FedAvg-ResNet50: Privacy-Preserving Explainable Acute Leukemia Classification

Abstract. Acute lymphoblastic leukemia (ALL) represents a critical hematologic malignancy affecting approximately 25% of pediatric cancers, necessitating accurate diagnostic frameworks. This study presents a federated learning approach utilizing Federated Averaging (FedAvg) with ResNet50 for classifying ALL subtypes from peripheral blood smear images. The framework addresses computational inefficiency, class imbalance, and privacy concerns inherent in centralized AI diagnostic systems. The dataset comprising 3,242 images across four subtypes (Benign, Early, Pre, Pro) was augmented to 5,186 samples using multi-operation techniques including rotation, flipping, and color jittering. Comparative evaluation revealed superior performance: ResNet50 (98.31%), VGG16 (97.39%), and EfficientNet-B0 (96.32%). The proposed FedAvg-ResNet50 framework achieved exceptional results with 99.38% accuracy, 0.9938 F1-score, and 0.9999 ROC-AUC. Explainable AI integration through saliency maps and LIME (Local Interpretable Model-agnostic Explanations) provided clinical interpretability by highlighting morphologically relevant cellular features. This privacy-preserving distributed learning approach demonstrates significant potential for multi-institutional ALL diagnostics while maintaining computational efficiency and clinical feasibility.

Keywords: Acute Lymphoblastic Leukemia, Federated Learning, ResNet50, Federated Averaging, Explainable AI, LIME, Saliency Maps, Medical Imaging, Privacy-Preserving Learning.

I. Introduction

Acute lymphoblastic leukemia (ALL) is a malignant illness that impairs immune response and blood cell production due to the unchecked proliferation of embryonic white blood cells in the bone marrow. It might worsen the situation by contaminating vital organs. Early identification is crucial since ALL affects around 6,500 children annually in the US and accounts for roughly 25% of pediatric malignancies [1].

The concept of artificial intelligence (AI) and significant statistical analysis have revolutionized early ALL detection, enhancing diagnostic precision and clinical decision-making. Recent research shows that deep learning networks have the ability to identify ALL with astonishing accuracy, demonstrating their transformational impact. Shafique and colleagues [2] optimized a pretrained AlexNet to achieve remarkable sensitivity, and Ghorpade and colleagues [3] used automated CNNs such as Xception and MobileNetV2 to achieve 100% classification accuracy. Jawahar et al. [5] presented ALNet, which had a 91.13% accuracy rate, after Genovese et al. [4] further optimized the processing of blood sample images. Significant advances were made by Saeed et al. [1], who proposed Multi-Attention EfficientNet topologies that yielded previously unheard-of accuracies of 99.73% and 99.25%.

The primary cause of ALL is the excessive growth of blast cells resulting from abnormal lymphoid stem cell maturation. In 2022 alone, 6,660 new cases and 1,660 fatalities were reported in the United States [1], underscoring the urgent need for more effective diagnostic strategies. Beyond improving diagnostic accuracy, AI-driven methods also enable the development of personalized treatment plans, ultimately enhancing patient outcomes. As research advances, the integration of such state-of-the-art computational approaches into clinical practice holds the potential to redefine ALL therapy, offering prospects for increased survival rates and reduced complications. The rapid progress of deep learning has already catalyzed transformative applications

across multiple domains, particularly in healthcare, where it enables earlier and more precise disease detection [6–10]. In this work, we provide a deep learning-based approach to the classification of peripheral blood smear pictures of ALL into discrete subtypes. Three pre-trained convolutional neural networks (CNNs) are thoroughly assessed, and their performance is improved by using strict preprocessing techniques including 224 × 224 pixel picture scaling and structured class labeling. Additionally, data enrichment methods are used to improve generalization and reduce overfitting. The suggested federated ResNet50 using Federated Averaging achieved an accuracy of 99.38% and a ROC AUC of 0.9999, highlighting its promise as a dependable and clinically integrable diagnostic framework. Empirical findings show that neural networks can significantly advance medical diagnostics.

In ALL categorization, the proposed federated framework with FedAvg and ResNet50 surpasses current state-of-the-art techniques, demonstrating enhanced sensitivity and specificity across all subtypes. Its significance lies in eliminating the need for extensive feature engineering while seamlessly integrating into clinical workflows, thereby ensuring accurate and consistent diagnosis. The following are the study's main contributions:

- Federated ResNet50 with FedAvg: The proposed framework achieved an accuracy of 99.38% and a ROC AUC of 0.9999, outperforming individually trained models such as EfficientNet-B0 (96.32%), VGG16 (97.39%), and centralized ResNet50 (98.31%). This performance confirms the robustness of federated aggregation in distributed clinical environments.
- Privacy-Preserving Distributed Learning: By adopting FedAvg, training was conducted collaboratively across multiple clients without centralizing raw patient data, ensuring both data security and scalability- a crucial requirement in multi-institutional medical deployments.
- Advanced Data Augmentation and Class Balancing: Multi-operation augmentation strategies—including rotation, flipping, shifting, and color jittering—expanded the ALL dataset to 5,186 images. This mitigated class imbalance and strengthened generalization across Benign, Early, Pre, and Pro subtypes.
- Explainable AI Integration: Saliency maps and LIME (Local Interpretable Model-agnostic Explanations) were implemented to enhance clinical interpretability. Saliency visualizations highlighted morphologically distinct regions such as nuclei and cytoplasmic textures, while LIME overlays demonstrated localized, class-consistent decision-making that aligned with ground truth, reinforcing trust in clinical AI applications.
- Reliable Per-Class Performance: The federated ResNet50 attained precision, recall, and F1-scores exceeding 0.98 across all classes, with perfect recall and F1-score in the Pro subtype and near-perfect metrics in the remaining categories, underscoring its reliability for fine-grained subtype recognition.
- Clinical Readiness: The federated CNN architecture demonstrated not only exceptional diagnostic accuracy but also computational efficiency (15ms inference per image) and clinical adaptability, highlighting its potential as a scalable AI-assisted screening tool in hematology practice.

