FEW-SHOT LEARNING FOR AML CELL CLASSIFICATION USING PROTOTYPICAL NETWORKS

I Gde Eka Dirgayussa[1], Kevin Elfancyus Herman[2], Doni Bowo Nugroho[3], Sekar Asri Tresnaningtyas[4], Meita Mahardianti[5], Nurul Maulidiyah[6], Rafli Filano[7], Rudi Setiawan[8], Muhammad Artha Jabatsudewa Maras[9], Yohanssen Pratama [10]

[1,2,3,4,5,6,7,8,9] Department of Biomedical Engineering, Faculty of Industrial Engineering, Institut Teknologi Sumatera, Lampung, 35365, Indonesia
[10]Nara Institute of Science and Technology, Nara, Ikoma, Takayamacho, 8916, Japan

[1]i.dirgayussa@bm.itera.ac.id, [2]kevin.121430057@student.itera.ac.id, [3]doni.nugroho@bm.itera.ac.id, [4]sekar.tresnaningtyas@bm.itera.ac.id, [5]meita.mahardianty@bm.itera.ac.id, [6]nurul.maulidiyah@bm.itera.ac.id, [7]rafli.filano@bm.itera.ac.id, [8]rudi.setiawan@bm.itera.ac.id, [9]muhammad.maras@bm.itera.ac.id, [10]yohanssen.pratama.yl0@is.naist.jp

ABSTRACT

Accurate blood cell classification is crucial for diagnosing Acute Myeloid Leukemia (AML) but limited medical data poses challenges for traditional machine learning models. This study presents a Few-Shot Learning (FSL) framework utilizing a Prototypical Network architecture with a ResNet-34 backbone to classify AML blood cell types from microscopic images. In this study, we utilize datasets consisting of 15 morphologically distinct cell classes. A 15-way, 5-shot, 5-query episodic setup was adopted to simulate data-scarce conditions. Evaluation via 5-fold cross-validation yielded strong performance, with an average accuracy of 97.76%, precision of 98.78%, recall of 96.55%, and F1-score of 97.76%. FSL training times were consistent (4.22–4.26 minutes per fold), and t-SNE along with confusion matrices confirmed the model's ability to distinguish similar cell types. To validate the approach, its performance was compared with a conventional supervised CNN using the same ResNet-34 backbone. The FSL model outperformed the CNN across all metrics such as accuracy (98.32% vs. 77.25%), precision (98.55% vs. 76.87%), recall (98.31% vs. 78.66%), and F1-score (98.33% vs. 75.26%), while also requiring far less training time (~4.24 min/fold vs. ~420 min total). These results highlight the promise of FSL based methods for accurate, efficient, and scalable hematologic diagnostics in data limited settings.

***Keywords*- Few shots learning, Acute myeloid leukemia, Prototypical networks, Classification**

I. INTRODUCTION

Acute Myeloid Leukemia (AML) is an aggressive hematologic malignancy that originates in the bone marrow and is characterized by the uncontrolled proliferation of immature myeloid precursor cells. These abnormal cells fail to differentiate into functional blood components, leading to their accumulation in the bone marrow, peripheral blood, and extramedullary tissues. As a result, normal hematopoiesis is disrupted and manifesting clinically as anemia, increased susceptibility to infections, and bleeding complications[1]. Uncontrolled proliferation of leukemic cells results in their infiltration into the bone marrow, peripheral blood, and extramedullary tissues. This infiltration disrupts normal haematopoiesis, manifesting clinically as anemia, heightened infection risk, and bleeding complications [2]. Statistically, AML is most frequently diagnosed with adults, with its incidence rising significantly with age [3]. Timely and accurate classification of AML cells is critical in clinical practice due to the marked heterogeneity in their morphology, immunophenotype, and genetic profile. These diverse characteristics substantially influence treatment response and patient prognosis. Therefore, identifying specific subtypes of AML cells can facilitate the development of more personalized and effective therapeutic strategies[4], [5]. In this context, a reliable classification system plays a pivotal role in supporting early diagnosis and enabling targeted treatment planning.

To support diagnosis, automated technologies for AML cell classification have been actively explored [6], [7], [8], [9], [10]. By leveraging such technology, the analysis of medical images becomes faster, more accurate, and less prone to human errors caused by fatigue or inattention[11]. Recent studies have demonstrated the potential of machine learning models in performing diagnostic tasks related to AML [11], [12], [13], [14]. However, the development of automated classification systems for AML still faces several major challenges[15]. A key challenge lies in the limited availability of annotated training data, which restricts model generalization across varied disease presentations and patient populations. A major limiting factor is the scarcity of accurately labelled microscopic images, especially for rare AML subtypes, which compromises generalization across cases [16]. Additionally, the issue of class imbalance where certain cell types are underrepresented, introduces bias toward the majority class. As a result, the model often overlooks minority classes, leading to suboptimal performance in those categories [17]. Although data augmentation and transfer learning have been widely used to mitigate these issues, they are often insufficient to capture the variability of images across different domains in real-world clinical settings [18]. Another persistent challenge is the limited interpretability of deep learning models, which impedes clinicians from understanding the decision-making process behind automated classification outcomes[19]. Therefore, a new approach is needed that can overcome data limitations, especially in the context of complex and

diverse AML cell classification. In response to these challenges, this study explores an AML classification framework using Few-Shot Learning (FSL) with a Prototypical Network and ResNet-34 backbone[20]. FSL has emerged as a promising approach to overcome the limitations associated with conventional supervised learning, particularly in domains where collecting large-scale annotated datasets is impractical or costly. FSL enables machine learning models to extract meaningful patterns and generalize from only a small number of labelled samples per class, thereby significantly reducing the dependency on extensive manual annotations. This characteristic makes FSL particularly advantageous in medical imaging tasks, where expert annotation is labour-intensive, time-consuming, and especially challenging for rare disease subtypes with limited representation [21].

This approach is particularly relevant for rare AML subtypes, which are challenging to collect in large quantities. Several studies have highlighted the effectiveness of FSL in improving performance in medical image classification tasks. Notably, FSL architectures have demonstrated competitive accuracy and strong adaptability to varied class distributions, making them an ideal choice for clinical applications such as AML diagnosis. [21], [22]. FSL architecture has exhibited competitive accuracy and strong adaptability to diverse class distributions [23]. Beyond minimizing reliance on large-scale data annotation, FSL also holds the potential to accelerate diagnostic workflows and enhance interpretative consistency among medical professionals. Thus, the development of an FSL based automated classification system constitutes a pivotal advancement toward the broader and more equitable integration of artificial intelligence in clinical hematology. By enabling robust model performance from limited annotated samples, FSL directly addresses one of the most pressing challenges in medical imaging. This is particularly relevant for rare AML subtypes, which are inherently underrepresented in available datasets due to their low incidence and the high cost of expert annotation. Moreover, recent studies have shown that FSL models not only achieve competitive accuracy under data-constrained conditions but also exhibit superior adaptability to imbalanced and heterogeneous class distributions[24], [25], [26], [27], [28].

