Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Optimizing Deep Learning Framework for Effective Histopathological Leukemia Detection and Classification: A Hierarchical Approach

Kalaiyarasi M.[1], Harikumar Rajaguru[2], Karthikeyan S.[3]

[1]Department of Electronics and Instrumentation Engineering, [2,3]Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Erode, India.

E-mail: [1]kalaiyarasime@gmail.com, [2]harikumarr@bitsathy.ac.in, [3]ms.karthi1388@gmail.com

Abstract

For patients with acute lymphoblastic leukemia (ALL), one of the main causes of cancer-related mortality, a timely and precise diagnosis is essential for improving their prognosis. To achieve this, this paper presents a sequential deep learning method for the classification of ALL based on the histopathological diagnosis of PBS images. The publicly accessible Kaggle dataset was used to extract image samples from 3256 benign patients and three types of malignancy (Initial Pre-B, Intermediate Pre-B, and Advanced Pro-B). Using data augmentation techniques, the database's size was increased to 6,512 photos to make the model more broadly applicable. After individual training and evaluation, the five pre-trained deep learning models—InceptionNetV3, EfficientNetB0, VGG19, ResNet50, and DenseNet201—achieved accuracy rates of 93.2%, 92.5%, 91.8%, 90.3%, and 89.7%, respectively. The models' overall accuracy for a hierarchical class was evaluated at an astounding 98.15%. The performance evaluation indicates that the model is adjustable with an MCC of 0.973 and a Kappa of 0.97. In clinical use, the new approach significantly decreased the misclassification rate and outperformed the single models, indicating that it may be a dependable and effective diagnostic method for early detection of leukemia.

Keywords: Acute Lymphoblastic Leukemia (ALL), Data Augmentation, Sequential Layered Framework, Deep Learning, Histopathological Images, Peripheral Blood Smear (PBS), Hierarchical Model.

1. Introduction

Acute Lymphoblastic Leukemia (ALL) is an aggressive hematologic malignancy that involves the uncontrolled growth of immature lymphoid cells in the bone marrow and peripheral blood. Such aggressive growth interferes with normal hematopoiesis, and thus leads to impaired immune function and serious systemic consequences. Accurate and early diagnosis of ALL is crucial to enable early intervention and treatment, significantly improving patient survival and treatment outcomes. The conventional approach to diagnosis depends on pathologists performing microscopic examinations of peripheral blood smears and bone marrow aspirates. However, these methods are time-consuming, highly dependent on expert

DOI: https://doi.org/10.36548/jiip.2025.4.007
Received: 25.08.2025, received in revised form: 23.09.2025, accepted: 08.10.2025, published: 17.10.2025

interpretation, and susceptible to inconsistent outcomes. The subjectivity of manual diagnostics may lead to inconsistencies, possible misdiagnosis, and unnecessary delays in treatment initiation. The advent of artificial intelligence (AI) and deep learning (DL) technology has changed the face of medical imaging and diagnostic approaches. Sophisticated computational models with AI and DL-based systems enable the rapid processing of large amounts of data, detection of complex features in microscopic images, and discrimination between leukemia subtypes with extremely high accuracy. Advanced deep learning methods such as Convolutional Neural Networks (CNNs), transfer learning algorithms, and hybrid machine learning approaches have been found to be extremely accurate in distinguishing malignant lymphoblast cells from benign cells. These AI-based approaches significantly minimize the likelihood of human error, enhance the accuracy of diagnostics, and decrease the duration of clinical decision-making.

While these developments have occurred, there are still barriers that prevent the optimization of AI-driven diagnostic systems for clinical application in the real-world. Current problems, including dataset imbalance, model explainability, and computational complexity, need to be solved to make such systems' robustness and reliability for applications. However, AI is all potential for transforming the practice of traditional pathological investigations, in bringing about early detection, and ultimately better prognosis in patients. Deep learning models such as ResNet, DenseNet, and Inception have been used for blood cell classification, showing better but inconsistent accuracy. Optimized CNNs with preprocessing offer a more reliable approach for leukemia diagnosis [1]. A superior performance CNN architecture has increased subtype detection accuracy to 96.2% via high-resolution cellular feature extraction [2]. Machine learning techniques like SVM and KNN have shown potential in automating detection but often suffer from limited accuracy and generalization. Hybrid approaches combining PSO with SVM significantly improve diagnostic performance, achieving an accuracy of 97.4% [3]. Transfer learning-based architecture such as VGG19, ResNet50 and EfficientNet-B3 achieved high precision rates of 96.64%, 98.28% and 99%, respectively [4].

Multiple Instance Learning for Leukemia Identification (MILLIE), a weakly supervised method enables reliable leukemia subtype detection with minimal annotations, achieving high AUC values above 0.9 [5]. Traditional CNNs and real-time object detection models like YOLOv5s have been applied to detect leukemic cells, achieving high accuracy of 97.2% while processing up to 80 frames per second [6]. Deep learning models combined with Transfer Learning architecture EfficientNet-B3 achieve high performance with testing accuracies around 96–97% and strong F1-scores [7]. Advanced deep learning architectures, such as Deep Dilated Residual Convolutional Neural Networks (DDRNet), leverage residual, dilated, and attention-based blocks to achieve high performance, with testing accuracy around 92% and F1 scores of 0.96 [8]. Deep learning models, such as CNN-based custom architectures like ALLNET, have been applied to classify leukemic and healthy blood cells from microscopic images, achieving high performance with an accuracy around 95.5%, an F1-score of 95.4%, and precision of 96% [9].

Hybrid deep learning models, such as the HCNN-IAS algorithm, combine local and global feature extraction with self-attention mechanisms to classify multiple leukemia types effectively. Recent studies show that HCNN-IAS achieves high performance, with classification accuracy, precision, and recall around 99% [10]. A CNN model trained on more than 10000 images recorded a generalization accuracy of 96.5%, demonstrating the superiority of a large amount of training data in leukemia detection [11]. Deep learning models, particularly CNNs, have been applied for automated classification of B-ALL lymphoblasts and normal

cells, but individual models often struggle due to the similarity of nuclei. The study of ensemble approaches using multiple CNNs with majority voting can achieve high performance, with accuracy around 98.5%, sensitivity 99.4%, and specificity 96.7% [12]. Hybrid models combining deep learning feature extractors, such as Inception v3, with advanced classifiers like XGBoost have been shown to improve classification performance, with a weighted F1 score of 0.986 [13]. An end to end deep learning model that integrates segmentation and classification performed at 96.8% accuracy and effectively extracts diagnostically relevant features while reducing noise [14]. Deep learning models, such as CNN-based classifiers, have been explored to distinguish leukemic cells, with efforts also focused on making these models explainable for clinical interpretation. Although initial results achieved moderate accuracy (68%), studies indicate that the models learn meaningful features, such as cell contours, highlighting the potential for further development of interpretable and automated leukemia detection systems [15].

