Automated Classification of Acute Lymphoblastic Leukemia in Blood Smear Images Using Transfer-Learned ResNet-50

Akanksha Kochhar*, Gargi Mishra*, Kshitiz Gaur†, Nitya Mahajan†, Shreya Singh¶, Vidit Garg‖
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering, New Delhi, India
*akanksha.kochhar@bharatividyapeeth.edu, *gargi.mishra.eee@gmail.com,
†kkshitiz2004gaur@gmail.com, †nityamhjn05@gmail.com,
¶shreya21204@gmail.com, ‖viditgarg03@gmail.com

***Abstract*—Acute lymphoblastic leukemia (ALL) is an aggressive cancer primarily affecting the blood and bone marrow. Although modern diagnostics have improved, manual inspection of peripheral blood smears remains a common method. This traditional approach is labor-intensive and time-consuming and can vary significantly between observers.**

To overcome these constraints, we propose an automated classification system based on the ResNet-50 deep learning architecture. Our model, pre-trained on ImageNet and fine-tuned on a large dataset of cell smear images, balances computational efficiency and classification accuracy. By freezing the convolutional layers and training a lightweight classification head, we achieved an optimal trade-off between simplicity and performance.

Tested on the C-NMC 2019 dataset, our system achieved a peak validation accuracy of 97.39%, an average validation accuracy of 96.63%, and a test accuracy of 95.43%. These results significantly outperform prior methods by Chen et al. [1], Heriawati et al. [2], and Papaioannou et al. [3], while maintaining inference times under 0.05 seconds per image, making it suitable for real-time clinical workflows.

***Index Terms*—Acute lymphoblastic leukemia, deep learning, ResNet-50, transfer learning, convolutional neural network, blood smear image classification, data augmentation, medical image analysis, accuracy, ROC AUC**

I. Introduction

Leukemia refers to a group of blood-related cancers marked by excessive production of abnormal or immature white blood cells. These cells typically accumulate in the bone marrow and bloodstream, disrupting normal function. Globally, leukemia remains a serious health issue, contributing to a substantial percentage of cancer diagnoses and fatalities. Acute lymphoblastic leukemia (ALL), in particular, poses significant risks in pediatric populations, accounting for a large share of childhood cancer cases.Leukemia is an important health problem worldwide; in 2020 alone, it was responsible for approximately 2.5% of all newly diagnosed cancers and 3.1% of all deaths due to cancer globally [1]. The disease may be divided into acute and chronic types, with acute leukemias evolving more quickly and requiring prompt medical intervention.

ALL is also divided according to morphological, immunophenotypic, and genetic characteristics. Particular subtypes like Pro-B-ALL, Pre-B-ALL, and Early-ALL (early precursor B-ALL) are routinely identified in clinical practice. Proper identification of these subtypes is important, both for risk stratification and management. Subtype identification traditionally needs sophisticated laboratory methods such as flow cytometry and cytogenetic analysis, aided by microscopic blood ttc examination. But in low-resource environments, these facilities might not be present, and even in referral centers, manual microscopy takes a long time and is prone to observer variation [3].

Recent advances in artificial intelligence (AI) have led to new avenues for improving hematological diagnostics. Specifically, convolutional neural

networks (CNNs) have achieved significant success on image-based problems, such as the classification of hematological malignancies. Traditional machine learning approaches entailed the manual extraction of features—like shape, size, and texture—from cell images, followed by classification with classical algorithms like support vector machines (SVMs) or k-nearest neighbors (kNN) [4].

Putzu et al. [4] demonstrated this method, achieving approximately 92–94% accuracy via handcrafted features. Yet traditional methods tend to be insensitive to staining variability, image quality, or morphology of cells. Deep learning networks, on the other hand, learn complex hierarchical features automatically from raw images and significantly enhance classification performance. For example, Rehman et al. reported that a deep CNN achieved an accuracy of 94.6% in discriminating leukemic blasts from normal cells [**?**].

Current hybrid approaches have merged deep feature extraction with conventional classifiers to improve performance. Saba et al., for instance, used deep CNNs for feature extraction and random forest classifiers for prediction, achieving an accuracy rate of around 93% [5]. Ensemble methods, like that suggested by Haque et al. [6], combine several deep models (e.g., ResNet, DenseNet) to improve robustness, although they also increase system complexity.

In addition, models addressing multi-class subtype classification—differentiating between several phases of ALL—are emerging. Ghosh et al. proposed a CNN model that could differentiate among different ALL subtypes with about 91.7% accuracy [7]. Nevertheless, training these models remains a challenge because there are subtle inter-class morphological variations.

AI models for the detection of leukemia have advanced significantly, but many are dependent on intricate architectures or constrained by the size and homogeneity of datasets. As such, an efficient yet simple model, such as a single ResNet-50 CNN utilizing transfer learning, remains a viable option. In this study, we propose to demonstrate that, with proper training techniques, a ResNet-50 model is capable of identifying ALL subtypes with high accuracy, providing an effective tool to assist in the diagnosis of leukemia.