The novelty of this study lies in the implementation of a customized FSL framework specifically designed for AML cell classification under data-scarce conditions. This framework addresses two major challenges: the scarcity of annotated data and class imbalance in hematologic imaging. Our contributions include a tailored Prototypical Network with ResNet-34 backbone, comprehensive evaluation using per-class metrics and visualization tools (t-SNE, ROC), and analysis of computational efficiency. This study provides not only methodological advancement in medical image classification but also a potential diagnostic aid for under resourced clinical environments. In particular, the model is engineered to tackle two critical challenges in hematologic image analysis: the limited availability of annotated samples and the pronounced imbalance across cell subtypes, especially in rare disease categories. By combining efficient feature extraction with meta learning, the proposed approach is capable of robust generalization under minimal supervision. This study contributes to the expanding body of knowledge on data-efficient deep learning in hematology by offering a comprehensive evaluation of FSL performance in classifying diverse AML cell types. Empirical results demonstrate performance that surpasses conventional deep learning baselines in few-shot scenarios. Beyond technical outcomes, the societal impact of this work is noteworthy. By enabling high-performing classification with minimal data, the framework provides a scalable solution for hematologic diagnostics in under-resourced regions. As a result, AML can be detected more quickly and precisely. The approach also promotes wider adoption of AI in medical settings.

II. METHODS

A. DATASETS

This study utilized a publicly available dataset from The Cancer Imaging Archive (TCIA), specifically the AML Cytomorphology collection[29]. The dataset comprises a total of 18,365 high-resolution with expert labelled single-cell images collected from peripheral blood smears of 200 patients. The data were curated at Munich University Hospital between 2014 and 2017 as part of a comprehensive morphological assessment initiative. Figure 1 presents examples of the fifteen morphologically distinct blood cell types used in this study.

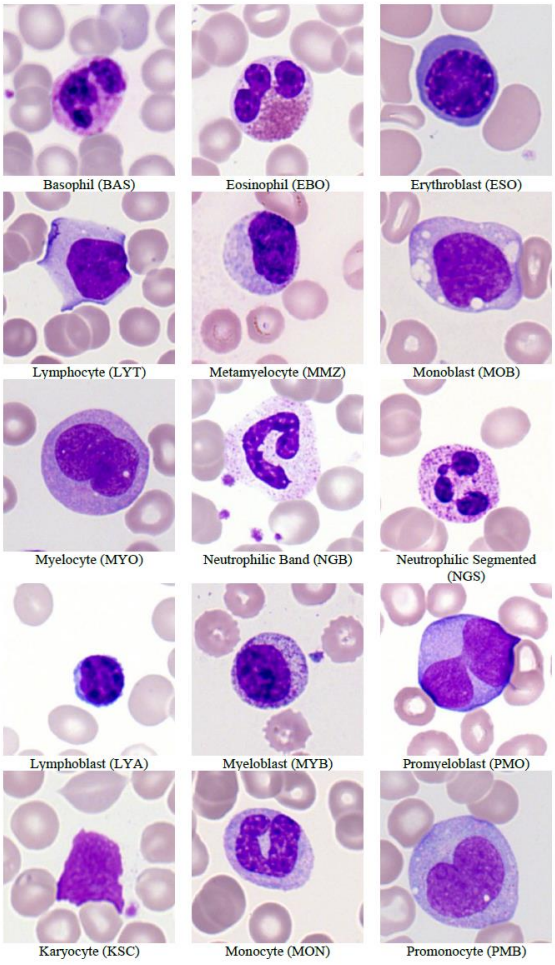

Figure 1. Examples of fifteen distinct blood cell types associated with Acute Myeloid Leukemia (AML) used in this study.

Dataset image acquisition was conducted using the M8 digital microscope/scanner (Precipoint GmbH, Freising,

Germany) under 100x optical magnification with oil immersion. Each cell image was manually annotated by trained hematology experts and categorized into 15 morphologically distinct leukocyte classes. The classification schema reflects a standardized diagnostic protocol used in clinical hematopathology, thereby increasing the clinical relevance of the dataset. To evaluate annotation consistency, a subset of cell images was relabeled up to two times, allowing for the assessment of both inter-rater and intra-rater variability. This aspect adds a level of annotation robustness that enhances the credibility of the dataset for training machine learning models. The dataset's high-quality annotations and fine-grained morphological categories have made it a benchmark resource for hematologic image classification research. Prior studies have used this dataset to train convolutional neural networks for multi-class classification, highlighting its potential in automated diagnostics. However, the limited sample count for certain rare cell types presents a unique opportunity to evaluate data-efficient learning paradigms, such as FSL. In this study, we leverage these characteristics to investigate how well FSL models can learn under realistic clinical constraints involving limited and imbalanced data availability. To ensure systematic data handling, all images were organized into subdirectories based on their corresponding cell classes. The dataset was hosted in a cloud-based storage environment and accessed via Google Collaboratory, enabling GPU-accelerated training and ensuring reproducibility of experiments. This setup facilitated efficient data loading, real-time image augmentation, and seamless integration into the FSL framework adopted in this study.

B. DATA PREPROCESSING AND AUGMENTATION

To improve model generalization and ensure compatibility with pretrained architectures, we applied a set of preprocessing and augmentation procedures to the image data. Each image was first converted into a three-channel grayscale format and resized to 224 × 224 pixels, aligning with the input size typically expected by convolutional neural networks. During training, we introduced variability through random horizontal flips, rotations of up to ±20 degrees, and moderate colour jittering, adjusting brightness and contrast by a factor of 0.2. After these transformations, the images were converted into tensor format and normalized using the mean and standard deviation values from ImageNet. These steps aimed to enhance robustness and reduce the risk of overfitting, especially given the limited sample size.

C. FEW-SHOT LEARNING CONFIGURATION

We adopted a 15-way, 5-shot, 5-query episodic learning framework to simulate realistic low-resource classification conditions. To further ensure the robustness of the model under low-resource constraints, we repeated the episodic training across multiple randomized splits. Each training episode consisted of a randomly sampled support set and query set, closely simulating deployment scenarios where only a few annotated samples may be available for rare AML subtypes. This process strengthens the ability of the model to generalize under varying intra-class and inter-class variations, thus improving its clinical applicability. This configuration follows the standard FSL evaluation protocol. In each training episode, 15 distinct classes were randomly sampled from the available dataset. For each class, 5 support images and 5 query images were selected, resulting in a total of 150 images per episode. This episodic configuration allows the model to repeatedly practice the task of learning new categories from limited examples, thereby promoting generalization to unseen classes. A total of 100 episodes were generated during the training phase to provide sufficient task diversity and avoid overfitting specific label patterns. To ensure rigorous and unbiased performance evaluation, we employed a 5-fold cross-validation strategy. The dataset was partitioned into five subsets, with each subset serving once as the validation fold while the remaining four were used for training. Average classification accuracy and other performance metrics were computed across all folds to obtain a stable estimate of model performance. Prior to episodic construction, a standardized label encoding scheme was implemented to convert descriptive class names into abbreviated numerical codes. This step ensured consistency in label handling during model training and evaluation, particularly when interfacing with PyTorch-based data loaders. Additionally, the dataset was systematically organized into class-specific folders, which streamlined the sampling process by enabling efficient indexing and retrieval of support and query sets for each episode. This organization not only improved load efficiency during training but also contributed to the reproducibility of experimental results across different environments.

D. PROTOTYPICAL NETWORK ARCHITECTURE WITH RESNET-34 BACKBONE

For the embedding function in our FSL framework, we adopted Prototypical Network architecture[30] which has been demonstrated to be highly effective for low data classification problems across various domains.

