

AI-Driven Detection and Staging of Acute Lymphoblastic Leukemia from Gene Expression and Morphological Characteristics

Senthil Vadivu M¹, Mani Shankar M² and Kavitha R³

¹⁻³Statistics and Data Science, Christ University, Bengaluru, India

Email: senthil.vadivu@christuniversity.in, mani3shankar3@gmail.com, kavitha.rajamani@christuniversity.in

Abstract— Acute Lymphoblastic Leukemia (ALL) is a rapidly progressing type of blood cancer, and early diagnosis plays a key role in improving patient care. This work introduces a two-phase approach that combines data from gene expression profiles and blood smear images to support both detection and staging of ALL. In the first phase, gene data is analyzed using Random Forest and Logistic Regression models to identify whether a sample is ALL-positive. The second phase focuses on classifying disease progression by analyzing microscope images into four stages: Benign, Early, Pre-leukemic, and Progressive. This is achieved through three strategies: a custom Convolutional Neural Network (CNN), a fine-tuned ResNet-50 model, and a hybrid model where deep features from ResNet are used with a Support Vector Machine (SVM). Among these, the hybrid method showed the most consistent performance. The integration of molecular and image data provides a reliable tool that may help enhance ALL diagnosis in clinical settings.

Index Terms— Acute Lymphoblastic Leukemia, gene expression analysis, blood smear classification, machine learning, deep learning, Random Forest, logistic regression, convolutional neural network, ResNet-50, SVM.

I. INTRODUCTION

Acute Lymphoblastic Leukemia (ALL) is a fast-moving cancer that affects the blood and bone marrow, often leading to a buildup of immature lymphoid cells. Though it can develop at any age, children are most frequently diagnosed [6, 7]. The urgency of early detection in ALL is well known—it can significantly affect how well patients respond to treatment. To identify the disease, doctors typically examine blood smears, run molecular tests, or use immunophenotyping [1, 2]. These techniques have been standard for years and are quite effective. However, they are not always quick, and the results can depend heavily on who is interpreting them. Some hospitals may also lack the resources needed to carry out these tests efficiently [9][12].

In the last decade, there has been a big shift toward collecting more complex data in medicine. This includes both gene expression data and medical images, which can now be stored and processed at scale [8]. As a result, researchers have been exploring ways to use computational tools like machine learning and deep learning to make sense of this information and assist in medical decisions [6, 10]. This research works explore a diagnostic method that brings together gene and image data to detect and classify ALL. The process happens in two stages.

First, gene expression profiles are used with machine learning models like Random Forest and Logistic Regression to check if the sample suggests ALL [1, 10]. This step includes selecting important genes and simplifying the data. If ALL is detected, the next step is to look at blood smear images to see how far the disease has progressed. These are sorted into four categories: Benign, Early, Pre-leukemic, and Progressive. Three different models are considered for this research: a custom CNN [11], a fine-tuned ResNet-50 [5], and a hybrid approach that takes features from ResNet and feeds them into an SVM [5]. By combining both molecular and visual inputs, this system is designed to make ALL diagnosis more consistent, especially in places where traditional lab testing is limited or time-consuming.