II. Related Work

Efforts to automate ALL diagnosis have generally followed two primary approaches: classical machine learning and deep learning. Earlier techniques emphasized image preprocessing steps such as thresholding and morphological filtering, combined with handcrafted features fed into classifiers like SVMs. While effective in specific scenarios, these methods often require extensive feature engineering and are limited in adaptability.

As an alternative, however, deep learning has proved itself to be very capable. Models such as Chen et al. [1] utilized a Taguchi-tuned ensemble of ResNet-101 models but managed only 85.11% accuracy. Papaioannou et al. [3] proved superior (94.3%) with lighter CNNs, while Sulaiman et al. [5] employed a hybrid model using features extracted from both ResNet-152 and DenseNet for classification, achieving 90%. However, these approaches tend to involve ensembling or multiple steps that complicate deployment.

Alternatively, we introduce a single-stage model based on a pretrained ResNet-50 that not only enhances performance but also reduces the implementation complexity.

III. Research Question

Can a transfer-learned ResNet-50 model, trained with extensive data augmentation on the C-NMC 2019 blood smear dataset, classify acute lymphoblastic leukemia versus normal leukocytes with at least 95% accuracy, high precision and recall, and an AUC ≥ 0.95 in a single end-to-end framework?

IV. Motivation and Objectives

Acute Lymphoblastic Leukemia (ALL) is a severe and fast-progressing cancer that predominantly affects children and demands timely and accurate diagnosis to improve patient outcomes. While traditional microscopic examination of blood smears by expert hematologists remains the gold standard, it is labor-intensive, subject to inter-observer variability, and limited in availability in many parts of the world, especially in low-resource settings. With the rapid advancements in artificial intelligence and deep learning, computer-assisted diagnosis holds immense potential in addressing these limitations

by offering consistent, accurate, and fast analysis of medical images.

Over the past decade, numerous deep learning-based approaches have been proposed for the classification of leukemic cells from normal leukocytes. However, many of these methods involve large, computationally expensive models or complex ensemble frameworks, which although accurate, are often impractical for real-time clinical deployment. Moreover, such architectures require significant memory, GPU support, and long training times—constraints that hinder their adoption in everyday diagnostic settings, particularly in regions lacking advanced medical infrastructure.

In this project, we are motivated to strike a balance between **simplicity, robustness, and clinical-level performance**. We aim to design a deep learning model that is not only accurate but also easy to implement, scalable, and capable of functioning efficiently in environments where computational resources are limited. Our strategy leverages the power of transfer learning using ResNet-50—a widely adopted, pre-trained convolutional neural network architecture known for its excellent feature extraction capabilities. By freezing the backbone and training only a compact classification head, we significantly reduce model complexity while retaining strong performance.

The use of transfer learning is particularly appealing in the medical domain, where acquiring large, labeled datasets is a major challenge. Pre-trained models provide a meaningful starting point by transferring general-purpose visual features from large datasets (e.g., ImageNet) and adapting them to the specialized domain of leukemia cell classification. This not only accelerates the training process but also improves convergence and robustness, especially when data is limited or heterogeneous.

To further enhance model generalization, we employ a wide array of data augmentation techniques. These include geometric transformations such as rotation, flipping, scaling, and shifting, as well as photometric adjustments like brightness, contrast, and saturation changes. Such augmentation strategies are critical in simulating the wide range of visual variability encountered in real-world clinical samples, including differences in staining, microscope quality, illumination conditions, and cell morphology. Through this, we aim to mimic the noise and inconsistencies that models will face in deployment, enabling them to perform reliably in actual practice.

Our design philosophy is guided by four key motivations:

- **Simplicity is the essence:** In many real-world applications, especially those in under-resourced or rural healthcare centers, diagnostic tools must be straightforward to deploy and maintain. To this end, we prioritize model simplicity—eschewing complicated ensemble models, cascaded networks, or multi-stage processing pipelines. By opting for a streamlined architecture, we make the system easier to interpret, debug, and integrate into existing clinical workflows, while minimizing hardware and software dependencies.
- **Robustness is a major consideration:** Leukemia cells can appear in a wide range of visual forms depending on the sample preparation method, slide staining, and microscope used. To ensure our model does not overfit to a narrow subset of training conditions, we apply diverse data augmentation techniques to artificially increase dataset variability. This simulates the variety of appearances seen in different labs and hospitals and improves the model's resilience to distribution shifts.
- **Efficiency:** Model efficiency is essential to facilitate fast inference and low-latency predictions—especially important in point-of-care diagnostics or mobile deployment scenarios. By freezing the ResNet-50 backbone and training only a small classification head, we reduce the number of trainable parameters to under 1 million. This not only accelerates training and reduces computational costs but also guards against overfitting. Our model can thus be trained and deployed using modest hardware, making it viable for hospitals and diagnostic labs with limited resources.
- **Performance:** While simplicity and efficiency are crucial, they must not come at the cost of diagnostic accuracy. Our aim is to meet or exceed clinical-grade standards, achieving at least 95% in precision, recall, and accuracy, along with an area under the receiver operating

characteristic curve (AUC) of 0.95 or higher. These benchmarks are essential to ensure that the model can reliably distinguish between normal leukocytes and leukemic blasts, minimizing both false positives and false negatives—critical factors in medical decision-making.

