LeukemiaCellNet: A Hybrid CNN-Transformer Architecture for Accurate Classification of Leukemia Blood Cells

by

Nazmul Hasan
20301250
Syeda Rubaiya Nahar
20241008
Rakin Mostafa
20101084

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
January 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nazmul Hasan
20301250

Syeda Rubaiya Nahar
20241008

Rakin Mostafa
20101084

Approval

The thesis/project titled "LeukemiaCellNet: A Hybrid CNN-Transformer Architecture for Accurate Classification of Leukemia Blood Cells " submitted by

1. Nazmul Hasan (20301250)
2. Syeda Rubaiya Nahar (20241008)
3. Rakin Mostafa (20101084)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 25, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Jannatun Noor Mukta
Associate Professor,CSE
Director, BSDS
United International University

Co-Supervisor:
(Member)

Md. Tanzim Reza
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi
Associate Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Timely and precise leukemia diagnosis remains a challenge in healthcare. This study improves accuracy utilizing powerful AI technologies, with a focus on deep learning models. This study examines CNN architectures, including VGG., Xception, Inception, ResNet, and DenseNet, together with transformer-based models like ViT, DaViT and MaxViT. A hybrid model merging ViT and ResNet50 was proposed, utilizing CNNs for image feature extraction and transformer networks for feature representation.

Analysis using a medical-image dataset assessed the F1 score, recall, and precision of the model. Meanwhile, the model ResNet acquired an accuracy of 54.61% but the transformer models, MaxVit and Davit achieved 53%. However, ResNet50-ViT(Hybrid model) excelled all models, by attaining 92.87% accuracy with well-balanced recall and precision.

The findings illustrate the promise of fully automatic leukemia diagnosis as potentially disruptive, with the possibility of using hybrid AI models to address healthcare diagnostic problems in an accurate and timely manner. The combination of state-of-the-art convolutional and attention mechanisms provides adequate solutions to the complexities associated with leukemia, setting a new standard for AI in hematology and opening new opportunities for the development of new automated diagnostic systems.

Keywords: Leukemia classification, Hybrid CNN-Transformer, Deep learning, Medical image analysis, Automated diagnostic

Dedication

This thesis is dedicated to all the cancer patients whose sickness goes undetected due to a lack of proper detection tools. We hope our work, though small but a promising effort, can aid in making a difference with progressive early detections and thus contribute to saving lives.

Acknowledgement

We humbly acknowledge the Almighty Allah for guiding us and giving us patience, wisdom, and capability to complete our thesis with proper dedication and success. We extend our heartfelt thanks to our supervisor Dr. Jannatun Noor Mukta for being patient with us and allowing us to work under her guidance and, our co-supervisor Md. Tanzim Reza for providing us with his invaluable time to support us. We also owe deeply to everyone who helped us when we faced difficulties in our thesis. We hope we have done justice to everybody who provided their help and support and whose contribution led to the success of completing our thesis journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Acute lymphoblastic leukemia (ALL) is a blood cancer that is common among children though it can also arise in adults [1]. This type of cancer occurs as a result of the excessive proliferation of lymphoblasts, which are immature white cells, in the bone marrow leading to disturbance in the processes of blood cell generation and functioning. ALL is aggressive and in case no treatment is given in time, it can be fatal; hence early intervention is needed. Cytological diagnosis usually includes visual examination of cytological smears for the presence of B-lymphoblast cells which are malignant blasts and this procedure is sometimes difficult for pathologists. Often the microscopic differences between the malignant B-lymphoblasts and normal B-lymphoid precursors are non-prominent and this may lead to misinterpretation which in turn may cause delays in cancer recognition [32]. Because of the established association of timely and correct diagnosis with treatment success, there is an increasing demand for the use of more advanced diagnostic aids that would enable pathologists to obtain timely and accurate diagnoses.

1.2 Background Information

Recent changes in the field of medical imaging include developments that artificial intelligence (AI) and deep learning promise, in this case, automated imaging [8].

1.2.1 CNN in Medicine

Convolutional neural networks (CNNs) as deep learning models, have brought innovation in image classification tasks because machines can learn and extract features of the image without any assistance or guidance. In the field of medicine, this has been very beneficial. Today, CNNs were shown to be applicable in predicting pneumonia from radiographs, breast cancer using ultrasound images, and skin lesions from histopathological samples. For example, Nahid and his colleagues used CNNs to detect pneumonia from chest radiographs[21]. Similarly, Daoud and his team worked with pathologists and applied different approaches using CNN-based ultrasound imaging for breast images, achieving much better diagnostic accuracy than previously possible[13]. Such applications exemplify the capability of deep learning to be employed in fast-tracking and improving the precision of diagnostics.

1.2.2 ViT and CNN comparison

Vision Transformers(ViTs) divide images into patches and process them, tending each patch as a token. Patches are not kernels, and therefore, conventional Convolutional Neural Networks (CNNs) will not be able to process images and patches similarly. ViTs however, look at images in patches and focus on treating each patch as a 'token-cardboard.' These tokens then go through several layers of self-attention that relate different aspects of the same image, regardless of how far apart those aspects are. This makes ViTs particularly powerful in capturing global context and long-range dependencies, which are sometimes challenging for CNNs due to their inherently localized processing nature.

1.2.3 Hybrid models of ViT-CNN

The combination of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) into an ensemble model presents a novel strategy for medical image interpretation problems like Acute Lymphoblastic Leukemia (ALL) detection [40]. This composite model is based on the fact that both ViTs and CNNs come with their different strengths. What has been associated with CNNs' capabilities for a long is the propensity to extract local features in images due to the spatial internal hierarchy and local patterns. This identification of the detailed structures and features at multiple spatial scales is particularly useful for medical images where the smallest difference is already an important diagnostic sign. A hybrid model that integrates ViTs and CNNs takes advantage of their particular attributes and offers a more complete approach to ALL detection. CNN layers allow local feature maps to be extracted, concentrating on certain areas and gathering key details that are important for the recognition of the morphological features of leukemia cells. Such local features once extracted are fed into the ViT layers which add global context by recognizing dependencies and spatial relationships across large regions of the image. This two-layer structure also enables the model to optimally utilize both local and global information at the same time to improve decision-making in every aspect. This hybrid architecture sets the stage for a formidable challenge where ALL detection is concerned. Leukemia cells show quite several morphological changes that can sometimes be very small and hard to appreciate. Slight variations in geometry, volumetry, or the texture of the cells may reflect pathogenic changes which can only be appreciated by a model that views the cells at both the micro and macro levels. The diagnostic capabilities of the ensemble model are enhanced because the CNNs are used for the identification of specific internal structures such as the cell granularity, shape, or boundaries, and the ViTs are used to evaluate the distribution and relationships of these internal structures concerning the image as a whole. In the present research, the authors propose a new hybrid model called ViT-CNN which is an amalgamation of CNN and Vision Transformer architectures. The Integrated Model can use CNNs' spatial resolution and Vision Transformers' global resolution at once. This method is useful in the case of detecting leukemia as it uses complex features of the tumor cells including morphometry and the relative position of the features. The blended approach attempts to address the drawbacks of the existing image classification techniques, leading to more effective and correct discrimination, with the prospect of producing more dependable ALL diagnostic results. Furthermore, it is well known that many ensemble models exhibit much better

resilience to noise, as well as a capacity to generalize across diverse data domains [2]. Medical images are not free from artifacts, staining heterogeneity, or structural incoherence which pose challenges to conventional models. The hybrid ViT-CNN model should more effectively deal with such changes since the model structure of the ViT includes a self-attention mechanism that focuses on the important context and ignores irrelevant information, while every important detail is captured by the CNN.

1.3 Challenges

One of the major issues in medical imaging is focused on diseases such as ALL where one comes across the need to handle imbalanced datasets and more importantly, even for a single disease, this is a common problem. For instance, medical datasets are known to possess images containing an unbalanced number of other classes, too many images for healthy cells when juxtaposed against ones with malignant cells. Surge in imbalance may create overstated reliance upon the majority class, resulting in poor identification of and classification of instances belonging to the minority classes. There is a need for novel data enhancement methods that add more of the minority class without disrupting the nature of the data. This study also offers a ViT-CNN ensemble model along with a new technique of data enhancement called Difference Enhancement-Random Sampling (DERS). In medical image datasets, a reenactment of class strategies where images comprising one class outnumber more than a few other classes as often holds back the efficiency of deep learning models. The DERS methodology is focused on this problem by extending the majority classes so that a balanced dataset is used for the model training. With DERS, the model is expected to use multi-scale features across all classes to reduce bias and improve the overall performance of generalization. This method is especially helpful in the case of medical images because even though data with many images per category can be extremely useful for a good model but is not easy to attain.

1.3.1 Different Enhancement - Random Sampling (DERS)

The Difference Enhancement-Random Sampling (DERS) technique is an augmentation strategy specially developed with class imbalance challenges in mind in the context of medical data. DERS acts under two assumptions: focusing on difference enhancement and controlled random sampling. All these mechanisms enhance the ability of the model to understand useful patterns from the minority classes and make more balanced predictions.

Difference Enhancement

The "Difference Enhancement" facet of DERS aims at improving the depictions and contrast of discernible features in the images of minority features. This involves utilizing several image processing techniques that enhance key morphological features, often constituted by edges, textures, and colors that are less seen in images of minority classes. For instance, in leukemia cell images, certain features such as nuclear contour or cytoplasmic granules contribute to the identification of healthy cells and malignant cells. By improving these features, DERS helps the model identify and

learn from the differences existing in the minority class images in an easy way which helps the model to differentiate the minority class images when classifying it from the majority class.

Random Sampling

The second component, referred to as "Random Sampling" suggests a solution to the imbalance by resampling images from the minority class with greater variability. DERS does not adopt simple oversampling which may cause overfitting when the model is shown the same images, rather Controlled randomness in DERS is used to slightly modify the augmented samples to create more variety. This may involve random rotations, shifts, or zooms, which in turn introduces a new image set for the model to train on which in the end increases the perception of a larger dataset. Random sampling method in this case also contributes to an improved model by making it less reliant on few samples and more generalized.

1.3.2 DERS Role in ViT-CNN Hybrid Model

The combined effects of the Difference Enhancement and the Random Sampling components of the DERS make it a convenient approach to deal with the problem of class imbalance. DERS meets this objective by heating crucial factors and showing different queer of classes images so that the model has a more uniform and representative training set. Classification not only enhances the accuracy of the classifier but also weakens the paradigm shift toward the main class and allows more precise and reliable forecasts for all classes. The role of DERS in the sphere of our hybrid ViT-CNN model deserves attention as well. As it is known, ViTs and CNN have different approaches in information processing and there is a reason DERS comes as a natural fit for the dual structure of the model since both minority class images' local feature and global dependencies are augmented and well-maintained during the training process. When the minority samples are being augmented and a local feature is being concentrated on by the CNN layers, more details can be captured by the layers of the network. In contrast, the ViT layers enable comprehension of the spatial arrangements and interrelations between these features by utilizing a wider context defined by random images of aligned features. To cut a long story short, the DERA method is a step forward in solving the problem of training a model with an imbalanced medical dataset. Thanks to DERS combining Difference Enhancement to highlight the minority class and Random Sampling to supply the diversity of training data, the model achieves balanced learning, and the overall strength and robustness of the ViT-CNN combination model is improved. It guarantees that the model does not only achieve an accurate detection of ALL cells but does this across different class distributions, which makes it a powerful device for clinical applications.

1.4 Research Objective

In Chapter 2, we will learn about prominent works in the field of medical technology, their working methods, and limitations. Especially in the field of blood cell-related works, most of them are limited at some point. Our research aims to overcome those

limitations. If we restate and make an overview of the most significant limitations, they go like this:

- Many studies focused on classifying leukemia cells but did not include methods to segment or localize specific cell types.
- Traditional models such as SVM, Decision Trees, or even standalone CNNs struggled to capture complex morphological and contextual features of leukemia cells, particularly for challenging datasets with noise or heterogeneity.
- Many studies were limited to binary classification (leukemic vs. healthy cells) without addressing subtype classification or segmentation of complex cell structures.
- Insufficient Use of Attention Mechanisms. Many CNN-based approaches could not model global dependencies effectively.
- Several studies relied on small, imbalanced, or unimodal datasets, which restricted the generalizability and scalability of their models.
- Basic augmentation techniques, such as resizing and flipping, were often used leading to suboptimal performance when dealing with class imbalance.
- Many approaches lacked explainability, making them less reliable for clinical practitioners.

Our study fulfills the absence of such important aspects by focusing on the following research objectives.

- We will integrate local and global feature extraction.
- Unlike studies limited to binary classification, this paper addresses multi-class classification, distinguishing between various leukemia subtypes.
- Improved Handling of Class Imbalance.
- Will provide specific WBC identification, localization, and segmentation.
- The hybrid model is designed to be scalable for larger, diverse datasets, ensuring its applicability beyond the scope of small, curated datasets.
- Will employs self-attention mechanisms to capture long-range dependencies and relationships in the images.
- The combination of CNN and Transformer will improve model interpretability, making it more reliable and understandable for clinical practitioners.
- Will provide more accuracy in multi-class classification.

1.5 Research Contributions

The contributions of this research can be summarized as follows:

1. **Introduction of a ViT-CNN ensemble model:** The integration of Vision Transformers and Convolutional Neural Networks gives rise to a new hybrid model. CNNs are great for local features whereas ViT models the global aspects, so ALL detection can be implemented with this comprehensive model architecture.

2. **Development of the Difference Enhancement-Random Sampling (DERS) method:** This data augmentation technique works on the hurdles caused by the decision bias of having an imbalanced medical dataset which would otherwise affect the performance of the model while classifying images belonging to the majority or the minority classes.

3. **Detailed model evaluation:** The primary purpose of the study conducted here is to present a comprehensive evaluation and analysis of the aforementioned model along with context to how it synthesizes with other models that are built around the traditional CNN or even other ensemble models. This comparison is crucial to corroborate the efficacy of the model and also to prove its feasibility in actual clinical use.

4. **Comprehensive analysis of methodology:** The study features a detailed assessment of the available dataset, pre-processing techniques, architecture of the model, and approaches to conformation for the setup used for the study. This kind of clarity in methodology is going to benefit future researchers and practitioners who wish to test the same or a similar methodology in the application of other medical images.