II. RELATED WORK

Many researchers in recent years have worked on using computational methods to support leukemia diagnosis. These efforts have generally focused on gene expression analysis, image-based approaches, or both—each offering unique strengths in different aspects of detection and classification. A key early study by Golub et al. [1] showed that gene expression profiles from microarray data could successfully separate leukemia subtypes. This work marked an important shift toward using supervised learning in cancer classification. Later, Dudoit et al. [2] compared a variety of classifiers and emphasized the importance of feature selection when dealing with high-dimensional biological data. More recent methods like the one proposed by Ilyas et al. [10] used linear programming with built-in feature selection. Their technique achieved high accuracy while being computationally lightweight. Similarly, the MD- ALL platform by Wu et al. [8] demonstrated the use of RNA-seq data for the robust molecular diagnosis of ALL, combining multiple omics layers for subtype prediction. On the imaging side, CNN-based techniques have seen substantial growth. Said et al. [3] and Dasariraju et al. [4] applied deep learning to classify leukemia from blood smear images. Despite working with relatively small datasets, they achieved strong performance using transfer learning strategies. Shafique and Tehsin [9] customized the AlexNet architecture to distinguish between ALL subtypes and healthy cells, although they did not include staging. Togacar et al. [5] developed a hybrid method that combined deep features from ResNet-50 with an SVM classifier—a strategy that influenced part of our stage two approach. Ramesh and Thouti [11] applied CNNs with data augmentation but focused only on binary classification, omitting stage-wise differentiation. In terms of broader reviews, Saba et al. [7] provided an overview of recent deep learning applications in ALL diagnosis, outlining common challenges and directions for improvement. Ghaderzadeh et al. [6] proposed an end-to-end deep learning model tailored to classify leukemic cells in microscopic images, emphasizing the clinical potential of AI in hematopathology. Recent surveys, such as Jha and Dutta [13], have highlighted the rising importance of hybrid frameworks—especially combinations of CNNs with traditional classifiers like SVM—for improving diagnostic accuracy in medical imaging. Additionally, Khan et al. [15] showed that coupling CNNs with SVM classifiers provides enhanced robustness and interpretability in leukemia detection tasks. Beyond individual modalities, Yu et al. [14] demonstrated the power of multimodal learning to integrate both imaging and genomic data, setting a new direction for holistic cancer diagnosis frameworks. While these studies have contributed significantly, most focus on either gene expression or image data, and few attempt to combine them. Our work builds on these foundations by integrating both molecular and morphological analysis in a two-stage pipeline to detect ALL and assess its progression more comprehensively.

III. METHODOLOGY

The proposed framework is structured as a two-stage diagnostic system for the identification and progression analysis of Acute Lymphoblastic Leukemia (ALL). This system integrates molecular data analysis and image-based evaluation to provide a more comprehensive approach to leukemia diagnosis. As illustrated in Figure 1, the first stage focuses on analyzing gene expression data. This involves collecting gene-level information from patient samples, followed by data preparation and filtering steps. The goal of this stage is to identify whether the sample is indicative of ALL. A general classification model is applied to distinguish ALL-positive cases from other leukemia types or normal conditions.

If a case is classified as ALL-positive, it progresses to the second stage. Here, a corresponding microscopic image of the patient’s blood smear is used to assess the progression of the disease. This visual data undergoes preprocessing and is then passed through a separate analysis pipeline. The task at this stage is to assign the image to one of four diagnostic categories: Benign, Early, Pre-leukemic, or Progressive.

The two stages work together in a conditional sequence. Molecular-level data is first used to screen for ALL, and only positive cases proceed to image-based stage classification. This design helps optimize computational

efficiency while also leveraging the strengths of both data types. The overall aim of the methodology is to create a diagnostic tool that is reliable, interpretable, and adaptable to real-world clinical settings.

IV. EXPERIMENTATION

The overall process in this study is split into two stages as in Fig 1. The first stage focuses on identifying ALL using gene expression data, while the second deals with determining the disease stage based on blood smear images.

A. Stage 1: Gene Expression-Based Detection

The first stage begins with gene expression data collected from three leukemia types: ALL, AML, and CLL. Since the gene sets were not identical across the datasets, we worked only with the genes they had in common. Each sample was formatted so that genes became columns, and each patient made up a row. To get the data ready for modeling, we applied the following steps:

- Genes that did not vary much across patients were removed using a variance filter. A Random Forest was trained to rank the remaining genes by importance. We kept the top 10. The selected data was scaled, and PCA was used to reduce the features to just one dimension. We then trained two models to classify samples: Random Forest and Logistic Regression. Both were tested using five-fold cross-validation to check consistency.

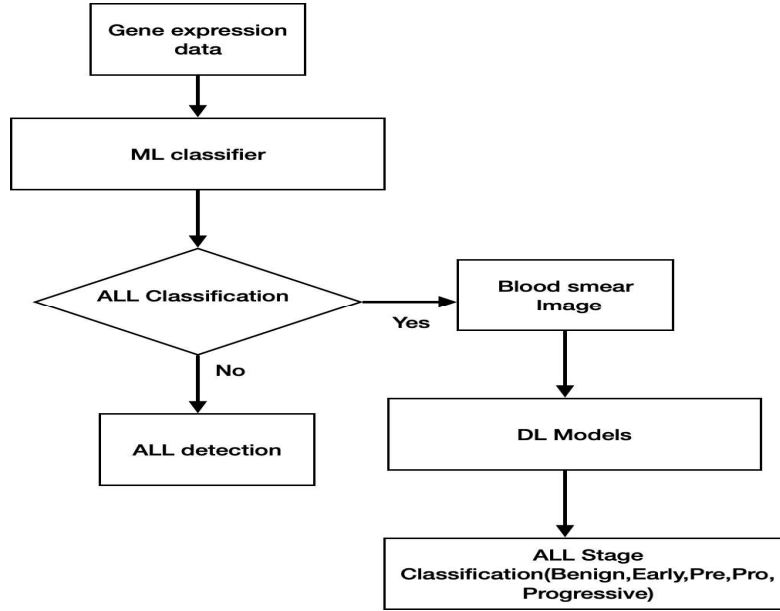


Figure 1: General framework for ALL detection and staging.