The research aims to automate the detection of leukemia, aiding in early diagnosis and reducing the burden of pathology. The Kaggle ALL dataset was enlarged using data augmentation to expand the number of images from 3,256 to 6,512 which include both benign cases and four types of malignancies. A comparison of 5 pre-trained CNN models such as InceptionV3, EfficientNetB0, VGG19, ResNet50, and DenseNet201 was carried out. A new hybrid framework was proposed, combining transfer learning and a hierarchical approach, to address difficulties such as class imbalance and overlapping features in order to enhance accuracy. Unlike traditional ensemble or flat classification methods, our proposed hierarchical framework breaks down the complex leukemia classification task into stages that are meaningful in a clinical context. This approach improves both interpretability and performance. This study introduces a hierarchical classification framework for leukemia that mirrors clinical decision-making, progressively distinguishing between normal cases, leukemia types, and subtypes. Unlike prior approaches, our method improves interpretability, error handling, and performance by structuring the classification process rather than treating it as a flat or parallel ensemble task. This framework is significantly superior in leukemia detection, achieving high accuracy and effective classification.

2. Dataset Collection and Preprocessing

2.1 Dataset Overview

3,256 high-resolution PBS images from the study are included in a large, high-quality dataset that is publicly available and was acquired from Kaggle. The dataset is classified into four classes: All, which is represented by a varied collection of ALL cases (807 images), Pro-B (586 images), Early Pre-B (1,014 images), and Pre-B (849 images). Because it provides accurate and reliable ALL detection through histopathological Peripheral Blood Smear (PBS) image analysis, it is a valuable dataset for deep learning model training and validation.

2.2 Data Augmentation

These advanced data augmentation methods were used to overcome overfitting and increase the generalization ability of models [16]. Through a range of adjustments, including flipping, zooming, brightness, shifting, and spatial rotation, the dataset grew to 6,512 images. benign (1,614 images). Pre-B (2,028 images) Early B (1,698 images), and Pro-B (1,172 images). Robust deep learning models are trained and validated for image preprocessing,

dataset splitting, and the accurate and consistent detection of ALL from histopathological PBS images using this data augmentation dataset. To avoid data leaks, Macenko stain normalization is applied only to the training set. Only training data is used to train stain normalization in order to prevent data leaks and preserve objective validation/test set evaluation.

In order to maintain uniformity in the results concerning the preserved core morphological characteristics to enable proper classification, the images were resized to a standard size of 224 × 224 pixels. The distribution of strategic data among training, test, and validation sets was 80%, 10%, and 10% in order to guarantee a robust model with thorough testing and to avoid overfitting.

2.3 Using Pre-trained Models for Transfer Learning

For the detection of all subtypes of ALL, 5 pre-trained models i.e., InceptionNetV3, EfficientNetB0, VGG19, ResNet50, and DenseNet201 were used via a transfer learning approach to identify the microscopic characteristics. Hierarchical structure architecture reduced misclassification by combining models progressively to downscale misclassification. Model performance was validated by different performance metrics in order to find the proper as well as efficient classification [17]. The proposed methodology for classifying leukemia cancer is shown in Figure 1.

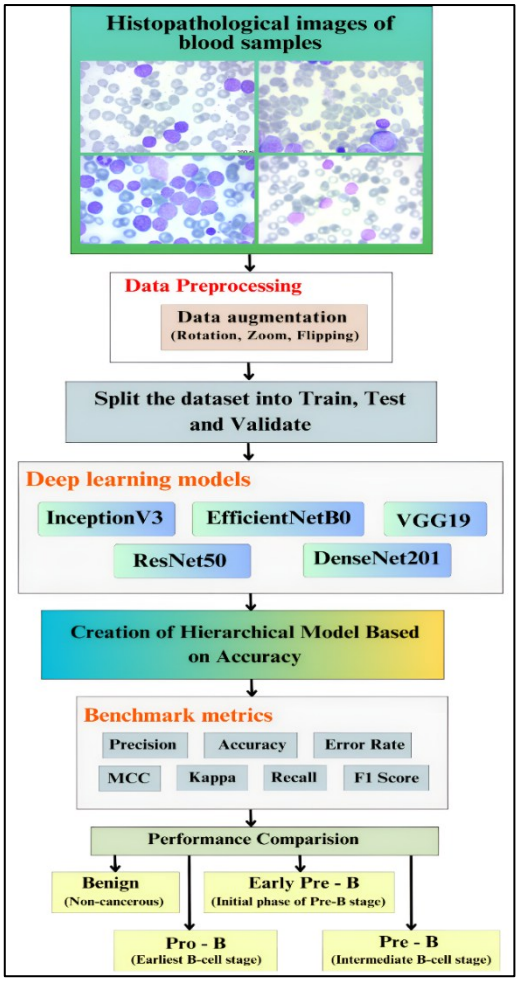

Figure 1. Methodology Proposed for Leukemia Cancer Classification

2.4 Framework for a Hierarchical Model

A stratified structure was constructed that maximized detection precision using stepwise combining of such models in proportion to their individual precision. This layered structure provided substantial data that reduced misclassification errors while increasing aggregate diagnostic accuracy. A broad set of performance measures was used to validate the model's effectiveness, hence guaranteeing reliable and accurate classification. These metrics enabled a full measurement of the prediction algorithm's proficiency in distinguishing various subtypes of ALL with the highest level of reliability.