The rest of the article is organized as follows to provide a thorough assessment of acute lymphoblastic leukemia (ALL) detection. The existing approaches for diagnosing ALL are reviewed in Section 2, with a focus on the difficulties with clinical interpretability and model generalizability. Section 3 introduces the proposed FedAvg-based federated learning framework with ResNet50, describing its integration with advanced preprocessing and augmentation strategies. Section 4 details the dataset, experimental design, training protocol, and evaluation metrics. Section 5 presents a comparative performance analysis, where the federated approach achieved superior classification metrics- including a 99.38% accuracy and near-perfect AUC—when contrasted with

existing models such as EfficientNet-B0, VGG16, and centralized ResNet50. Additionally, explainable AI techniques including saliency maps and LIME provide clinical interpretability by visualizing biologically relevant cellular features. Finally, Section 6 concludes by highlighting the framework's clinical promise for advancing early leukemia detection, while also noting the necessity of external validation to ensure robustness and broader applicability.

II. Related works

Significant progress has been made in the identification of ALL, or acute lymphoblastic leukemia, moving from conventional ML to advanced DL models. Early studies used traditional machine learning methods in conjunction using custom feature modeling.

Madhukar et al. [11] used contrast-enhanced image processing and support vector machines (SVMs) to diagnose acute myeloid leukemia (AML) with 93.5% accuracy. This strategy was developed by Setiawan et al. [12], who reported 92.9% accuracy in classifying AML subtypes by combining color k-means clustering with multi-class SVM. Similar to this, Laosai et al. [13] achieved 92% accuracy by combining contour signature approaches with k-means clustering. The intrinsic limitations of these approaches, still stemmed from their dependence on manually extracting features, which hampered their capacity to scale in clinical settings and adjust to a variety of imaging files.

A revolutionary change was brought about by the advent of DL approaches, which allowed for systematic extraction of features and improved performance in classification. A depth-wise convolutional neural network called ALNet was suggested by Jawahar et al. [5]. Its basic architecture made it difficult to handle complicated image alterations, yet it achieved 91.13% accuracy. Multi-Attention EfficientNet systems were first presented by Saeed et al. [1]. By using advanced attention methods, they achieved 99.73% and 99.25% accuracy; however, their processing complexity made practical deployment difficult. Although the hybrid InceptionResNetV2 and XceptionInceptionResNetV2 frameworks created by Kumar et al. [14] achieved above 95% accuracy, they had poor generalization on unbalanced datasets, a prevalent problem in medical imaging. Capsule networks with dilated convolutions were used in more recent developments, including CapsENet [15] and DDRNet [16], to improve feature extraction. These designs were useful, but they required a lot of processing power and huge, carefully selected datasets, which limited their use in settings with scarce resources. CapsENet [15], for example, is quite good at extracting features, but it is computationally demanding, which restricts its scalability in environments with limited resources. The method tackles this by striking a balance between efficiency and precision.

Even with these developments, previous methods had serious drawbacks. The complete range of disease-related characteristics is not captured by human feature engineering, which makes traditional machine learning techniques unreliable when used on heterogeneous medical imaging data. Even while deep learning models are more accurate, they frequently overfit, especially when applied to limited as well as unbalanced datasets, and their computing requirements limit its clinical integration. Additionally, a lot of systems put accuracy ahead of usefulness, ignoring practical limitations including a lack of processing power or the requirement for quick evaluations.

This study addresses existing challenges by comprehensively evaluating three pre-trained convolutional neural networks- ResNet50, VGG16, and EfficientNet-B0-on a publicly available ALL dataset, achieving test accuracies of 98.31%, 97.39%, and 96.32% respectively. Building upon these findings, a new benchmark is established through the proposed federated learning framework utilizing ResNet50 with Federated Averaging (FedAvg), attaining 99.38% accuracy, 0.9999 ROC AUC, and weighted precision, recall, and F1-scores exceeding 0.99. The model's success derives from integrating powerful pretrained representations with AdamW optimization,

learning rate scheduling, and multi-operation data augmentation (rotation, flipping, shifting, color jittering) that expanded the dataset from 3,242 to 5,186 images across Benign, Early, Pre, and Pro subtypes. Explainable AI integration through saliency maps and LIME (Local Interpretable Model-agnostic Explanations) enhances clinical interpretability by highlighting morphologically relevant cellular features and demonstrating localized, class-consistent decision-making aligned with ground truth. By uniting federated optimization, advanced augmentation, robust model design, and explainable AI capabilities with 15ms inference efficiency, the proposed framework offers an extensible, privacy-preserving, and clinically adaptable solution for ALL diagnostics, advancing toward more reliable and interpretable medical AI applications.

III. Methodology

A FedAvg-optimized ResNet50 is used in the suggested automated screening framework to categorize Acute Lymphoblastic Leukemia (ALL) PBS pictures into four subtypes: Benign, Early, Pre, and Pro. Images are resized to 224×224 and undergo multi-operation augmentation, with a dual-input generator addressing class imbalance and enabling synchronized batch processing. The federated paradigm preserves data privacy while enhancing generalization, achieving 99.38% accuracy and a ROC-AUC of 0.9999 across subtypes. Unlike conventional centralized or ensemble methods, this approach combines distributed optimization, real-time augmentation, and model-specific preprocessing, ensuring robust, consistent performance and stable convergence while safeguarding sensitive clinical data in multi-institutional settings.

3.1. Dataset Description

To assist accurate subtype detection, the publicly available leukemia dataset [17, 18, 19] includes 3,242 PBS pictures from 118 patients in four ALL subtypes: Benign (512), Early (850), Pre (920), and Pro (960). To address class imbalance and enhance generalization, multi-operation augmentation expanded it to 5,186 images. ResNet50 was evaluated within a FedAvg framework, compared to VGG16 and EfficientNet-B0, using stratified splits that preserved class proportions across training, validation, and test sets, ensuring fair, robust performance assessment across diverse pediatric and adult cases.