There are mentioned benefits, however, limitations are also present, including the strength of the model training within the hybrid architecture. Increasing the number of tasks assigned to each, the model has a better chance to optimize the performance of each one separately thereby enhancing convergence speed and possibly decreasing the quantity of annotated data required to train the model efficiently. This offers an advantage, especially in medical imaging where the volume of annotated data is always in short supply, hence time-consuming to obtain.

To sum up, for the above models, the interpretability nonetheless is greatly improved because the physicians can see how local features could influence the outcome prediction with other contextual features in situ.

1.6 Thesis Structure

This section gives a fast overview of what each chapter expounds upon in the thesis:

- **Chapter 1: Introduction**
  - Contains the problem statement and research objectives.
  - Contains starters on the topics surrounding CNNs, Vision Transformers (ViTs), and hybrids in medical imaging.

  - Challenges faced in leukemia classification are highlighted.

- **Chapter 2: Literature Review**
  - A review of any work that is available in existence regarding deep learning applied in medical image analysis.
  - This works to compare different methods, including CNN-based and transformer-based models, with their viability.
  - This section identifies possible gaps left by previous studies that this research seeks to cover.

- **Chapter 3: Methodology**
  - Describes the dataset, preprocessing, and augmentation techniques.
  - Describes the models implemented in this work: individual CNNs, transformers, and the proposed hybrid model.
  - Explains the evaluation metrics used to measure performance.

- **Chapter 4: Results and Analysis**
  - Summarizes the performance metrics of the various models in terms of accuracy, F1 score, precision, and recall.
  - A comparison of the models was made between the CNNs and transformers and hybrid ResNet50-ViT.
  - Discussions of important findings, limitations, and implications are made.

- **Chapter 5: Conclusion and Future Work**
  - Provides an overview of the contributions and findings from the research.
  - Built into our suggestions for improvements based on larger datasets, further augmentation, and real-world clinical applications.
  - Future perspectives of the AI-based leukemia diagnosis will also be discussed.

Chapter 2

Literature Review

In this section, we review and analyze several foundational studies related to automated leukemia diagnosis, focusing on the approaches, data sets, methodologies, and limitations of previous work. The majority of researchers in this area focus on classification techniques to improve the accuracy of leukemia diagnosis. Key factors impacting accuracy include the quality of data (particularly noise levels), the features extracted, and the methods used for classification.

A prominent study by Hsieh et al. [4] utilized a Support Vector Machine (SVM) approach, enhanced with information gain (IG-SVM), to classify leukemia cells. This technique achieved a notable 98.10% accuracy in distinguishing leukemic cells, demonstrating the potential of SVM in hematological classification tasks. However, the study did not incorporate an object detection model, thus limiting its utility in applications requiring the segmentation or localization of specific cell types.

Similarly, the research conducted by Hariprasath et al. [14] explored the detection of Acute Lymphocytic Leukemia (ALL) using morphological features extracted from blood cell images in the ALL-IDB dataset. Their methodology used SVM classifiers for noisy data, achieving an accuracy of 90.33%, and KNN for noise-free data, resulting in an accuracy of 91.5%. Although their work classified malignant cells effectively, it did not advance the analysis to detect or segment leukemia cells within a larger object detection framework.

In another study, Khaled A. S. Abu Daqqa et al. [7] focused on prediction and diagnosis using multiple classification algorithms, including Decision Tree, SVM, and KNN. The Decision Tree algorithm demonstrated the highest accuracy at 77.30%, while SVM and KNN achieved 76.82% and 70.15%, respectively. Although the study presented comparative insights into algorithmic performance, it primarily emphasized prediction accuracy and did not explore image processing techniques or object detection methods.

Subhash Rajpurohit et al. [10] developed a dual-module system combining image processing with machine learning classifiers, including SVM, FNN, and KNN, to identify acute lymphoblastic leukemia in microscopic images. Their model, which used the ALL-IDB dataset, achieved a high classification accuracy of 98.33% with a Convolutional Neural Network (CNN). While the authors successfully classified

each image with notable accuracy, the study did not delve into the segmentation or identification of specific white blood cell (WBC) components, as seen in other studies.

Further, Shitong and Min [3] proposed an automated method using Fuzzy Cellular Neural Networks (FCNN) for blood cell recognition in microscopic images, achieving a 99% accuracy rate. Their work showcased the potential of fuzzy algorithms in handling image noise and complexity; however, like other studies, it lacked an object detection approach for WBC identification, focusing instead on cell classification.

The study by Esti Suryani et al. [6] utilized a fuzzy rule-based system to recognize types of acute leukemia, including ALL and AML M3, by analyzing the morphology of WBCs, achieving an accuracy of 83.65%. Their system generated rules for leukemia-type recognition but did not incorporate object detection to identify individual cell components within an image.

Roy et al. [5] took a novel approach by capturing shadow images of blood samples, transmitting them to a smartphone, and processing the data with a customized algorithm, achieving over 95% accuracy. While the study contributed a unique approach to mobile-based processing, it did not include object detection and thus lacked precise WBC component identification.

Research by S. Alagu, A. Priyanka N., Kavitha G., and Bhoopathy Bagan K. [25] produced a strong system for recognizing Acute Lymphoblastic Leukemia (ALL). The methodology first used U-Net to separate white blood cell nuclei while subsequently employing deep features from AlexNet together with GoogleNet and SqueezeNet. Feature extraction techniques were used to produce fused features while Mutual Information, Minimum Recursive Maximal Relevance (MRMR), and Recursive Feature Elimination (RFE) methods implemented feature selection procedures. Feature fusion methods produced data with better diagnostic efficiency that reached a 98.2% accuracy level when tested on the ALL-IDB2 database. ANOVA statistical testing ensured an evaluation of selected feature discriminatory power in the research analysis. The conducted research brought major progressive results but faced important restrictive elements. The research used a limited dataset consisting of 198 images thereby restricting the approach's ability to scale beyond its current capacity. The approach requires unimodal input images only and lacks attention circuits which reduces its capacity to perceive intricate data relationships.

A study led by M.T. Haque Khan et al.[37] developed an automated Acute Lymphoblastic Leukemia subtype detector for minutesular blood smear images by implementing four deep neural network models including MobileNetV2, ResNet50, ConvNet, and VGG19. 3,256 peripheral blood smear images formed the basis for this analysis which separated samples between benign and malignant cells while subclassifying malignancies into Early Pre-B to Pro-B acute lymphoblastic leukemia subtypes. MobileNetV2 demonstrated the best testing accuracy rate of 97.42% while surpassing ResNet50 with 85.26% accuracy and ConvNet at 91.28%. Although the research demonstrated significant advancements in classification performance, limitations remain. The dataset size, while sufficient for initial experiments, lacks the

diversity and scalability required for robust real-world applications. Additionally, the use of standard preprocessing techniques, such as resizing and augmentation, may fail to capture intricate morphological variations present in real-world datasets.

In their study Mallick et al., [38] employed a five-layer DNN to classify leukemia from gene expression microarray samples leading to a testing accuracy of 98.21% [33]. The study used high-dimensional data efficiently yet its limited 72-sample dataset created challenges for scalability. The clinical distinction between ALL and AML testing in the research narrowed its practical application boundaries.

The work of Tran et al. performed a White Blood Cell image evaluation using the InceptionV3 Model on the C-NMC database to achieve a best-in-class performance level of 99.6% [39]. The model displayed outstanding performance in TWO category detection but failed to meet standard metrics for evaluating unbalanced datasets which included both F1 scores and recall. The model's use of one-dimensional images and absence of transformer-based contextual analysis revealed weaknesses since it failed to handle multiple blood cancer subtypes or various medical data types.

Building on the work of Riaz Ullah Khan et al., [42], researchers applied RCNN for RBC detection reaching a 91.21% success rate. The research showed success in noise reduction as well as cell segmentation but this progress was limited by its exclusive focus on one cell type with only 12,500 images from which augmented samples were derived leading to restricted diagnostic scalability.

In addition, research by Tran et al., [12] introduced LeukemiaNet, a custom CNN that achieved 97.2% accuracy on an augmented dataset of 1,676 images. While effective for classifying ALL, AML, and normal cells, the study's reliance on basic augmentation techniques and small datasets hindered its scalability and robustness.

The study by Revanda et al., [36] used enhanced Mask R-CNN for instance segmentation, achieving 83.72% accuracy on a dataset of 301 images. Although the study excelled in segmenting individual lymphoblasts, it struggled with subtype sensitivity (e.g., L2: 54.55%) and limited dataset size. Manual annotation further limited scalability.

Ahmed S. Negm et al.,[9] did research and achieved an impressive accuracy of 99.517% using a combination of K-means clustering for segmentation, feature extraction techniques, and classifiers such as neural networks and decision trees. Two datasets containing 757 samples served as input to the system for blast, myelocyte, and segmented cell detection. Previous research showed that the developed algorithm reached high levels of sensitivity and specificity although its success depended heavily on both preprocessing and segmentation components for extracting features correctly. The technological approach with K-means clustering in this study shows limitations when dealing with complex cellular visual patterns and overlapping cell categories. This algorithm faces limitations in effectively modeling both spatial and global contextual features because it lacks an integrated CNN-Transformer architecture.

The study by Ahmad Badruzzaman and Aniati Murni Arymurthy,[41] evaluated ResNet, ResNeXt, and EfficientNetV2 architectures for the detection of blast cells in microscopic images, achieving a maximum accuracy of 79.45% for ResNet 18 in blast cell classification. The study highlighted the efficiency of residual learning in ResNet and ResNeXt for faster convergence but also noted that these models were prone to overfitting compared to the EfficientNetV2 architecture. However, the dataset limitations, such as class imbalance and a lack of extensive augmentation, constrained the models' ability to generalize across all leukocyte classes. The study also focused solely on image-based modalities without exploring hybrid or multimodal frameworks.

The paper by S. Rezayi et al. [30] investigated the use of ResNet-50, VGG-16, and a custom convolutional neural network (CNN) for binary classification of leukemic versus healthy cells. The models achieved validation accuracies of 81.63%, 84.62%, and 82.10%, respectively, using a dataset of 12,528 microscopic images. The study also employed six machine learning classifiers, with the highest accuracy of 81.72% obtained using Random Forest. The custom CNN was noted for its simpler architecture and reduced computational overhead compared to pre-trained models, making it suitable for resource-constrained environments. Despite its contributions, the study is limited by its exclusive focus on binary classification and reliance on single-modality data (microscopic images), which restrict its applicability to more complex diagnostic scenarios. The dataset used was also imbalanced, which, despite weighting techniques, may affect model performance in real-world applications.

Despite the progress made, a gap remains in leukemia detection using object detection techniques. Existing studies predominantly rely on classification methods, limiting their applicability in scenarios where detailed cellular analysis is required. In our work, we address this gap by implementing Faster-RCNN, an advanced object detection model that enables the identification and localization of specific blood components by matching them with trained data. This approach provides a more granular understanding of the blood sample, allowing for efficient and accurate detection of leukemic cells.

Deep learning's transformative impact has permeated various sectors, notably in medical diagnostics, security systems, and object detection, where it has demonstrated impressive capabilities in interpreting complex data and achieving high accuracy levels. For instance, significant research has focused on applications in brain tumor detection [31], intrusion detection, and the detection of multiple fused objects [28], [35], [34], [15]. An ensemble method combining NASNetLarge and VGG16, as proposed by Kasani et al. [19], was particularly notable for classifying cancerous and healthy cells, applying pixel normalization (0-1 range) to enhance model accuracy and utilizing data augmentation to counter dataset imbalance, leading to an accuracy of 96.58%.

Further exploration in blood cell analysis was conducted by Ullah et al. [24], who developed a state-of-the-art VGG16 model with Efficient Channel Attention (ECA) to identify healthy and blast cells from blood smears. This model, which incorporated semantic feature learning to highlight critical image regions, reached an accuracy

of 91%. Similarly, a highly specialized VGG16-based model combined with transfer learning techniques was applied by Genovese et al. [26] to detect Acute Lymphoblastic Leukemia (ALL), achieving robust results with microscopic blood images from the ALL-IDB dataset.

In cardiovascular diagnostics, a groundbreaking application of deep neural networks for coronary image analysis was introduced with the TreeVes-Net model [23]. Designed to interpret Fractional Flow Reserve (FFR) values from CT angiography images, TreeVes-Net combines coronary geometry with fluid dynamics, leveraging a recurrent neural network (RNN) to analyze extensive datasets, including 13,000 simulated and 180 clinical cases, achieving AUC scores of 0.92 and 0.93.

Progressive Sequential Causal GANs (PSGAN) presented by researchers [16] offered another innovative approach, specifically aimed at creating Late Gadolinium Enhancement (LGE)-equivalent images for diagnosing specific tissue areas. PSGAN demonstrated a progressive framework with sequential causal learning and unique loss functions for synthesis and segmentation, yielding a segmented accuracy of 97.17% and a correlation coefficient of 0.96 for scar detection.

In coronary imaging, the Progressive Medical Modality (PMD) [17] was developed to assist in Visual-Based Diagnostic Imaging (VBDI) across diverse intracoronary modalities, utilizing a structure-deformable neural network with a bidirectional pyramidal network to adapt to variations in vessel size. This advanced system expanded the learning dataset for intracoronary imaging, significantly enhancing vessel adaptability, as proven in extensive experiments.

A further contribution to blood cell classification was proposed by Ghaderzadeh et al. [33], who developed a majority voting ensemble of InceptionV3, ResNet-V2, Xception, and DenseNet121 models to classify normal and blast cells within the ISBI-2019 dataset. By applying data preprocessing and augmentation techniques, this ensemble model reached 98.5% accuracy. Additionally, the use of hybrid approaches in blood cell analysis was demonstrated by Sahlol et al. [22], combining VGGNet with the Salp Swarm Algorithm (SASSA) for feature selection and noise reduction. This hybrid model, coupled with a Support Vector Machine (SVM) classifier, achieved accuracies of 96.11% on the ALL-IDB2 dataset and 87.9% on ISBI-2019, showing the potential of SVM classifiers in distinguishing normal from abnormal cells.