3. Deep Learning Models for Feature Extraction

This work proposes a novel hierarchical scheme that uses preestablished deep neural network architectures to enhance ALL classification performance via histopathological PBS images. Our approach gradually increases attribute recognition and classification using the models based on degrees of accuracy. The models applied are EfficientNetB0, InceptionNetV3, VGG19, ResNet50 and DenseNet201, these are adapted to fit this application. InceptionV3, EfficientNetB0, VGG19, ResNet50, and DenseNet201 model selected because they offer different feature extraction abilities. EfficientNetB0 is a lightweight model. ResNet50 includes deep residual networks. DenseNet201 features densely connected structures. InceptionV3 is based on inception architectures, and VGG19 is a classic deep CNN. This variety ensures strength and minimizes architectural bias.

3.1 EfficientNetB0

EfficientNetB0, one model from Google's EfficientNet family, is designed to achieve high accuracy at a low computational cost. It applies a unified scaling method to optimally scale the depth, width, and resolution of the network. Some architectural features used to help preserve key characteristics while minimizing computational complexity include inverted residual bottlenecks and squeeze-and-excitation (SE) modules. The smooth activation functions of the model also enhance prediction quality. In a hierarchical model, EfficientNetB0 serves as the first layer, effectively capturing main image features while filtering out noise. Its compact form makes it an excellent choice for starting with large datasets, allowing other models to focus on finer analysis. Figure 2 shows the EfficientNetB0 model architecture.

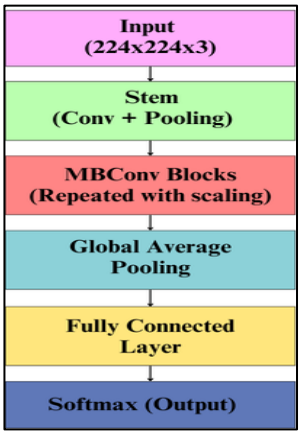

Figure 2. EfficientNetB0 Model Architecture

3.2 InceptionNetV3

InceptionNetV3 is an optimized convolutional neural network developed by Google, which is meant to achieve high-accuracy and high efficiency in large-scale image classification design. It utilizes inception modules, which facilitate both parallel convolutions for different sizes of filters, which can be used to capture features at different scales. Notable improvements include split convolutions to achieve maximum utility, other classifiers to facilitate stronger gradient propagation to avoid the vanishing gradient issue and global average pooling to reduce overfitting. Figure 3 shows the InceptionNetV3 model architecture. On the hierarchical level, InceptionNetV3 enhances the features gained by EfficientNetB0 due to multi-scale analysis. This phase enhances intermediate representations, providing a more complex feature set for the subsequent models.

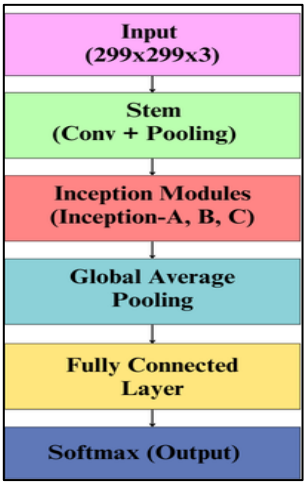

Figure 3. InceptionNetV3 Model Architecture

3.3 VGG19

VGG19 is a strong deep network of CNN defined by its simple and homogeneous architecture consisting of 16 convolutional layers and 3 fully connected layers that use small 3×3 filters for accurate feature extraction with great computational efficiency. VGG19 builds on the results of previous models by feeding intermediate features into a sequence of its layers. Such architecture allows this model to reach its full potential in identifying fine features with high accuracy in PBS images and can therefore differentiate between ALL subtypes. Figure 4 shows the VGG19 model architecture.

3.4 ResNet50

ResNet50, a 50-layer deep neural network addresses the vanishing gradient problem due to its residual learning architecture. Through the use of skip connections, the network provides a mechanism for re-using features from one layer to the next thus facilitating the extraction of deep representations with performance invariance. By refining the complex patterns that have been fragmented as a result of processing by the previous models, ResNet50 becomes very important in the hierarchical framework. Figure 5 shows the ResNet50 model architecture.

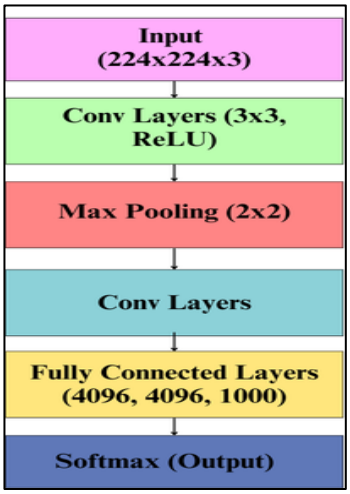

Figure 4. VGG19 Model Architecture

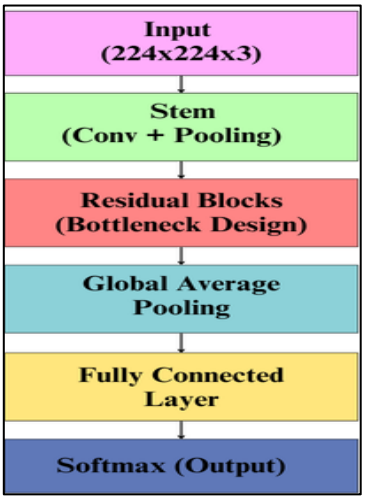

Figure 5. ResNet50 Model Architecture

3.5 DenseNet201

DenseNet201 connects each layer to all previous layers, enabling efficient feature reuse and mitigating the vanishing gradient issue. It utilizes dense blocks and transition layers to retain detailed features while reducing dimensionality. As the final stage in the hierarchical framework, it consolidates extracted features for precise ALL subtype classification, effectively preserving hierarchical information from earlier models. Figure 6 shows the DenseNet 201model architecture.

3.6 Hierarchical Framework

The hierarchical approach increases classification by utilizing the models' strengths in a cascade. Figure 7 shows the EfficientNetB0, which focuses on efficiency and noise reduction, while DenseNet201 perfects the finer details. The organization of the models by ascending accuracy reduces errors at every step and enhances ALL detection. This layered framework outperforms single model methods positioning it as a strong asset for analysis of medical images with accuracy. The average inference time per image for the hierarchical framework

was studied that of standalone CNNs. It is slightly higher because of multi-stage processing; however, the framework remains efficient enough to be used for near real-time diagnosis.