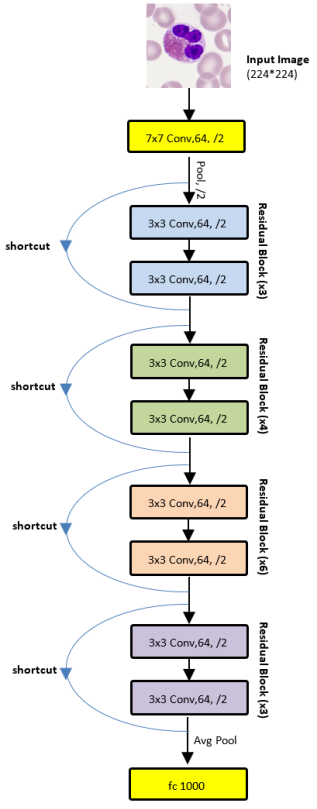

Figure 2. ResNet-34 Architecture

The Prototypical Network model works by forming a metric space where classification is carried out based on the proximity of the feature to the centroid vector (prototype) of each class. The prototype vector is obtained from the average embedding vector of the entire instance in the *support set* for each class. The pseudocode for the FSL algorithm using the Prototypical Network is shown in Table 1. In each *training episode*, the model codes both *support* and *query sets* into the feature space being studied. The classification process is performed by calculating the Euclidean distance between the embedding of the *query* sample and the nearest prototype vector of the available class.

Table 1. Pseudocode of Few-shot Learning Algorithm using Prototypical Network in this paper

Input:

Support dataset $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x^n}, \boldsymbol{y^n})\}_{n=1}^{N}$, where $\boldsymbol{N} = \boldsymbol{n}$-way × $\boldsymbol{k}$-shot
Query dataset $\boldsymbol{Q} = \{(\boldsymbol{x^m}, \boldsymbol{y^m})\}_{m=1}^{M}$
Maximum iterations $\boldsymbol{T}$, embedding network $\boldsymbol{f_\phi}$

1. Initialize parameters $\boldsymbol{\phi}$
2. **for** t = 1 to $\boldsymbol{T}$ **do**
3. Shuffle the dataset $\boldsymbol{\mathcal{D}}$
4. Sample $\boldsymbol{n}$ classes with $\boldsymbol{k}$ support and $\boldsymbol{q}$ query images
5. Form support set:

$$\boldsymbol{X^s} = \{\boldsymbol{x_i^s}\}_{i=1}^{nk}, \quad \boldsymbol{Y^s} = \{\boldsymbol{y_i^s}\}_{i=1}^{nk}$$

6. Form query set:

$$\boldsymbol{X^q} = \{\boldsymbol{x_j^q}\}_{j=1}^{nq}, \quad \boldsymbol{Y^q} = \{\boldsymbol{y_j^q}\}_{j=1}^{nq}$$

7. Compute embeddings:

$$\boldsymbol{Z^s} = \boldsymbol{f_\phi}(\boldsymbol{X^s}), \quad \boldsymbol{Z^q} = \boldsymbol{f_\phi}(\boldsymbol{X^q})$$

8. Compute prototype:

$$\boldsymbol{p_c} = \frac{\boldsymbol{1}}{\boldsymbol{k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{f_\phi}\,(\boldsymbol{x_i^s}), \quad \text{such that } \boldsymbol{y_i^s} = \boldsymbol{c}$$

9. Compute distances to prototypes:

$$\boldsymbol{d_{j,c}} = \| \boldsymbol{f_\phi}(\boldsymbol{x_j^q}) - \boldsymbol{p_c} \|^2$$

10. Compute softmax prediction:

$$\boldsymbol{\hat{y}_j} = \text{softmax}(-\boldsymbol{d_{j,:}})$$

11. Compute cross-entropy loss:

$$\boldsymbol{\mathcal{L}} = -\sum_{\boldsymbol{j=1}}^{\boldsymbol{nq}} \boldsymbol{\log \hat{y}_{j,y_j^q}}$$

12. Update $\boldsymbol{\phi}$ using gradient descent:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \boldsymbol{\alpha \nabla_\phi \mathcal{L}}$$

13. **end for**

Output: Trained embedding function $\boldsymbol{f_\phi}$

for each class k, the prototype μ_k vector is calculated as:

$$\mu_k = \frac{1}{k} \sum_{i=1}^{k} f\left(x_i^k\right) \quad (1)$$

Where $f\left(x_i^k\right)$ denotes the encoder function and x_i^k represents i^{th} support sample from class k, with K samples per class in the support set. Each prototype μ_k serves as the central representation of its respective class in the embedding space. Once the prototype is obtained, each query sample d is then evaluated by calculating the Euclidean distance to each prototype which is calculated as:

$$d(q, \mu_k) = \|f(q) - \mu_k\|_2 \quad (2)$$

Samples will be classified into the class with the prototype that has the least distance. The probability of the classification is calculated through the SoftMax function against negative distance to ensure a distributed probabilistic result based on the equation:

$$p(y = k \mid q) = \frac{\exp\left(-d(f(q), \mu_k)\right)}{\sum_{k'} \exp\left(-d(f(q), \mu_{k'})\right)} \quad (3)$$

To construct the feature embedding space, we utilized a ResNet-34 backbone as the encoder network. ResNet-34 is a deep convolutional neural network consisting of 34 weighted layers and approximately 3.6 billion floating-point operations (FLOPs)[20], [31]. This architecture is especially known for its use of residual connections identity shortcut paths that allow gradients to bypass certain layers during backpropagation thereby mitigating the vanishing gradient problem. These residual connections facilitate the optimization of deeper models by ensuring that gradient signals remain strong even in early layers, thus enabling faster convergence and improved generalization. As shown in Figure 2, the ResNet-34 architecture comprises an initial convolutional layer with a 7×7 kernel and a stride of 2. It followed by a max pooling layer, which together serves to extract and compress low-level spatial features from the input image. The core of the network consists of four sequential stages of residual blocks with increasing depth and feature richness at each stage. Specifically, the first stage includes three residual blocks, the second has four, the third has six, and the fourth concludes with three. Each residual block contains two 3×3 convolutional layers, each followed by batch normalization and a ReLU non-linearity. The network also employs a hierarchical design in which the number of feature channels doubles with each stage (e.g., from 64 to 128 to 256 to 512), while the spatial dimensions are halved via stride-2 convolutions. This mechanism enables the model to progressively learn higher-level abstractions. In the final stage of the ResNet-34 encoder, global average pooling is applied to compress the spatial dimensions of the final feature maps into a fixed-size vector. Subsequently, the architecture includes a fully connected layer which in its original ImageNet configuration, the outputs predictions across 1,000 distinct categories. In our implementation, we modified this classification layer to match the number of AML cell types defined in our dataset. A SoftMax activation function was then applied to generate a normalized probability distribution over the predicted classes. The combination of global average pooling and fully connected layers enables the network to retain class-relevant semantic information while maintaining parameter efficiency.

E. TRAINING STRATEGY

The model was trained episodically using the train protonet function with input images resized to 224×224 pixels to conform with standard convolutional network requirements. Each episode simulated a 15-way 5-shot 5-

query classification task and generating a total of 75 support images and 75 query images. Within each episode, class prototypes were computed by averaging the embedding vectors of the respective support images. The classification of query images was subsequently performed by calculating the negative Euclidean distance between their embeddings and the corresponding prototypes. The training process was optimized using the Adam algorithm[32], with loss and accuracy metrics monitored at each episode to assess convergence. To evaluate the model's ability to generalize to previously unseen classes, a 5-fold cross-validation procedure was employed and the average performance metrics across all folds were reported. Through this episodic training paradigm, the model was able to construct a robust and discriminative embedding space, thereby enhancing its capability to accurately recognize novel cell types under low resource labelling conditions.