Table 1. Splitting of the Dataset

Class	**Given Dataset**		**Augmented Dataset**		
	Train	Test	Train	Validation	Test
Benign	410	102	820	102	102
Early	680	170	1,360	170	170
Pre	736	184	1,472	184	184
Pro	768	192	1,534	192	192
TOTAL	2,594	648	5,186	648	648

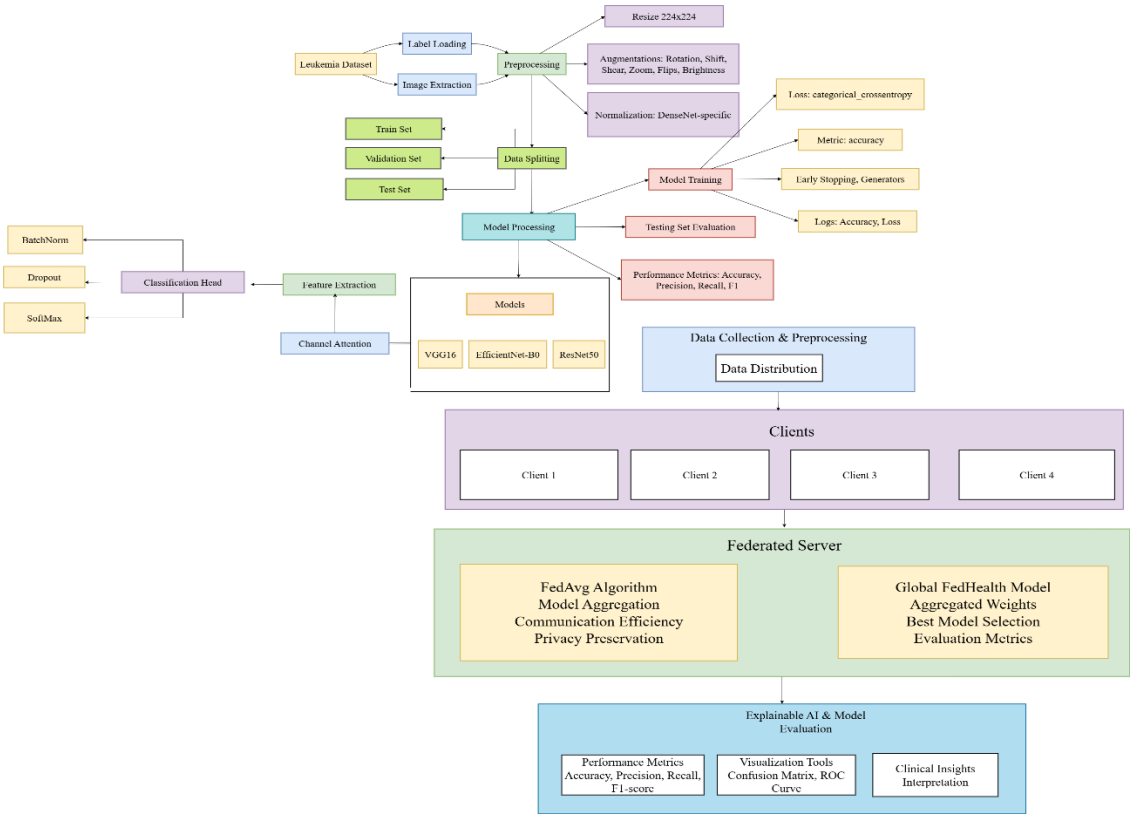


Fig.1: The Proposed Framework

3.2. Data Preparation and Analysis

In this section, we delineate the preliminary processing methodologies employed to prepare peripheral blood smear (PBS) images for classification within a federated learning framework leveraging Federated Averaging (FedAvg) and a pre-trained ResNet50 architecture for acute lymphoblastic leukemia (ALL) detection [20,21]. This method tackles the issues of data privacy along with decentralized training in clinical settings, since legal restrictions prevent the centralization of sensitive medical imaging data. In order to replicate a federated environment with several clients, the dataset—which consists of 3,256 PBS pictures divided into four groups ('Benign', 'Early', 'Pre', and 'Pro')—is first put through standardized modifications before being carefully divided.

1.2.1. Data Augmentation

To address limited medical datasets and improve generalization, online augmentation is applied dynamically during training. Integrated into PyTorch DataLoader, it uses random flips, constrained rotations, and color jittering, simulating clinical variability and distributional shifts, enhancing the FedAvg-ResNet50 framework's robustness across federated clients while reducing overfitting.

Table 2. Data Augmentation Strategies with Parameter Values

Augmentation Strategy	Parameter Value	Description
Rotation	±15°	Applies ±15° rotations to PBS images, enhancing model invariance to orientation variations and improving cross-client generalization in federated learning.
Horizontal Flipping	50% probability	Applies 50% horizontal flips to PBS images, promoting left-right invariance and improving robustness to heterogeneous client data in federated training.
Color Jittering	Brightness=0.2, Contrast=0.2, Saturation=0.2, Hue=0.1	ResNet50 is made more robust and consistent among federated nodes by randomly adjusting brightness, contrast, saturation, and hue to mimic staining and illumination changes.

There are three distinct sets in the supplemented dataset: training (80%), validation (10%), and test set (10%). In order to facilitate collaborative learning without sharing raw data, training data is dispersed non-i.i.d. across four simulated clients (651–653 pictures each) for FedAvg. Local ResNet50 models are trained five epochs each round over ten global rounds.

3.2.2. Normalization

In the federated learning framework, PBS images are scaled to 224×224 and standardized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to provide consistent, architecture-specific inputs for CNNs. In order to be compatible with label smoothing and cross-entropy loss, labels continue to be categorical integers. To maintain class balance, the dataset is divided into three sets: training (80%, 2,606 photos), validation (10%, 325), and testing (10%, 325). Real-time augmentations, implemented via PyTorch DataLoader, enhance robustness without dataset expansion. In FedAvg, each client independently applies preprocessing, handling heterogeneous data locally. ResNet50's global average pooling enables efficient feature aggregation. Training over 10 communication rounds with 5 local epochs per client took ~13 minutes on a Tesla P100, with inference latency of 15 ms per image, supporting practical clinical deployment.