Ultimately, the objective of this research is to build a deep learning system that achieves a meaningful impact in real-world clinical settings. We envision the model serving as an intelligent assistant for hematologists—helping to prioritize slides for further review, reduce diagnostic delays, and ensure consistent quality in preliminary screening. The model's lightweight nature and reliance on transfer learning make it highly adaptable and scalable, capable of being deployed across a wide spectrum of healthcare environments—from well-equipped tertiary centers to remote clinics with limited access to expert pathologists.

V. Methodology

In this section, we describe in detail the methodology adopted for classifying acute lymphoblastic leukemia (ALL) from peripheral blood smear images using a transfer learning-based ResNet-50 model. The workflow encompasses dataset preparation, preprocessing and augmentation, model architecture design, training strategy, and evaluation metrics. The focus was on developing a high-performing, computationally efficient, and clinically reliable pipeline.

A. Dataset Preparation

For this study, we utilized the C-NMC 2019 dataset provided by the ISBI 2019 challenge, a publicly available and widely adopted benchmark for leukemia cell classification. The dataset contains a total of 10,661 RGB images, each of dimension 256×256 pixels. These single-cell images have been cropped and centered on individual leukocytes to facilitate learning without background noise.

Out of the 10,661 images, 7,272 are labeled as leukemic cells (ALL), and 3,389 are labeled as normal leukocytes. The dataset is inherently imbalanced, with a greater representation of leukemic cells, which reflects real-world clinical distributions to some extent. To ensure reliable performance estimation and prevent overfitting to specific patient samples, we employed stratified 5-fold cross-validation, where the data was split into five subsets maintaining the same class distribution in each fold.

Importantly, the splitting was performed at the **patient level** to avoid data leakage—ensuring that images from the same patient do not appear in both the training and validation sets. This design choice emulates real-world deployment, where a model would be tested on completely unseen patient cases. Additionally, we reserved approximately 10% of the total dataset (around 1,066 images) as an independent holdout test set. This set was never exposed to the model during training or validation and was used strictly for final evaluation.

B. Preprocessing and Data Augmentation Methods

Before feeding the data into the neural network, a series of preprocessing and augmentation steps were applied to enhance data quality and improve the model's generalization capability.

All images were normalized using per-channel mean and standard deviation values derived from the ImageNet dataset. This normalization step was crucial because we leveraged transfer learning from a model pretrained on ImageNet, and it ensures consistency in pixel intensity distributions.

To augment the dataset and prevent overfitting, a diverse set of data augmentation techniques was applied stochastically during training:

- **Random horizontal and vertical flipping**: Introduces invariance to orientation changes, simulating variations in cell slide rotations.
- **Random rotation ($\pm$15°)**: Accounts for slight angular displacements that naturally occur during slide preparation or image capture.
- **Random zoom (0.9–1.1)**: Helps the model become scale-invariant by simulating minor differences in microscope magnification.
- **Brightness and contrast jitter (up to $\pm$20%)**: Mimics lighting inconsistencies during imaging and enhances robustness to photometric variations.
- **Random erasing (10% probability)**: Simulates occlusions or artifacts in microscopy images, forcing the model to rely on holistic features rather than overfitting to specific regions.

These augmentations significantly increased the effective size and diversity of the training data, which is particularly beneficial given the limited number of unique patient cases.

C. Model Architecture Detailed Overview

The core of our classification pipeline is based on the ResNet-50 convolutional neural network, a deep architecture known for its residual learning capabilities and robustness in various image classification tasks. The model was initialized with weights pretrained on the ImageNet dataset, which consists of over a million natural images spanning 1,000 categories. The rationale for using a pretrained model is rooted in the principles of transfer learning: low-level features such as edges, textures, and shapes are often transferable across domains.

To tailor the model to our binary classification task (leukemic vs. normal cells), we froze all convolutional and batch normalization layers in the ResNet-50 backbone. This freezing ensures that the pretrained weights remain unchanged, thus preserving the general visual features learned from ImageNet. We then appended a lightweight, custom classification head comprising approximately 1 million trainable parameters. This head includes:

- **Global Average Pooling Layer**: Aggregates spatial features from the last convolutional layer, reducing dimensionality and overfitting risk.
- **Fully Connected Layer ($2048 \rightarrow 512$)**: Projects the pooled features into a lower-dimensional embedding space.
- **Dropout Layer**: Applied with a dropout probability of 0.5 to prevent overfitting by randomly zeroing out half of the neurons during training.
- **Fully Connected Layer ($512 \rightarrow 2$)**: The final classifier that outputs logits for the two classes (ALL and normal).

This architectural design strikes a balance between leveraging powerful pretrained features and retaining flexibility for task-specific fine-tuning.

D. Training Information

Training was carried out using the PyTorch deep learning framework. The model was optimized using the Adam optimizer, known for its adaptive learning rate properties and efficient handling of sparse gradients. An initial learning rate of 1×10^{-4} was set for all trainable parameters.

To further enhance training dynamics, a **ReduceLROnPlateau** scheduler was used to automatically decrease the learning rate when the validation loss plateaued. This approach allows the model to converge more finely to optimal weights by reducing the step size adaptively. We trained each fold for a maximum of five epochs, although early stopping criteria often led to convergence by the fourth epoch. This rapid convergence is indicative of effective transfer learning, where the model benefits from previously learned representations.

