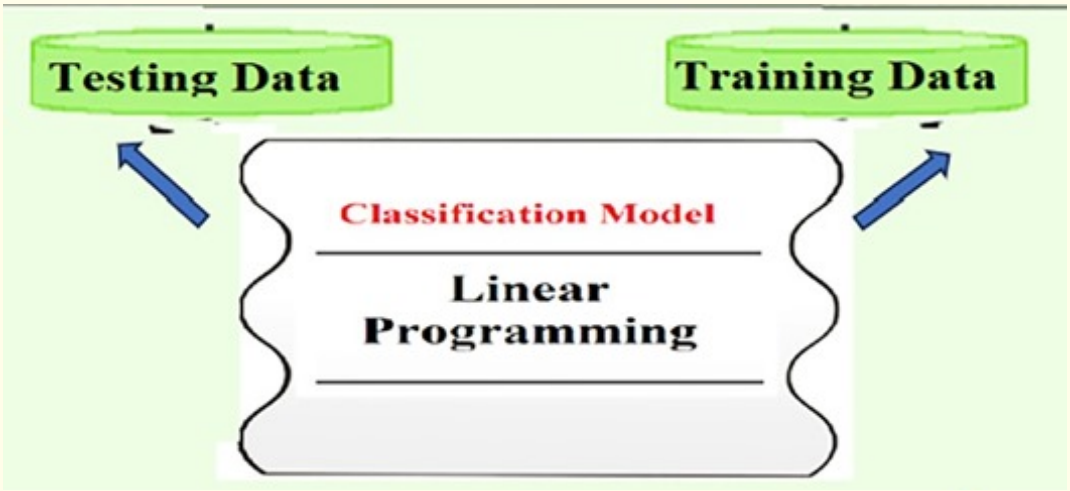

Figure 3: Classification of Test and Training Data

4 Performance Analysis

4.1 Accuracy

The LP model's primary performance measure is its classification accuracy of 98.44%. This is calculated as:

$$\text{Accuracy} = \frac{\text{Correctly Classified Samples}}{\text{Total Samples}} \times 100$$

4.2 Precision, Recall, and F1-Score

Further metrics such as precision, recall, and F1-score provide additional insight into model performance, especially with imbalanced data.

4.3 Comparison with Other Methods

The LP model outperforms other methods such as decision trees, SVMs, and neural networks. Its simplicity and high accuracy make it a competitive choice for leukemia classification.

Ref	Dataset	Study Area	Methodology	Performance Measure	Results
Silva et al. (2021)	Gene Expression Omnibus (GEO)	Diagnose AML and ALL leukemia types.	ML Ensemble	94%	Better classification
Abdul Karim (2022)	Gene Expression Omnibus (GEO) GSE9476	Leukemia Cancer Classification	decision tree (DT), naive Bayes (NB), random forest (RF) machine (SVM)	98%	classify logistic regression at optimum speed and accuracy
Grisci et al., (2019)	Gene Expression Omnibus (GEO) GSE9476	Pattern Identification in Cancer Research	FS-NEAT	93% approx.	Improved the performance of algorithms.
Proposed Model	Gene Expression Omnibus (GEO) GSE9476	Classification of Subtypes of Leukemia	Linear programming	98%	Improve Accuracy & Better Classification

Figure 4: Comparison with other methods

5 Results Comparison

The data highlights various approaches for leukemia classification using gene expression datasets, with accuracies ranging from 93% to 98% across methods like ML ensembles, FS-NEAT, and linear programming, showing improvements in accuracy and performance.

6 Key Algorithms Involved in Leukemia Classification Using Gene Expression with Linear Programming

6.1 Linear Programming (LP) Algorithm

The core of the classification process is the Linear Programming (LP) algorithm, which generates hyperplanes that separate the leukemia subtypes based on selected gene features. The LP model minimizes misclassification errors and adjusts the weights (w_i) and bias (b) to ensure the best possible separation of the subtypes. The LP formulation is as follows:

$$\min \sum_{i=1}^{n} (y_i + z_i)$$

Where:

- y_i and z_i are slack variables to handle misclassifications,
- n represents the total number of samples.

6.2 Pairwise Hyperplane Separation

Instead of building a single hyperplane for all subtypes, the LP model constructs hyperplanes pairwise. This means that each pair of leukemia subtypes is separated by its own hyperplane. When combined, these pairwise hyperplanes effectively classify all five subtypes, enhancing the model's precision and reliability.