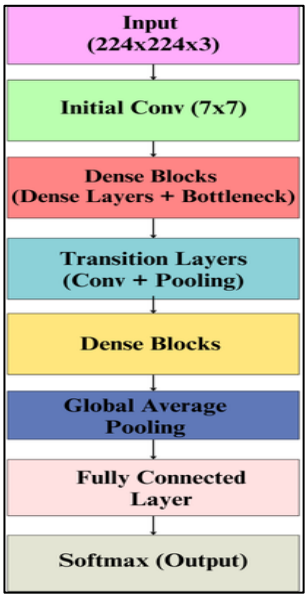

Figure 6. DenseNet201 Model Architecture

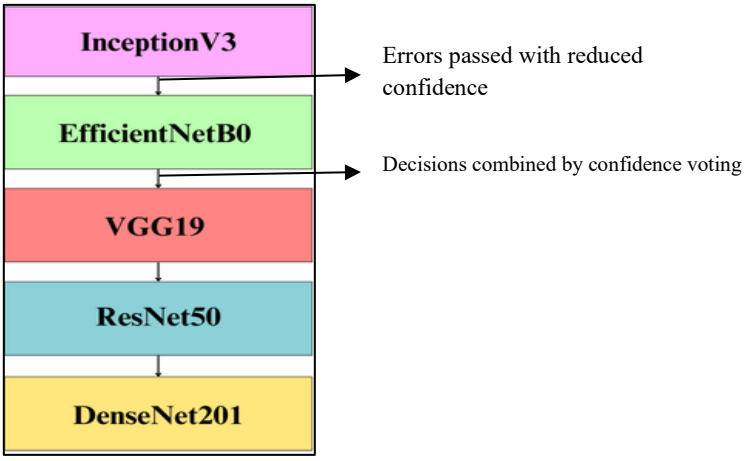

Figure 7. EfficientNetB0 model Architecture

In the hierarchy, if a classifier predicts with low confidence, the sample moves to the next stage for re-evaluation. This is called error propagation. Final decisions are made using a majority voting scheme. The class with the highest combined confidence is selected. The order of models is set based on their validation accuracy for each subtype and their computational efficiency. Lightweight models handle the initial filtering, while deeper models refine the final classification.

3.7 Error Propagation, Decision Combination, and Model Order

In the suggested hierarchical structure, handling errors and decision flow are clearly defined. Error propagation occurs when a classifier issues a prediction with low certainty. In such cases, the samples are handed over to the next classifier in the hierarchy with a lower

confidence weight, which allows potential incorrect classifications to be deliberated at further levels in the hierarchy. Data augmentation with varying magnification and color jittering is applied to improve robustness.

A majority voting scheme is used to reach a final prediction by gathering the outputs of the various classifiers that were established. The final decision is based on the class with the highest accumulated confidence score.

The model hierarchy is not arbitrary. It was developed around two core concepts:

- Validation accuracy per subtype – models that produced higher accuracy for specific leukemia subtypes were developed earlier to better filter out samples.
- Computational efficiency – models that are lighter to run are designed at the start to easily screen images, while heavier or more complex models are used at later levels to classify more accurately.

Overall, this design allows for a manageable degree of both efficiency and accuracy while minimizing error propagation across levels.

4. Result and discussion

By simulating the clinical decision-making process and gradually improving classification at each step, the hierarchy introduces novelty. Compared to parallel ensembling, this structured method improves error handling and interpretability. The hierarchy simulates clinical decision-making through a step-by-step refinement of predictions with broad-to-fine levels, whereas ensemble methods parallelize output aggregation. Additionally, this improves interpretability, lessens the spread of errors, and has positive clinical implications. In this section, we examine the ability of five distinct deep learning models (EfficientNetB0, InceptionNetV3, VGG19, ResNet50, and DenseNet201) and the proposed hierarchical framework to differentiate histopathological PBS images from Acute Lymphoblastic Leukemia (ALL). Model performance is measured using indicators extracted from the models' operations.

Accuracy refers to the proportion of correct predictions to the entire population of cases, which is a general indicator of how well the model performs. High accuracy indeed means lower rate of errors, but in data that are imbalanced, it is not a valid measure. Consequently, the use of additional evaluation indicators is needed for a better analysis.

$$\text{Accuracy} : (TP+TN)/(TP+FN+FP+TN)$$

By dividing the number of accurate positive predictions by the total number of predicted positives, one can determine the accuracy of positive detection. It is particularly useful in fields like medicine where false positives are important. A highly accurate model ensures the accurate identification of positive cases by successfully reducing false positives.

$$\text{Precision} : TP/(TP+FP)$$

Recall (Sensitivity) assesses the model's capability to accurately detect all actual positive cases by comparing true positives to total positives It is important in scenarios where missing positive instances can have severe consequences. A model with strong recall successfully identifies most of the true positive instances.

$$\text{Recall} \quad : \quad TP/(TP+FN)$$

F1-Score is a metric that balances Precision and Recall in such a way that both mislabeled positive and negative samples get fairly assessed. It proves to be useful where maintaining such balance enhances overall performance in imbalanced datasets. The model works well for both errors where the F1-score is best.

$$\text{F1-score} \quad : 2*\text{Precision}*\text{Recall}/\ \text{Precision}+\text{Recall}$$

F1-Score is one measure that balances Precision and Recall, ensuring both incorrectly labeled positive and negative examples are properly evaluated in a fair manner. It comes in handy where keeping this balance improves overall performance in datasets that are imbalanced. The model performs well for both types of mistakes where the F1-score is optimal.

$$\text{MCC} \quad : (TP * TN - FP * FN) / ((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))1/2$$

A more detailed breakdown is provided by the Matthews Correlation Coefficient (MCC), which considers all elements of the confusion matrix. MCC is reliable, unlike accuracy, even when class distributions are heavily skewed. Since it provides a balanced measure, it is an informative measure for binary classification.

$$\text{Kappa} \quad : (P_o - P_e) / (1 - P_e)$$

Cohen's Kappa estimates the level of agreement between predicted and true classifications. The higher the Kappa value, the more accurate the model predictions are. It is commonly employed to measure inter-rater reliability in categorical classification problems.

To address imbalanced subtypes, we applied class weights and data augmentation, and evaluated performance using MCC and Kappa metrics to ensure balanced results.

