AML vs ALL Classification Using Machine Learning and Graph Neural Networks

Abrar Ahmad, Mehar Ali, Wahaj Aziz, Hamza Farooq
Department of Computer Science
Information Technology University

***Abstract*—Classifying acute leukemia subtypes from peripheral blood smear images is essential for quick diagnosis and treatment planning. While convolutional neural networks perform well in analyzing leukemia images, they mainly focus on pixels and often do not effectively capture interactions between individual cells, which are crucial for hematological diagnosis. In this work, we suggest a graph-based approach for leukemia classification. Each microscopy image is represented as a graph, where nodes represent individual cells with morphological and categorical features, and edges show spatial connections between neighboring cells.**

We conduct an extensive study of various graph neural network (GNN) [1] architectures, including Graph Isomorphism Networks (GIN), Graph Attention Networks (GAT) [2], GraphSAGE, Transformer-based GNNs, and several improved versions that include Jumping Knowledge [3], virtual nodes, and attention-based pooling mechanisms. All models are tested using a consistent training and preprocessing process, with strict patient-level data separation to avoid data leakage.

Experimental results on the H_100X_C2 dataset show that attention-based message passing is especially effective for this task. The GAT model achieves a test accuracy of 95.0% and a sensitivity of 99.26% for Acute Myeloid Leukemia (AML), reflecting a very low false-negative rate for clinically significant cases. Our findings underscore the need to model cell-to-cell relationships using adaptive attention methods and demonstrate that graph neural networks offer a powerful and clear framework for leukemia classification based on cell-level annotations.

***Index Terms*—Leukemia, AML, ALL, Graph Neural Networks, Medical Imaging, Machine Learning**

I. Introduction

Acute leukemia is a serious blood cancer [4] that needs quick and precise classification to determine treatment options. The main types, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL), have different prognoses, treatment protocols, and clinical outcomes. Traditionally, diagnosing leukemia relies on experts examining images of blood or bone marrow smears. Pathologists look at individual cell shapes, how they are arranged, and how they relate to each other. However, this manual analysis can take a lot of time, is subjective, and depends on the availability of specialists. This situation has led to the need for automated and supportive diagnostic systems.

Recent developments in deep learning, especially convolutional neural networks (CNNs), have produced encouraging results for classifying leukemia from microscopy images. These methods usually work on the pixel or patch level, learning visual features directly from raw images. While they are effective, these approaches often treat images as unstructured grids and do not specifically capture the relationships between individual cells. In contrast, diagnosing blood disorders relies not just on how cells look alone but also on their spatial arrangement, co-occurrence of cell types, and the overall patterns seen in the smear. Overlooking these interactions may reduce the accuracy and reliability of models based only on pixels.

Graph-based representations offer a clear and logical way to model these relationships. By depicting each cell as a node and using edges to represent spatial closeness or biological similarity, graphs allow for explicit modeling of how cells interact with each other. Graph Neural Networks (GNNs) [1] build on deep learning techniques for non-Euclidean settings and have shown good results in tasks where understanding relationships is crucial, such as predicting molecular properties, analyzing social networks, and increasingly in biomedical [5] settings. In histopathology [6] and cytology, GNNs can merge cell-level features with spatial context, similar to how human experts reason during diagnosis.

Despite this promise, using GNNs for leukemia classification is still not widely explored. Current studies often focus on just one architecture or fail to compare various message-passing methods, pooling strategies, and architectural modifications in a systematic way. Additionally, the effects of attention mechanisms, network depth, and global context modeling in cell-level leukemia graphs have not been thoroughly examined across a consistent experimental setup.

In this study, we fill these gaps by providing a thorough evaluation of multiple GNN architectures for binary leukemia classification [7] (AML vs. ALL) using cell-level annotations from microscopy images. Each image is turned into a graph where nodes represent individual cells with morphological and categorical features, and edges show spatial relationships created through a k-nearest neighbor approach. We test various GNN models [1], including Graph Isomorphism Networks (GIN), Graph Attention Networks (GAT), GraphSAGE, and Transformer-based GNNs, alongside enhanced versions that integrate Jumping Knowledge [3] connections, virtual nodes, and attention-based pooling methods. All models use the same preprocessing pipeline and a strict patient-level data division to avoid information leakage.

Our experimental results on the H_100X_C2 dataset show that attention-based message passing is particularly effective for classifying leukemia. The GAT [2] model reaches a test accuracy of 95.0

The main contributions of this study are summarized as follows:

- We introduce a cell-level graph representation for leukemia classification that combines morphological, categorical, and spatial information from microscopy images.
- We conduct a systematic and unified comparison of key GNN architectures [1] and enhancements for graph-level leukemia classification.
- We show that attention-based GNNs [1] perform better than fixed aggregation methods, achieving state-of-the-art results on the H_100X_C2 dataset.
- We provide a detailed analysis of pooling strategies, network depth, and global context modeling, offering practical insights for future graph-based medical imaging systems [5].