F. EVALUATION PROTOCOL

Model performance was evaluated using the evaluate fold function on a distinct set of test episodes, employing the same 15-way 5-shot 5-query configuration as used during training. In each episode, class prototypes were derived from the support set and utilized to classify query samples by computing Euclidean distances within the learned embedding space. The predicted labels were then compared with the corresponding ground truth to assess classification accuracy. Equations (4) through (7) define the core evaluation metrics, using the following standard abbreviations: TP denotes true positives, TN refers to true negatives, FP indicates false positives, and FN corresponds to false negatives. These metrics form the basis for computing accuracy, precision, recall and F1-score, which collectively provide a comprehensive assessment of the model's classification performance. Accuracy is defined as the percentage of instances that are correctly classified out of all instances considered in the evaluation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision is defined as the ratio of true positive predictions to the total number of instances classified as positive by the model

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall reflects the classifier's sensitivity in detecting the positive class, computed by comparing correctly predicted positives to all actual positive instances, including false negatives

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1 score represents a single metric that combines precision and recall using their harmonic mean, particularly useful when evaluating performance on imbalanced datasets

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \quad (7)$$

To gain deeper insights into the model's classification performance across all categories, a confusion matrix was generated to visualize the distribution of correct predictions and misclassifications among the 15 distinct AML related cell types. For consistency, class labels within the confusion matrix were represented using the same abbreviation scheme employed during training and evaluation. This comprehensive evaluation framework provided valuable insights into the model's generalization capability under FSL constraints, particularly in contexts with limited annotated data. To support model's performance, we use a receiver operating characteristics (ROC) curve as a tool for visualize a classifier's performance and to analyse classifier behaviour depending on its performance. The Area Under ROC Curve commonly referred to as AUC, serves as a standard metric for evaluating classifier performance. Since it represents a portion of the unit square, its value inherently ranges between 0 and 1. The potential AUC values range from 0.5 (no diagnostic capabilities) to 1, with higher AUC values indicating better categorization performance[33]

G. COMPUTATIONAL HARDWARE AND SOFTWARE

All experimental procedures in this study were carried out using Google Collaboratory. Model training and evaluation were performed on virtual machines provisioned with either NVIDIA Tesla T4 or V100 graphics processing units (GPUs). These systems were configured with memory capacities ranging between 12 GB and 16 GB. The implementation of the proposed method was executed in Python version 3.10, utilizing the PyTorch framework (version 2.0) for neural network construction and optimization. Image preprocessing and augmentation pipelines were developed using the torchvision and Pillow libraries that enabling efficient data transformation and variability induction. To promote experimental reproducibility, all random number generators were initialized with fixed seed values. Furthermore, the entire experimental pipeline was executed within a consistent virtualized runtime environment to eliminate variability arising from system-level differences.

III. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the experimental results obtained from the implementation of the FSL framework. The objective is to evaluate the framework's effectiveness in accurately classifying diverse types of blood cells associated with AML, despite the constraints posed by limited annotated data. The analysis encompasses both quantitative and qualitative assessments to provide a holistic view of model performance. Specifically, this section details the classification metrics obtained across multiple folds of cross-validation, enabling an evaluation of consistency, generalizability, and robustness across different data partitions. In addition, the class-wise performance metrics are explored to highlight the model's ability to differentiate between morphologically similar cell types.

A. OVERALL PERFORMANCE OF THE FEW-SHOT CLASSIFICATION MODEL

Table 2 presents a detailed summary of the performance metrics achieved by the FSL framework, which was rigorously evaluated through 5-fold cross-validation under a 15-way, 5-shot, 5-query episodic training scheme. The model consistently exhibited outstanding classification capability, attaining an average

accuracy of 98.32%, precision of 98.55%, recall of 98.31%, and F1-score of 98.33% across all folds. Notably, the standard deviation for each metric remained low (≤ 0.47). These consistently high results affirm the model's generalizability and robustness in identifying subtle differences among 15 morphologically diverse AML related blood cell types, even under highly constrained data conditions. The ability to maintain strong performance despite being trained with only five annotated samples per class per episode emphasizes the practical utility of FSL for biomedical classification tasks in settings with limited dataset.

Table 2 The average performance metric

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	98.13	98.16	98.13	98.13
2	97.73	98.78	97.73	97.73
3	98.80	98.83	98.80	98.80
4	98.80	98.83	98.80	98.80
5	98.13	98.14	98.13	98.13
Average	98.32	98.55	98.31	98.33
Deviation	0.47	0.36	0.47	0.47

To assess the learning behavior and convergence stability of the proposed Prototypical Network model, we analyzed the training loss and accuracy curves across five cross-validation folds as shown in Figure 3. The evolution of these metrics over 100 training episodes provides critical insights into model generalization, optimization

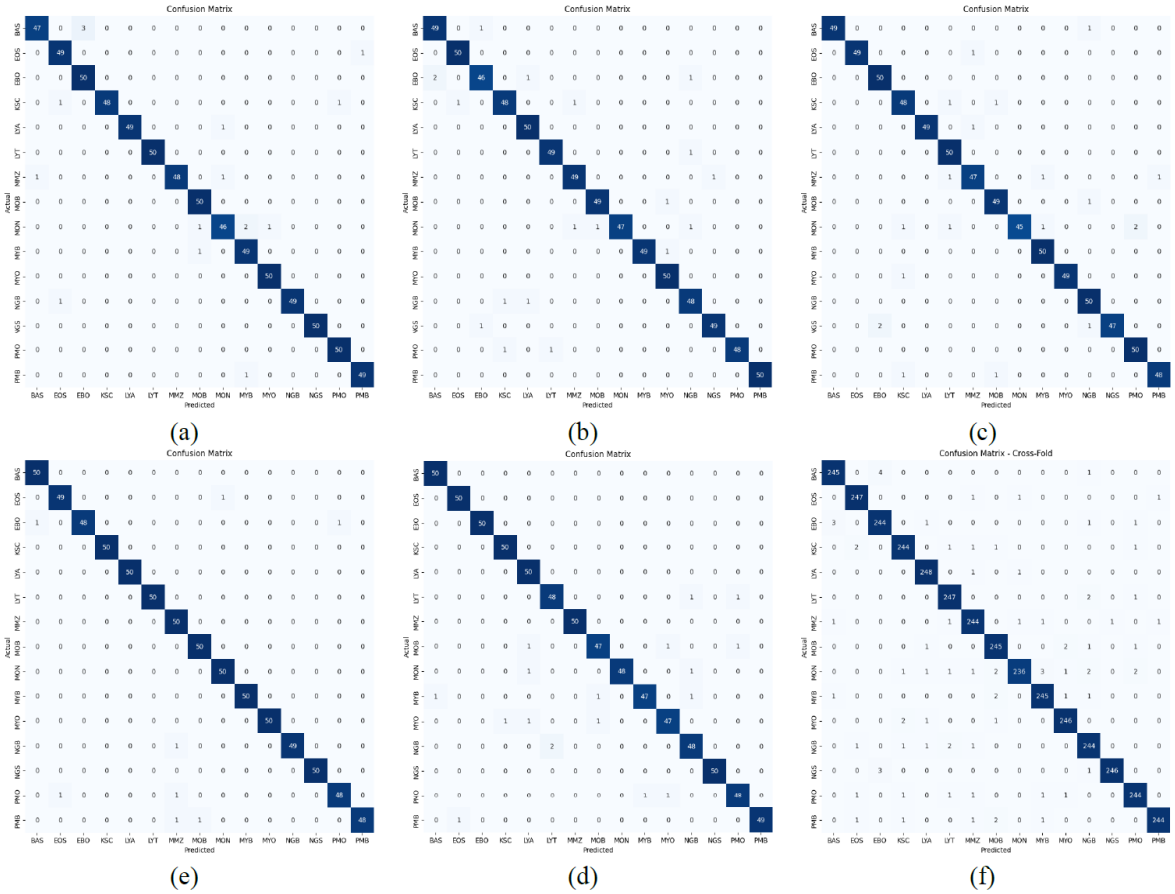

Figure 3. Performance evaluation based on training and testing: (a) Fold 1, (b) Fold 2, (c) Fold 3, (d) Fold 4, (e) Fold 5 and (f) ROC Curve