To address the complexity of white blood cell (WBC) nucleus segmentation, recent research has favored U-Net and U-Net++ architectures, yielding superior results in distinguishing normal from blast cells. Data enhancement approaches, such as data augmentation and transfer learning with models like AlexNet [11], [20], have also been vital to addressing data limitations in the context of ALL detection, where enhanced datasets significantly boost model performance.

The work by Jing et al. [29] is also noteworthy, as they introduced the VIT-CNN ensemble, combining EfficientNetB0 and a vision transformer model for b-lymphoblast detection. With data sampling enhancements and normalization, their model at-

tained a 99.03% test accuracy. HistoCNN and HistoNet, introduced by Genovese et al. [27], provided CNN-based models (VGG-16, ResNet-18) employing k-means clustering, c-means, and advanced thresholding for WBC segmentation, performing robustly on the ALL-IDB dataset.

Lastly, ensemble methods continued to outperform individual models in recent studies, as evidenced by models such as ResNet101-9 [18] and weighted ensembles [32], achieving accuracies of 85.11% and 88.3% respectively. These results underscore the superiority of ensemble approaches for complex medical image classification tasks. Such diverse applications and technical advancements highlight the immense adaptability and potential of deep learning across a spectrum of diagnostic challenges.A literature comparison has been shown in Table 2.1

Table 2.1: Comparison Table with Related Literature

Study	**Methodology**	**Accuracy**	**Limitations**
IG-SVM for Leukemia Classification [4]	IG-SVM (SVM with Information Gain)	98.10%	No object detection; limited to classification
ALL Detection using Morphology [14]	SVM for noisy data, KNN for noise-free data	90.33% (SVM); 91.5% (KNN)	No object detection or segmentation
Multiple Classification Algorithms [7]	Decision Tree, SVM, and KNN	77.30% (DT); 76.82% (SVM); 70.15% (KNN)	No image processing or object detection
Dual-Module Image Processing [10]	Image processing + CNN	98.33%	Limited segmentation and specific WBC identification
Blood Cell Recognition using FCNN [3]	Fuzzy Cellular Neural Network (FCNN)	99%	Focused on binary classification, not WBC localization
Rule-based Leukemia Identification [6]	Fuzzy rule-based system	83.65%	No object detection for individual cell components
Shadow Imaging for Blood Cells [5]	Lens-free shadow imaging and mobile processing	¿95%	No object detection and precise WBC identification
VGG16 for Healthy vs. Blast Cells [24]	Attention-based CNN (Efficient Channel Attention)	91%	Limited to classification

Continued on next page

Continued from previous page

Study	Methodology	Accuracy	Limitations
VGG16 and Transfer Learning [26]	Adaptive unsharpening with VGG16	Not specified	No segmentation details
Blood Cell Classification Ensemble [33]	Ensemble: InceptionV3, ResNet-V2, Xception, DenseNet121	98.50%	No individual cell component detection
Hybrid Model with SVM [22]	VGGNet + Salp Swarm Algorithm (SVM)	96.11% (ALL-IDB2); 87.9% (ISBI-2019)	Limited segmentation and detection
U-Net for ALL Detection [25]	U-Net for WBC nuclei separation, AlexNet, GoogleNet, SqueezeNet	98.2%	Small dataset (198 images); no attention circuits
ALL Subtype Detection with DNNs [37]	MobileNetV2, ResNet50, ConvNet, VGG19	97.42% (MobileNetV2)	Lacks dataset diversity; preprocessing limitations
DNN for Leukemia Classification [38]	Five-layer DNN for microarray leukemia classification	98.21%	Small dataset (72 samples); limited scope (ALL/AML)
InceptionV3 for WBC Detection [39]	InceptionV3 for WBC detection using C-NMC dataset	99.6%	Weak for unbalanced data; lacks transformer-based analysis
RCNN for RBC Detection [42]	RCNN for RBC detection and segmentation	91.21%	Focus on RBC only; dataset limits scalability
LeukemiaNet for ALL and AML [12]	Custom CNN for ALL, AML, and normal cell classification	97.2%	Small dataset (1,676 images); lacks advanced augmentation
Mask R-CNN for Lymphoblasts [36]	Enhanced Mask R-CNN for lymphoblast segmentation	83.72%	Poor subtype sensitivity; small dataset (301 images)
K-means for Leukemia Classification [9]	K-means clustering + neural networks	99.517%	Overlapping cells; lacks CNN-Transformer integration
ResNet for Blast Cell Detection [41]	ResNet, ResNeXt, EfficientNetV2 for blast cell detection	79.45% (ResNet 18)	Dataset imbalance; overfitting; lacks hybrid approaches
CNN for Leukemia Classification [30]	Custom CNN + pretrained models (ResNet-50, VGG-16)	84.62% (VGG-16)	Binary classification only; dataset imbalance

Chapter 3

Methodology

This section explains in detail the data preprocessing, models used, and the evaluation criteria selected for the validity and clarity of this study. The overall workflow diagram of this study has been shown in Figure 3.1.

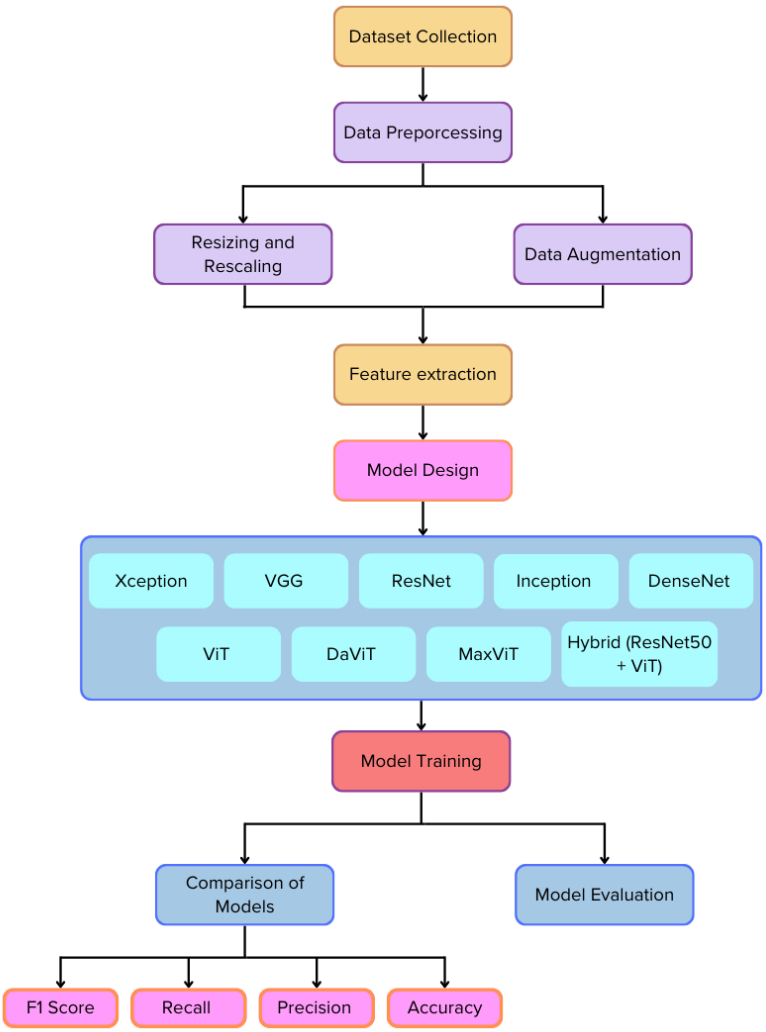

Figure 3.1: Workflow diagram of the proposed study

3.1 Dataset Description

The dataset applied in the study is known as the AML Cytomorphology dataset and was obtained from The Cancer Imaging Archive. It consists of more than 81 thousand images of blood cells. This dataset consists of numerous cases, depicting images of blood smears from confirmed cases of leukemia, and is intended specifically for cytomorphological investigations of leukemia. This dataset comprises four prevalent AML subtypes with defining genetic abnormalities and typical morphological features according to the WHO 2022 classification: (i) APL with PML::RARA fusion, (ii) AML with NPM1 mutation, (iii) AML with CBFB::MYH11 fusion (without NPM1 mutation), and (iv) AML with RUNX1::RUNX1T1 fusion, as well as a control group of healthy stem cell donors.[43]

This AML Cytomorphology dataset is significant in building and validating artificial intelligence models that are for the diagnosis of leukemia as it contains a wide variety of cell morphological features in leukemia. A total of 189 peripheral blood smears from the Munich Leukemia Laboratory (MLL) database from the years 2009 to 2023 were digitized. First, all blood smears were scanned with 10x magnification and an overview image was created. Using the Metasystems Metafer platform, cell detection was performed automatically using a segmentation threshold and logarithmic color transformation. Further analysis regarding the quality of the region within the blood smear was performed automatically. Per patient, 99-500 white blood cells were then scanned in 40x magnification via oil immersion microscopy in .TIF format, corresponding to 24.9m x 24.9m (144x144 pixels). For this, a CMOS Color Camera from MetaSystems with a resolution of 4096x3000px and a pixel size of 3.45m x 3.45m was used. Four pixels were binned into one, leading to a size of 6.9m x 6.9m, and a resolution of 6.9m / 40 (1px = 0.1725m). Additional information about the patient's age, sex, and blood counts are provided in a separate .csv file.[43]

Images depict minor and major clinical features that are crucial for diagnosis, thus the AITTA dataset is suitable for the training of deep learning models that identify and classify leukemia. More specifically, the AML Cytomorphology dataset targets single-cell images whose morphology needs to be assessed accurately. This type of image within the volume of this dataset depicts distinct cellular traits like nuclear contour, chromatin distribution, texture of the cytoplasm, and some other important structural components. The availability of this depth of knowledge enables the training of the machine learning models on the features that enable the discrimination between leukemic and normal cells. The expert pathologists' morphological classifications are also part of the dataset which serves as model training and validation guidance. All the annotations significantly enhance the performance of such algorithms as they serve as absolute reference points for model predictions which consequently improve classification metrics.

The sample dataset has been shown in Figure 3.2.

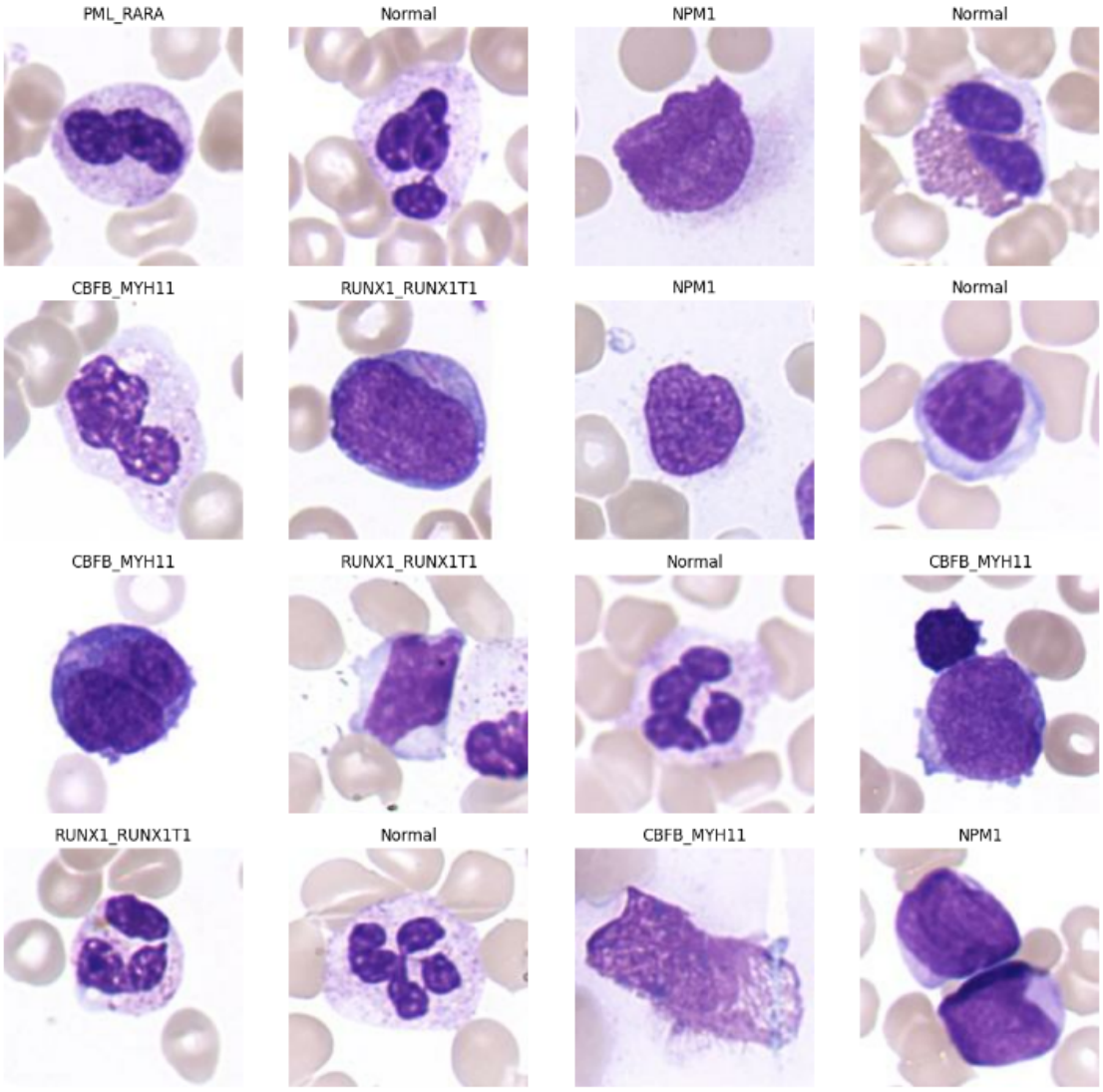

Figure 3.2: Dataset Sample

3.2 Data Pre-Processing

To prepare the dataset for training and testing, multiple pre-processing steps were employed to standardize and augment the images, thereby improving the robustness and generalization capabilities of the CNN-Transformer model.