II. Related Work

A. Deep Learning for Leukemia and Blood Cell Image Analysis

Automated analysis of peripheral blood and bone marrow smear images has been an active area of research, especially with the rise of deep learning. Early methods depended on handcrafted features that described cell shape, texture, and color. These were followed by traditional classifiers like support vector machines or random forests. Recently, convolutional neural networks (CNNs) have become the main approach, performing well in tasks such as leukocyte detection, cell segmentation, and leukemia subtype classification. CNN-based models have been used for both patch-level and whole-image classification, taking advantage of their ability to learn visual patterns directly from raw pixel data.

Several studies have shown high accuracy in identifying leukemia subtypes using CNNs trained on microscopy images. However, these methods generally view images as regular grids. They focus on local visual patterns and struggle to capture relationships among individual cells fully. Some works do include multi-instance learning or attention mechanisms at the image or patch level, but the relationships between cells are often modeled implicitly. This may affect how understandable and robust the models are, especially in cases where spatial arrangement and patterns of cell co-occurrence are important for diagnosis.

B. Graph-Based Methods in Medical Image Analysis

Graph-based representations have emerged as a powerful method for modeling structured and relational information in medical imaging. In histopathology and cytology, graphs are often built by representing cells or tissue components as nodes, with edges that show spatial closeness or biological links. These representations have been applied to tasks such as tumor grading, tissue classification, and disease prognosis, allowing for explicit modeling of spatial context and intercellular interactions [5] [8].

Graph Neural Networks (GNNs) [1] adapt message-passing techniques for graph-structured data and have demonstrated strong performance in biomedical tasks [5] where understanding relationships is crucial. Previous research has looked into using GNNs [1] for analyzing histopathological images, often showing better results than purely convolutional methods by utilizing interactions at the cell level. However, many existing studies concentrate on one GNN architecture [1] or a particular pooling method, making it hard to draw general conclusions about the effectiveness of different message-passing strategies in medical [5] contexts.

C. Graph Neural Network Architectures and Pooling Strategies

Many GNN architectures [1] have been put forward, each varying in how they gather neighborhood information. Graph Isomorphism Networks (GIN) are built to be maximally expressive under the Weisfeiler–Lehman test, using learnable multi-layer perceptrons for gathering data. GraphSAGE incorporates neighborhood sampling and fixed aggregation functions, which enables scalable learning on larger graphs. Graph Attention Networks (GAT) utilize learnable attention mechanisms to assign different weights to neighbor contributions, which allows the model to focus on the most informative relationships.

Beyond message passing, pooling strategies are vital for graph-level classification. Global pooling methods like sum, mean, or max aggregation offer straightforward and efficient graph representations. More advanced techniques, such as attention-based pooling and Jumping Knowledge [3] connections, aim to capture information at multiple scales or in a more global context. Transformer-based GNNs further enhance attention methods by adding edge-aware transformations.

Despite the increasing variety of GNN architectures and pooling techniques, their comparative behavior in cell-level medical [8] image graphs is still not well understood. Specifically, the advantages of attention-based aggregation, adaptive depth selection, and global context modeling for leukemia classification have not been thoroughly assessed. Our work aims to fill this gap by providing a comprehensive experimental comparison of multiple GNN architectures [1] and enhancements within a consistent preprocessing and evaluation framework.

Contributions

This work makes the following key contributions:

- We propose a cell-level graph representation for leukemia classification [9]. Individual cells extracted from microscopy images are modeled as nodes with morphological and categorical features. Spatial relationships are represented through weighted edges.
- We present an empirical evaluation of major Graph Neural Network (GNN) architectures. This includes Graph Isomorphism Networks (GIN), Graph Attention Networks (GAT) [2], GraphSAGE, and Transformer-based GNNs, all under a consistent preprocessing, training, and evaluation process.

- We analyze the impact of architectural improvements such as Jumping Knowledge [3] connections, virtual nodes, and attention-based pooling strategies on graph-level leukemia classification performance.
- We show that attention-based message passing performs much better than fixed aggregation methods. The proposed GAT model achieves a test accuracy of 95.0 percent and a sensitivity of 99.26 percent for Acute Myeloid Leukemia (AML) on a patient-level split.
- We provide detailed experimental analysis and practical guidance for implementing graph-based learning methods in medical [8] image analysis tasks that involve cell-level relational structure.