Random Forest

This model was kept intentionally simple: just 10 trees, a depth limit of 3, and thresholds set to ensure splits only occurred with at least 40 samples, and no fewer than 20 in any leaf node. Besides classification, the Random Forest also helped with identifying the most informative genes.

Logistic Regression

Here, we used a more constrained version with L2 regularization and a small C value (0.001) to prevent overfitting. It worked well especially once PCA was applied despite its simpler structure. To better understand the model behavior, we also plotted a learning curve for Random Forest to observe how performance changed with more training data.

B. Stage 2: Classification Using Blood Smear Images

For patients identified as ALL-positive in the first stage, the next step is to analyze their blood smear images to determine how far the disease has progressed. These images are categorized into four stages: Benign, Early, Pre-leukemic, and Progressive. To perform this classification, we explored three separate deep learning approaches.

Custom CNN Model

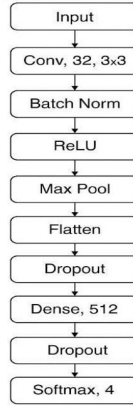


Figure 2: Custom CNN architecture used for stage classification.

The first approach involved building a Convolutional Neural Network (CNN) from scratch. The network included three convolutional layers with 32, 64, and 128 filters. After each convolutional layer, we added batch normalization, ReLU activation, and max pooling to manage feature scaling and dimensionality. The output of these layers was flattened and passed into a fully connected layer with 512 neurons and dropout. Finally, the network ended with a softmax layer to classify the image into one of the four stages as shown in Figure 2. Training was done using CrossEntropyLoss and the Adam optimizer. This model served as a baseline for comparison.

Transfer Learning with ResNet-50

In the second method, we used a pretrained ResNet-50 model. The final classification layer was replaced with a new fully connected layer tailored to output four classes. To retain useful low-level features from the pretrained model, the early layers were frozen, and only the deeper layers were fine-tuned on our dataset as shown in Figure 3. By starting with pretrained ImageNet weights, this model reached convergence faster and showed better generalization on unseen data.

Hybrid: ResNet Features + SVM Classifier

Finally, we built a hybrid model that combined the strengths of both CNNs and traditional machine learning. After training the ResNet-50 model, we extracted deep feature vectors from its penultimate layer. These high-level representations were then flattened and used to train a separate linear Support Vector Machine (SVM) as shown in Figure 4. This setup allowed us to pair powerful deep feature extraction with the decision-making strength of SVMs. Among the three approaches, this hybrid pipeline delivered the most accurate results.

C. Dataset Sources

Stage 1: Gene Expression. The gene data used in Stage 1 was sourced from the NCBI GEO repository. It included samples from three leukemia types: Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia

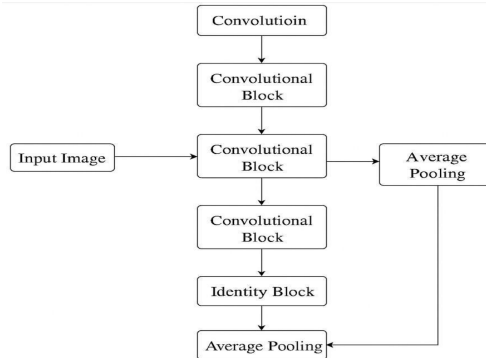


Figure 3: Adapted ResNet-50 model used for ALL stage classification

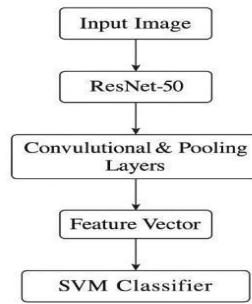


Figure 4: Hybrid ResNet + SVM classifier for stage prediction using deep features.

(AML), and Chronic Lymphocytic Leukemia (CLL). Since the gene lists across these datasets were not identical, we focused only on the genes that were shared by all three. This led to a final set of 45 common genes that were used in the analysis.