Across all folds (Figures 4, a–e), a rapid decrease in training loss was observed during the initial 10 to 20 episodes, indicating an efficient initial learning phase where the model quickly adapted to the prototypical structure of the embedding space. Correspondingly, training accuracy exhibited a steep incline within the same interval, typically surpassing 85% before episode 30. The synchronous and stable reduction in training loss, alongside a steady increase in accuracy, reflects the model's progressive learning behaviour. This pattern suggests that the embedding function gradually improved its ability to discriminate between class prototypes during the early stages of training. More specifically, Fold 1 and Fold 3 demonstrated a smooth and monotonic decline in loss, with final values stabilizing near zero and accuracy plateauing at approximately 98–100%, reflecting high intra-fold consistency and a well-formed embedding space. Fold 2 and Fold 5 displayed slightly noisier trajectories in loss curves after episode 50, although without significant degradation in accuracy. This may reflect minor fluctuations in episodic task difficulty or class variance, but not indicative of overfitting or instability.

Further performance analysis was conducted using the ROC curves, as depicted in Figure 4.f. This curve evaluates the discriminative capacity of the model across all 15 classes. Remarkably, every class yielded an Area Under the Curve (AUC) of 1.00, indicating near-perfect separation between true positives and negatives across all categories. This result reinforces the framework's capability to handle fine-grained classification tasks and

supports its suitability for clinical deployment, where high sensitivity and specificity are paramount. Collectively, these findings validate the efficacy of the proposed FSL architecture in learning robust representations that facilitate accurate classification under low-resource conditions, highlighting its potential for real-world hematology diagnostics, especially in environments where data annotation is labor-intensive or infeasible. All classes achieved an AUC of 1.00, indicating the model's excellent ability to discriminate between positive and negative classes. An ideal ROC curve lies in the upper left corner, where the True Positive Rate (TPR) approaches one and the False Positive Rate (FPR) nears zero [34]. The curves in this visualization closely approximate this ideal, suggesting minimal misclassification during testing. These results further validate the robustness of the Prototypical Network combined with ResNet-34 in handling multi-class classification tasks under few-shot settings. These findings confirm that the model not only provides accurate probabilistic predictions but also generalizes consistently across test data.

However, its final accuracy was still competitive, exceeding 97%, with minimal residual loss. The convergence behavior across all folds confirms the model's robustness under few-shot constraints, where only limited labeled exemplars per class were provided during each episode. The low final loss values (< 0.1 in most cases) suggest that the network successfully minimized intra-class distances while maximizing inter-class separability in the embedding space, a critical goal in metric-based FSL. Importantly, the absence of divergence or overfitting phenomenon with 100 training episodes.

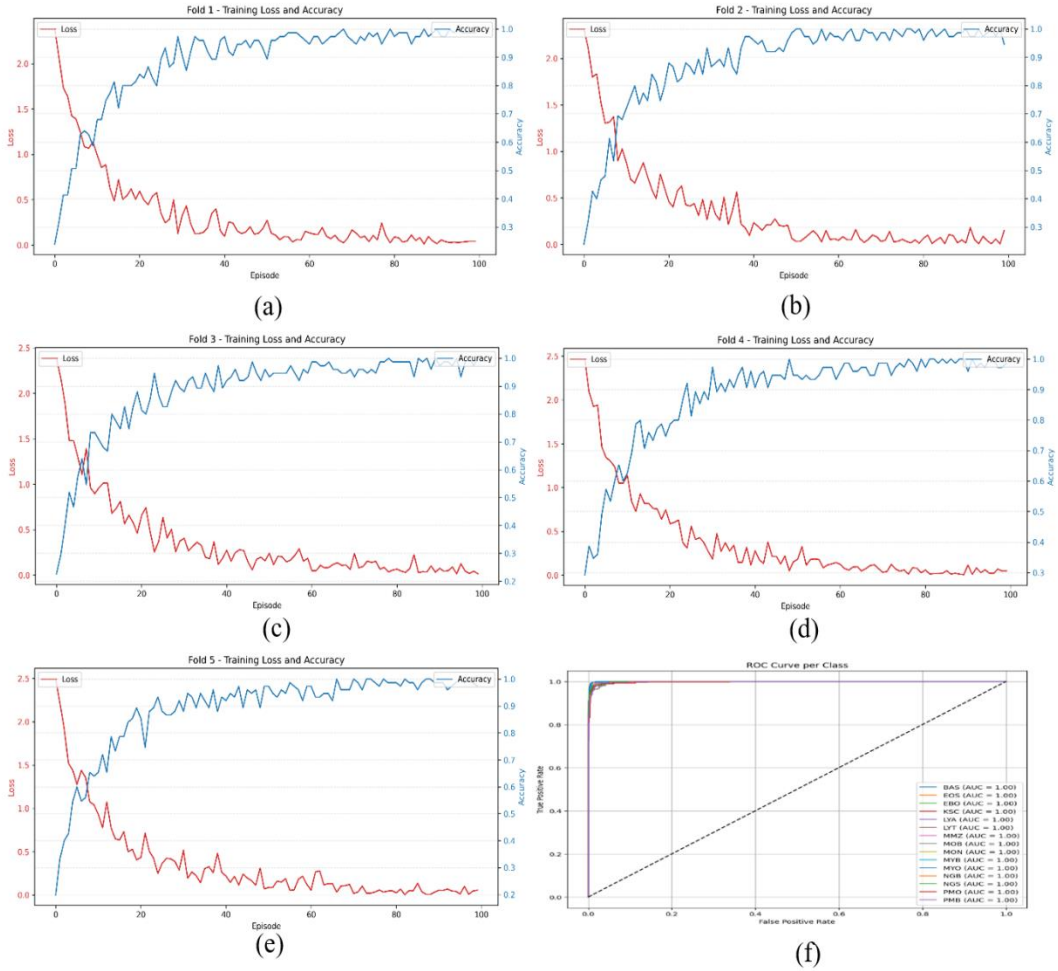

Figure 4. Confusion Matrix Comparison: (a) Fold 1, (b) Fold 2, (c) Fold 3, (d) Fold 4, (e) Fold 5 and (f) Final Evaluation Across All Folds

This process demonstrates that the episodic training framework effectively regularized the learning process. This stability is particularly notable given the morphological complexity and inter class visual similarities characteristic of the hematological cell types in our dataset. In summary, the training dynamics observed across folds highlight both the effectiveness and reliability of the Prototypical Network approach for fine-grained classification in low-data regimes. The consistent performance profiles across partitions underscore the suitability of episodic training and distance-based classification in biomedical few-shot learning scenarios.