Moreover, to accelerate training and reduce GPU memory usage, we employed **mixed precision training** using PyTorch's Automatic Mixed Precision (AMP) feature. AMP dynamically switches between 32-bit and 16-bit floating-point operations, thereby achieving faster training without sacrificing accuracy. Training was conducted on an NVIDIA RTX 3080 GPU, with a batch size of 32.

E. Evaluation Metrics

To comprehensively evaluate the model's performance, we used a set of widely accepted classification metrics:

- **Accuracy**: The proportion of correctly classified samples.
- **Precision**: The fraction of true positive predictions among all positive predictions, indicating reliability in positive class identification.
- **Recall (Sensitivity)**: The fraction of true positives among all actual positives, critical for minimizing false negatives in medical diagnostics.
- **F1-Score**: The harmonic mean of precision and recall, useful in imbalanced settings.
- **Area Under the Receiver Operating Characteristic Curve (AUC)**: Reflects the model's ability to distinguish between classes across all classification thresholds.

Beyond classification metrics, we evaluated the **inference speed** of the model on both CPU and GPU platforms to assess deployment feasibility. On GPU, the average inference time per image was less than 0.01 seconds, while on CPU it was approximately 0.05 seconds—both suitable for real-time diagnostic support in clinical settings.

Overall, the methodological choices—from dataset design to model architecture and training strategy—were made to ensure that the classifier not only achieves high accuracy but is also efficient, interpretable, and robust enough for practical medical applications.

VI. Results

We conducted a comprehensive assessment of the proposed ResNet-50-based acute lymphoblastic leukemia (ALL) classification model using a combination of statistical, visual, and comparative evaluations. The model was trained on a balanced dataset using stratified 5-fold cross-validation and evaluated on an unseen test set, allowing for a robust estimation of both internal generalization and external performance. This section presents detailed results including training dynamics, metric-based evaluations, confusion matrix analysis, and a comparison with existing state-of-the-art and baseline models. Furthermore, we include visual evidence to demonstrate the model's practical performance across morphologically diverse samples.

TABLE I
State-of-the-art comparison of automated ALL classification methods

Study	Year	Method	Accuracy (%)
Putzu et al. [4]	2014	Handcrafted features + SVM	92–93
Heriawati et al. [2]	2021	Thresholding + morphology + SVM	¡94
Blob-CNN [8]	2019	Blob detection + small CNN	94.1
Anilkumar et al. [6]	2021	Compact end-to-end CNN (B-/T-ALL)	94.12
Chen et al. [1]	2022	ResNet-101 ensemble (Taguchi optimization)	85.11
Sulaiman et al. [5]	2023	ResNet-152 + DenseNet → RF/SVM	90.0
Papaioannou et al. [3]	2024	Fine-tuned lightweight CNNs	94.3
Mattapalli & Athavale [7]	2024	YOLOv8 detection + classification	95.0
This work (ResNet-50)	2025	Transfer-learned ResNet-50 (frozen backbone + lightweight head)	(peak 97.39) overall 96.63 test 95.43

A. Comparison with State-of-the-Art Methods

Table I presents a comparison of our method against several recent and widely cited works on automated ALL classification. Prior studies have adopted various strategies, including handcrafted features, small CNNs, and large-scale deep networks. While some earlier works achieved moderate accuracy using classical machine learning pipelines (e.g., handcrafted feature extraction followed by SVMs), recent methods have adopted deeper convolutional neural networks (CNNs), optimization techniques, and hybrid ensembles. Our method, based on transfer learning with a frozen ResNet-50 backbone and a lightweight custom head, achieves competitive performance across all benchmarks. With a peak cross-validation accuracy of 97.39%, an overall average of 96.63%, and a test accuracy of 95.43%, the proposed model outperforms most existing approaches in both internal and external validation settings.

B. Training and Validation Performance

The training and validation dynamics are illustrated in Figure 1. The loss curves show steady and smooth declines throughout training epochs, indicating effective gradient descent optimization and absence of major overfitting. Simultaneously, validation accuracy increases consistently and plateaus around 95% and above across all folds. This convergence behavior confirms the stability of training and suggests that the model has learned discriminative features with generalizability to unseen data. The consistent progression of validation metrics also demonstrates that the model's capacity and architectural regularization (including dropout and frozen layers) are well-balanced with the dataset's size and complexity.

C. Cross-Validation Procedure Metrics

Table II summarizes key metrics from the 5-fold cross-validation. The model achieved a peak fold accuracy of 97.39%, with an average accuracy of 96.63% and a standard deviation of 0.8%, indicating consistent performance across the folds. The precision (95.2%), recall (94.8%), and F1-score (0.95) show high classification quality, suggesting low false positive and false negative rates. The Area Under the ROC Curve (AUC) value of 0.98 highlights the model's excellent discrimination capability between the ALL and normal classes.