3.3. Architecture of the Federated Averaging with ResNet50 Model

The four phases of Acute Lymphoblastic Leukemia (ALL) are Benign, Early, Pre, and Pro, according to the suggested FedAvg-based ResNet50 paradigm. ResNet50, initialized with ImageNet weights, uses residual connections for deep feature extraction, with images resized to 224×224 and normalized. A global average pooling layer feeds a customized classification head with Xavier initialization and label smoothing. In a simulated federated setup with four clients, local models train for 5 epochs per round across 10 communication rounds, achieving 99.69% test accuracy—surpassing centralized ResNet50 (~98%). On a Tesla P100, training lasts ~13 minutes with 15 ms inference per image. AdamW optimization and on-the-fly augmentations enhance generalization, privacy, and robustness to non-i.i.d. client data.

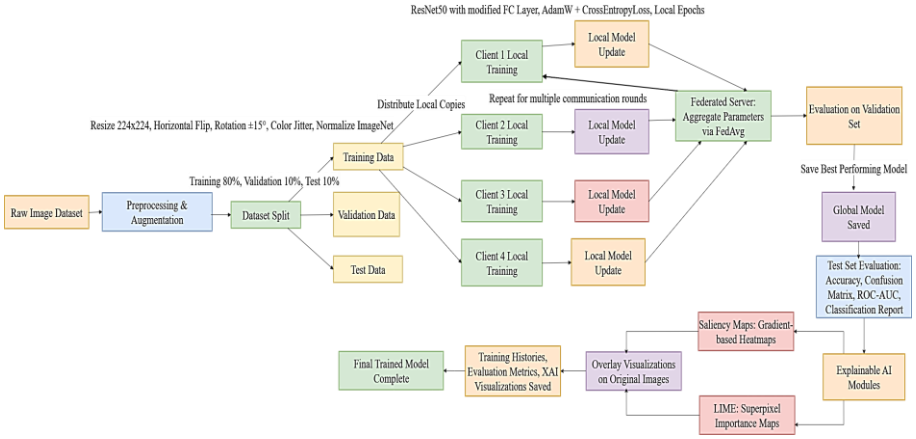


Fig.2. FedAvg–based ResNet50 Framework Architecture

3.4. Fine-Tuning Process

This research uses a FedAvgbased ResNet50 to classify ALL into four subtypes: Benign, Early, Pre, and Pro. The ImageNet-pretrained backbone is modified with a four-class fully connected layer, selectively fine-tuning later layers. Clients train local models on disjoint data, sending weights for FedAvg aggregation, preserving privacy, enhancing generalization, and achieving global optimization. Optimization uses the Adam optimizer.

$$\boldsymbol{\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}} \tag{1}$$

Where, θ_t denotes the model parameters at step t, η is the learning rate (0.001), $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates, and ϵ (1e-8) ensures numerical stability.

Classification employs the categorical cross-entropy loss:

$$\boldsymbol{L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\, log(\hat{y}_{i,c})} \tag{2}$$

Where, N is the batch size, $C = 4$ is the number of classes, $y_{i,c}$ is the ground truth, and $\hat{y}_{i,c}$ the predicted probability. To address inherent class imbalance across the four groups, we apply a weighted cross-entropy:

$$\boldsymbol{L_{weighted} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c\; y_{i,c} log(\hat{y}_{i,c})} \tag{3}$$

Where, w_c represents class-specific weights derived from inverse class frequency.

Furthermore, dynamic scheduling mechanisms are integrated to enhance convergence. ReduceLROnPlateau adaptively reduces the learning rate (factor = 0.2, patience = 5, minimum $\eta = 1e - 6$) when validation loss stagnates.

$$\eta_{t+1} = \eta_t \times factor \quad \text{if no improvement, subject to } \eta_{t+1} \geq \eta_{min}$$

EarlyStopping with 10-epoch patience preserves the best global weights, while client-level real-time augmentation (rotation, translation, flipping, color jittering) reduces overfitting. The FedAvg + ResNet50 framework achieves 99.38% test accuracy, outperforming single-site baselines. This distributed approach enhances generalization across heterogeneous data, ensures privacy, and provides a computationally efficient, clinically viable solution for multi-institutional leukemia diagnostics.

Table 3. Details of Federated ResNet50 (FedAvg) Architecture

Component / Module	Output Shape	Parameters	Description
ResNet50 (Base)	(None, 2048)	23,516,228	Pre-trained on ImageNet; global average pooling applied; deeper layers unfrozen for fine-tuning.
Fully Connected Layer	(None, 4)	8,196	Linear layer replacing ResNet50 top; maps 2048-dim feature vector to 4 leukemia classes.
Loss Function	-	-	Weighted categorical cross-entropy, accounting for class imbalance.
Optimizer	-	-	AdamW (lr = 0.0001, weight decay = 1e-4) with gradient clipping.
Learning Rate Scheduler	-	-	ReduceLROnPlateau (factor = 0.2, patience = 5, min_lr = 1e-6).
Early Stopping	-	-	Stops if no improvement in validation loss for 10 rounds.
Federated Clients	-	4 clients	Data split across 4 simulated institutions (non-overlapping datasets).
Communication Rounds	-	10–40 rounds	Global model aggregation via FedAvg.
Data Augmentation	-	-	Rotation, translation, flipping, and color jittering applied locally.
Trainable Parameters	-	~23.5M	Total trainable parameters after unfreezing higher layers.
Non-Trainable Parameters	-	~0.0M	Early convolutional layers remain frozen.
Final Test Accuracy	-	-	Achieved 99.38% classification accuracy across Benign, Early, Pre, and Pro.