III. Dataset and Graph Construction

A. Dataset Source

The dataset used in this study is the *Blood Cancer Dataset (Leukemia)* from the Intelligent Machines Lab at Information Technology University (ITU) related to MICCAI 2024. The dataset is publicly accessible through the official repository linked in the project materials. The Intelligent Machines Lab collected and organized all microscopy images and annotations. This study uses the dataset only for academic research [9].

B. Data Acquisition and Imaging Conditions

The dataset includes peripheral blood smear microscopy images taken under various imaging conditions to mirror real-world diagnostic differences. Images were collected at three magnification levels (100×, 40×, and 10×) using three different setups: a high-quality microscope camera, a low-cost microscope camera, and a mobile phone camera (Honor 9X Lite). These variations lead to differences in spatial resolution, lighting, color distribution, and noise levels.

Each image follows a standardized naming system:

`H/L_100X/40X/10X_C1/C2`,

where `C1` indicates mobile phone camera acquisition, `C2` indicates microscope camera acquisition, and `H/L` differentiates between high-quality and low-quality camera setups.

C. Class Composition and Task Definition

The dataset includes samples from various leukemia subtypes [9]. However, most images are of Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Minority subtypes like Chronic Myeloid Leukemia (CML) have very few samples and were excluded to avoid class imbalance and potential labeling errors. Thus, this work focuses on a binary classification task between AML and ALL at the image (graph) level.

D. Selected Subset

All experiments in this study use the `H_100X_C2` subset, which includes images captured at 100× magnification with a high-quality microscope camera. This subset provides the richest cellular detail, making it ideal for detailed cell-level analysis. Images are divided at the patient level based on predefined `train.json` and `test.json` files to prevent information leakage across groups.

E. Cell-Level Annotations

Each image in the selected subset comes with structured JSON annotation files that provide information at the cell level. For each annotated cell, the dataset includes:

- **Cell centroid coordinates** (x, y) in image space.
- **Cell category label**, which indicates the type of annotated cell based on expert labeling.
- **Morphological attributes** collected during the dataset preparation, such as cell area, perimeter, and shape descriptors.

Cells without valid annotations after preprocessing are removed. Images with no valid annotated cells are excluded from further analysis.

F. Graph Construction

Each microscopy image is represented as an undirected graph $G = (V, E)$, where each node $v_i \in V$ corresponds to an individual annotated cell. Node features are created by combining categorical and numerical information provided in the dataset annotations. Specifically, categorical cell-type labels are represented using one-hot encoding, while numerical morphological attributes are adjusted to have a zero mean and unit variance.

Edges are formed based on the proximity of cells. For each node, edges connect it to its k nearest neighboring cells in Euclidean space using centroid coordinates. Edge weights are defined as the inverse of the Euclidean distance between connected cell pairs. This setup allows the graph to represent both adjacency and relative spatial closeness.

The resulting graph captures local cellular interactions and broader tissue organization patterns within each microscopy image. Each graph is labeled according to the leukemia subtype (AML or ALL) associated with the source image.

IV. Methodology

This section explains the graph-based learning framework for classifying leukemia. We start by outlining the problem setup, then discuss the graph neural network architectures we used and the training process.

A. Problem Formulation

We have a set of microscopy images. Each image is turned into a graph $G = (V, E)$ as described in Section III. Each graph represents a single patient sample and has a label $y \in \{0, 1\}$, indicating either Acute Myeloid Leukemia (AML) or Acute Lymphoblastic Leukemia (ALL). The goal is to create a graph-level classifier $f(G; \theta)$ that predicts the leukemia subtype based on the relationships and attributes of the labeled cells [9].

B. Graph Neural Network Framework

Let $h_i^{(l)}$ be the feature representation of node i at layer l. Graph neural networks update node representations by repeatedly gathering information from neighboring nodes according to

$$h_i^{(l+1)} = \phi\Big(h_i^{(l)},\ \text{AGG}\big(\{h_j^{(l)} : j \in \mathcal{N}(i)\}\big)\Big),$$

where $\mathcal{N}(i)$ represents the neighbors of node i, $\text{AGG}(\cdot)$ is an aggregation function that does not depend on the order of inputs, and $\phi(\cdot)$ is a transformation we can learn.

We test several GNN architectures within this single framework. These include Graph Isomorphism Networks (GIN), Graph Attention Networks (GAT) [2], GraphSAGE, and Transformer-based graph models. All architectures use the same graph inputs for a fair comparison.