3.2.1 Resizing and Rescaling

From the image preprocessing pipeline, the first component is the restructuring and rescaling of images to a standard format. Resizing addresses the problem of all images in the dataset being irregular and varied with the values fixed by $IMAGE\_SIZE$, which should be the case. This is important as it is difficult for deep learning networks particularly, convolutional neural networks to work on the data whose dimensions are not the same. Given the nature of the input data, images for this instance, images that do not meet the required input size of the model are resized. Normalization will therefore be performed by the Rescaling layer, which changes pixel values to the fraction of 0 and 1, by dividing each pixel value by 255. This is because normalizing each pixel will always have a value between 0 and 255. As for neural networks, Normalization is critical since it averts problems such as slow convergence during training, as well as ensures that all the input features are such that they are reasonably close to each other across a given scale, thus improving the model's performance.

3.2.2 Data Augmentation

Data augmentation entails creating modified instances of original images to synthesize additional images which would otherwise be very difficult to gather such as naturally inclined images of objects. Rotation for example is used to combine randomization of positional modifications of an image while maintaining the object intact. The goal of this is to increase the resiliency of the model against various distortions the image may undergo as it would be rotated in the process of training. These are all examples of how augmentation works, as it introduces a certain amount of noise to the dataset the model is being trained with to increase its efficacy. Seed 42 handles specifically the field of version determination. Data augmentation is a necessary part of training deep learning models because it enables in a sense more variable data without having to collect other actual data.

3.2.3 Application of Data Augmentation for training Dataset

To enhance the model diversification and generalization properties, the training dataset was augmented with flipping and random rotation techniques. They form the type of variability that may be encountered in the real world, such as in the case of the same object but being viewed at different angles. Random rotation resulted in rotating the image to a maximum of 20% of its dimensions to allow the model to see an object in different positions. Moreover, some images were horizontally flipped so that objects do not always appear in the same position. A fixed seed was used to control these transformations (seed=42) to achieve repeatability. In this case, 20% of the training data had augmentation performed so that the model could be exposed to a wider variety of images but without distorting the meaning embedded

in the labels. Such a procedure lessens the probability of overfitting and improves the performance of the model in new data.

3.3 Implemented Models

3.3.1 Xception

The building blocks of this architecture include depthwise separable convolutions which are extensions of the Inception model's components. There exists a crucial difference between Xception and Inception when it comes to viewing structuring the convolutions. In the conceptional design of this architecture, the volumetric weight kernel is separated into two: spatial, and channel. This results in the reduction of the parameters and computation cost while maintaining a given level of performance. Xception's architecture was constructed with the premise of using depthwise separable convolutions to make Inception blocks better, thus allowing them to have a slimmer design and still maintain great performance on large image datasets. Its effectiveness in image classification tasks has been made known, hence its frequent application in computer vision and transfer learning tasks.

3.3.2 VGG(Visual Geometry Group)

The Visual Geometry Group at the University of Oxford developed a deep convolutional neural network type of architecture which is referred to as VGG (Visual Geometry Group). The model is characterized by its depth and simplicity and the two most common variants are VGG16 and VGG19 which only have a variation in the number of layers they possess. VGG implements a uniform structure where all the convolutional layers use 3x3 filters along with 2x2 max pooling layers. However, this simplicity comes at the cost of having a high parameter count which makes VGG very expensive in terms of computation and memory. Notwithstanding these disadvantages, VGG has been widely used for several applications such as image recognition and even for feature extraction. It is good enough to be considered a baseline model for deep learning and can be found in a wide variety of research as a transfer learning model, that is when researchers want to use large models with more complexity.

3.3.3 ResNet

ResNet is a deep learning model that has gained attention for its architecture in training Very deep neural networks. The main advancement of ResNet over other approaches is the use of residual connections, or skip connections in the model. In a neural network, these skip connections instead of predicting the output directly, predict the error from the desired output. In the presence of a senior network as the aim, optimal direct outputs are pre-trained, which makes the networks able to evolve in much deeper networks without reducing the ability of networks. ResNet has been dominant in several domains ranging from image classification to object recognition and separation. The Universe of variants includes ResNet50, ResNet101, and ResNet152. These networks vary across architecture depth and size characteris-

tics. Due to its strength and cross-fit, the ResNet architecture has been a revolution in the world of deep learning.

3.3.4 Inception

Inception is a term that refers to the convolutional neural network design that was proposed by Google and that proved to be revolutionary in image classification tasks. Central to Inception's architecture is the so-called "Inception modules," which compute three different sizes of convolutions (1x1, 3x3, 5x5) within a single layer. This strategy enables the network to learn spatial features at various levels within the same layer of the network. Also, the inception structure helps to expand the receptive field without heavy computation by applying 1x1 convolutional layers before applying larger kernels. Such an approach enables improving level feature representation with lower efforts in computation. In particular, several types of inception were created, Inception-v3 is the most popular among them, especially for image classification problems. Its modular architecture, ease of scaling, and high efficacy turned it into a very efficient architecture, especially in scenarios that require rapid and accurate analysis of images.

3.3.5 DenseNet(Densely Connected Convolutional Networks)

DenseNet is a deep learning architecture that takes the idea of dense connections between the layers one step further. In this case, every layer receives every other layer's output through a feed-forward mechanism which decreases the amount of work that has to be done by this network. This dense connectivity is said to be beneficial in two main aspects: it helps to decrease the number of parameters and enhances gradient backflow during backpropagation. One of the concepts put forth in DenseNet is that a layer should always be able to accept inputs from all the layers that are before it. This promotes better usage of features and consequently better representations of learning. The efficiency and performance of the Densenet architecture are promising and enable it to surpass the accuracy and efficiency of normal CNNs making it suitable for cases where time for training and parameter usage are of concern. The architecture has been applied to various domains such as medical image analysis, object detection, and image segmentation.

3.3.6 DaViT(Deformable Attention Vision Transformer)

The Deformable Attention Vision Transformer was able to combine deformation and attention things to the Vision Transformer architecture that was already in place and improve its capacity on more complex spatial relationships via attention. The original attention mechanism used in the ViT allows every patch to be treated the same, whereas DaViT implements a deformable attention mechanism. The attention lets the model choose the areas of an input image that are relevant for the task, likely improving performance on irregularly shaped and fine details of images and allowing the model to perform better. From the time when DaViT was introduced to the world, it has consistently outperformed computer vision models specializing in object detection and segmentation by this time transforming even objects with finer details to the patches that were the most relevant.

3.3.7 MaxViT(Maximal Vision Transformer)

MaxViT is a new architecture of a transformer design that is aimed at solving the computation limits of Vision Transformers (ViTs). The most important feature of MaxViT is its ability to incorporate local and global attention mechanisms in a single framework. This approach captures both short-range and long-range dependencies. MaxVit combines a hierarchical attention system to cut down the number of computations required for self-attention while keeping efficiency intact. This arrangement allows MaxViT to process and analyze the data on a large scale as compared to the traditional ViTs. Hence key tasks that involve the use of large and sophisticated images become a key focus area for MaxViT. Compared with other transformer-based models, MaxViT also outperforms other models in terms of broader-scale vision tasks such as image classification, object localization, and detection.

3.3.8 ViT(Vision Transformer)

Vision Transformer (ViT) is a new model in deep neural networks that uses a transformer architecture which was first implemented in natural language to the computer vision domain. The image for ViT is first split into fixed patches, each of the patches is first linearly embedded as a vector before being sent to the transformer encoder. Unlike the use of convolutional neural networks (CNN) which looks at the image in successive layers, ViT uses the attention mechanism in transformers to look at the whole image at a go. ViT has displayed outstanding efficiency on large sets such as ImageNet and earned recognition for its worldwide information processing. Thus, we see that while ViT does require a great deal of data to function well, it has broken records in most image classification works, as well as a few aspects of vision. Consequently, in the recent developments in deep learning, this is one of the most advanced models.

3.3.9 Hybrid Model (ResNet50 + ViT)

A hybrid model that incorporates ResNet50 in addition to a Vision Transformer utilizes a combination of both architectures' strengths. ResNet50 acts as a backbone for feature extraction, employing its deep residual blocks to capture hierarchical and spatial information. The features extracted from ResNet50 are subsequently passed on to a Vision Transformer for the establishment of long-term and contextual relationships. This hybrid concept benefits from reasonable feature extraction of ResNet50 and the global attention mechanisms of Vision Transformer resulting in improved image classification, object detection, segmentation, and other model performance metrics. By merging the local feature extraction capabilities of CNNs with the global context processing abilities of transformers, this hybrid model can efficiently model detailed and large-scale dependencies in images, hence, making it suitable for various challenging tasks in computer vision.

Mechanism of ResNet50+ViT combination

The roles of ResNet50 overlap in its focused features concerning images, such as the texture, shape, and border of the cells, and our study benefits from its approach.

Its residual connections help the model retain detailed spatial information needed to recognize patterns more effectively. ViT performs the study at the global level, where it pieces the image with fixed-sized patches, applying self-attention to analyze long-range dependencies and spatial relationships throughout the image. With a fusion of the two ideas, the hybrid model thus enjoys the benefits of fine-grained details, at the same time viewing the structures on a macro scale that supports the classification of cellular structures more accurately.

One important breakthrough in this hybrid approach is in accessing both local and global features. ResNet50 gives detailed insight and concentration for specific spatial areas in the images; ViT, on the other hand, provides understanding within a larger context for spatial relationships. This dual representation sequence allows the hybrid model to more effectively identify and classify the leukemia cells as compared to the standalone use of these architectures. An example of an enhancement is the detection of slight morphological changes in cells that have escaped detection by standalone CNNs, using an attention mechanism to gaze at the whole picture. The architecture of the model is shown in Figure 3.3.

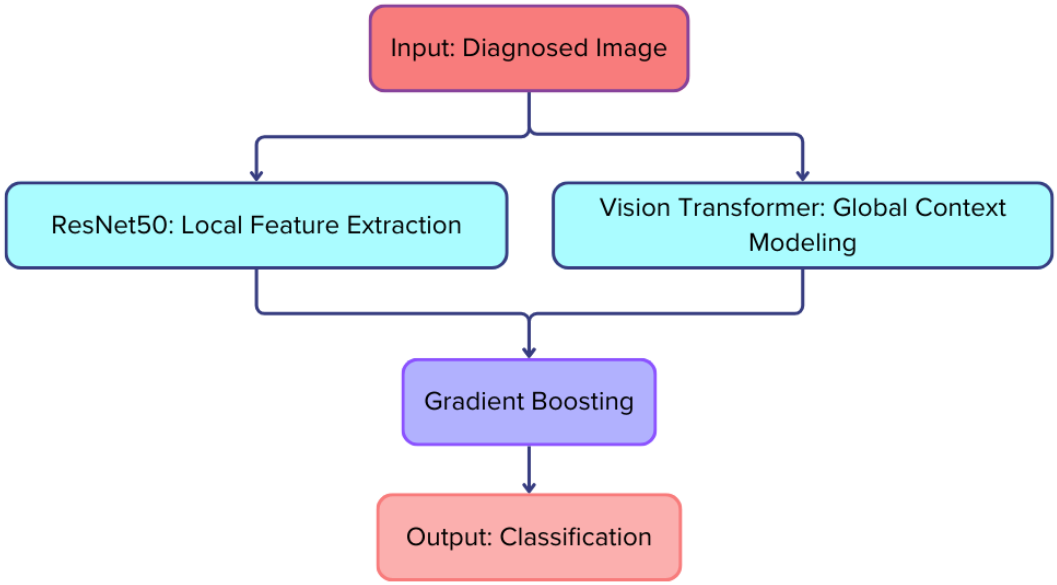

Figure 3.3: Hybrid Model Architecture (ResNet50 + ViT)

3.4 Evaluation Metrics

Accuracy

Accuracy is often referred to as the main basis for evaluation of the model adoption

on the prediction. Its weight in the overall effectiveness of the model is substantial, though for imbalanced datasets its value is not so high as it ignores different types of errors.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.3.1}$$

F1 Score

F1 ideal implements the combination of two important features of the effectiveness metric, even when the datasets are not evenly distributed among the available classes. F1 is stated as a combination of precision and recall, which appropriately partially addresses the issue of false positive and false negative. When the F1 score is high it means that a model is robust and can quickly make decisions without much confusion even when there is a lot of noise present in the environment.

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{5.3.2}$$

Precision

How correct the positive predictions are provided – that is to say, precision. High precision suggests that a given model has been good at cutting back on false positives which is a good thing if it costs too much to be spending time and resources going through false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.3.3}$$

Recall

Recall is a measure of how many of the actual positives the model successfully captures meaning the ratio of true positives to a total number of actual positives. A high recall rate is desirable in situations where false alarms are acceptable even as all the positives must be classified. Sensitivity is a way to assess the probability of finding positive cases by a test or a model. For instance, in medical diagnostics, it can be explained as the ability of a test that seeks the condition to find the condition among those who have it.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3.4}$$

Explanation of Terms

Where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

Chapter 4

Result and Analysis

4.1 Analysis of Model Performance

The evaluated models, including Xception, VGG, ResNet, Inception, DenseNet, DaViT, MaxViT, ViT, and a Hybrid model combining ResNet50 and ViT, are state-of-the-art deep learning architectures commonly employed for image classification tasks. These models leverage advanced feature extraction capabilities, convolutional layers, and attention mechanisms to identify patterns within datasets effectively.

Xception

The Xception model improves the performance of a deep neural network by building on top of depthwise separable convolution and can achieve modular representations by separating spatial and channel information. Despite the introduction of newer architecture, the performance of the model was still quite low achieving an accuracy of about 0.4084, F1 score of 0.3949, and recall and precision equal to 0.4084 and 0.4061 respectively. Such scores are indicative that the model struggled greatly in regards to generalizing to the dataset likely due to the model's ability to learn complex patterns being inadequate. The following results are provided in Figures 4.1, 4.2, and 4.3 in the form of the confusion matrix, the Accuracy/loss curve, and the inference results together with four figures whose confidence scores are indicated.