B. CONFUSION MATRIX ANALYSIS

The confusion matrices across folds showed a strong dominance of diagonal values. These results indicate a high rate of correct classifications across most cell types. Notably, classes such as Basophil (BAS), Eosinophil (EOS), Monocyte (MON), and Segmented Neutrophil (NGS) consistently achieved more than 240 correct predictions out of 250 test samples. This result reflects the

model's strong ability to recognize and generalize the distinctive morphological characteristics of these cell types with high consistency. The confusion matrix analysis from the 5-fold cross-validation shows that the model performed excellently in distinguishing 15 AML-related blood cell types (see Figure 5f). A small number of misclassifications were observed, primarily among cell classes with closely resembling morphological characteristics. The most frequents errors occurred between Promonocytes (PMO) and Monoblast (MOB), as well as between Lymphoblasts (LYA) and Lymphocytes (LYT). These errors likely stem from overlapping visual features attributable to adjacent stages of hematopoietic differentiation differences that are challenging even for experienced human experts to distinguish. Additional misclassifications were noted between Myeloblasts (MYB) and Myelocytes (MYO), likely reflecting subtle morphological transitions during myeloid maturation. These classification errors appeared sporadic and non-systematic, exerting minimal impact on the overall performance metrics. The consistent distribution of correct classifications across all folds further supports the model's robustness and generalization capacity.

C. CLASS EVALUATION

Although the overall performance metrics in FSL scenario demonstrated high effectiveness, further evaluation is necessary to gain deeper insight into the model's behavior on a per-class basis. Such an evaluation is crucial for assessing the consistency of the model in distinguishing between different blood cell types.

In this study, key evaluation formula such as precision, recall, and F1-score were calculated for each individual blood cell class to assess the reliability of the classification results. Figure 5 presents a class wise comparison of performance metrics in the form of a bar chart, illustrating the stability of scores across all hematopoietic cell categories. The evaluation results revealed that most blood cell classes achieved precision, recall, and F1-score values exceeding 95%. These findings indicate that the model successfully captures subtle morphological characteristics that distinguish one cell type from another, despite being trained with only a few examples. The model also demonstrated effectiveness in constructing an embedding space capable of clearly separating each class. During each training episode, the model was exposed to This training strategy encouraged the model to focus on the most discriminative features, leading to strong generalization performance on unseen data.

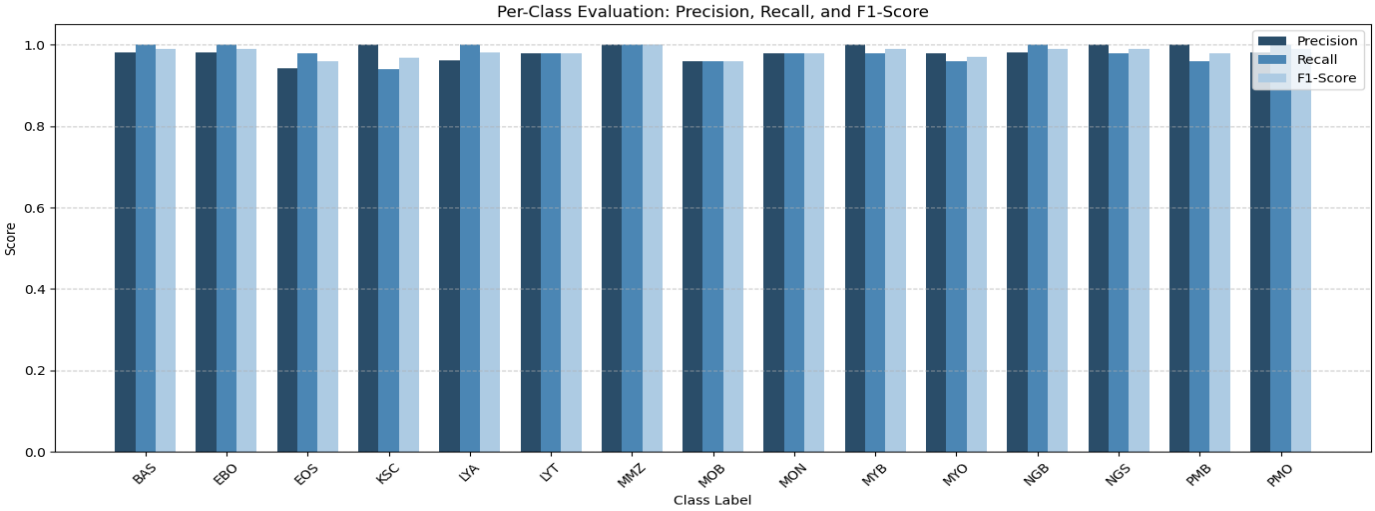

Figure 5. FSL model performance across all classes based on three key metrics: Precision, Recall, and F1-Score.

As a result, the model was not only capable of accurately identifying majority classes but also maintained high performance on minority classes or classes with limited data representation. Furthermore, the analysis showed that the model was able to distinguish between cell types with high visual similarity, such as Lymphoblast (LYA) and Karyocyte (KSC), which often cause ambiguity in microscopic diagnosis. This capability highlights the model's strength in recognizing complex morphological patterns commonly encountered in hematological analysis. Overall, these findings underscore the robustness and reliability of the FSL approach, particularly for applications in clinical settings where expert-annotated data is scarce. Consistent performance across classes supports the potential of this model to be further developed as a reliable decision support system in the diagnosis of hematological diseases, such as AML.

D. t-SNE EMBEDDING VISUALIZATION ANALYSIS

To gain deeper insight into how the model organizes and interprets learned representations, a visual analysis of the extracted features was conducted using t-distributed Stochastic Neighbour Embedding (t-SNE). This dimensionality reduction technique is widely utilized to project high-dimensional data into a lower-dimensional space typically two dimensions to facilitate intuitive visualization of complex feature distributions [35]. In this study, the 512-dimensional feature vectors generated by the ResNet-34 encoder were projected into a two-dimensional plane using t-SNE, enabling visual assessment of how the model clusters samples according to their learned representations. Figure 6 presents the resulting t-SNE projection obtained from the test set embeddings of one-fold in the cross-validation process. Each point in the plot represents a single blood cell image,

with colors corresponding to the respective AML related cell classes. The visualization shows that the model learned a feature space that organizes samples into compact and coherent clusters. These clusters generally align with the correct class labels.

However, limited overlap is observed between certain classes such as Monocytes (MON) and Neutrophilic Band cells (NGB), as well as Karyocytes (KSC) and Lymphoblasts (LYA) reflecting areas of ambiguity. These overlaps can be attributed to subtle morphological similarities or transitional features between cell types, which are common challenges in hematological image analysis. Cells undergoing early differentiation stages or belonging to the same hematopoietic lineage may exhibit overlapping visual features that complicate accurate classification.

Importantly, these visual findings are consistent with the quantitative evaluation metrics reported earlier, including high class-wise precision, recall, and dominance of correct predictions in the confusion matrix. The t-SNE visualization not only confirms the model's capacity to learn discriminative embeddings but also provides diagnostic insight into which class pairs are more susceptible to misclassification. This makes the t-SNE projection a valuable supplementary tool for model interpretation and refinement.