Stage 2: Blood Smear Images. For the second stage, we worked with a Kaggle dataset¹ that provides blood smear images labeled according to leukemia progression. These images fall into four categories: Benign, Early, Pre-leukemic, and Progressive.

D. Stage 1 Dataset Structure and Preprocessing

Originally, the gene expression data came in a format where gene names were rows and samples were columns. To make things consistent, the datasets were transposed so that each patient became a row and the genes became features. Once the format was aligned, we needed to clean the data before feeding it into the model. Any missing gene values were filled with the average value for that gene. Then, to eliminate features that did not contribute much, we filtered out genes with very low variance — setting the threshold at 0.1. To figure out which genes were the most useful for classification, we trained a basic Random Forest and used it to rank gene importance. From that, we picked the top 10 genes. At this point, we had a much smaller and more informative feature set. We normalized the data using z- score scaling to make sure all values were on the same scale, and then applied PCA to reduce everything to one main component for modeling.

This version of the dataset — filtered, scaled, and compressed — was used as input for the classifiers in Stage 1.

E. Preprocessing and Feature Selection (Stage 1)

After reformatting the gene expression data, the next step involved narrowing down the feature space. A variance threshold of 0.1 was applied first, which helped remove genes that showed very little change across patient samples. This step filtered out redundant information and helped highlight more informative features.

To further refine the dataset, a Random Forest classifier was trained to score the remaining genes based on their predictive contribution. From this ranking, the top 10 genes were chosen for modeling. Finally, to simplify the dataset and reduce noise, Principal Component Analysis (PCA) was used. This reduced the feature set to a single principal component that captured the essential variation in the data.

F. Model Evaluation

We tested two classification models during this stage: Random Forest and Logistic Regression. Both were assessed using stratified 5-fold cross-validation to ensure that class distributions were balanced in every fold.

Effect of Variance Threshold

To find the right cutoff for low-variance genes, we tested several thresholds ranging from 0.01 to 0.3. The results are shown in Figure 5. While Random Forest performed well across all thresholds, Logistic Regression showed more sensitivity to the choice of features and scaling. Its performance dropped when noise was not effectively removed.

Based on these tests, we settled on a threshold of 0.1. It struck a good balance, eliminating unhelpful features without sacrificing accuracy, and leaving enough variation for meaningful classification and dimensionality reduction.

Effect of Number of Selected Features

We also tested how the number of input genes affected performance. Using the feature importance rankings from Random Forest, we trained both models on the top 10, 20, and 50 genes. The accuracy trends are shown in Figure 6.

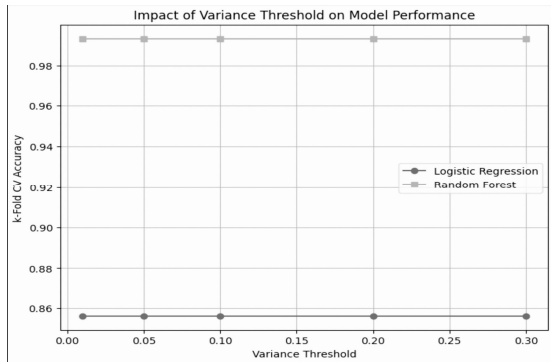


Figure 5: Impact of variance threshold on model performance

Random Forest remains robust, while Logistic Regression is more affected by feature variability. Random Forest remained strong across all gene counts, though its accuracy peaked slightly when using 10 features. Logistic Regression stayed consistent but didn’t reach the same level of performance, likely due to its linear assumptions and limited capacity to model complex relationships.

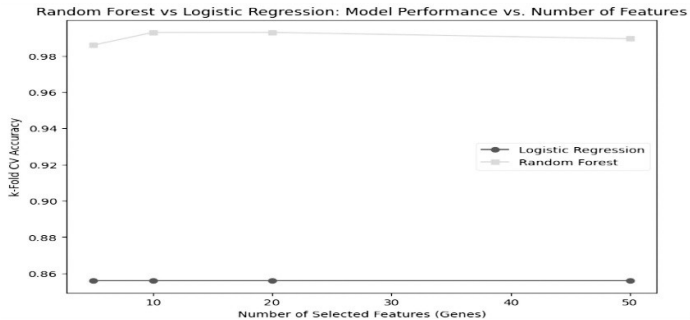


Figure 6: Effect of number of selected common genes on model accuracy

Random Forest consistently performs better than Logistic Regression.