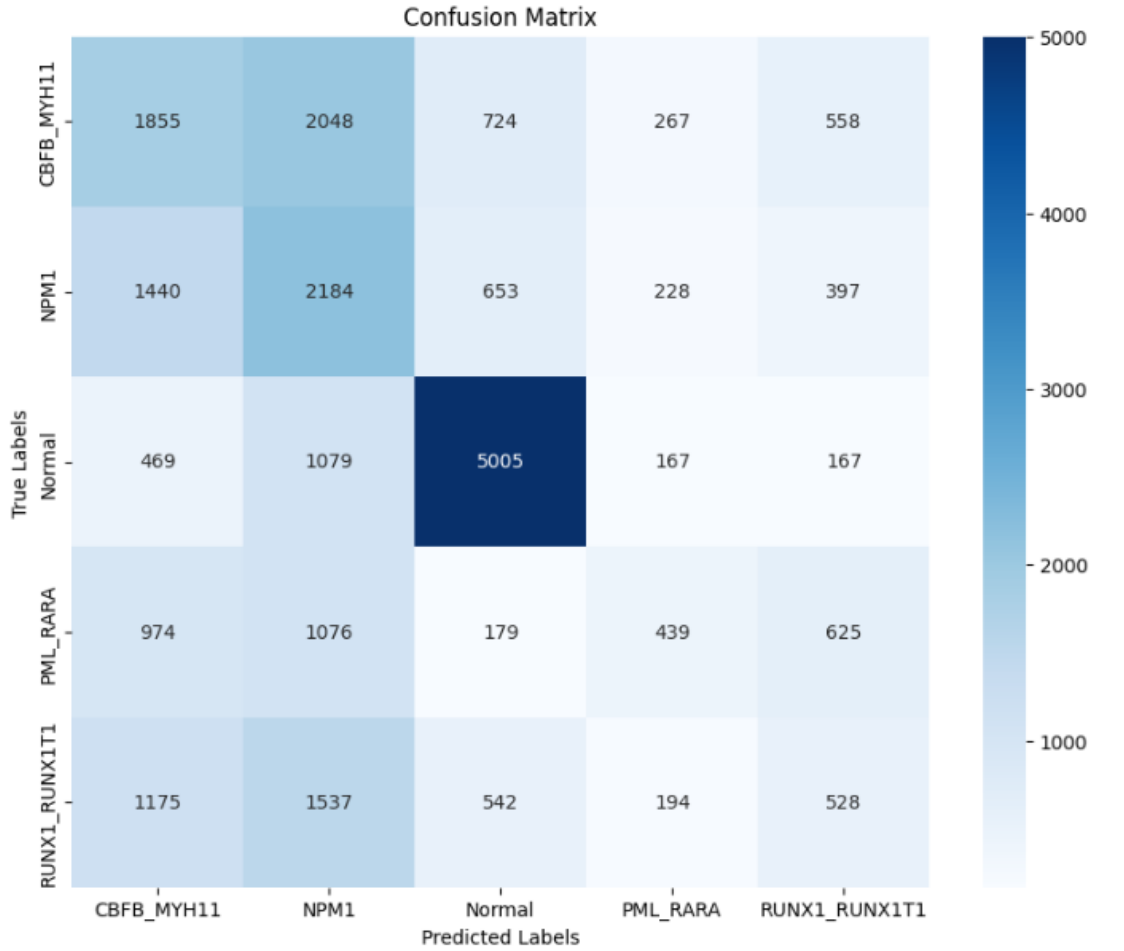

Figure 4.1: Confusion Matrix of Xception Model

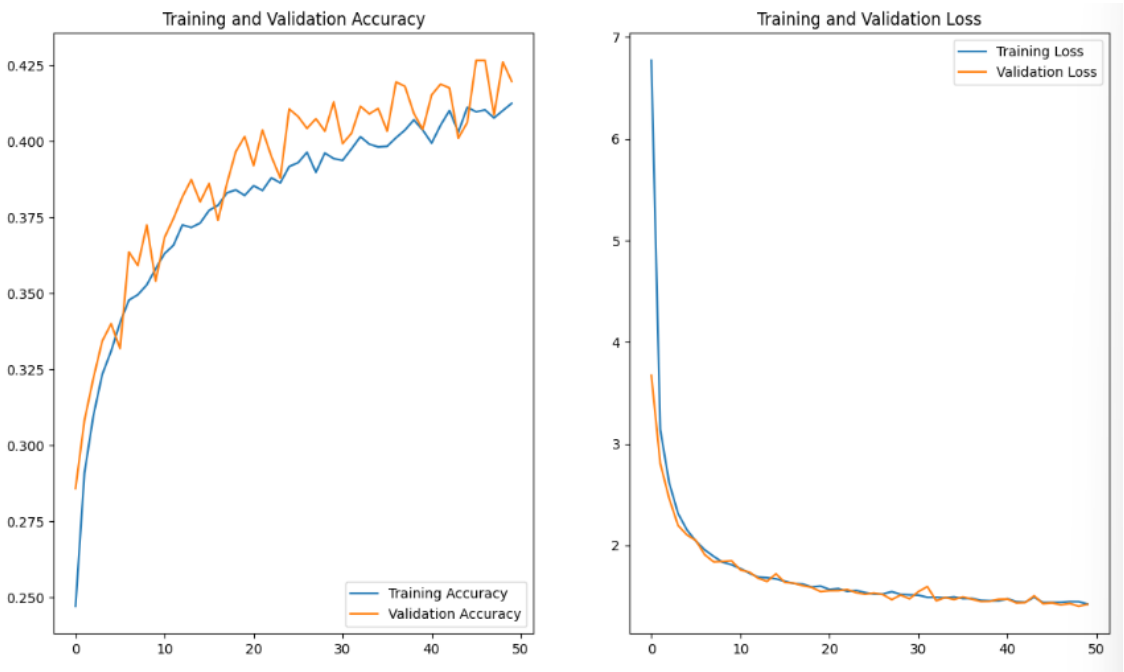

Figure 4.2: Training, Validation Accuracy and Loss of Xception Model

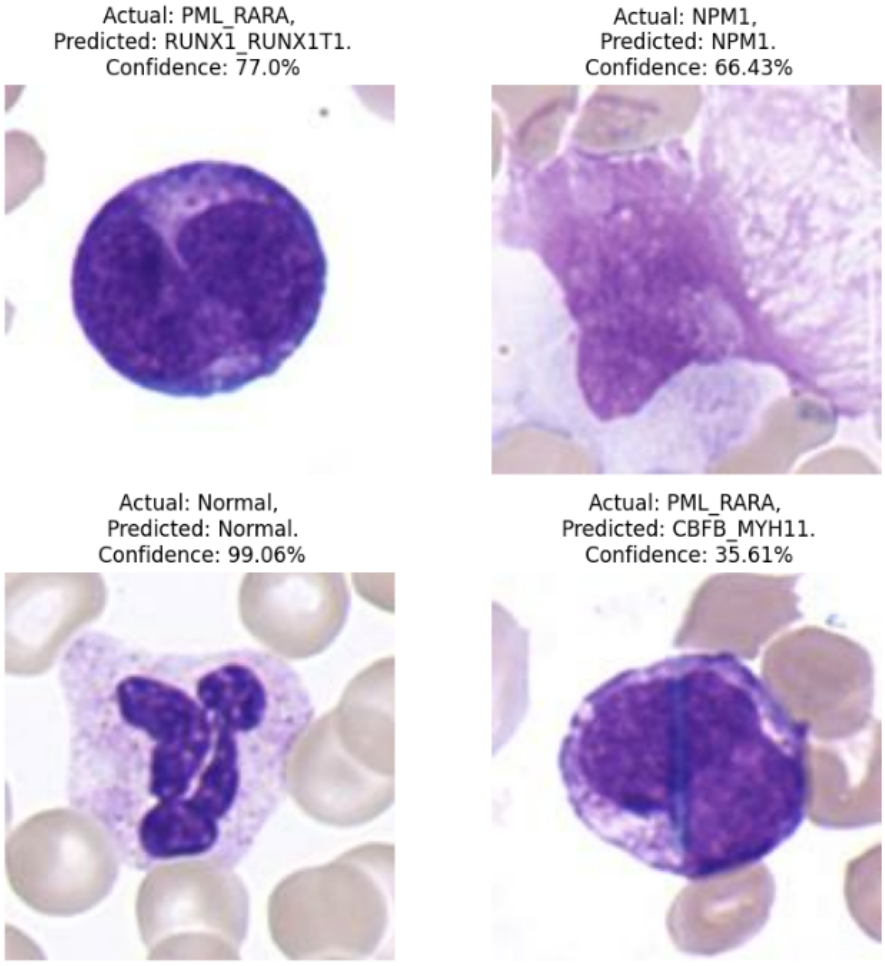

Figure 4.3: Sample Inference of Xception Model with confidence levels

VGG

The VGG model, coupled with its popularity owing to a sequential deep convolution architecture with a compact structure proved to perform reasonably well in that its accuracy was 0.4969, F1 score of 0.4893, and recall and precision of 0.4969 and 0.4867 respectively. Even though it did better than Xception, its ability to model complex structures and sophisticated datasets was greatly limited. The confusion matrix, loss curve, and the inference graphs are presented in Figures 4.4, 4.5, and 4.6 which can provide further understanding of the predictions and behavior of the model.