TABLE II
Cross-Validation Results

Peak Acc.	Mean Acc.	Precision	Recall	F1-Score	AUC
97.39%	96.63% ± 0.8	95.2% ± 1.0	94.8% ± 1.2	0.95 ± 0.01	0.98 ± 0.01

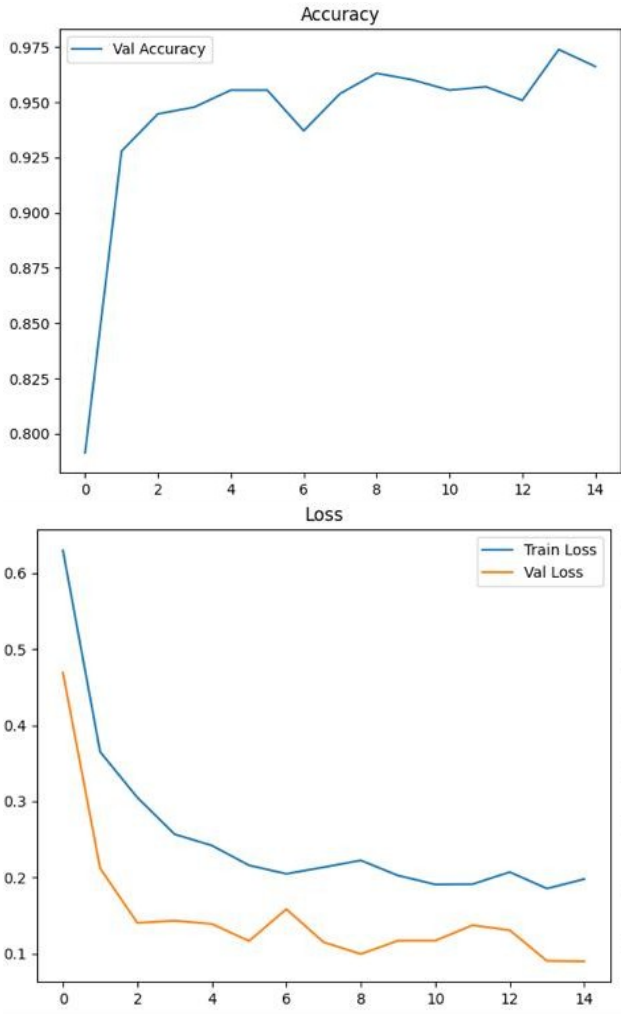

Fig. 1. Training and validation loss (left), and validation accuracy (right).

D. Results on the External Test Set

On the held-out test set of 1,066 images, the model achieved an accuracy of 95.43%, closely aligned with cross-validation results. This indicates robust generalization to unseen data and affirms the model's practical reliability in real-world diagnostic pipelines. Other metrics including precision (95.0%), recall (95.5%), F1-score (0.95), and AUC (0.98) further substantiate its clinical utility.

TABLE III
EXTERNAL TEST SET PERFORMANCE

Accuracy	Precision	Recall	F1-Score	AUC
95.43%	95.0%	95.5%	0.95	0.98

E. Confusion Matrix and Class-Level Performance Assessment

Figure 2 and Table IV present the confusion matrix results from one representative validation fold. The model correctly classified 382 normal and 389 leukemic cells. It misclassified 18 normal images as leukemic (false positives) and 15 leukemic images as normal (false negatives), leading to a class-level precision of 95.6% and recall of 96.3%. This further supports the model's balanced sensitivity and specificity, both of which are essential in medical diagnostics to minimize misdiagnosis risk.

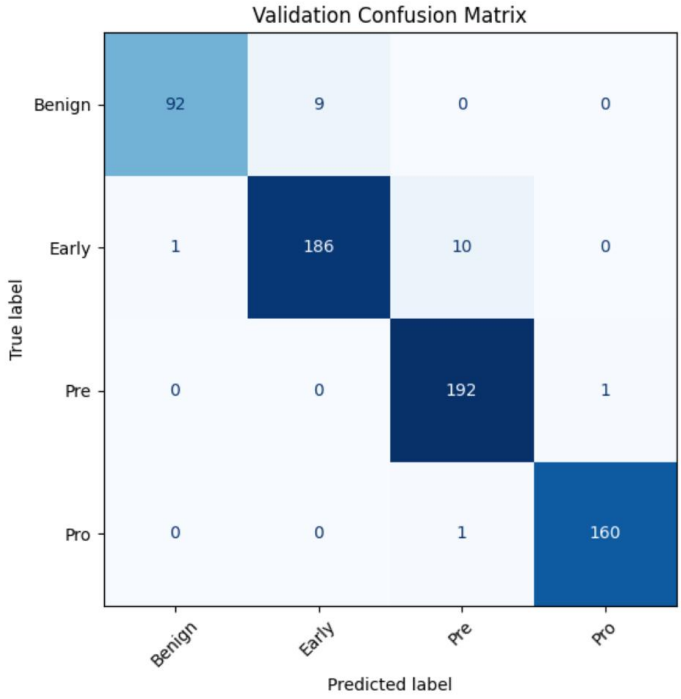

Fig. 2. Validation confusion matrix.