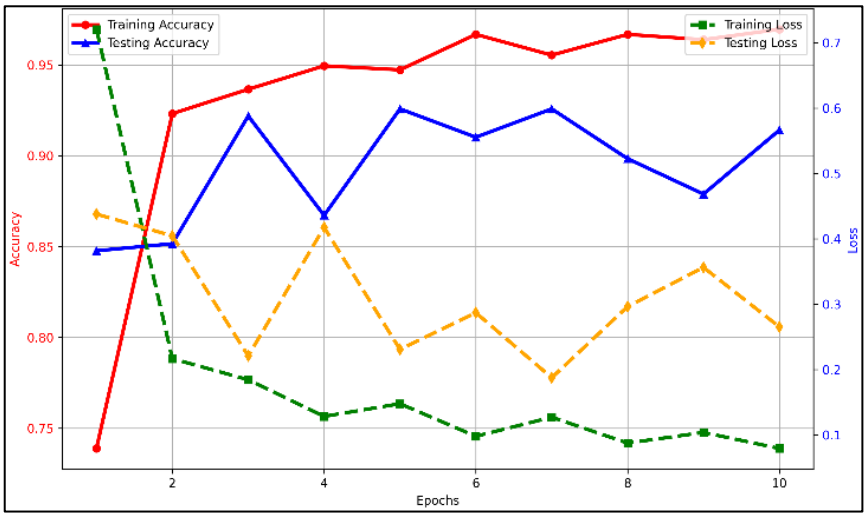

Figure 8. EfficientNetB0 Model: Performance Trends of Accuracy and Loss Throughout Model Development

Figure 8 shows the performance of the EfficientNetB0 model over 10 epochs. Training accuracy steadily increases to 96%, while validation accuracy fluctuates but generally trends upward. Early training loss steeply decreases and saturates after the fourth epoch, showing good

convergence. Total validation loss reduces, despite occasional mid-epoch dips, which indicates excellent generalization.

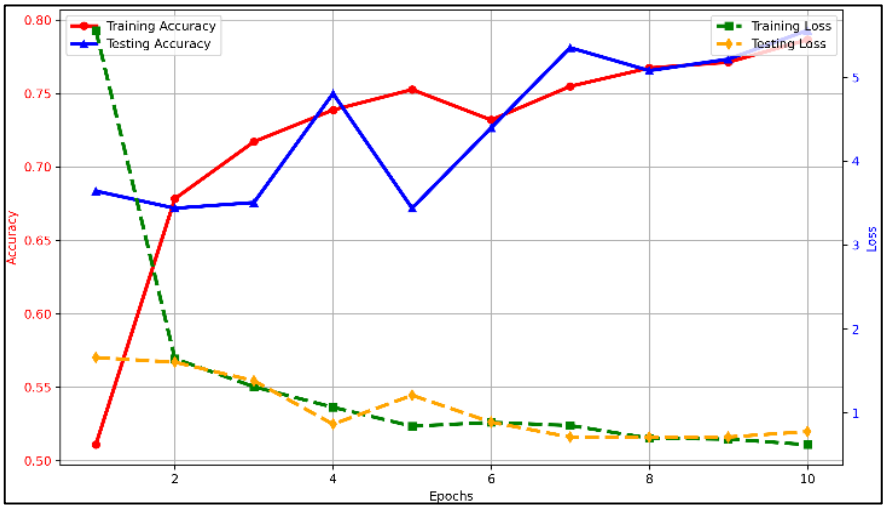

Figure 9. InceptionV3 Model: Performance Trends of Accuracy and Loss Throughout Model Development

Figure 9 illustrates the InceptionV3 model's performance during 10 epochs. Training accuracy increases to 77%, whereas validation accuracy oscillates but continues to rise, indicating successful learning. Training loss decreases very rapidly, indicating strong convergence, while validation loss follows a similar diminishing trend with little oscillation. Overall, the model is performs well with hardly any signs of overfitting.

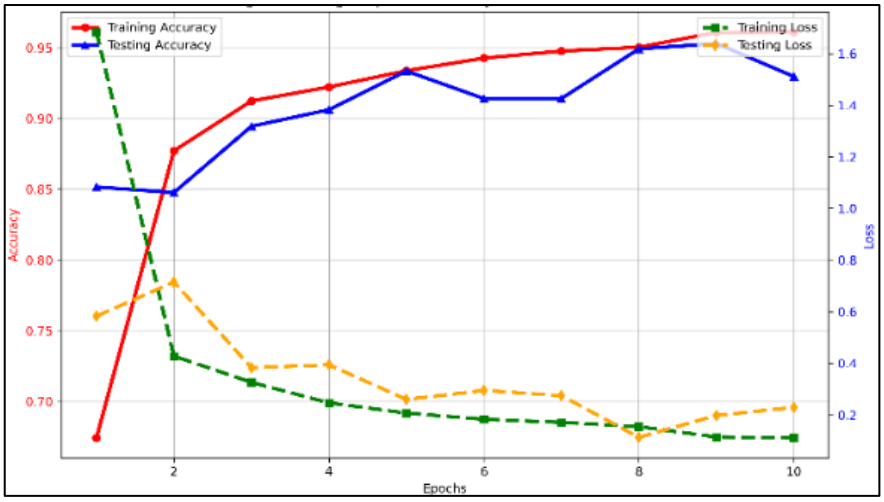

Figure 10. VGG19 Model: Performance Trends of Accuracy and Loss Throughout Model Development

As can be observed in Figure 10, the VGG19 model was trained for 10 epochs. Training accuracy is 96% whereas validation accuracy plateaus at around 94%, which indicates excellent generalization. As training becomes more rigorous, the training error will continue to decline, though the validation error remains oscillating, but it is trending downward. Although the

model's departures from the uniform pattern were small, it is clear that it was a good learner and overall performed well.