C. Attention-Based Message Passing

In attention-based models, we perform neighborhood aggregation using learned attention coefficients that determine how much influence neighboring cells have. For nodes i and j, we calculate the attention coefficient α_{ij} as

$$\alpha_{ij} = \frac{\exp\big(a(h_i, h_j)\big)}{\sum_{k \in \mathcal{N}(i)} \exp\big(a(h_i, h_k)\big)},$$

where $a(\cdot)$ is a function that we can learn. This approach helps the model focus on important cellular interactions while downplaying less relevant neighbors.

D. Graph-Level Representation

After L rounds of message passing, we gather node embeddings into a fixed-dimensional graph representation through a global pooling operation. We explore both mean pooling and pooling based on attention. Attention-based pooling produces a weighted sum of node embeddings, allowing the model to concentrate on key cell groups that have the greatest impact on the final diagnosis.

The pooled graph representation goes through a fully connected classifier to make the final prediction.

E. Training Objective

We optimize the model parameters by reducing the binary cross-entropy loss between the predicted probabilities and the actual labels. We use the Adam optimizer for training, with early stopping based on validation results. All models are trained under the same optimization conditions to maintain consistency across experiments.

V. Models and Results

This section presents a detailed comparison of fifteen GNN models regarding leukemia classification. The GNN models being evaluated are categorized into three main architectures: Graph Isomorphism Networks (GIN), GraphSAGE, and Graph Attention Networks (GAT) [2] with their significant variants. To make a legitimate comparison, all models are subjected to training and evaluating with the same experimental conditions, namely the same graph construction pipeline and dataset splits.

A. Graph Isomorphism Networks (GIN)

We discuss Graph Isomorphism Networks first because they have a strong theoretical background when it comes to distinguishing between different graph structures. The basic architecture of GIN comprises three GINConv layers, each of which is determined by a two-layer MLP with batch normalization and ReLU activations. Then, graph-level representations are generated through global sum pooling.

The three-layer baseline GIN reached a test accuracy of 0.9152. The reduction of the depth to two layers makes the generalization easier, hence the result obtained is a test accuracy of 0.9304. This actually means that the really deep model may lead the training set to overfit because the dataset size is not large enough [10].

There are quite a few different architectural modifications considered in this context. The addition of a virtual node aimed at enabling the flow of global information results in an overall decrease in the quality of validation and test performance, which indicates that the imposed global context is not a friend to this task. The attention pooling approach, on the other hand, provides very little improvement over the just met baseline but it does not surpass simpler pooling techniques either. At the same time, the employment of Jumping Knowledge (JK) connections has led to a marked performance gain by countering the node representation at different scales, and finally, a test accuracy of 0.9413 was reached.

B. GraphSAGE

GraphSAGE models are subsequently assessed because of their inductive neighborhood aggregation mechanism, which is perfectly compatible with scalable biomedical graph learning. The three-layer GraphSAGE model, which serves as a baseline, reaches a test accuracy of 0.9130. Likewise, in the case of GIN, reducing the depth to two layers enhances generalization and results in higher test accuracy of 0.9174.

Which includes the improved versions, GraphSAGE with Jumping Knowledge connections is the only one to always perform better than the rest and reaches a test accuracy of 0.9326. Attention-based pooling gives only small improvements, while the virtual node variant again leads to lower accuracy. These findings indicate that it is better to use JK connections to maintain hierarchical neighborhood representations than to use virtual nodes to impose global context.

C. Graph Attention Networks (GAT)

Through the evaluation of Graph Attention Networks, the efficacy of learnable attention mechanisms in the modeling of cell-to-cell interactions is assessed. The baseline three-layer GAT [2] model, having good validation performance and little overfitting, is able to attain the highest single-model test accuracy of 0.9500.

When the depth is dropped down to two layers, there is a slight performance loss, which points to the fact that the deeper attention-based aggregation is helpful for this task. The performance GAT [2] is not consistently improved by the incorporation of Jumping Knowledge [3] connections and