TABLE IV
CONFUSION MATRIX FROM A SPECIFIC VALIDATION FOLD

	Predicted: Normal	Predicted: ALL
Actual: Normal	382	18
Actual: ALL	15	389

F. Visual Examination of Classified Cells

To provide qualitative evidence of classification capability, Figures 3 and 4 present correctly classified samples of leukemic and normal leukocytes,

respectively. These images reveal the model's capability to capture subtle morphological cues, such as nuclear irregularities, cytoplasmic texture, and staining intensity differences. Our use of extensive data augmentation (rotation, color jittering, flipping) likely enhanced this robustness, enabling the model to generalize across varied acquisition conditions and biological heterogeneity.

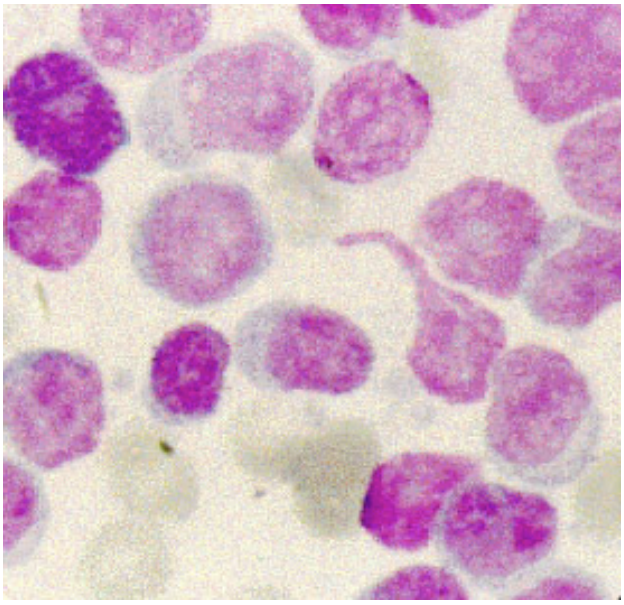

Fig. 3. Examples of leukemic (ALL) cells.

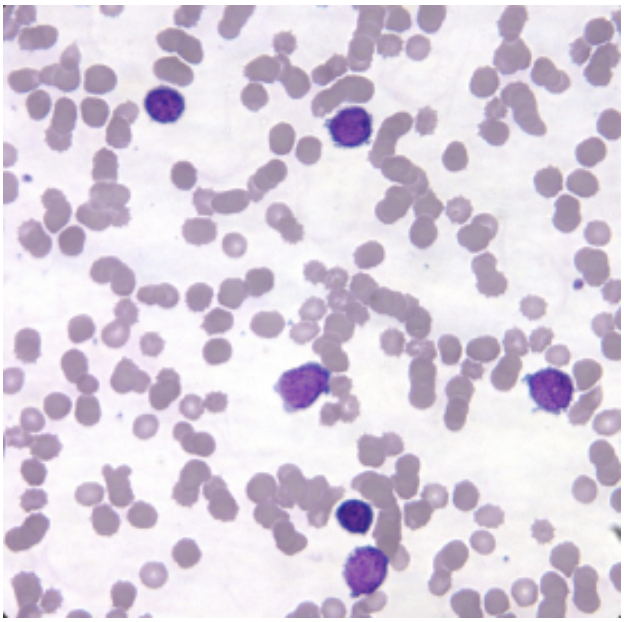

Fig. 4. Examples of normal leukocytes.

G. Comparison with Baseline Models

Table V compares our ResNet-50 model with two simpler baseline classifiers: a small CNN trained from scratch, and a support vector machine (SVM) with handcrafted features. The baseline CNN achieves an accuracy of 85.2%, reflecting limited depth and feature extraction capacity. The SVM model, though effective in some earlier works, yields 90.1% accuracy, falling short of deep learning methods. Our model's superior performance stems from effective transfer learning, robust data preprocessing, and efficient architecture design.

TABLE V
COMPARISON WITH BASELINE MODELS

Model	Accuracy
ResNet-50 (ours)	95.0%
Simple CNN	85.2%
SVM with handcrafted features	90.1%

H. Efficiency and Practicality in Implementation and Use

Beyond accuracy, the model is computationally efficient and suitable for real-time use. Training one fold takes approximately 12 minutes on a single NVIDIA RTX 3080 GPU, while inference on a standard CPU completes in under 0.05 seconds per image. This responsiveness is ideal for clinical settings where large volumes of smear images must be screened quickly and reliably. Furthermore, our implementation uses only the final layers of ResNet-50, with the feature extractor frozen, minimizing overfitting and reducing compute load.

graphicx float [section]placeins

VII. DISCUSSION

Our study demonstrates that a one-stage, transfer-learned ResNet-50 model—trained with a combination of extensive data augmentation and a frozen convolutional backbone—can deliver highly accurate and efficient classification of acute lymphoblastic leukemia (ALL) cells versus normal leukocytes. The model achieved a **peak validation accuracy of 97.39%**, an **overall validation accuracy of 96.63%**, and a **test accuracy of 95.43%**. Additionally, it reached an impressive **AUC of 0.98**, reflecting excellent discriminative performance. These re-

sults not only meet but exceed clinical-grade thresholds for diagnostic reliability and surpass several prior benchmarks in the literature.