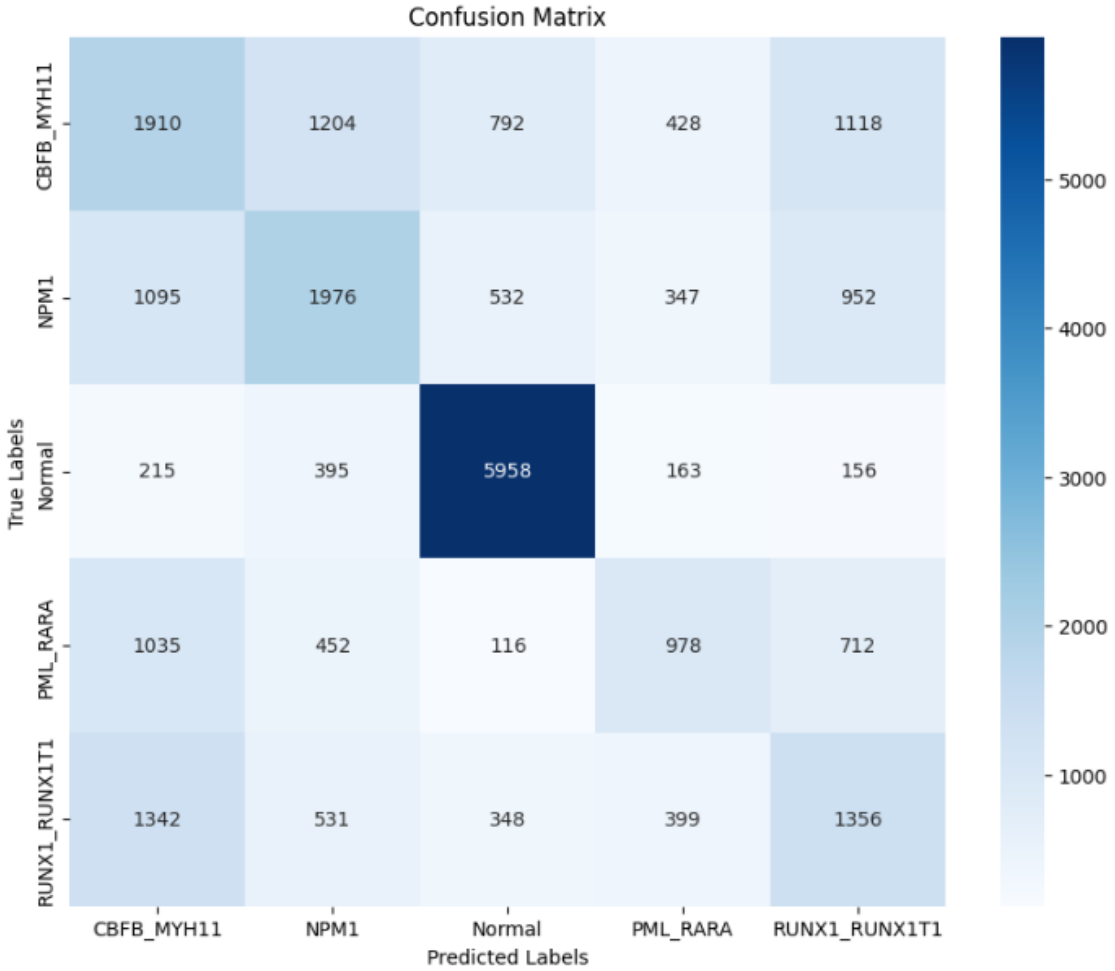

Figure 4.4: Confusion Matrix of VGG model

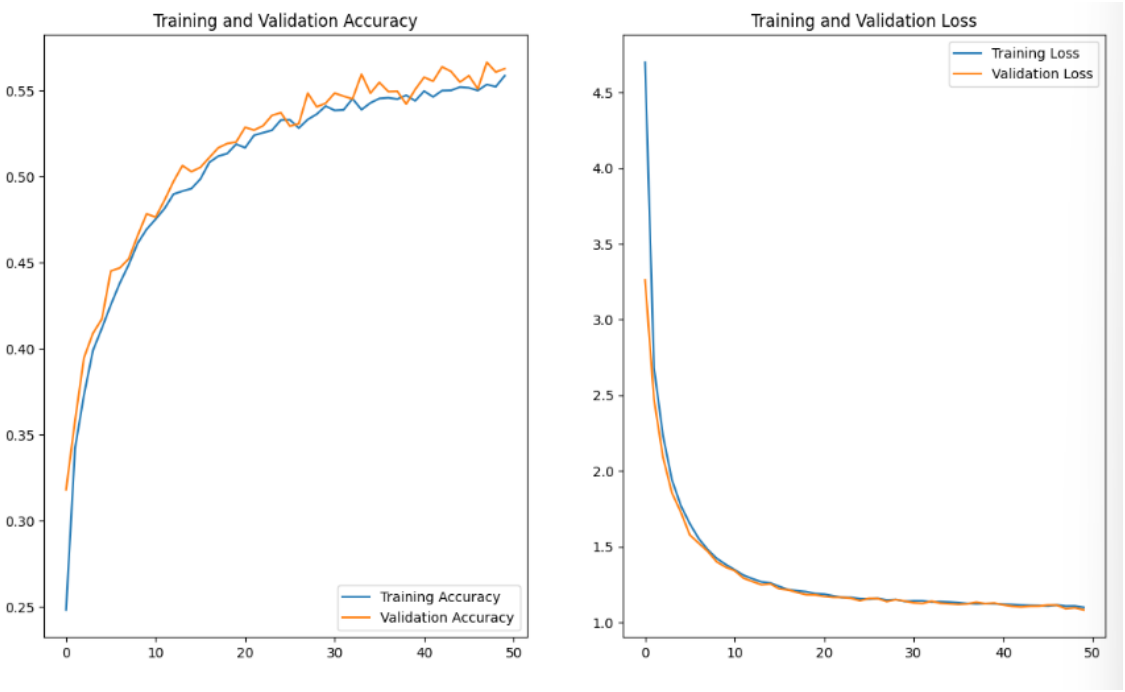

Figure 4.5: Training, Validation Accuracy and Loss of VGG model

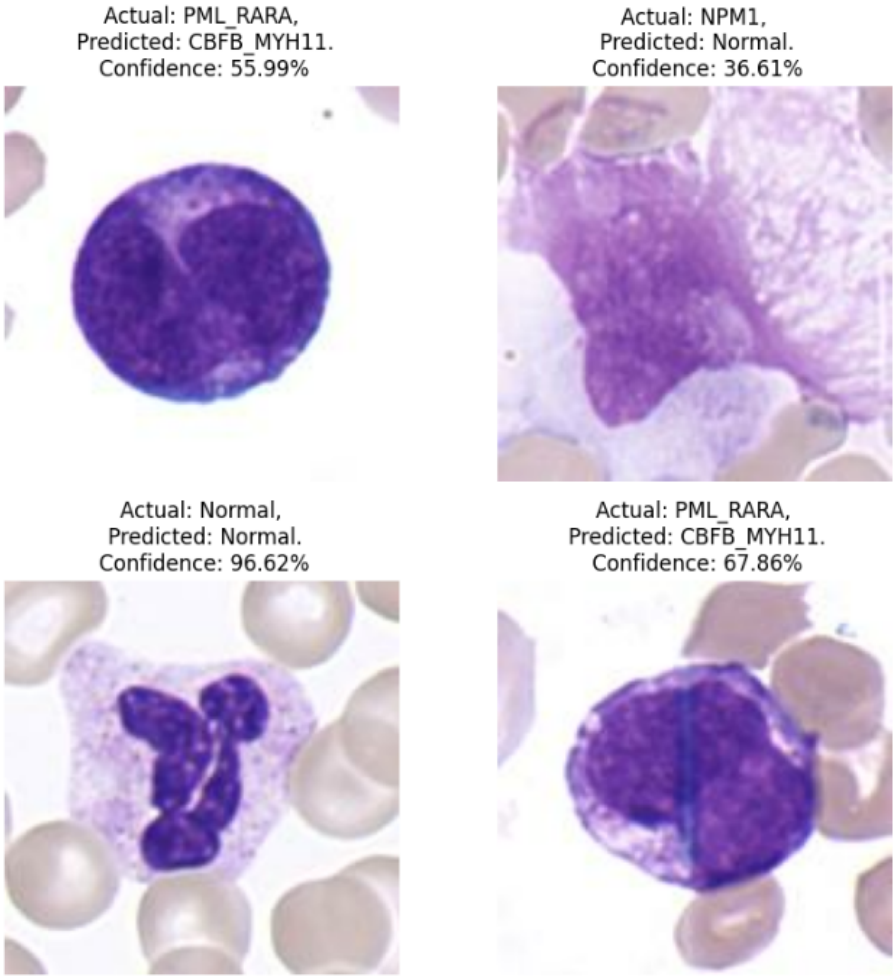

Figure 4.6: Sample Inference of VGG model with confidence levels

Resnet

ResNet, which makes use of residual connections to address deep learning vanishing gradient problems, produced better results than simpler models. It achieved an accuracy of 0.5461, F1 of 0.5369, recall of 0.5461, and precision of 0.5356. These results emphasize its capability of learning more layers and the ability to keep the gradient well-behaved during the course of training. While the outcome was respectable, there are still things that can be better. The confusion matrix, loss curve, and inference results are included in Figures 4.7, 4.8, and 4.9 respectively to offer more details on its patterns.

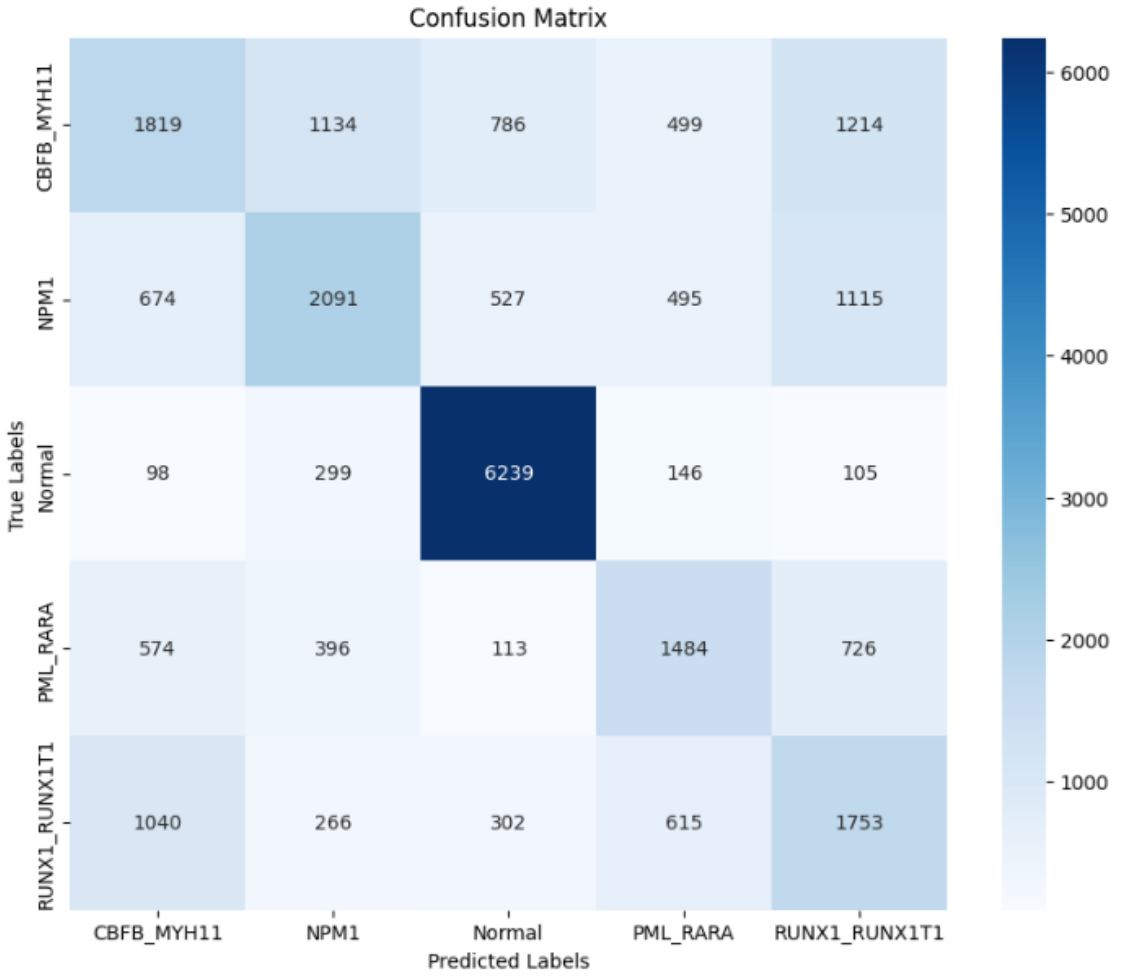

Figure 4.7: Confusion Matrix of ResNet model

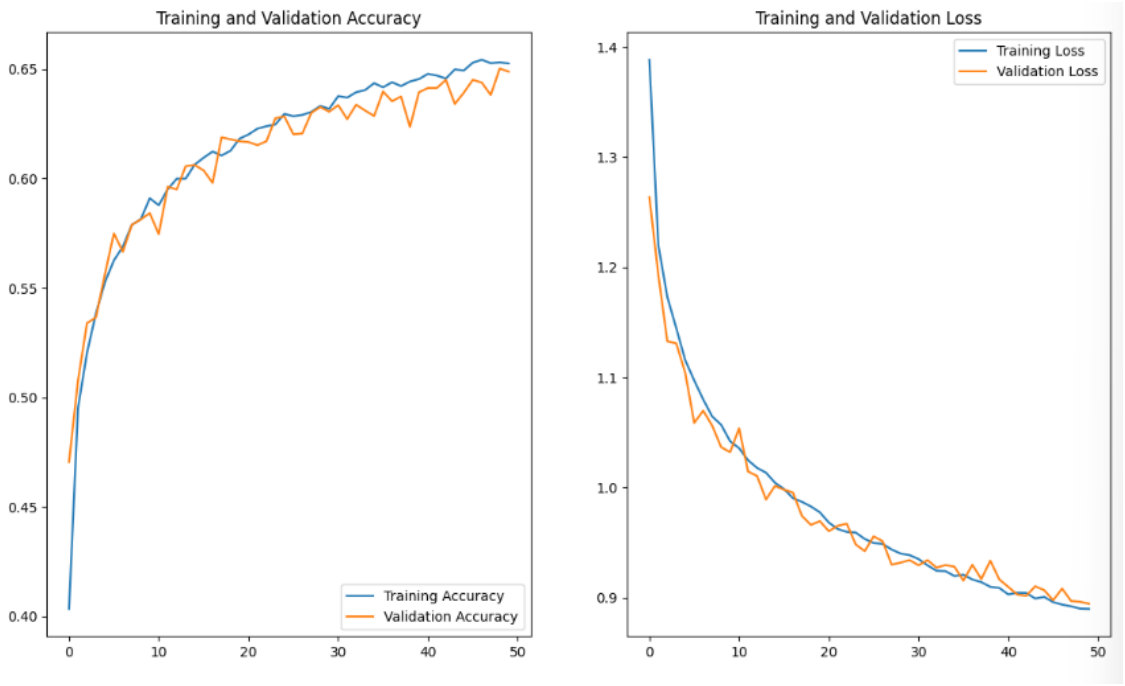

Figure 4.8: Training, Validation Accuracy and Loss of ResNet model

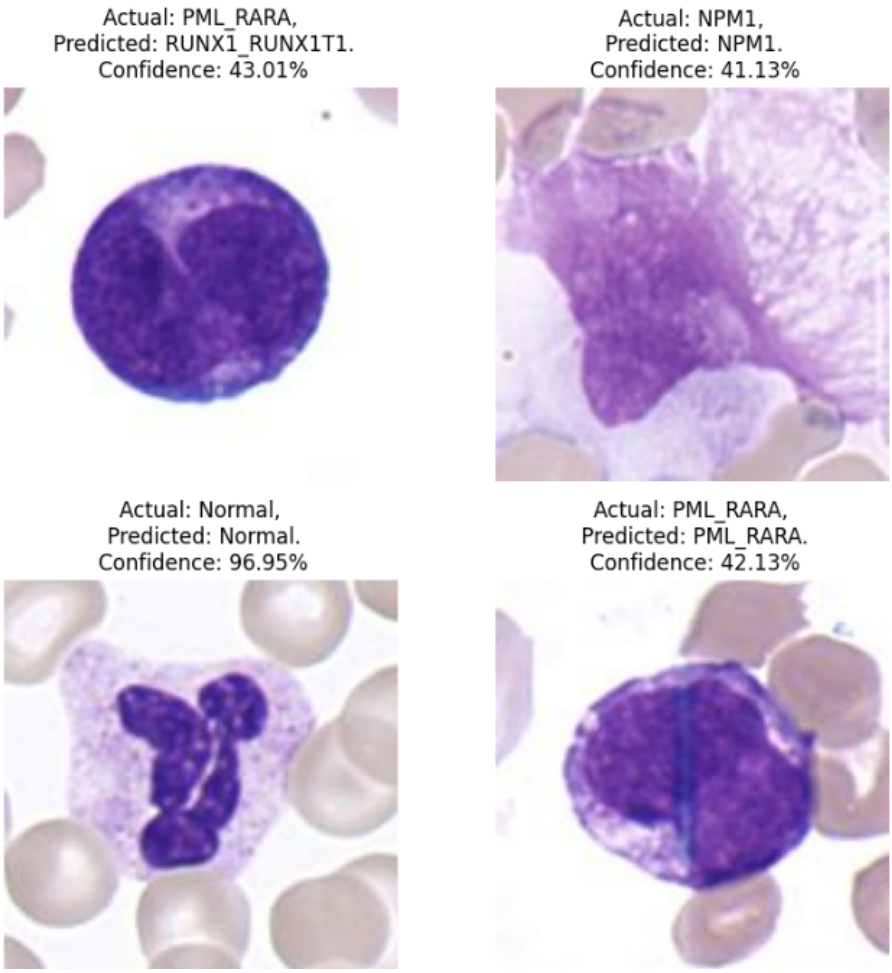

Figure 4.9: Sample Inference of ResNet model with confidence levels

Inception

The Inception model, characterized by its availability of inception blocks which aids in multi-dimensional robust extraction of features, yielded the lowest performance result on this architecture set. It performed at an accuracy of 0.3485, F1 score of 0.2954, recall of 0.3485, and precision of 0.3893. These results, in turn, point to weaknesses that would be expected to be encountered, either because of the depth of the dataset or the structure that was used. Figures 4.10, 4.11, and 4.12 show the confusion matrix, a loss curve, and inference images respectively to summarize the overall results.