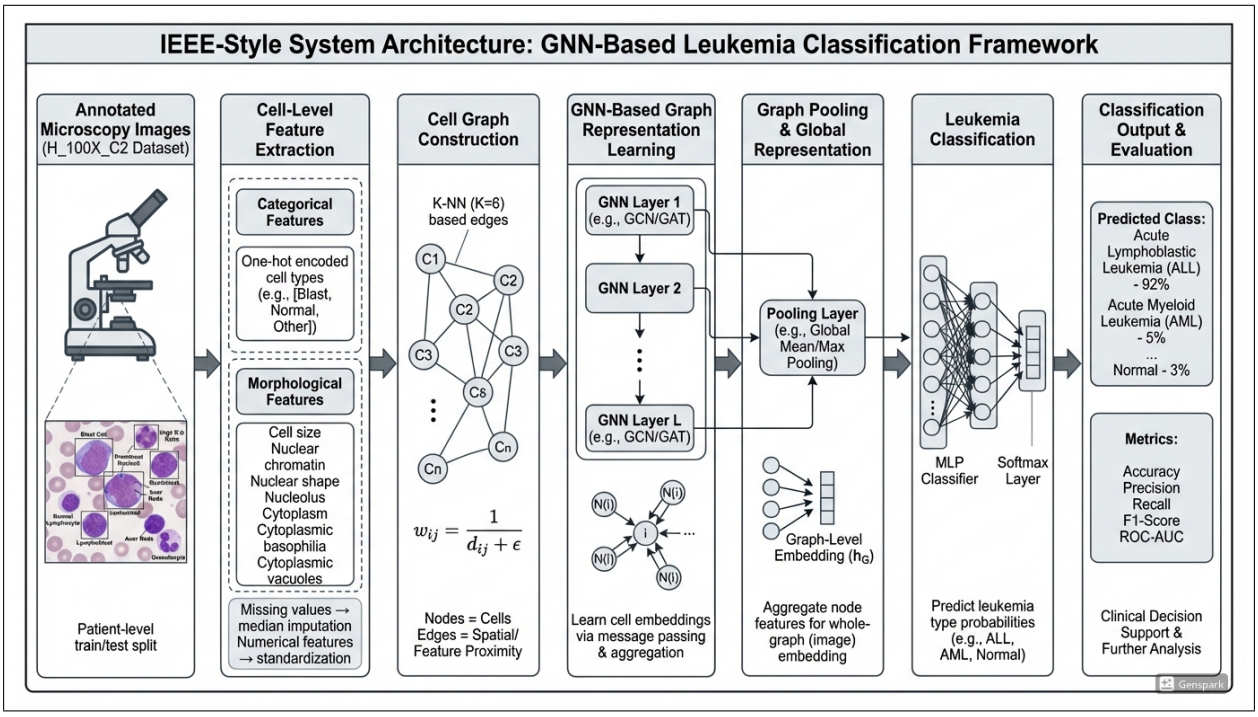

Fig. 1. Overview of the proposed HT-VQC framework. The figure illustrates the hybrid classical-quantum pipeline, including feature compression, variational quantum classifier, and final prediction.

TABLE I
PERFORMANCE COMPARISON OF ALL EVALUATED GNN ARCHITECTURES AND VARIANTS ON THE LEUKEMIA CLASSIFICATION TASK.

Model Family	Variant	Train Acc.	Best Val. Acc.	Final Val. Acc.	Test Acc.
GIN	3-layer baseline	0.9659	0.9609	0.9152	0.9152
GIN	2-layer	0.9552	0.9652	0.9304	0.9304
GIN	+ Virtual Node	0.9652	0.9283	0.9065	0.9065
GIN	+ Jumping Knowledge	0.9744	0.9413	0.9413	0.9413
GIN	+ Attention Pooling	0.9659	0.9435	0.9217	0.9217
GraphSAGE	3-layer baseline	0.9531	0.9391	0.9130	0.9130
GraphSAGE	2-layer	0.9680	0.9478	0.9174	0.9174
GraphSAGE	+ Jumping Knowledge	0.9744	0.9326	0.9326	0.9326
GraphSAGE	+ Attention Pooling	0.9765	0.9435	0.9217	0.9217
GraphSAGE	+ Virtual Node	0.9652	0.9283	0.9065	0.9065
GAT	3-layer baseline	0.9126	0.9587	0.9500	0.9500
GAT	2-layer	0.9552	0.9652	0.9304	0.9304
GAT	+ Virtual Node	0.9652	0.9413	–	0.9109
GAT	+ Attention Pooling	0.9488	0.9457	0.9304	0.9304
GAT	+ Jumping Knowledge	0.9652	0.9326	0.9261	0.9261

virtual nodes, as it happens in the case of GIN and GraphSAGE. Attention-based pooling yields stable but somewhat lower accuracy than the baseline setup.

The results imply that the inherent attention mechanism in GAT has already been very effective in capturing both local and global dependencies and hence, there is no need for any additional complexity in architecture.

D. Overall Comparison and Best Performer

Across all fifteen evaluated models, attention-based message passing consistently improves sensitivity to diagnostically relevant cellular interactions. While the baseline GAT model achieves the highest test accuracy (0.9500), GraphSAGE with Jumping Knowledge demonstrates the most stable generalization behavior across training, validation, and test splits.