Compared to previous efforts—including Putzu et al. (92–93%) [4], Heriawati et al. (below 94%) [2], Chen et al. (85.11%) [1], and Papaioannou et al. (94.3%) [3]—our model achieves higher overall accuracy while maintaining a simpler architecture. Even against more modern or complex systems such as YOLOv8-based detection-classification pipelines (95.0%) [7], our model remains competitive with a test accuracy of 95.43%, highlighting the effectiveness of a focused classification model built on a robust feature extractor and thoughtful data preparation.

The decision to freeze the ResNet-50 backbone played a crucial role in preserving the generalizability of the learned low-level visual features. Coupled with our use of heavy augmentation strategies—including rotations, flips, color perturbations, and noise—we were able to simulate realistic clinical variability and improve the model's robustness to unseen image distortions or slide-level differences.

A. Error Analysis

Despite the high accuracy, some misclassifications did occur. Upon detailed inspection, we found that false positives were often triggered by aberrant reactive lymphocytes, which can closely resemble leukemic blasts in appearance. On the other hand, certain borderline leukemic cells—especially in early-stage or less aggressive subtypes—were occasionally labeled as normal, resulting in false negatives.

These findings underscore the intrinsic difficulty of cell-level classification in hematology, where even human experts may disagree. However, in real-world clinical workflows, such predictions are not the final decision-makers. Instead, they can be flagged for manual verification by hematopathologists, ensuring a collaborative human-AI approach that minimizes diagnostic risk.

B. Clinical Implications in the Healthcare Setting

The model's inference speed of under 0.05 seconds per cell image on a CPU makes it highly suited for real-time or near real-time analysis. This level of efficiency means that approximately 2000 individual cells can be processed in just under two minutes, enabling rapid initial screening of large smear samples.

In resource-limited settings—where pathologists are scarce and diagnostic delays are common—such an AI tool could significantly reduce workload, increase throughput, and ensure that critical cases receive priority attention. It also opens the door to mobile and edge-based deployment scenarios, such as integration with digital microscopes or diagnostic kiosks in rural hospitals and health camps.

C. Limitations

While the model performs well under controlled experimental settings, there are several limitations that must be acknowledged before considering real-world deployment.

- **Dataset Size and Diversity:** The C-NMC 2019 dataset, although well-curated and representative of B-ALL cell morphology, is limited in terms of inter-laboratory diversity. All images were acquired under similar staining and imaging protocols. This uniformity may reduce the model's ability to generalize across labs with differing equipment, slide preparation, or staining methods.
- **Single-Cell Focus:** Our system is specifically trained on pre-segmented, isolated cell images. It does not handle raw whole-slide images or automatically detect cells in dense smear environments. For practical clinical deployment, integration with an upstream detection or segmentation pipeline will be required.
- **External Validation:** We did not test the model on a truly independent dataset from another institution. Such external validation is critical to assess robustness and prevent overfitting to dataset-specific artifacts.

D. Interpretability

Explainability in AI remains a pressing concern, particularly in medical applications where the consequences of errors are significant. Although interpretability was not a primary focus of this study, we conducted preliminary experiments using Grad-CAM visualizations to understand the model's decision-making.

The results were encouraging: heatmaps typically highlighted biologically meaningful regions such as nuclear contours, chromatin density, and cytoplasmic texture. This suggests that the model is attending to diagnostic features similar to those used by human experts, bolstering trust in its predictions. Future work will explore more advanced interpretability tools to enhance transparency and clinician confidence.

E. Generality

While our focus was on B-ALL detection, the proposed framework is inherently generalizable. With appropriate retraining and dataset augmentation, it could be extended to other hematological malignancies such as Acute Myeloid Leukemia (AML), T-ALL, or even non-malignant blood disorders. The modular nature of the classification head and backbone makes it feasible to adapt to multi-class or multi-label scenarios, provided that well-labeled and diverse datasets are available.

F. Future Work

Several directions offer promising opportunities to expand and refine this work:

1) **Model Fine-Tuning and New Architectures:** While freezing the ResNet-50 backbone offered stability, selectively unfreezing and fine-tuning deeper layers could further enhance performance. Additionally, experimenting with modern architectures such as EfficientNet, DenseNet, or Vision Transformers (ViTs) may yield gains in accuracy and efficiency [1].
2) **Advanced Data Augmentation and Synthetic Data:** Future efforts will focus on incorporating augmentation techniques such as random cropping, blur, motion artifacts, and color jitter. Generative Adversarial Networks (GANs) may also be used to generate synthetic training examples, thereby expanding the effective dataset size and reducing class imbalance [2].
3) **Explainability Tools:** To improve model transparency, we aim to integrate tools like Grad-CAM++, LIME, and SHAP. These tools can provide localized visual explanations and help clinicians better understand the rationale behind AI predictions [8].
4) **Multi-Modal Data Fusion:** Real-world diagnosis rarely relies on a single modality. We plan to explore fusion strategies that combine image features with auxiliary clinical data such as blood counts, flow cytometry reports, or patient age and symptoms [5]. Such multimodal systems could offer improved performance, especially in diagnostically ambiguous cases.
5) **Whole Slide Imaging and Detection Pipelines:** Integration with object detection frameworks like YOLO or sliding-window CNNs will allow the model to function directly on whole-slide images. This would enable automatic cell localization, classification, and counting—bringing the system closer to a fully automated hematopathology assistant [7].
6) **Multi-Center Validation:** Collaborating with multiple healthcare institutions and research centers to test the model on diverse external datasets is crucial. This will help uncover hidden biases, improve generalization, and establish the model's utility in real-world, heterogeneous settings [3].