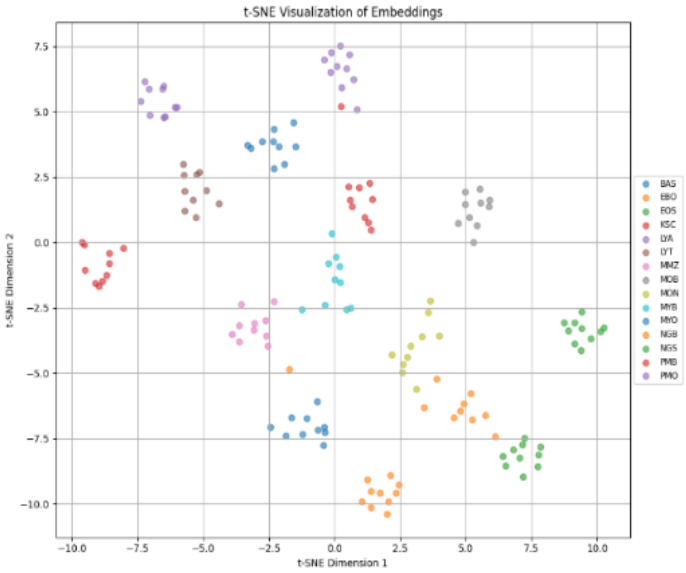

Figure 6. t-SNE plot of ResNet-34 embeddings showing distinct clusters of AML cell types. Each point represents a cell and is color-coded by class

For instance, the identified areas of class overlap can inform targeted strategies for enhancing class-specific augmentation, refining embedding sensitivity, or incorporating auxiliary information in future model iterations. Overall, the t-SNE projection provides a compelling visual complement to the statistical evaluation, reinforcing the model's effectiveness in learning meaningful representations. It also highlights opportunities for further optimization, particularly in distinguishing morphologically similar or underrepresented AML cell types. The strong performance metrics achieved by the model suggest that the embedding space learned through episodic training captures essential morphological features of each AML cell type. Furthermore, the t-SNE visualization underscores the model's ability to group semantically similar cell types while maintaining inter-class separability.

E. COMPUTATIONAL EFFICIENCY AND MODEL COMPLEXITY

In developing the proposed FSL-based blood cell classification system, training time was evaluated as a key performance metric to assess practical feasibility in resource-constrained medical environments. The model demonstrated exceptional temporal efficiency, with training durations consistently ranging from 4.22 to 4.26 minutes per fold, yielding an average of 4.24 minutes and a minimal standard deviation of 0.02 minutes.

Table 3. Training Time for Each Fold

Fold	Time (minutes)
1	4.23
2	4.22
3	4.23
4	4.26
5	4.26
Average	4.24
Deviation	0.02

This stability underscores the robustness of the model architecture and the optimization of the training pipeline, including efficient preprocessing, episodic sampling, augmentation, and GPU memory transfer. The Prototypical Network's reliance on class embeddings and Euclidean distance rather than heavily parameterized final layers combined with episodic training on small subsets (e.g., 15-way 5-shot) further reduced computational demands. Such consistency and efficiency position the proposed framework as a practical solution for point-of-care biomedical applications where both accuracy and speed are critical.

F. COMPARISON WITH CONVENTIONAL CNN-BASED CLASSIFIER

To further validate the effectiveness of the FSL framework, a comparative experiment was conducted using a conventional convolutional neural network (CNN) trained under standard supervised learning as shown in Table IV. This baseline model utilized the same ResNet-34 backbone architecture to ensure a fair comparison in terms of representational capacity. However, instead of episodic meta-learning, the conventional CNN was trained using a traditional mini-batch stochastic gradient descent on the same dataset. Moreover, the full-class label supervision applied all training samples.

The CNN baseline was trained using a stratified 80:20 split for training and validation, without applying the few-shot episodic structure. All 15 classes were included during training. All image preprocessing steps and data augmentation strategies were kept consistent with the FSL framework. The training process used cross-entropy loss and the Adam optimizer, and the model was trained for 100 epochs to ensure convergence.

Table 4. Performance Comparation between FSL and Conventional CNN

Metric	FSL (ProtoNet + ResNet-34)	CNN (Supervised + ResNet-34)
Accuracy (%)	98.32	77.25
Precision (%)	98.55	76.87
Recall (%)	98.31	78.66
F1-Score (%)	98.33	75.26
Training Time	~4.24 min/fold	~420.35 min total

The results clearly indicate that the FSL framework consistently outperformed conventional CNN across all key evaluation metrics. Notably, the FSL model achieved a significantly higher recall and F1-score, particularly in underrepresented classes. This demonstrates the superior generalization ability of metric-based meta-learning in low-data regimes. In contrast, the conventional CNN exhibited signs of overfitting to majority classes and struggled to correctly classify minority cell types. Moreover, the training time of the FSL framework was substantially lower per fold, benefiting from the reduced episodic input and efficient class prototype computation. Conventional CNN required longer training cycles due to its reliance on dense classification layers and larger training sets. These findings reinforce the suitability of FSL for biomedical classification tasks where annotated data is limited and class distribution is imbalanced. The metric-learning approach of the Prototypical Network, combined with episodic training, provides a more robust solution in comparison to conventional deep learning models.

IV. CONCLUSION

This study demonstrates that a Few-Shot Learning (FSL) approach based on a Prototypical Network with a ResNet-34 backbone can accurately classify blood cells associated with Acute Myeloid Leukemia (AML) even with limited training data. Using episodic training and 5-fold cross-validation, the model achieved an average accuracy of 98.32% with a low standard deviation, indicating strong stability and generalization capability. The t-SNE visualization confirmed clear inter-class separability, supporting the quantitative findings. The method's key strengths lie in its computational efficiency and its ability to address minority class challenges, reducing the reliance on large, fully annotated datasets. This framework holds potential for extension to other hematologic malignancies and application in resource-limited clinical settings. Despite the promising results, the study has several limitations, including the use of data from a single institution, the "black box" nature of the model, and the assumption of clean annotations. Future work should incorporate multi-institutional datasets to evaluate domain robustness, integrate Explainable AI techniques such as Grad-CAM or SHAP to improve interpretability and address noisy label handling to better reflect real-world diagnostic conditions. sing directions for applied impact in low- and middle-income countries.

V. REFERENCES

[1] P. Han Yu, Z. Yan Zhang, Y. Yuan Kang, P. Huang, C. Yang, and H. Naranmandura, "Acute myeloid leukemia with t(8;21) translocation: Molecular pathogenesis, potential therapeutics and future directions," *Biochem Pharmacol*, vol. 233, p. 116774, Mar. 2025, doi: 10.1016/J.BCP.2025.116774.

[2] Z. Koolivand, F. Bahreini, E. Rayzan, and N. Rezaei, "Inducing apoptosis in acute myeloid leukemia; mechanisms and limitations," *Heliyon*, vol. 11, no. 1, p. e41355, Jan. 2025, doi: 10.1016/J.HELIYON.2024.E41355.

[3] R. F. Schlenk, "Acute myeloid leukemia: introduction to a series highlighting progress and ongoing challenges," Feb. 01, 2023, *Ferrata Storti Foundation*. doi: 10.3324/haematol.2022.280803.

[4] D. Chen, S. Bansal, A. Mitchell, A. M. Tierens, A. G. X. Zeng, and J. E. Dick, "Isomarker: An Explainable Machine Learning Framework for Identification of Cell State Markers in Acute Myeloid Leukemia," *Blood*, vol. 144, no. Supplement 1, p. 1542, Nov. 2024, doi: 10.1182/BLOOD-2024-210986.

[5] D. Malani *et al.*, "Implementing a Functional Precision Medicine Tumor Board for Acute Myeloid Leukemia," *Cancer Discov*, vol. 12, no. 2, pp. 388–401, Feb. 2022, doi: 10.1158/2159-8290.CD-21-0410.