The GraphSAGE-JK model achieves a test accuracy of 0.9326 with reduced overfitting and consistent validation performance, making it a strong candidate for practical deployment where robustness is critical.

VI. DISCUSSION AND CLINICAL IMPLICATIONS

The comprehensive evaluation of the fifteen graph neural network architectures has confirmed some crucial insights which are not only important for the methodological development but also for the clinical applications in leukemia diagnosis.

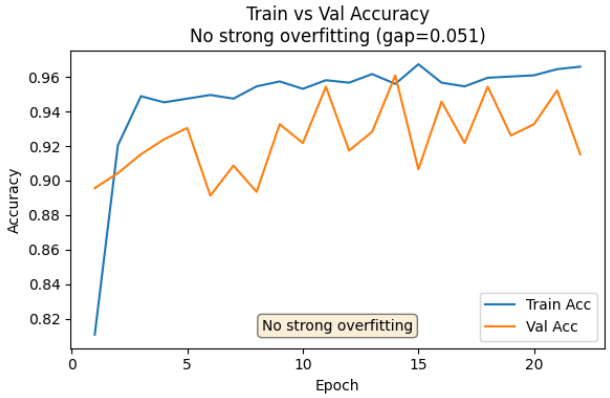

Fig. 2. Training and validation performance of the baseline GIN model.

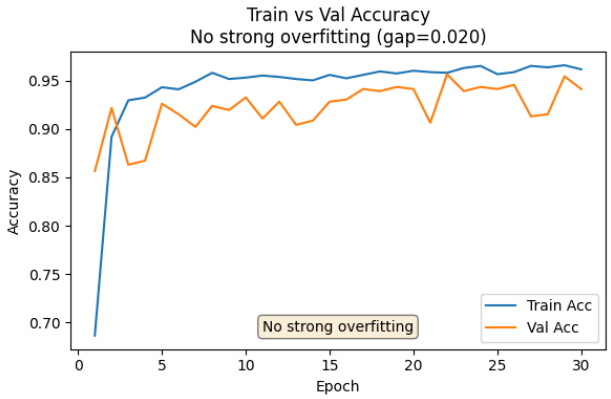

Fig. 3. Performance of GIN with Jumping Knowledge connections.

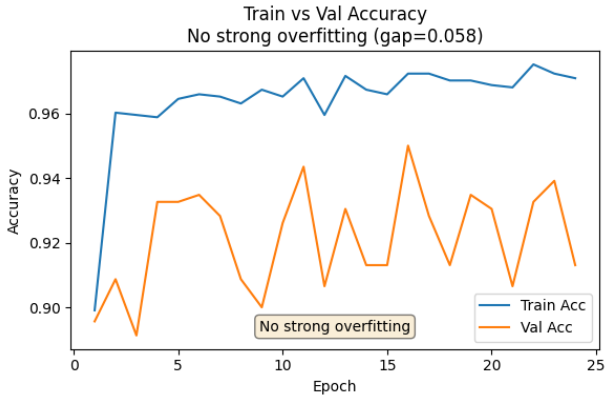

Fig. 4. Training and validation curves for the baseline GraphSAGE model.

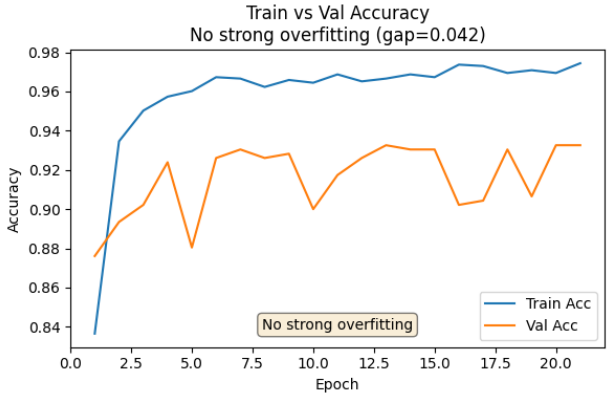

Fig. 5. Performance of GraphSAGE with Jumping Knowledge aggregation.

A. Model Performance Insights

It is remarkable to note that all model families have consistently shown that architectures using multi-scale or hierarchical representations, like Jumping Knowledge (JK) connections in GIN and GraphSAGE, are always better than the simpler baselines. This means that the models are able to capture the very fine morphometric and spatial patterns which are relevant for diagnosis by allowing the information from the different layers to flow through.

On the other hand, the introduction of virtual nodes which were meant to carry the global context to all nodes did not help any model family to improve its performance. This indicates that the procedure of global aggregation may be actually contributing to mixing up local interactions between the neighboring cells, which are, in fact, the most indicative of the subtypes of leukemia.