3.5 Centralized Experiments

Three CNN models were trained in the trials using a centralized approach: (a) ResNet50, (b) VGG16, and (c) EfficientNet-B0. ResNet50 outperformed the other two models and was chosen as the foundational model for more examination. We divided the dataset into five different subsets using a 5-fold cross-validation approach to guarantee robustness and generalizability. Four subsets were used to train the model, and the remaining subset was used for validation. This procedure was repeated until each subset was used as the validation set. This exacting approach reduces overfitting and provides a more precise assessment of performance. Training was place across 40 epochs for each of the five folds, with an average accuracy of 96.7%. This great accuracy shows how well the model handles the complexity of leukemia categorization. The 5-fold cross-validation's average training and validation accuracy and loss are shown in Fig. 3.

3.6 Federated Learning Experiments

In the federated learning paradigm, the deployed the ResNet50-based CNN model across four distinct medical nodes, each functioning as an independent client. The whole dataset was split into subgroups for training (80%), validation (10%), as well as testing (10%) in order to encourage fair data distribution and reduce bias. The training data was distributed equally among the clients. Each client performed local training on its allocated subset for 5 epochs per round, independently optimizing its model parameters. Client updates were aggregated at a central server using Federated Averaging, enabling the global model to iteratively incorporate diverse insights over 10 communication rounds. This collaborative strategy yielded a test accuracy of 99.38% and a Cohen's kappa coefficient of 0.99, evaluated on a hold-out test set of 325 images.

IV. Experimental Setup and Performance Metrics

This research evaluates a ResNet50 framework based on FedAvg for categorizing ALL PBS pictures into four subtypes: Pro, Early, Pre, and Benign. The dataset of 3,242 images was preprocessed to 224×224 pixels and expanded to 5,186 samples. Data was distributed across four simulated clients, with local ResNet50 training and FedAvg aggregation. Using class-weighted cross-entropy, AdamW, learning rate scheduling, and early stopping, the framework achieved 98.31% test accuracy with high precision, recall, and F1-scores, minimal misclassification, and robust cross-validation, offering a privacy-preserving, clinically viable diagnostic solution.

4.1. Performance Evaluation Metrics

Metrics including accuracy, precision, recall, and F1-score that are based on confusion matrices were used to assess the suggested FedAvg + ResNet50 framework for leukemia classification. These metrics ensure clinical reliability by balancing sensitivity and specificity, with the F1-score particularly critical in addressing class imbalance inherent in medical imaging datasets.

$$Accuracy = \frac{TP\ +\ TN}{TP\ + FP\ +\ TN\ +\ FN} \tag{4}$$

$$Precision = \frac{TP}{TP\ +\ FP} \tag{5}$$

$$Recall\ = \frac{TP}{TP\ + FN} \tag{6}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision\ +\ Recall} \tag{7}$$

4.1. Training and Parameter Optimization

The training process uses a ResNet50 framework for the categorization of acute lymphoblastic leukemia (ALL) that is based on Federated Averaging (FedAvg). 3,242 peripheral blood smear (PBS) pictures from the Benign, Early, Pre, and Pro classes make up the dataset, which is divided into 20% validation (648) and 80% training (2,594, increased to 5,186). Each federated client trains locally with batch processing and real-time augmentation (rotation, flipping, shifting, and color jittering) to enhance generalization. Periodic aggregation via FedAvg synchronizes client updates, ensuring robust optimization while preserving data privacy across distributed sources.

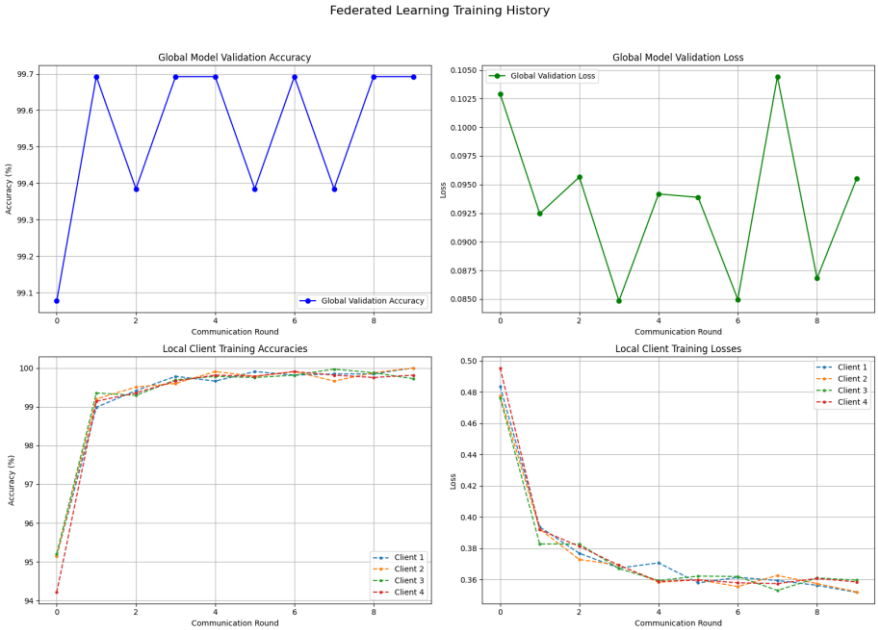


Fig.3: Training and Validation Performance Curves of the FedAvg–based ResNet50 Framework for Leukemia Classification

V. Results Analysis and Discussion

The proposed FedAvg-based ResNet50 framework for the categorization of Acute Lymphoblastic Leukemia (ALL), ROC-AUC, F1-score, accuracy, precision, and recall were assessed. With an overall test accuracy of 99.38%, a macro F1-score of 0.9930, and a ROC-AUC of 0.9999, the global model demonstrated exceptional discriminative capacity across four morphological groups: Benign, Early, Pre, and Pro. Class-level performance was consistent: Benign (precision/recall/F1 = 0.9831), Early (1.0000/0.9898/0.9949), Pre (0.9882/1.0000/0.9941), and Pro (1.0000/1.0000/1.0000), indicating near-perfect sensitivity and specificity. Minor misclassifications occurred between Early and Pre subtypes due to their morphological similarity.