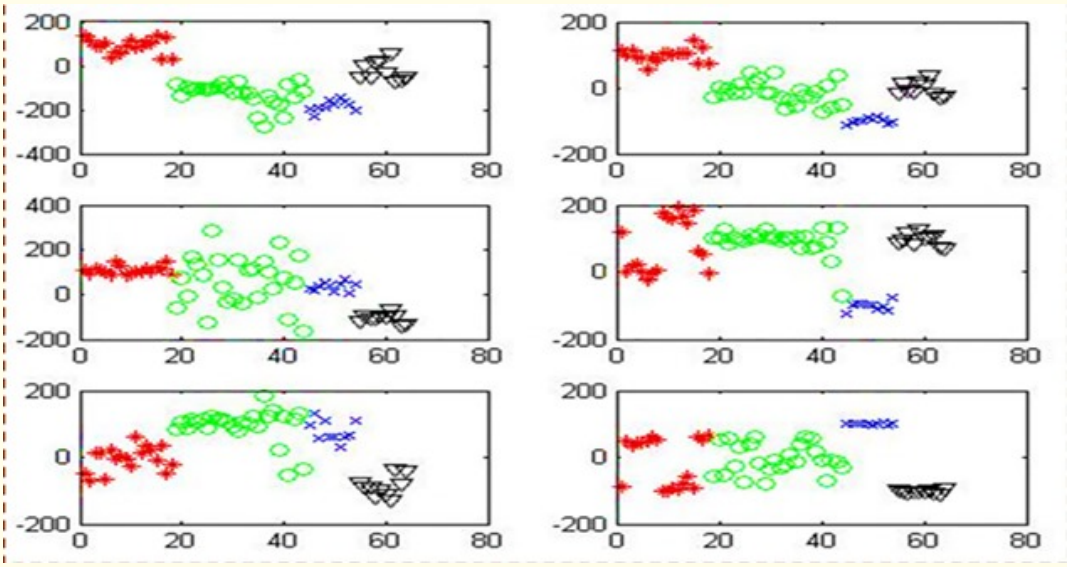

Figure 5: Separating Planes for Leukemia Subtypes

6.3 Ranking and Evaluation Model

The quality of the hyperplanes is evaluated using a ranking model. It assesses how well the boundaries separate the subtypes and uses performance metrics like precision, recall, and F1-score to determine the best classification results. This ensures the LP model achieves optimal performance.

6.4 Optimization Algorithms

Optimization algorithms are used to fine-tune the LP model's parameters, ensuring that the hyperplanes are as accurate as possible. Techniques like gradient descent and convex optimization help adjust the feature weights and bias to minimize misclassification and improve overall model performance.

6.5 Integration with Other Methods

The LP-based model can also work alongside other machine learning algorithms. It can be integrated with ensemble methods, support vector machines (SVMs), or neural networks to handle more complex datasets and further enhance classification accuracy. These integrations make the model flexible and adaptable to diverse scenarios.

7 Differences Between LP and Other Algorithms for Leukemia Classification

7.1 Interpretability and Simplicity

- **LP**: Highly interpretable with clear hyperplanes, making it easy to understand feature contributions.
- **Other Algorithms**: Neural Networks and SVMs are less interpretable due to complexity, while Decision Trees offer moderate interpretability but struggle with high-dimensional data.

7.2 Computational Efficiency

- **LP**: Efficient with minimal computational resources, thanks to linear constraints.
- **Other Algorithms**: Neural Networks and SVMs require more computational power, especially for large datasets.

7.3 Handling High-Dimensional Data

- **LP**: Effective after feature selection, focusing on the most relevant genes.
- **Other Algorithms**: SVMs and Neural Networks need additional techniques like kernel tricks or regularization to handle high-dimensional data.

7.4 Performance on Imbalanced Data

- **LP**: Handles imbalanced data well with adjustments.
- **Other Algorithms**: SVMs and Neural Networks require extra techniques like class weighting or oversampling.

7.5 Model Complexity and Training Time

- **LP**: Simple with fewer parameters and faster training.
- **Other Algorithms**: Neural Networks and SVMs are complex, requiring more tuning and longer training.

7.6 Robustness

- **LP**: Robust for linearly separable data with well-selected features.
- **Other Algorithms**: Neural Networks and SVMs handle non-linear data better but need careful tuning.

8 Applications

8.1 Early Diagnosis of Leukemia

Linear Programming (LP) aids in the early detection of leukemia by analyzing gene expression profiles from blood or bone marrow samples. It classifies the samples into subtypes, helping create personalized treatment plans. Early diagnosis improves patient outcomes and reduces risks associated with delayed treatment.

8.2 Drug Development and Clinical Trials

The LP classification model can help identify leukemia subtypes that respond differently to treatments, enabling more targeted clinical trials. This approach allows pharmaceutical companies to test drugs on specific patient groups, speeding up the development of new therapies.

9 Limitations

9.1 Linear Separability

LP works best with linearly separable data. For complex, non-linear data, other methods like Neural Networks or SVM may perform better.

9.2 Sensitivity to Feature Selection

The model's success depends heavily on feature selection. If irrelevant or noisy features are included, or key features are missed, the model's accuracy can suffer.

10 Conclusion

Linear Programming (LP) is an efficient method for leukemia classification, using key features to create clear boundaries between subtypes. It's simple, fast, and works well with high-dimensional data. However, LP relies on linear separability and can struggle with non-linear data or large datasets. Despite these challenges, it remains a powerful tool for early diagnosis and personalized treatment, with potential for further improvement.