G. Results Summary: Stage 1 Classification

TABLE I: STAGE 1 CLASSIFICATION ACCURACY (K-FOLD CV)

Model	Variance Threshold	Accuracy
Random Forest	0.1	0.991
Logistic Regression	0.1	0.857

Table 1 shows that Random Forest delivered the highest classification accuracy with minimal tuning. Logistic Regression, although simpler, maintained moderate performance throughout. Overall, Random Forest handled feature filtering and dimensionality reduction steps more effectively, making it a better fit for the gene expression data used in this stage.

H. Stage 2 Dataset Structure and Preprocessing

The dataset used in the second stage included RGB images of blood smears, categorized into four stages of leukemia progression: Benign, Early, Pre-leukemic, Progressive. Image quality and size varied across samples. To prepare the data, all images were resized to 224 × 224 pixels and converted to RGB where necessary. Standard normalization was applied using ImageNet’s mean and standard deviation values: mean = [0.5, 0.5, 0.5], std = [0.5, 0.5, 0.5]

To improve generalization, the training set was augmented with random horizontal flips and rotations. The dataset was then split into 80% training and 20% testing subsets using stratified sampling to maintain class balance.

I. Stage 2: Classification Results

Three deep learning models were used to classify the progression stage of ALL based on the blood smear images: a custom-built Convolutional Neural Network (CNN), a fine-tuned ResNet-50, and a hybrid model that combined ResNet-based deep features with a Support Vector Machine (SVM) classifier.

All models were trained and evaluated on the same data split. Table 2 provides a summary of their performance.

TABLE II: STAGE 2 CLASSIFICATION RESULTS (TEST SET, 652 SAMPLES)

Model	Accuracy (%)	F1-score (weighted)
Custom CNN	98.77	0.9877
ResNet-50	99.69	0.9969
Hybrid (ResNet + SVM)	100.0	1.0000

Among the three, the hybrid model produced perfect classification accuracy and F1-score on the test set. Both CNN and ResNet-50 also delivered strong results, with ResNet-50 slightly outperforming the custom model. These outcomes suggest that combining deep feature extraction with a classical SVM classifier can further enhance accuracy in medical image classification tasks. Analysis:

The custom CNN model delivered strong classification results, reaching 98.77% accuracy. Fine-tuning the ResNet-50 architecture led to further gains, pushing accuracy up to 99.69%. Among all models, the hybrid pipeline where deep features from ResNet were classified using an SVM, produced the highest scores, correctly classifying every sample in the test set. These outcomes suggest that pairing deep feature extraction with a traditional classifier like SVM can be particularly effective for medical image analysis.

Across all four stages—Benign, Early, Pre-leukemic, and Progressive, model predictions were stable, with minimal misclassification between classes.

V. DISCUSSION

This study introduced a two-stage diagnostic approach for Acute Lymphoblastic Leukemia (ALL), combining genetic and image-based information to improve classification accuracy. The results support the overall framework and provide insight into the strengths of each model and technique used.

A. Stage 1 – Gene Expression Classification

Random Forest outperformed Logistic Regression in identifying ALL from gene expression data. Its ability to handle feature variability and nonlinear patterns made it a more reliable choice for this task. Even with just 10 selected genes, Random Forest maintained nearly 99% accuracy, indicating that the selected features captured essential biological signals.

Feature Selection Observations

Restricting the input to the top 10 genes proved sufficient for classification, offering a balance between simplicity and predictive strength. Adding more features did not significantly enhance performance, suggesting that model efficiency can be preserved with a compact input set.

B. Stage 2 – Image-Based Staging

All three deep learning models performed well when tasked with staging ALL using blood smear images. The CNN achieved high accuracy, while the ResNet-50 model provided a modest improvement. However, the best results came from the hybrid approach, which combined ResNet feature extraction with SVM classification. This combination reached perfect scores on the test set, likely due to the contrast between the deep feature representation and the linear decision boundaries used by SVM.

Strengths and Limitations

The hybrid model's strong performance underscores the value of deep feature extraction. However, the small dataset size and lack of external validation remain limitations. Overfitting may be a concern despite augmentation and cross-validation, especially for clinical deployment. Further testing on independent datasets is needed to confirm the model's generalizability.