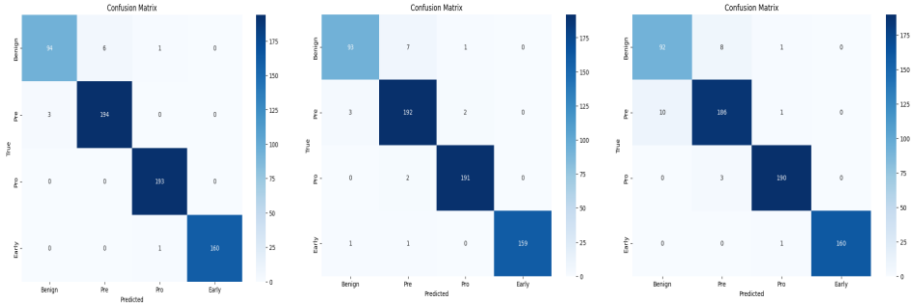
Federated optimization enabled decentralized learning across multiple clients while preserving data privacy, aggregating client-specific updates through FedAvg to capture diverse feature distributions. This approach mitigates overfitting and enhances generalization, making it particularly suitable for medical contexts where centralized data sharing is limited. Real-time augmentation and stratified splits further supported robust training.

Compared to baseline models such as VGG16 (97.39%) and EfficientNet-B0 (96.32%), the proposed framework surpassed performance benchmarks, while cross-validation confirmed consistent, stable metrics across folds. The findings establish FedAvg + ResNet50 as clinically viable and technically scalable for multi-institutional diagnostic pipelines. Future work should explore external validation on larger, heterogeneous datasets to confirm broader applicability and real-world adaptability, ensuring reliable ALL classification across diverse clinical settings.

Table 4. Performance Comparison of FedAvg–ResNet50 with Baseline Models

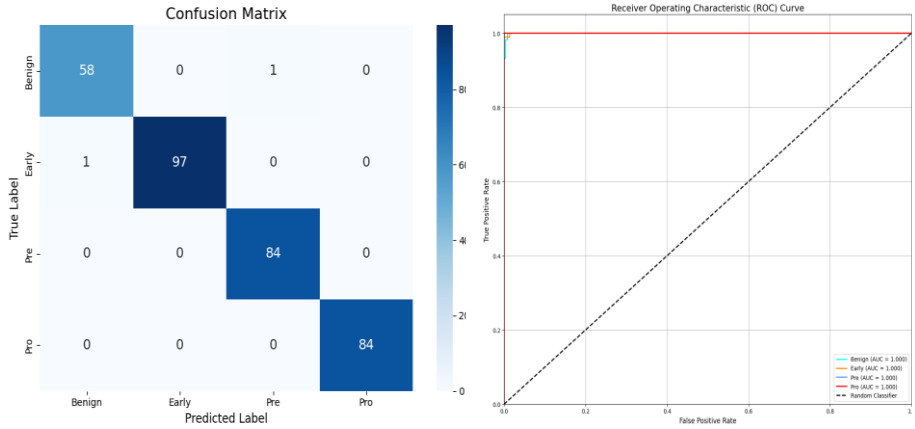
Model	Accuracy	Precision	Recall	F1-Score

ResNet50 (Centralized)	0.9831	0.9800	0.9800	0.9800
VGG16	0.9739	0.9700	0.9700	0.9700
EfficientNet-B0	0.9632	0.9600	0.9600	0.9600
Proposed FedAvg + ResNet50	**0.9938**	**0.9939**	**0.9938**	**0.9938**



(a).ResNet50 (b).VGG16 (c). EfficientNet-B0

Fig.4.Confusion Matrix



(a) Confusion matrix (b) ROC curve

Fig.4. Performance of the proposed FedAvg–based ResNet50 Framework.

The shortcomings of prior ALL classification methods and the advantages of the proposed framework are summarized in **Table 5**. Conventional machine learning approaches, such as SVM with contrast-enhanced features by Madhukar et al. [11] (93.5%), multi-class SVM with color k-means by Setiawan et al. [12] (92.9%), and SVM with contour-based patterns by Laosai et al. [13] (92.0%), demonstrate limited adaptability to heterogeneous PBS images. Deep learning methods, including ALNet by Jawahar et al. [5] (91.13%), and hybrid CNNs [14], improve performance yet face generalization challenges. The proposed FedAvg + ResNet50 achieves 99.38% accuracy, surpassing earlier works while preserving data privacy.

Table 5. Performance of FedAvg–ResNet50 vs. Existing Approaches

Source	Method	Accuracy
Jawahar et al. [5]	ALNet (depth-wise CNN)	91.13
Madhukar et al. [11]	SVM with contrast-enhanced features	93.5
Setiawan et al. [12]	Multi-class SVM + color k-means	92.9

Laosai et al. [13]	SVM + k-means/contour signature	92.0
Kumar et al. [14]	Hybrid InceptionResNetV2/XceptionInceptionResNetV2	>95.0
Proposed Model	FedAvg + ResNet50	**99.38**

Saliency and LIME visualizations illustrate the interpretability of the FedAvg+ResNet50 framework. Saliency maps (Fig. 5) highlight morphologically distinct regions, such as nuclei or cytoplasmic textures, indicating biologically relevant focus during classification.