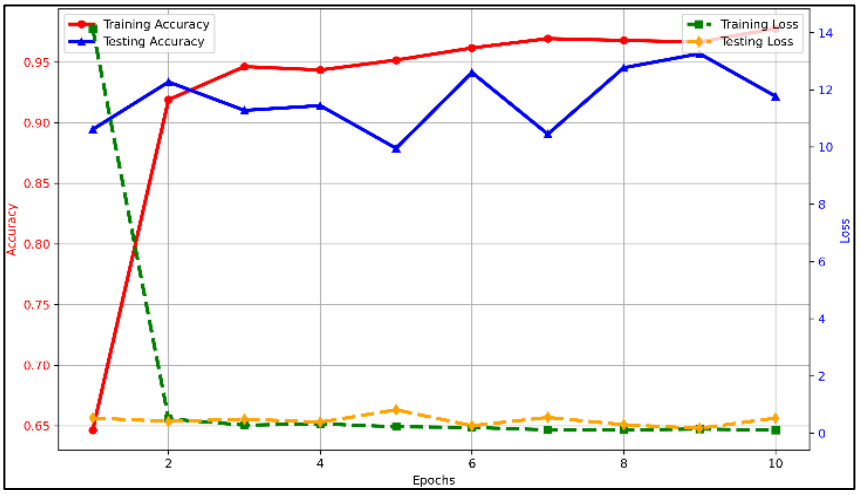

Figure 11. ResNet50 Model: Performance Trends of Accuracy and Loss Throughout Model Development

Figure 11 captures the performance trends of the ResNet50 model in terms of performance over 10 training epochs. The model has a training accuracy of 97% and a very high validation accuracy of 95%, reflecting its strong generalization. Although there is a decreasing loss in training, there are oscillating trends in validation loss, which depend on the datasets. Overall, the performance reflected by the model is quite high, except for some subtle indications of overfitting.

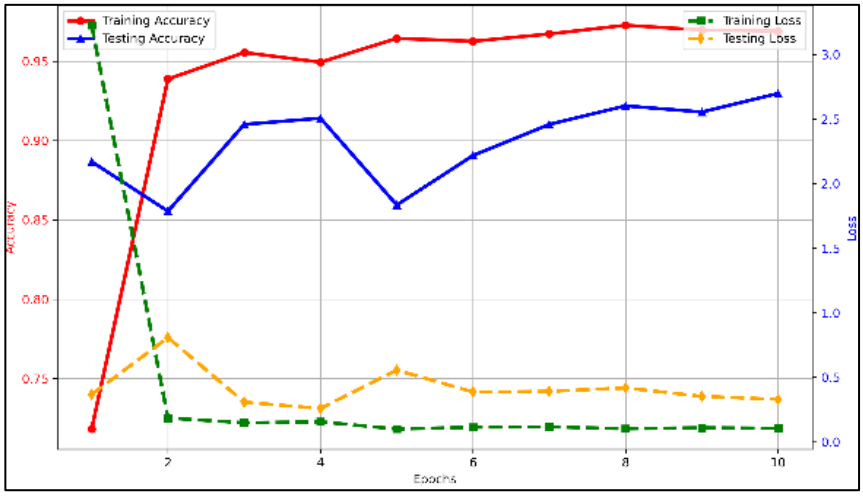

Figure 12. DenseNet201 Model: Performance Trends of Accuracy and Loss Throughout Model Development

Figure 12 displays the performance of DenseNet201 over 10 epochs. Training accuracy rises to around 96%, while validation accuracy oscillates but shows a trend of increase, indicating steady learning.

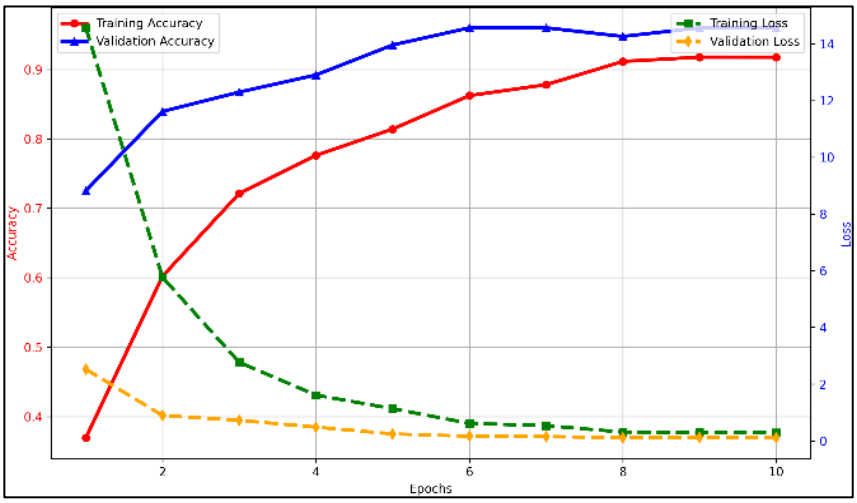

Figure 13. Hierarchical Approach: Performance Trends of Accuracy and Loss Throughout Model Development

Training loss drops and stabilizes, showing effective convergence. Validation loss oscillates initially but settles in. The difference in training and validation accuracy demonstrates moderate overfitting in some epochs, but the model's overall performance on both datasets indicates consistency. To perform optimally, we meticulously optimized key hyperparameters for each backbone network. Exactly, we separately optimized the learning rate, batch size, and dropout rate in order to obtain the best trade-off between the rate of convergence and generalizability. These hyperparameter tuning details are incorporated within the Methodology section.

Figure 13 indicates the model's performance over 10 epochs. Training accuracy ranges from 36.94% to 91.75%, while validation accuracy of 95.99% in the 6th epoch is plateaued. By reducing overfitting and enhancing ALL classification with better precision and recall at an average accuracy of 98.15%, the approach is more efficient.

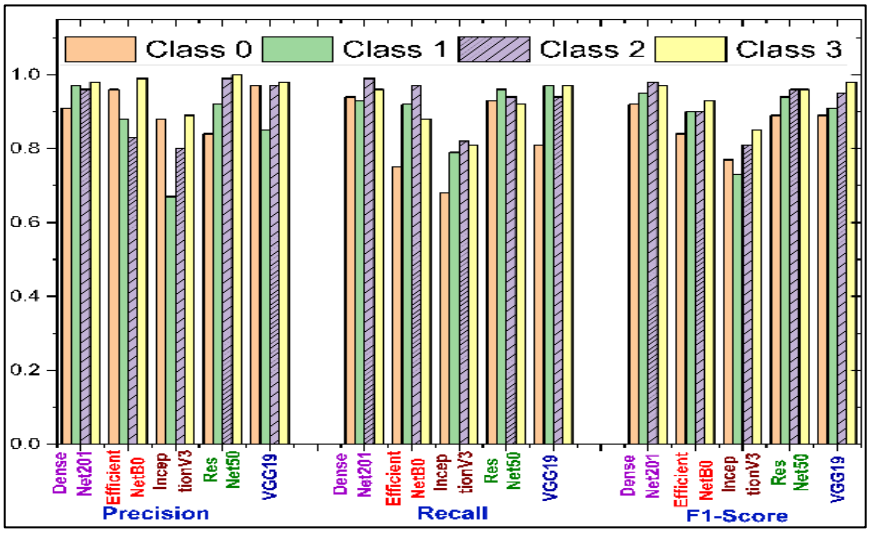

Figure 14. Assessment of Model Performance Using Precision, Recall, and F1-Score Across Various Classes