Comparison with Prior Work

Most existing studies tend to focus on either gene expression analysis or image-based classification when diagnosing ALL. For example, Golub et al. [1] used gene expression profiles to successfully distinguish between

ALL and AML using a limited set of genes. More recent efforts like Ghaderzadeh et al. [6] introduced convolutional neural networks (CNNs) for identifying leukemic cells from blood smear images. Similarly, Saba et al. [7] provided an overview of deep learning models applied in this domain, though without addressing staging. The present study builds on these approaches by combining both molecular and image-based data into a unified two-stage diagnostic pipeline. Unlike earlier works that treat detection and staging separately, this system offers an integrated workflow. In particular, the use of deep features from ResNet alongside an SVM classifier sets it apart, contributing to the improved performance. Additional support for multi-modal diagnostics is echoed by platforms such as MD-ALL [8], which emphasize the importance of integrating transcriptomic and genetic information for better classification of ALL subtypes.

VI. CONCLUSION

This work introduced a two-stage diagnostic framework for Acute Lymphoblastic Leukemia (ALL) that combines gene expression profiles and microscopic image data. In the first stage, gene expression was used to identify ALL using machine learning classifiers. Random Forest proved to be especially effective, delivering high accuracy with a compact set of features.

The second stage focused on staging the disease based on blood smear images. Among the models tested, a hybrid approach where deep features from ResNet-50 were passed to a Support Vector Machine offered the most reliable performance, achieving complete accuracy on the test set.

These results suggest that combining genetic and image-based inputs can improve diagnostic accuracy and offer more detailed insights into disease progression. While promising, this system would benefit from evaluation on larger and more diverse datasets. Future efforts may also include techniques such as Grad-CAM or attention-based visualization to enhance interpretability, especially in clinical environments where model transparency is critical and on expanding the dataset size, incorporating external validation cohorts.

REFERENCES

- [1] T. R. Golub 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] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.
- [3] Y. Said, A. Bensaid, and S. Mazouzi, "Deep learning for leukemia detection using blood smear images," *Computational Biology and Chemistry*, vol. 80, pp. 44–52, 2019.
- [4] S. Dasariraju and B. Johnson, "Automatic classification of leukemia from microscopic images using deep learning," *Biomedical Signal Processing and Control*, vol. 62, 102093, 2020.
- [5] M. Togacar, B. Ergen, and Z. Comert, "Classification of white blood cells using deep features obtained from ResNet50 model," *Applied Soft Computing*, vol. 97, 106810, 2021.
- [6] M. Ghaderzadeh et al., "Intelligent Deep Learning Approach for Acute Lymphoblastic Leukemia Detection Using Microscopic Images," *Healthcare*, vol. 10, no. 10, p. 1812, 2022.
- [7] T. Saba et al., "Recent Advances in Deep Learning for Acute Lymphoblastic Leukemia Diagnosis: A Review," *Journal of Personalized Medicine*, vol. 14, no. 1, pp. 1–18, 2024.
- [8] H. Wu et al., "MD-ALL: a platform for molecular diagnosis of B-cell acute lymphoblastic leukemia using RNA-seq data," *Haematologica*, vol. 109, no. 3, pp. 781–785, 2024.
- [9] S. Shafique and S. Tehsin, "Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Network," *IEEE Access*, vol. 9, pp. 146521–146531, 2021.
- [10] M. Ilyas et al., "Linear programming based computational technique for leukemia classification using gene expression profile," *Computers in Biology and Medicine*, vol. 132, 104322, 2021.
- [11] K. Ramesh and P. Thouti, "CNN-Based Leukemia Detection Using Microscopic Blood Images," *Biomedical Signal Processing and Control*, vol. 71, 103139, 2022.
- [12] Y. Zhang, X. Wang, and Q. Li, "Deep Learning Approaches for Cancer Prognosis Using Gene Expression Profiles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 2, pp. 880–889, 2022.
- [13] A. Jha and A. Dutta, "Hybrid Deep Learning Frameworks for Medical Image Classification: A Survey," *Computers in Biology and Medicine*, vol. 141, 105090, 2022.
- [14] L. Yu et al., "Multimodal Machine Learning for Integrating Genomic and Imaging Data in Cancer Diagnosis," *Briefings in Bioinformatics*, vol. 24, no. 1, pp. 1–16, 2023.
- [15] H. Khan, M. M. Islam, and S. Hussain, "Convolutional Neural Networks and Support Vector Machines for Leukemia Detection," *Artificial Intelligence in Medicine*, vol. 122, 102215, 2022.