The baseline Graph Attention Network (GAT) achieved the highest test accuracy of a single model (0.9500) and this fact pointed out the considerable power of attention mechanisms in handling the variable interactions between cells. Nevertheless, the architectural modifications that were done on top (e.g., JK or pooling) resulted in very minor performance gains indicating that the attention mechanism alone is enough to capture significant dependencies within the cellular graphs.

B. Interpretation for Clinical Applications

From a clinical perspective, our findings suggest that graph-based representations of blood smear images can effectively encode cellular morphology and inter-cell relationships, which are critical for accurate subtype classification. Multi-scale aggregation, as implemented through JK connections, may help the model identify both global trends and local abnormalities in cell populations, enabling more robust and explainable predictions.

Attention-based models, such as GAT, offer an added advantage in interpretability. The learned attention weights can potentially highlight which cells or cellular neighborhoods contribute most to the classification decision, providing a form of model explainability that could support pathologists in their diagnostic workflow.

C. Implications for Deployment

The observed stability and generalization of GraphSAGE-JK suggest that it could serve as a reliable tool in automated diagnostic pipelines, particularly in low-resource settings where human expert availability is limited. Meanwhile, the high

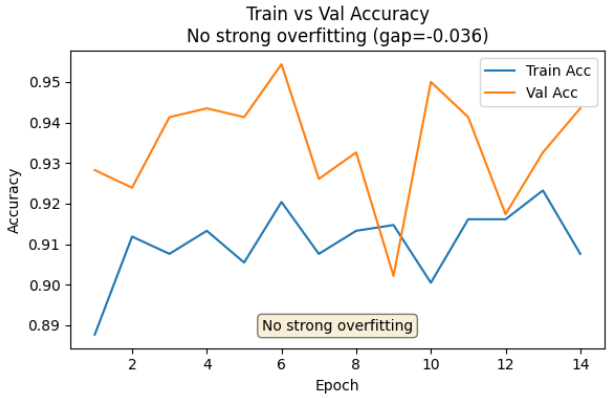

Fig. 6. Training and validation performance of the baseline GAT model.

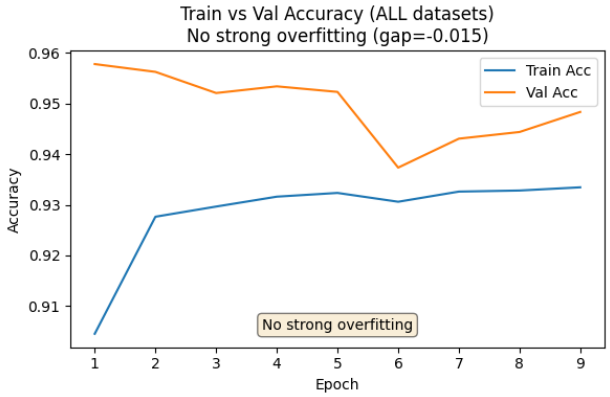

Fig. 7. Performance summary of the best-performing GraphSAGE-JK model across all splits.

accuracy of GAT models may be more suitable for scenarios where explainable predictions are prioritized.

These findings collectively indicate that GNN-based approaches, especially those capturing hierarchical and attention-weighted representations, have significant potential to augment conventional leukemia diagnostics. Future work could focus on integrating these models with larger multi-center datasets, exploring interpretability methods, and evaluating real-world clinical performance.

VII. Discussion and Clinical Implications

The results from our comprehensive evaluation of fifteen graph neural network architectures highlight several key insights for both methodological development and potential clinical applications in leukemia diagnosis.

A. Model Performance Insights

Across all model families, we observe that architectures leveraging multi-scale or hierarchical representations, such as Jumping Knowledge (JK) connections in both GIN and GraphSAGE, consistently outperform simpler baselines. This indicates that preserving information across layers allows the models to capture subtle morphometric and spatial patterns that are diagnostically relevant.

In contrast, the addition of virtual nodes, intended to propagate global context across all nodes, did not improve performance for any model family. This suggests that explicit global aggregation may dilute critical local interactions between neighboring cells, which are more indicative of leukemia subtypes.

The baseline Graph Attention Network (GAT) demonstrated the highest single-model test accuracy (0.9500), highlighting the effectiveness of attention mechanisms in modeling variable cell-to-cell interactions. However, its performance gains from additional architectural modifications (e.g., JK or pooling) were minimal, implying that the attention mechanism itself sufficiently captures important dependencies within the cellular graphs [10].