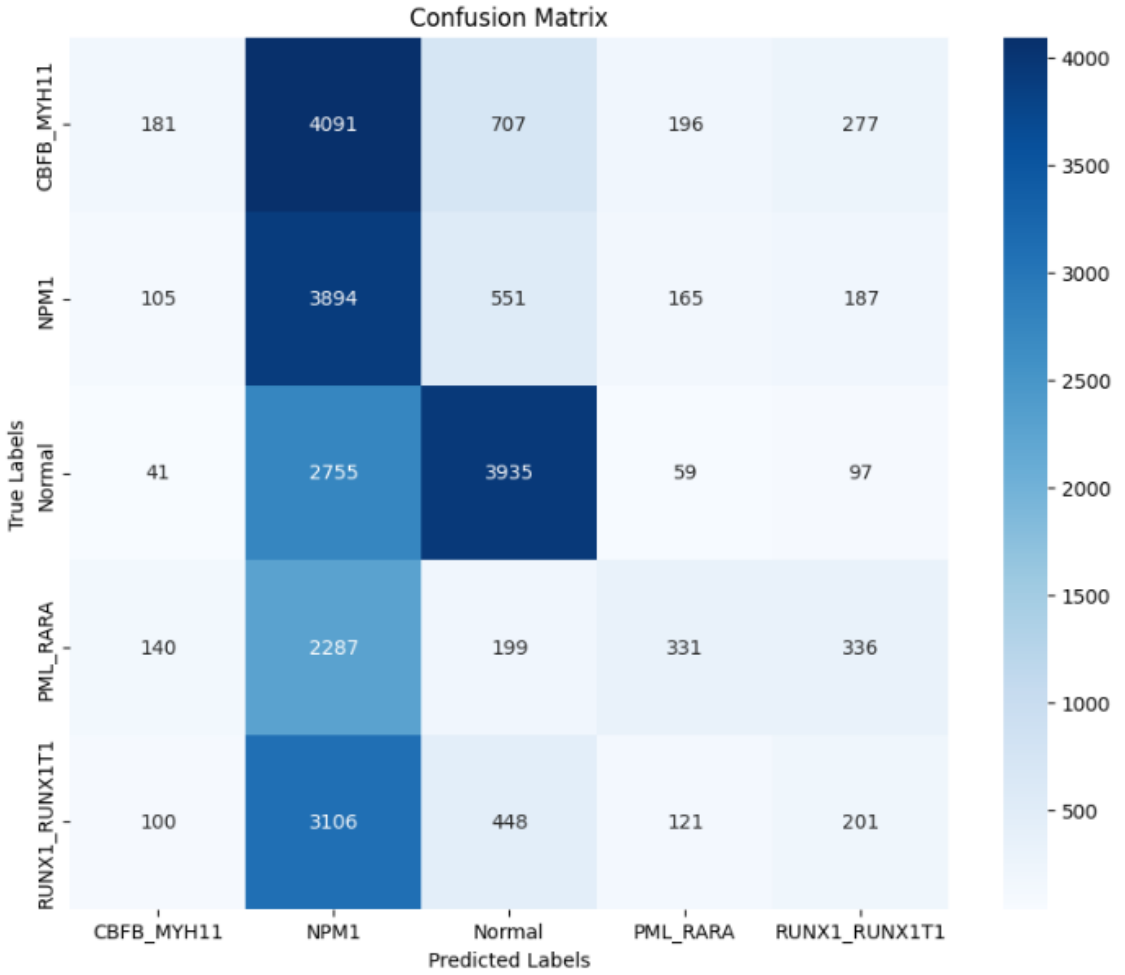

Figure 4.10: Confusion Matrix of Inception model

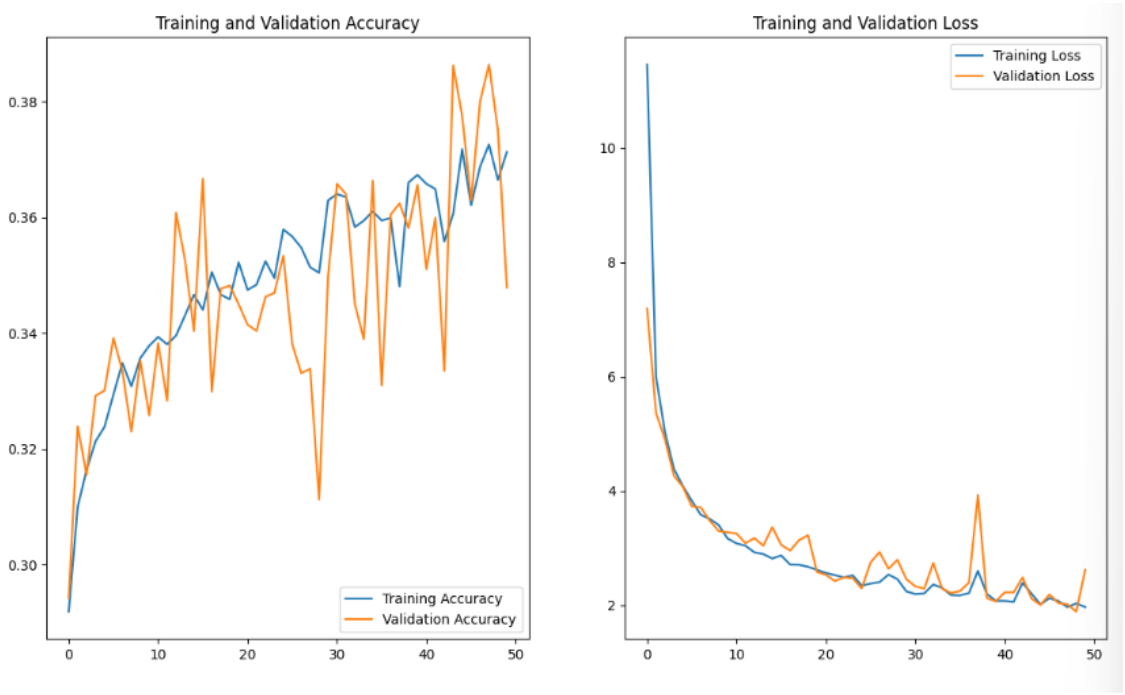

Figure 4.11: Training, Validation Accuracy and Loss of Inception model

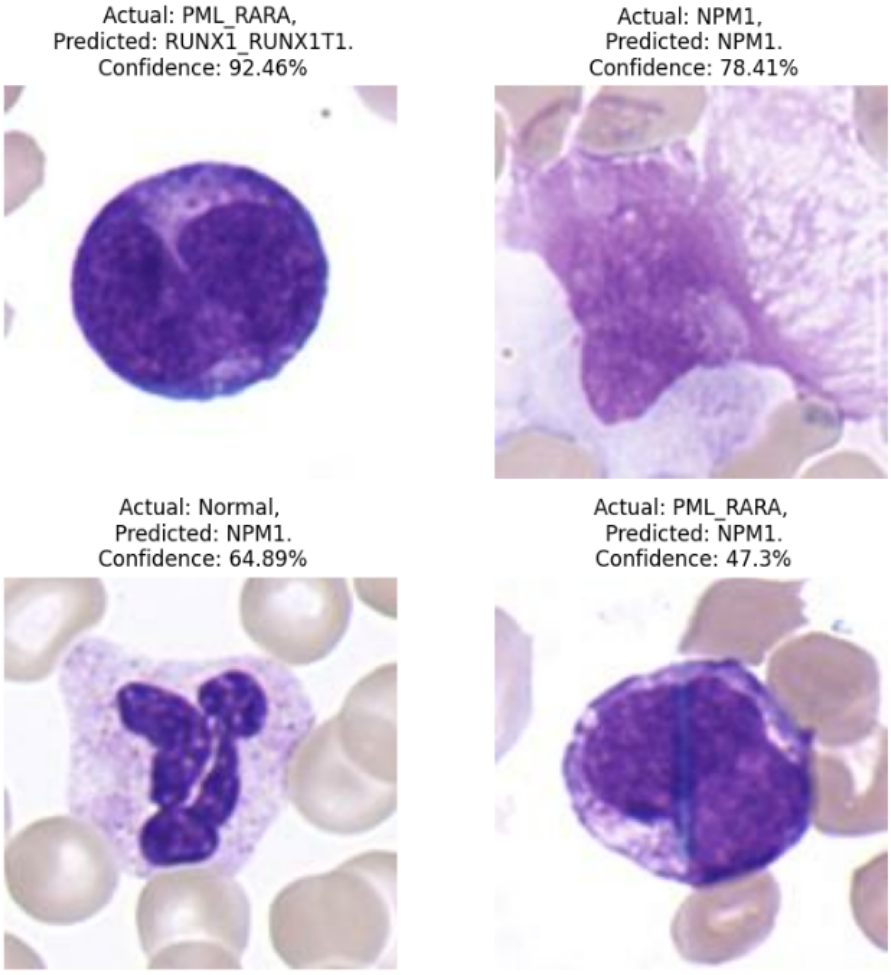

Figure 4.12: Sample Inference of Inception model with confidence levels

DenseNet

Maintaining a moderate performance with an accuracy of 0.4778, F1 score of 0.4628, recall of 0.4778, and a precision of 0.4778, DenseNet, which leverages dense connections to enhance feature reuse and improve the flow of gradients across layers, was at par with ResNet. It was, however, still able to provide a balanced performance due to its ability to transfer information through various layers. As seen in Figures 4.13, 4.14, and 4.15, DenseNet did perform decently in the given dataset as seen in the confusion matrix, loss curve, and visualization of inference.

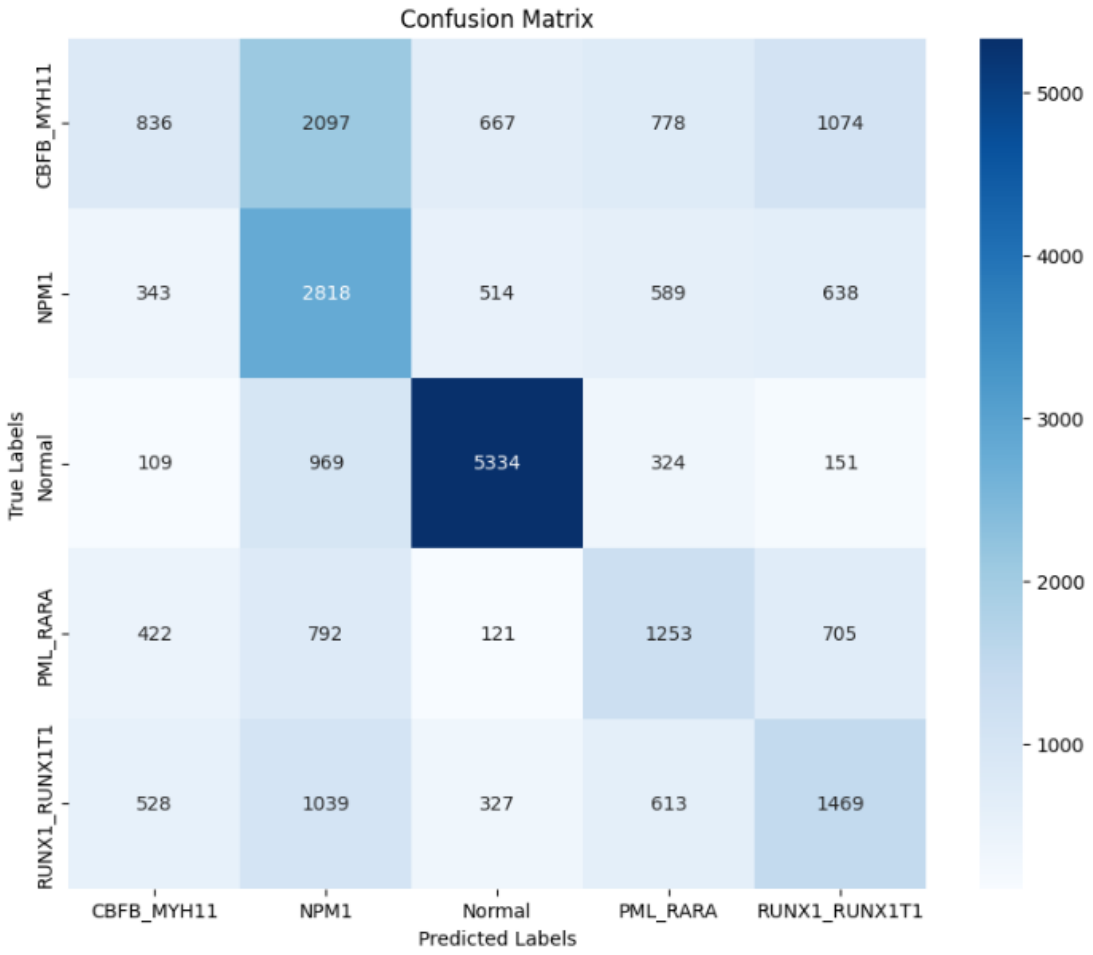

Figure 4.13: Confusion Matrix of DenseNet model

Figure 4.14: Training, Validation Accuracy and Loss of DenseNet model

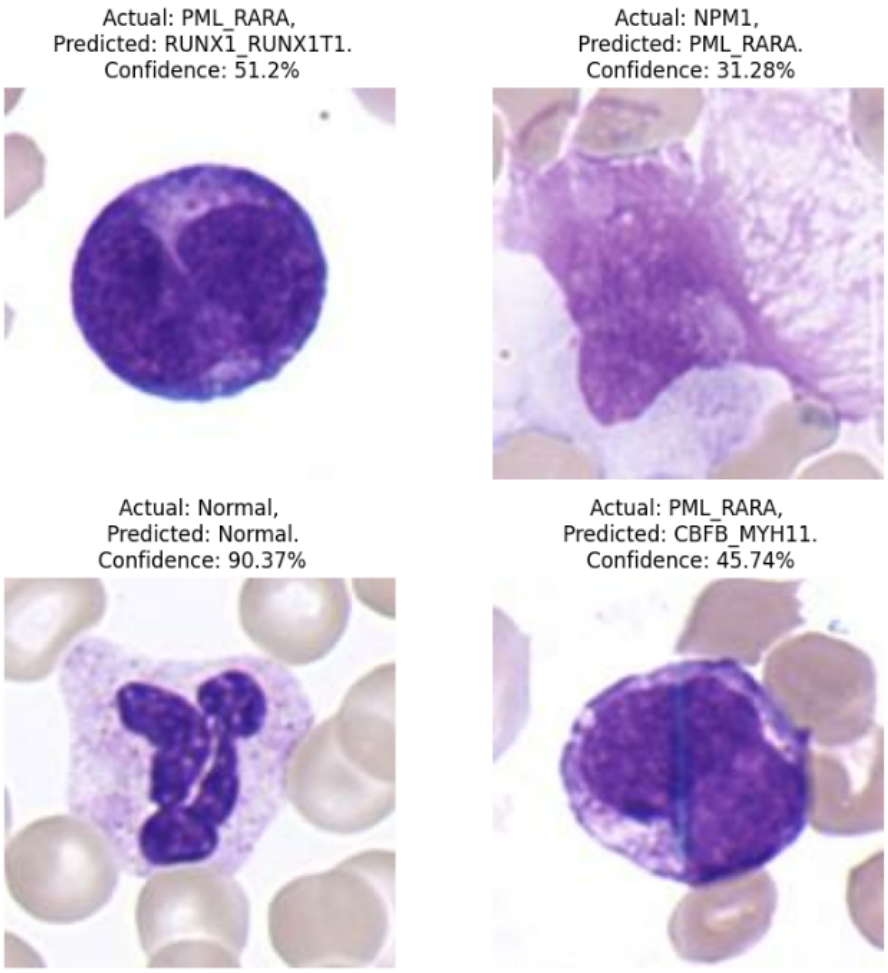

Figure 4.15: Sample Inference of DenseNet model with confidence levels

DaViT

Another model, DaViT, which is a dual attention vision transformer model has shown results with an accuracy of 0.5315, F1 score of 0.532, precision score of 0.5335, and recall of 0.5315 which are much better than some previous models. This model's metrics seem balanced in the sense that it appears able to predict well the most relevant features. Figures 4.16, 4.17, and 4.18, indicate the confusion matrix, loss curve, and visualization of inference for this model including confidence scores. This consistency is critical in ensuring that the model can process a variety of input features without losing its originality.