[6] M. S. Alim *et al.*, "Integrating convolutional neural networks for microscopic image analysis in acute lymphoblastic leukemia classification: A deep learning approach for enhanced diagnostic precision," *Systems and Soft Computing*, vol. 6, p. 200121, Dec. 2024, doi: 10.1016/J.SASC.2024.200121.

[7] B. Priscilla *et al.*, "AML-685 Leveraging Machine Learning for Rapid and Accurate Diagnosis of Acute Leukemia," *Clin Lymphoma Myeloma Leuk*, vol. 24, p. S330, Sep. 2024, doi: 10.1016/S2152-2650(24)01232-1.

[8] H. Yin *et al.*, "Classification of normal, AML, and ALL bone marrow smears based on deep learning and hyperspectral microscopic imaging," *Sens Actuators B Chem*, vol. 438, p. 137800, Sep. 2025, doi: 10.1016/J.SNB.2025.137800.

[9] M. A. Alsalem *et al.*, "A review of the automated detection and classification of acute leukaemia: Coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations," *Comput Methods Programs Biomed*, vol. 158, pp. 93–112, May 2018, doi: 10.1016/J.CMPB.2018.02.005.

[10] R. Saikia, R. Deka, A. Sarma, N. H. Singh, M. A. Khan, and S. S. Devi, "VNLU-Net: Visual Network with Lightweight Union-net for Acute Myeloid Leukemia Detection on Heterogeneous Dataset," *Biomed Signal Process Control*, vol. 107, p. 107840, Sep. 2025, doi: 10.1016/J.BSPC.2025.107840.

[11] S. Das *et al.*, "Marine Predators Algorithm with Deep Learning-Based Leukemia Cancer Classification on Medical Images," *CMES - Computer Modeling in Engineering and Sciences*, vol. 141, no. 1, pp. 893–916, Aug. 2024, doi: 10.32604/CMES.2024.051856.

[12] T. Zhang and G. Xue, "Fuzzy attention-based deep neural networks for acute lymphoblastic leukemia diagnosis," *Appl Soft Comput*, vol. 171, p. 112810, Mar. 2025, doi: 10.1016/J.ASOC.2025.112810.

[13] H. M. Rai *et al.*, "Deep Learning for Leukemia Classification: Performance Analysis and Challenges Across Multiple Architectures," *Fractal and Fractional*, vol. 9, no. 6, 2025, doi: 10.3390/fractalfract9060337.

[14] J. N. Eckardt *et al.*, "Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears," *Leukemia*, vol. 36, no. 1, pp. 111–118, Jan. 2022, doi: 10.1038/s41375-021-01408-w.

[15] J. E. Lewis, L. A. D. Cooper, D. L. Jaye, and O. Pozdnyakova, "Automated Deep Learning-Based Diagnosis and Molecular Characterization of Acute Myeloid Leukemia Using Flow Cytometry," *Modern Pathology*, vol. 37, no. 1, p. 100373, Jan. 2024, doi: 10.1016/J.MODPAT.2023.100373.

[16] H. M. Rai *et al.*, “Deep Learning for Leukemia Classification: Performance Analysis and Challenges Across Multiple Architectures,” *Fractal and Fractional*, vol. 9, no. 6, p. 337, May 2025, doi: 10.3390/fractalfract9060337.

[17] T. F. Isaka, J. Courtney, and C. Wynne, “Addressing Class Imbalance Issues in Haematological Images using Repeat Factor Sampling,” in *ICAAI 2024 - Conference Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, Association for Computing Machinery, Inc, Mar. 2025, pp. 21–27. doi: 10.1145/3704137.3704141.

[18] D. S. Depto, M. M. Rizvee, A. Rahman, H. Zunair, M. S. Rahman, and M. R. C. Mahdy, “Quantifying imbalanced classification methods for leukemia detection,” *Comput Biol Med*, vol. 152, p. 106372, Jan. 2023, doi: 10.1016/J.COMPBIOMED.2022.106372.

[19] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” Jul. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.media.2022.102470.

[20] O. Elharrouss, Y. Akbari, N. Almadeed, and S. Al-Maadeed, “Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision,” *Comput Sci Rev*, vol. 53, p. 100645, Aug. 2024, doi: 10.1016/J.COSREV.2024.100645.

[21] E. Pachetti and S. Colantonio, “A systematic review of few-shot learning in medical imaging,” *Artif Intell Med*, vol. 156, p. 102949, Oct. 2024, doi: 10.1016/J.ARTMED.2024.102949.

[22] A. Ouahab and O. Ben Ahmed, “ProtoMed: Prototypical networks with auxiliary regularization for few-shot medical image classification,” *Image Vis Comput*, vol. 154, Feb. 2025, doi: 10.1016/j.imavis.2024.105337.

[23] G. Işık and İ. Paçal, “Few-shot classification of ultrasound breast cancer images using meta-learning algorithms,” *Neural Comput Appl*, vol. 36, no. 20, pp. 12047–12059, Jul. 2024, doi: 10.1007/s00521-024-09767-y.

[24] H. M. Imran, M. A. Al Asad, T. A. Abdullah, S. I. Chowdhury, and M. Alamin, “Few Shot Learning for Medical Imaging: A Review of Categorized Images,” in *2023 IEEE 7th Conference on Information and Communication Technology (CICT)*, 2023, pp. 1–7. doi: 10.1109/CICT59886.2023.10455365.

[25] T. Dissanayake, Y. George, D. Mahapatra, S. Sridharan, C. Fookes, and Z. Ge, “Few-Shot Learning for Medical Image Segmentation: A Review and Comparative Study,” *ACM Comput. Surv.*, Jun. 2025, doi: 10.1145/3746224.

[26] Y. Ge, Y. Guo, S. Das, M. A. Al-Garadi, and A. Sarker, “Few-shot learning for medical text: A review of advances, trends, and opportunities,” *J Biomed Inform*, vol. 144, p. 104458, Aug. 2023, doi: 10.1016/J.JBI.2023.104458.

[27] T. Dissanayake, Y. George, D. Mahapatra, S. Sridharan, C. Fookes, and Z. Ge, “Few-Shot Learning for Medical Image Segmentation: A Review and Comparative Study,” *ACM Comput. Surv.*, Jun. 2025, doi: 10.1145/3746224.

[28] E. Pachetti and S. Colantonio, “A systematic review of few-shot learning in medical imaging,” *Artif Intell Med*, vol. 156, p. 102949, Oct. 2024, doi: 10.1016/J.ARTMED.2024.102949.

[29] C. , S. S. , M. C. , & S. K. Matek, “A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls [Data set],” *The Cancer Imaging Archive*, 2019, doi: https://doi.org/10.7937/tcia.2019.36f5o9ld.

[30] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4080–4090.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[32] S. Caltagirone and D. Frincke, “ADAM: Active Defense Algorithm and Model,” *Aggressive Network Self-Defense*, pp. 287–311, Jan. 2005, doi: 10.1016/B978-193183620-3/50014-2.

[33] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[34] K. Hajian-Tilaki, “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,” 2013.

[35] M. Akhavan Anvari, D. Rahmati, and S. Kumar, “t-Distributed stochastic neighbor embedding,” *Dimensionality Reduction in Machine Learning*, pp. 187–207, Jan. 2025, doi: 10.1016/B978-0-44-332818-3.00017-4.