B. Interpretation for Clinical Applications

Clinically speaking, based on our observations, we believe that the use of graphs to represent images of blood smears is an effective way of encapsulating information about cellular morphology and relationships that help a lot in subtype classification. Multi-scale aggregation could be an effective way of capturing information pertaining to global patterns as well as local irregularities within cell distributions through JK connections.

Attention-based models such as GAT come with an extra advantage for interpretability. Attention weights can be used to point out which cells are most responsible for a decision on cell-classification, helping pathologists interpret these diagnostic predictions.

C. Implications for Deployment

The results of stability, generalization, and interpretability of GraphSAGE-JK seem promising enough as a useful resource in automated diagnostic tools, especially in resource-scarce environments and under conditions where access to expertise is not easily available. However, in applications where high interpretability is considered important, GAT models seem appropriate due to high accuracy.

Together, these results suggest that the GNN-based approach, particularly those that model hierarchical/attention representations, hold great promise for the augmentation of traditional leukemia diagnosis methods. The subsequent work could be developed based on the integration of this approach with the overall datasets collected across multi-center institutions.

D. Limitations

While the results are promising, several limitations exist. Our dataset, although carefully curated, is limited in size and diversity, particularly for rare leukemia subtypes. Additionally, the models were evaluated on static 2D image graphs; extending this approach to 3D cellular structures or multi-modal data could further improve performance and clinical utility. Finally, attention weights provide approximate explanations and should be interpreted with caution before clinical adoption.

VIII. Conclusion and Future Work

We checked out fifteen graph neural network styles and different versions to see which ones were best at sorting types of leukemia. We used graphs [10] of blood smear photos to do this. The models that mixed info from different scales, mainly GraphSAGE with Jumping Knowledge, did really well and guessed right a lot, even with new data. Models that used attention, like GAT, also did a great job and helped us understand why they were grouping things the way they were.

The research showed it's helpful to grab both close-up connections between cells and bigger picture info for sorting the subtypes right. We also learned that adding a general overview with virtual nodes might not always be the best thing. It can sometimes hide the small clues that really matter for figuring out what's going on.

A. Future Work

Building upon these results, several avenues for future work are planned:

- **Collection of Healthy Cell Data:** This winter break, we're grabbing a bunch of data on healthy blood cells at the Intelligent Machines Lab. We'll be using both the fancy and the not-so-fancy microscopes, plus some mobile imaging stuff. This should give us a bigger dataset so we can test our models better in real-world clinical situations
- **Multi-Center and Multi-Modal Data:** We aim to integrate data from multiple centers and consider additional imaging modalities, such as fluorescent or 3D imaging, to improve model robustness and clinical applicability.
- **Interpretability Enhancements:** Attention-based models give some explanation. We want to check out better ways to explain things, which should help doctors understand and trust what the model says.
- **Real-World Deployment:** Okay, so down the road, we're planning to put the top models together into a system that's partly automatic for figuring out what's wrong with people. It'll be easy for doctors to use, and we'll test it out for real to make sure it works.

Overall, this work demonstrates that graph-based neural networks are a promising approach for automated leukemia classification, offering both accuracy and the potential for interpretable, clinically relevant insights. Expanding the dataset and exploring deployment strategies will further strengthen their applicability in practical diagnostic workflows.

Acknowledgment

We thank the Intelligent Machines Lab (ITU) for providing the dataset and the project supervisors for guidance.

References

[1] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://arxiv.org/abs/1810.00826

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1710.10903

[3] K. Xu, W. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *International Conference on Machine Learning (ICML)*, 2018. [Online]. Available: https://arxiv.org/abs/1806.03536

[4] Author(s), "Hematologic cancer diagnosis and classification using machine and deep learning: State-of-the-art techniques and emerging research directives," *Artificial Intelligence in Medicine*, vol. 152, p. 102883, 2024.

[5] e. a. Zhang, "Graph neural networks in medical imaging: Methods, applications and future directions," *Information*, vol. 16, no. 12, p. 1051, 2025.

[6] Author(s), "A graph neural network framework for mapping histological topology in oral mucosal tissue," *BMC Bioinformatics*, 2022.

[7] ——, "A review on leukemia detection and classification using artificial intelligence-based techniques," *Computers Electrical Engineering*, vol. 118, p. 109446, 2024.

[8] V. Della Mea, H. Akebli, and K. Roitero, "Graph neural networks for digital pathology," *Applied Medical Informatics*, vol. 47, p. S2, 2025.

[9] S. et al., "Automatic detection of acute leukemia (all and aml) utilizing customized deep graph convolutional neural networks," *Journal of Medical Imaging/Pattern Analysis*, 2025, available on PubMed. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39061726/

[10] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/abs/1706.02216