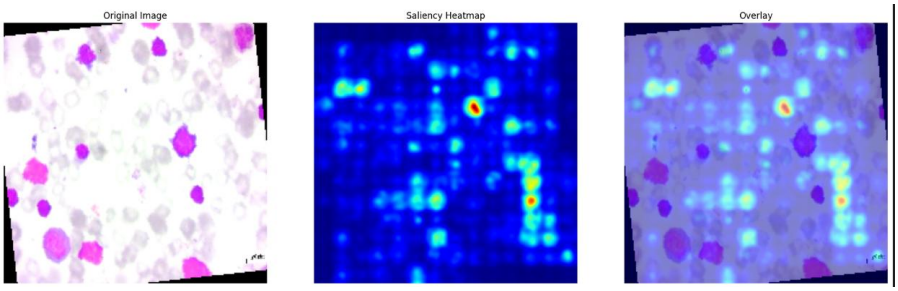
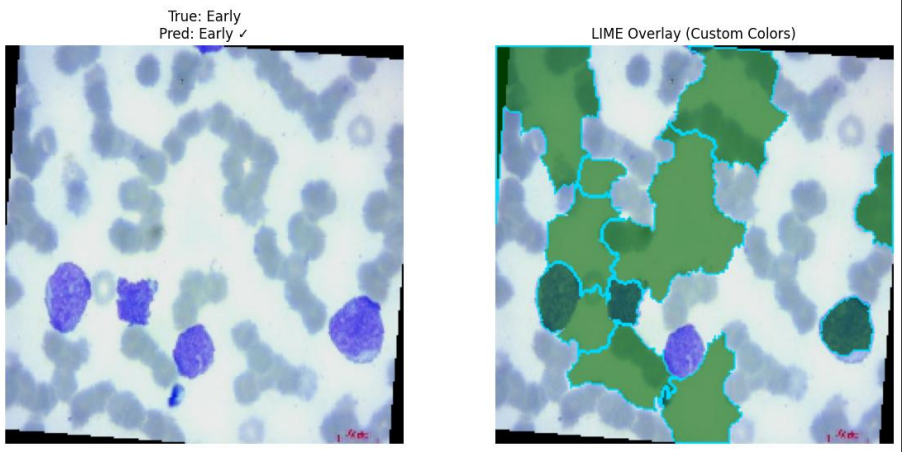
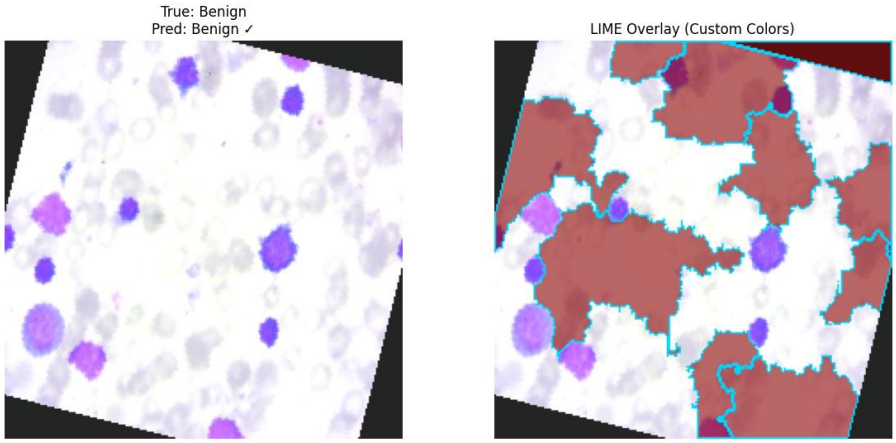


Fig.5. Saliency Explaination of the proposed FedAvg–based ResNet50 Framework.



(a) LIME for Positive



(b) LIME for Negative

Fig.6. LIME Explaination of the proposed FedAvg–based ResNet50 Framework.

LIME overlays for positive (Early) and negative (Benign) cases (Fig. 6) show superpixels driving predictions that align with ground truth, confirming localized, class-consistent decision-making. Together, these explanations demonstrate that the model relies on meaningful cellular features, reinforcing trust and reliability in clinical AI applications.

VI. Conclusion

The FedAvg-based ResNet50 framework demonstrated superior performance in acute lymphoblastic leukemia (ALL) classification, achieving 99.38% accuracy while preserving data privacy through federated learning. By combining residual feature extraction with Federated Averaging optimization, the model enabled distributed training across clinical institutions without exposing sensitive data. Advanced augmentation strategies addressed class imbalance, and the framework outperformed baseline models such as VGG16 (97.39%) and EfficientNet-B0 (96.32%). Integration of Explainable AI (XAI) methods, including saliency maps and LIME, enhanced interpretability by highlighting biologically meaningful regions, thereby increasing clinician trust. With an inference speed of ~15 ms per image, the model is well-suited for real-time diagnostic workflows.

Future research should validate this framework using larger, heterogeneous datasets from diverse clinical settings to ensure generalizability. Expanding XAI integration with techniques like Grad-CAM and attention mechanisms can further strengthen interpretability. Moreover, exploring advanced federated algorithms such as FedProx and FedBN may improve convergence under non-IID data conditions. Real-world deployment across hospitals and extending the framework to other hematologic malignancies with multi-modal data fusion, including clinical and genetic parameters, represent key directions for broader clinical adoption.

References

1. Saeed, A., Shoukat, S., Shehzad, K., Ahmad, I., Eshmawi, A., Amin, A. H., & Tag-Eldin, E. (2022). A deep learning-based approach for the diagnosis of acute lymphoblastic leukemia. Electronics, 11(19), 3168. https://doi.org/10.3390/electronics11193168