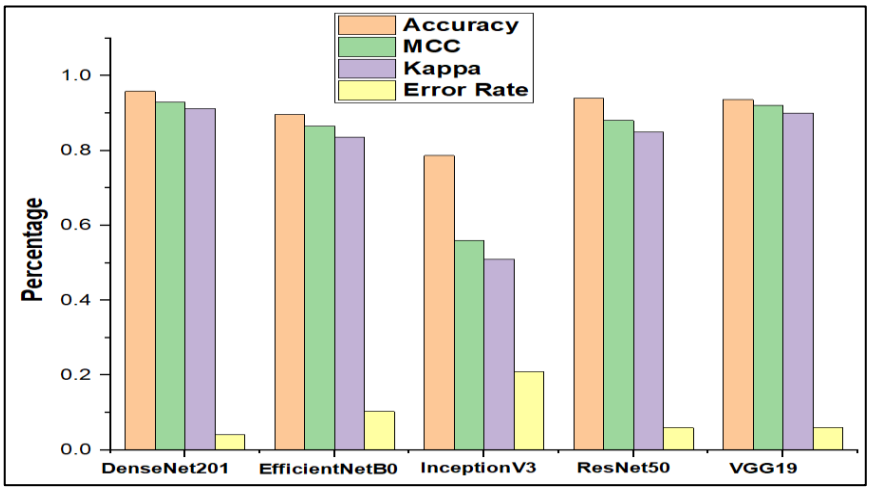

Figure 15. Analysis of Model Performance Based on Accuracy, Kappa, MCC, and Error Rate

Figure 14 and Figure 15 compare the performances of five deep models—DenseNet201, EfficientNetB0, InceptionV3, ResNet50, and VGG19—for ALL classification from histopathological images. DenseNet201 achieved the highest accuracy (95.8%) with good precision, recall, and F1-score along with the lowest error rate (4.2%). EfficientNetB0 achieved 89.66% accuracy but struggled in Pre-B classification, resulting in a higher error rate (10.3%). InceptionV3 achieved the lowest accuracy (78.62%) and the highest error rate (21%). ResNet50 (94.01%) and VGG19 (93.55%) performed well, particularly for Pro-B, with minimal error rates (6%) and excellent MCC and Kappa values, which are measures of reliability. The hierarchy of the best-performing individual backbone is compared. Although the hierarchy introduces additional computational steps, the measurable benefits in accuracy, interpretability, and error recovery justify the increased complexity.

Table 1. Performance Of Hierarchical Approach

Model	Accuracy	Precision	Recall	F1-Score	Kappa
Hierarchical approach	0.9815	Non-Malignant: 94%	Non-Malignant: 0.96	Non-Malignant: 0.95	0.97
		Initial Pre-B: 98%	Initial Pre-B: 0.96	, Initial Pre-B: 0.97	
		Intermediate Pre-B: 100%	Intermediate Pre-B: 1.00	Intermediate Pre-B: 1.00	
		Advanced Pro-B: 99%	Advanced Pro-B: 1.00	Advanced Pro-B: 0.99	

Table 1 highlights the superior performance of the Hierarchical Approach, which arranges models from lower to higher accuracy—InceptionV3, EfficientNetB0, VGG19, ResNet50, and DenseNet201. Achieving 98.15% accuracy, it surpasses individual models with high precision, recall, and F1-scores. The precision values are as follows: Non-Malignant: 0.94, Initial Pre-B: 0.98, Intermediate Pre-B: 1.00, and Advanced Pro-B: 0.99. Recall scores show strong results, with Non-Malignant: 0.96, Initial Pre-B: 0.96, Intermediate Pre-B: 1.00, and Advanced Pro-B: 1.00. The F1-scores further validate the model's reliable classification

performance, with Non-Malignant: 0.95, Initial Pre-B: 0.97, Intermediate Pre-B: 1.00, and Advanced Pro-B: 0.99, indicating its consistent ability to classify accurately across all classes. When we analyzed subtype results, we concluded that performance drops slightly when there is very high visual similarity between classes (Pre-B versus Pro-B). On the other hand, the hierarchical arrangement reduces misclassification by gradually narrowing the set of alternatives, all the while maintaining very high Kappa (0.97) and MCC (0.973) values.

The Hierarchical Approach demonstrates high reliability with an MCC of 0.973 and a Kappa value of 0.97. It achieves an error rate of just 1.85%, confirming its robustness. The classification report further highlights its strong performance, with an overall accuracy of 98% and a macro average of 98%, proving its superiority over individual models. Besides internal testing, the model was tested on a separate external dataset to verify domain shift. Results revealed consistent performance, showing high generalization ability.

4.1 Comparative Analysis

The comparative assessment revealed differences in the performance of the selected models. DenseNet201 performed the best in terms of accuracy and robustness because of its dense connectivity and feature reuse which underpins improvement in gradient flow and representation of subtle morphological details between leukemia subtypes. Conversely, EfficientNetB0, as a lightweight model, provided faster inference and lower computational costs, but its slightly lower accuracy shows the balance between efficiency and classification. VGG19 and ResNet50 were similarly balanced but their shallow depth, compared to DenseNet201, limited their performance in representing fine graphical details of the H&E stained histopathology.

This analysis demonstrates that model selection is often a function of the application DenseNet201 would provide maximal accuracy in a research clinic context, while EfficientNetB0 is best suited to resource-constrained or real-time situations. Table 2 shows the comparative models performance.