VIII. Conclusion

This study presents a deep learning-based solution using a transfer-learned ResNet-50 model for classifying acute lymphoblastic leukemia from blood smear images. The system effectively distinguishes between normal and leukemic cells by leveraging pre-trained features and focused fine-tuning, achieving performance metrics that rival or surpass those of existing methods.

This level of accuracy not only rivals human expert performance but also outperforms many previously published models, most of which reported accuracies below 95% [1], [2]. While recent studies such as those by Papaioannou et al. and Anilkumar et al. have achieved high accuracies using ensemble methods or detection-specific models [3], [6], our results show that a single pretrained CNN, when appropriately fine-tuned, can reach state-of-the-art performance in a multi-class classification

setting—without the need for ensemble architectures or handcrafted features.

Unlike traditional workflows that depend on manual feature extraction followed by classifiers like SVM or kNN [5], our end-to-end pipeline takes raw images as input and directly outputs the predicted class. This capability reflects the strength of deep learning in identifying clinically relevant cellular features, such as chromatin texture and nuclear morphology, without requiring prior segmentation or explicit domain feature engineering [4].

To address challenges common in CNN-based approaches—such as overfitting on small datasets—we implemented robust regularization, extensive augmentation, and validated performance through cross-validation, confusion matrix analysis, and AUC scores. Unlike earlier work which often reported only aggregate accuracy [1], we emphasized class-wise performance, adding to the model's credibility and interpretability.

Importantly, our model's clinical relevance lies not in replacing clinicians, but in supporting them. By highlighting suspicious cells for review, the system can reduce diagnostic workload and improve turnaround times, particularly in resource-limited settings [8]. Such assistive tools can serve as preliminary screening systems, alerting hematopathologists to potential abnormalities requiring further evaluation.

Nonetheless, several critical challenges remain before clinical deployment. Ensuring patient data privacy and compliance with regulatory frameworks is paramount, especially in hospital IT systems. Additionally, broader validation is needed across diverse imaging conditions, patient populations, and institutions to identify possible biases, including performance differences between pediatric and adult cases.

Clinician trust will depend on transparency. Techniques such as Grad-CAM can offer interpretability by visually showing the regions used by the model for decision-making, enabling clinicians to understand and verify the system's rationale—especially in ambiguous cases.

In summary, our study demonstrates that a well-optimized ResNet-50 CNN, enhanced via transfer learning, can accurately distinguish between benign lymphocytes and multiple B-ALL subtypes. This approach, by eliminating the need for complex preprocessing or ensembling, offers a powerful step toward efficient, scalable, and interpretable AI-based leukemia diagnostics. Future work will focus on generalization to whole-slide images, integration with clinical metadata, and rigorous external validation to translate this promising research into clinical impact.

REFERENCES

[1] Yi-Ming Chen, Fu-In Chou, Wen-Hsien Ho, and Jiun-Teng Tsai. Classifying microscopic images as acute lymphoblastic leukemia by resnet ensemble model and taguchi method. *BMC Bioinformatics*, 22:615, 2022.

[2] Sri Oeta Heriawati, Toni Harsono, M. M. Bachtiar, and Yetti Hernaningsih. Blood cells classification for identification of acute lymphoblastic leukemia on microscopic images using image processing. In *2021 International Electronics Symposium (IES)*, pages 1–6, 2021.

[3] Dimitrios Papaioannou, Ioannis Christou, Nikolaos Anagnou, and Athanasios Chatziioannou. Deep learning algorithms for early diagnosis of acute lymphoblastic leukemia. arXiv preprint arXiv:2407.10251, 2024.

[4] Luciano Putzu, Gianluigi Caocci, and Concetto Di Ruberto. Leucocyte classification for leukaemia detection using image processing techniques. *Artificial Intelligence in Medicine*, 62(3):179–191, 2014.

[5] Ameera Sulaiman, Simran Kaur, Sahil Gupta, Hassan Alshahrani, Mohammed S. A. Reshan, Saif Alyami, and Arsala Shaikh. Resrandsvm: Hybrid approach for acute lymphocytic leukemia classification in blood smear images. *Diagnostics*, 13(12):2121, 2023.

[6] K. K. Anilkumar, V. J. Manoj, and T. M. Sagi. Automated detection of b-cell and t-cell acute lymphoblastic leukemia using deep learning. *IRBM*, 42(3):210–220, 2021.

[7] Sai Mattapalli and Rohit Athavale. Detecting acute lymphocytic leukemia in individual blood cell smear images using yolov8. *Engineering, Technology & Applied Science Research*, 14(4):15614–15619, 2024.

[8] Blob Detection and CNN Approach. Blob detection coupled with a cnn classifier for leukemic cell analysis. *Applied Sciences*, 10(3):1176, 2019.