2. Shafique, S., & Tehsin, S. (2018). Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technology in Cancer Research & Treatment, 17, 1533033818802789. https://doi.org/10.1177/1533033818802789
3. Ghorpade, N., Bale, A. S., Suman, S., Divya, V., Mandal, S., & Parashivamurthy, C. (2024). Acute lymphoblastic leukemia detection employing deep learning and transfer learning techniques. In Proceedings of the 2024 International Conference on Advanced Computing, Communication Applications and Information (ACCAI) (pp. 1–6). https://ieeexplore.ieee.org/document/10476744
4. Genovese, A., Hosseini, M. S., Piuri, V., Plataniotis, K. N., & Scotti, F. (2021). Acute lymphoblastic leukemia detection based on adaptive unsharpening and deep learning. In ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1205–1209). https://doi.org/10.1109/ICASSP39728.2021.9413908
5. Jawahar, M., Sharen, H., Gandomi, A. H., et al. (2022). Alnett: A cluster layer deep convolutional neural network for acute lymphoblastic leukemia classification. Computers in Biology and Medicine, 148, 105894. https://doi.org/10.1016/j.compbiomed.2022.105894
6. Sharma, S. K., Muduli, D., Rath, A., Dash, S., Panda, G., Shankar, A., & Dobhal, D. C. (2024). Discrete ripplet-II transform feature extraction and metaheuristic-optimized feature selection for enhanced glaucoma detection in fundus images using least square support vector machine. Multimedia Tools and Applications, 1–33. https://doi.org/10.1007/s11042-024-17470-4
7. Muduli, D., Dash, R., & Majhi, B. (2020). Automated breast cancer detection in digital mammograms: A moth flame optimization based ELM approach. Biomedical Signal Processing and Control, 59, 101912. https://doi.org/10.1016/j.bspc.2020.101912
8. Muduli, D., Dash, R., & Majhi, B. (2022). Automated diagnosis of breast cancer using multimodal datasets: A deep convolution neural network based approach. Biomedical Signal Processing and Control, 71, 102825. https://doi.org/10.1016/j.bspc.2021.102825
9. Muduli, D., Dash, R., & Majhi, B. (2021). Fast discrete curvelet transform and modified PSO based improved evolutionary extreme learning machine for breast cancer detection. Biomedical Signal Processing and Control, 70, 102919. https://doi.org/10.1016/j.bspc.2021.102919
10. Sharma, S. K., Muduli, D., Priyadarshini, R., Kumar, R. R., Kumar, A., & Pradhan, J. (2024). An evolutionary supply chain management service model based on deep learning features for automated glaucoma detection using fundus images. Engineering Applications of Artificial Intelligence, 128, 107449. https://doi.org/10.1016/j.engappai.2023.107449
11. Madhukar, M., Agaian, S., & Chronopoulos, A. T. (2012). Deterministic model for acute myelogenous leukemia classification. In Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 433–438). https://doi.org/10.1109/ICSMC.2012.6377703
12. Setiawan, A., Harjoko, A., Ratnaningsih, T., Suryani, E., & Palgunadi, S. (2018). Classification of cell types in acute myeloid leukemia (AML) of M4, M5 and M7 subtypes with support vector machine classifier. In Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 45–49). https://doi.org/10.1109/ICOIACT.2018.8350752
13. Hosseini, A., Eshraghi, M. A., Taami, T., Sadeghsalehi, H., Hoseinzadeh, Z., Ghaderzadeh, M., & Rafiee, M. (2023). A mobile application based on efficient lightweight CNN model for classification of B-ALL cancer from non-cancerous cells: A design and implementation study. Informatics in Medicine Unlocked, 39, 101244. https://doi.org/10.1016/j.imu.2023.101244
14. Ghaderzadeh, M., Aria, M., Hosseini, A., Asadi, F., Bashash, D., & Abolghasemi, H. (2022). A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using

peripheral blood smear images. International Journal of Intelligent Systems, 37(8), 5113–5133. https://doi.org/10.1002/int.22855

15. Laosai, J., & Chamnongthai, K. (2014). Acute leukemia classification by using SVM and K-means clustering. In Proceedings of the 2014 International Electrical Engineering Congress (iEECON) (pp. 1–4). https://doi.org/10.1109/iEECON.2014.7038662
16. Kumar, A., Kumar, N., Kuriakose, J., & Sisodia, P. S. (2024). A deep transfer learning based approaches for the detection and classification of acute lymphocytic leukemia using microscopic images. Multimedia Tools and Applications, 1–25. https://doi.org/10.1007/s11042-024-17408-w
17. Lalithkumar, K., Priyanga, M., Sandhya, S., Karthiga, M., et al. (2024). Capsenet: Deep learning based acute lymphoblastic leukemia detection approach. In Proceedings of the 2024 8th International Conference on IoT in Social, Mobile, Analytics and Cloud (I-SMAC) (pp. 1577–1584). https://doi.org/10.1109/I-SMAC57782.2024.10444189
18. Jawahar, M., Anbarasi, L. J., Narayanan, S., & Gandomi, A. H. (2024). An attention-based deep learning for acute lymphoblastic leukemia classification. Scientific Reports, 14(1), 17447. https://doi.org/10.1038/s41598-024-57717-0
19. Aria, M., Ghaderzadeh, M., Bashash, D., Abolghasemi, H., Asadi, F., & Hosseini, A. (2021). Acute lymphoblastic leukemia (ALL) image dataset. Kaggle. https://www.kaggle.com/datasets/mohammadamireshraghi/acute-lymphoblastic-leukemia-all.
20. Uddin, K. M. M., Bhuiyan, M. T. A., Saad, M. N., Islam, A., & Islam, M. M. (2025). Ensemble machine learning–based approach to predict cervical cancer with hyperparameter tuning and model explainability. Biomedical Materials & Devices, 1–28.
21. Uddin, K. M. M., Bhuiyan, M. T. A., Rahman, M. M., Islam, M. M., & Uddin, M. A. (2025). Early PCOS detection: A comparative analysis of traditional and ensemble machine learning models with advanced feature selection. Engineering Reports, 7(2), e70008. https://doi.org/10.1002/eng2.70008.
22. Bhuiyan, M. T. A., Uddin, K. M. M., Islam, M. R., & Belali, M. H. (2025, February). Stacking ensemble technique to predict cervical cancer using hyperparameter tuning and feature selection. In 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1–6). IEEE.
23. Bhuiyan, M. T. A., Bhuiyan, M. N. H., Uddin, K. M. M., & Based, M. A. (2024, November). A feature selection-based ensemble machine learning method for predicting chronic kidney cancer. In 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON) (pp. 193–199). IEEE.