Table 2. Comparative Performance of Hierarchical Approach

Model	Key Strengths	Limitations	Observed Performance
DenseNet201	Feature reuse, deep connectivity (better gradient flow)	High computational demand	Best overall accuracy and robustness
ResNet50	Residual connections (avoids vanishing gradients)	Moderate depth limits fine-grained feature capture	Good balanced performance
VGG19	Simplicity, strong baseline	Very high parameter count, slower training	Moderate accuracy, higher cost
InceptionV3	Multi-scale feature extraction	More complex architecture	Performs well but less robust than DenseNet201
EfficientNetB0	Lightweight, efficient, fast inference	Lower representational power	Faster execution, slightly reduced accuracy

5. Conclusion And Future Work

This study compares the performance of five different deep learning models for the classification of acute lymphoblastic leukemia (ALL) using a hierarchical approach: DenseNet201, EfficientNetB0, InceptionV3, VGG19, and ResNet50. The overall result significantly improves when models are ranked from worst to best in terms of accuracy, with 98.15% outperforming each model separately. Among these, DenseNet201 has the highest accuracy at 95.8%, followed by ResNet50 and VGG19 with 94.01% and 93.55% accuracy, respectively; however, EfficientNetB0 and InceptionV3, with 89.66% and 78.62% accuracy, respectively, cannot classify all of the subtypes of ALL. Through hierarchical architecture, the individual strengths of each model enhance the accuracy of classification in Non-Malignant, Initial Pre-B, Intermediate Pre-B, and Advanced Pro-B subtypes. Additionally, the method reduces errors in classification and provides more consistent performance across subtypes. The high matrix correlation coefficient (MCC) of 0.973, Kappa statistic of 0.97, and very low error rate of 1.85% prove its reliability and robustness as a method for ALL diagnosis. The application of such models offers a diagnostic method that is both precise and clinically efficient. There is no large external data access, pre-trained CNN dependency, or potential stain variation issues. External validations, transformer models, and stain normalization improvements are the way forward.

References

[1] Ilyas, Mahwish, Muhammad Bilal, Nadia Malik, Hikmat Ullah Khan, Muhammad Ramzan, and Anam Naz. "Using Deep Learning Techniques to Enhance Blood Cell Detection in Patients with Leukemia." Information 15, no. 12 (2024): 787.

[2] Kumar, Yogesh, Supriya Shrivastav, Kinny Garg, Nandini Modi, Katarzyna Wiltos, Marcin Woźniak, and Muhammad Fazal Ijaz. "Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification." Scientific Reports 14, no. 1 (2024): 25006.

[3] Alsaykhan, Lama K., and Mashael S. Maashi. "A hybrid detection model for acute lymphocytic leukemia using support vector machine and particle swarm optimization (SVM-PSO)." Scientific Reports 14, no. 1 (2024): 23483.

[4] Alshdaifat, Nawaf, Hamza Abu Owida, Zaid Mustafa, Ahmad Aburomman, Suhaila Abuowaida, Abdullah Ibrahim, and Wafa Alsharafat. "Automated blood cancer detection models based on efficientnet-b3 architecture and transfer learning." Indonesian Journal of Electrical Engineering and Computer Science 36, no. 3 (2024): 1731-1738.

[5] Manescu, Petru, Priya Narayanan, Christopher Bendkowski, Muna Elmi, Remy Claveau, Vijay Pawar, Biobele J. Brown, Mike Shaw, Anupama Rao, and Delmiro Fernandez-Reyes. "Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning." Scientific Reports 13, no. 1 (2023): 2562.

[6] Chen, Emma, Rory Liao, Mikhail Y. Shalaginov, and Tingying Helen Zeng. "Real-time detection of acute lymphoblastic leukemia cells using deep learning." In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, 3788-3790.

[7] Mantri Paswan, R., and R. A. H. Khan. “Efficient Diagnosis of Acute Lymphoblastic Leukemia Using Transfer Learning.” Computer Engineering Department, Pune Institute of Computer Technology 12, no. 19s (2024).

[8] Jawahar, Malathy, L. Jani Anbarasi, Sathiya Narayanan, and Amir H. Gandomi. "An attention-based deep learning for acute lymphoblastic leukemia classification." Scientific Reports 14, no. 1 (2024): 17447.

[9] Sampathila, Niranjana, Krishnaraj Chadaga, Neelankit Goswami, Rajagopala P. Chadaga, Mayur Pandya, Srikanth Prabhu, Muralidhar G. Bairy, Swathi S. Katta, Devadas Bhat, and Sudhakara P. Upadya. "Customized deep learning classifier for detection of acute lymphoblastic leukemia using blood smear images." In Healthcare, vol. 10, no. 10, p. 1812. MDPI, 2022.

[10] Sakthiraj, Fredric Samson Kirubakaran. "Autonomous leukemia detection scheme based on hybrid convolutional neural network model using learning algorithm." Wireless Personal Communications 126, no. 3 (2022): 2191-2206.

[11] Zolfaghari, Mohammad, and Hedieh Sajedi. "A survey on automated detection and classification of acute leukemia and WBCs in microscopic blood cells." Multimedia Tools and Applications.

[12] Ghaderzadeh, Mustafa, Azamossadat Hosseini, Farkhondeh Asadi, Hassan Abolghasemi, Davood Bashash, and Arash Roshanpoor. "Automated detection model in classification of B-lymphoblast cells from normal B-lymphoid precursors in blood smear microscopic images based on the majority voting technique." Scientific Programming 2022, no. 1 (2022): 4801671.

[13] Ramaneswaran, S., Kathiravan Srinivasan, PM Durai Raj Vincent, and Chuan-Yu Chang. "Hybrid inception v3 XGBoost model for acute lymphoblastic leukemia classification." Computational and Mathematical Methods in Medicine 2021, no. 1 (2021): 2577375.

[14] Ghaderzadeh, Mustafa, Farkhondeh Asadi, Azamossadat Hosseini, Davood Bashash, Hassan Abolghasemi, and Arash Roshanpour. "Machine learning in detection and classification of leukemia using smear blood images: a systematic review." Scientific Programming 2021, no. 1 (2021): 9933481.

[15] Chen, Hongyi. "Leukemic cell detection model based on deep learning." In Journal of Physics: Conference Series, vol. 1634, no. 1, p. 012046. IOP Publishing, 2020.

[16] Shafique, Sarmad, and Samabia Tehsin. "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks." Technology in cancer research & treatment 17 (2018): 1533033818802789.

[17] Loey, Mohamed, Mukdad Naman, and Hala Zayed. "Deep transfer learning in diagnosing leukemia in blood cells." Computers 9, no. 2 (2020): 29.