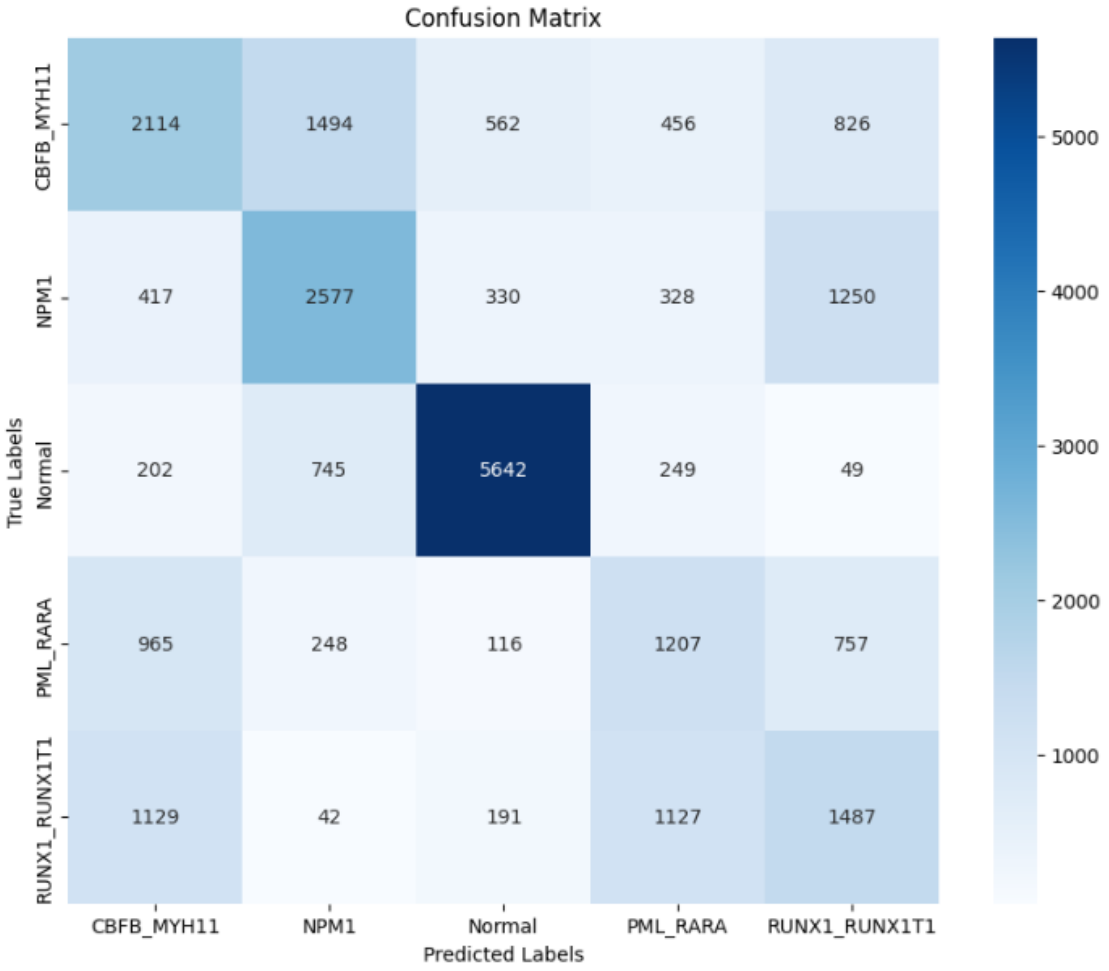

Figure 4.16: Confusion Matrix of DaViT model

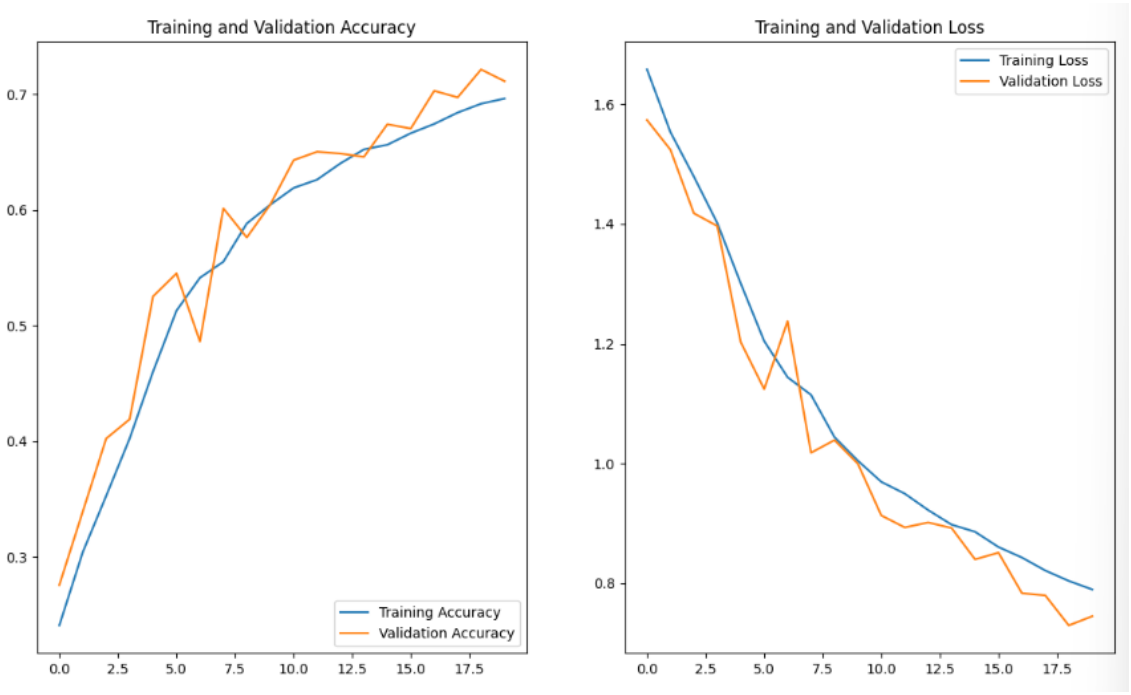

Figure 4.17: Training, Validation Accuracy and Loss of DaViT model

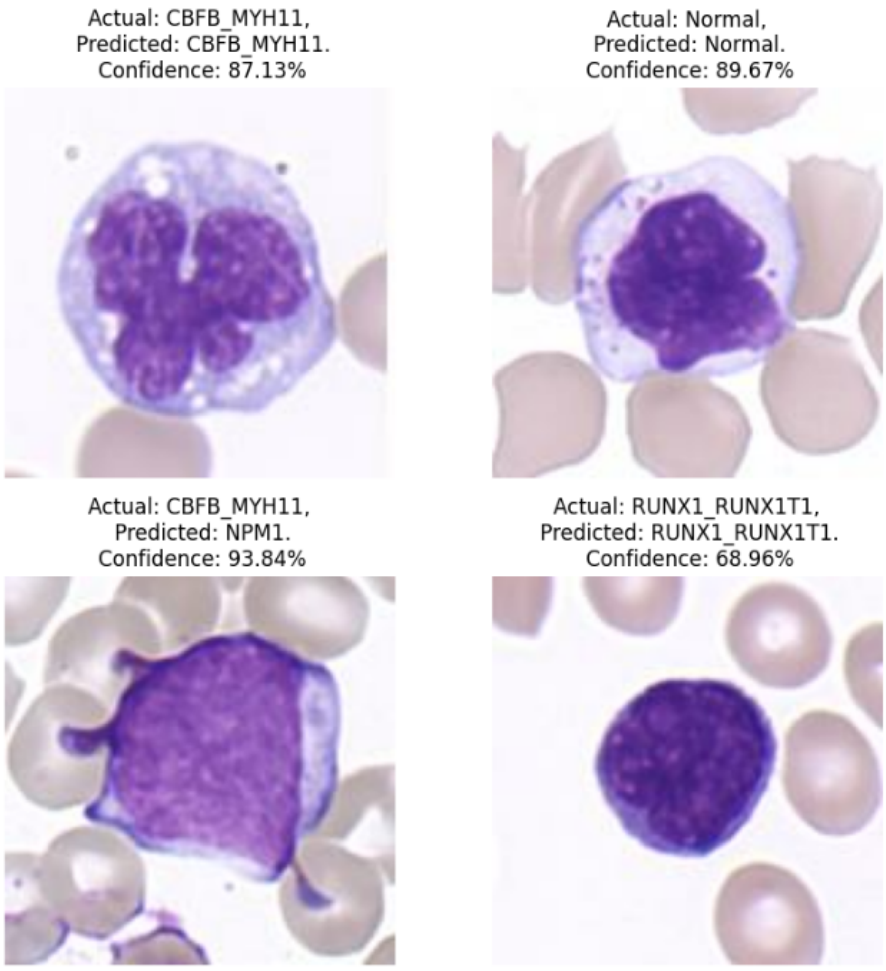

Figure 4.18: Sample Inference of DaViT model with confidence levels

MaxVIT

MaxViT, which utilizes multi-axis attention mechanisms to perform well in feature extraction, performed at the same level as DaViT with an accuracy of 0.5319, F1 score of 0.5294, recall of 0.5319, and precision of 0.5285. A good representation of the features was done through the attention mechanism-based architecture of the model, although much of its performance was capped in comparison to the Hybrid model. Figures 4.19, 4.20, and 4.21, detail the model performance which includes the confusion matrix, loss curve, and the outputs produced during inference such as predictions and confidence scores.

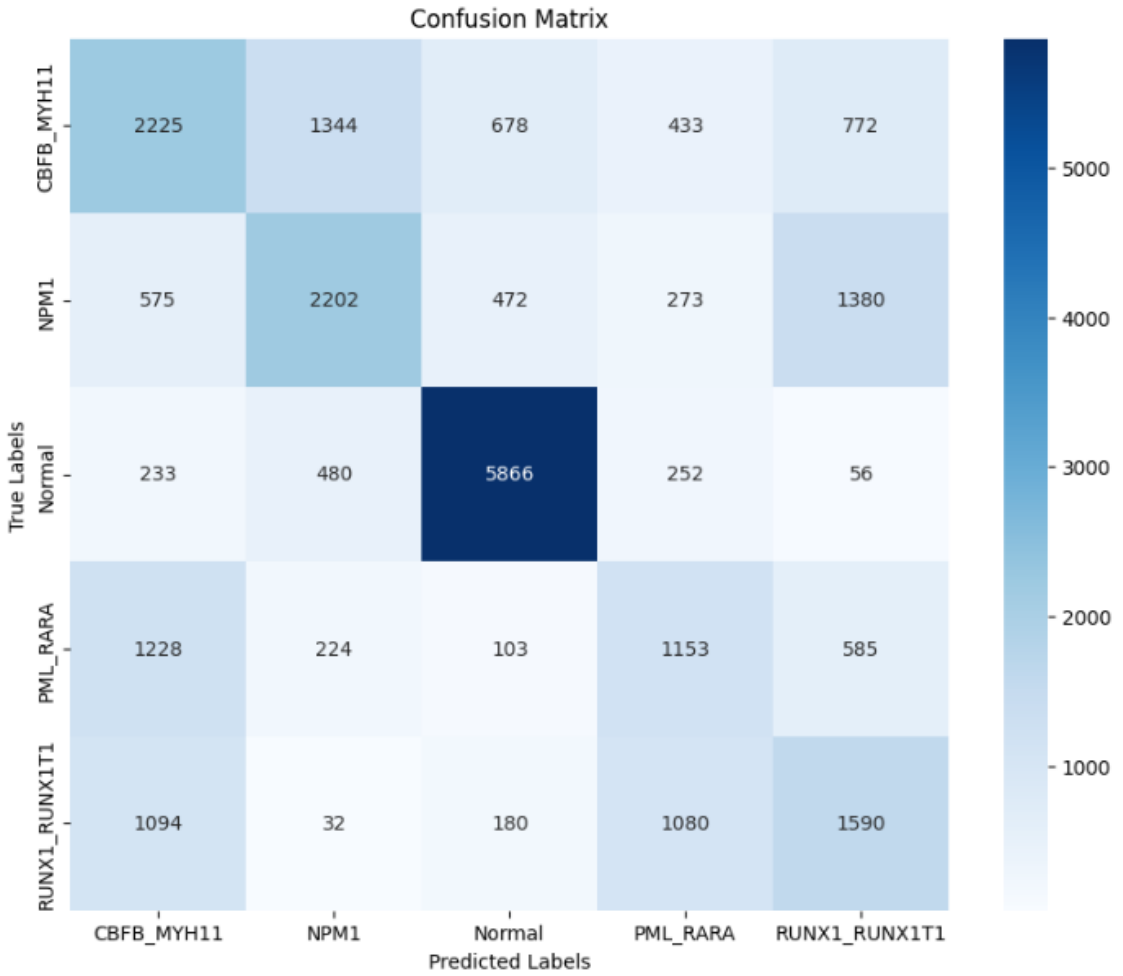

Figure 4.19: Confusion Matrix of MaxViT model

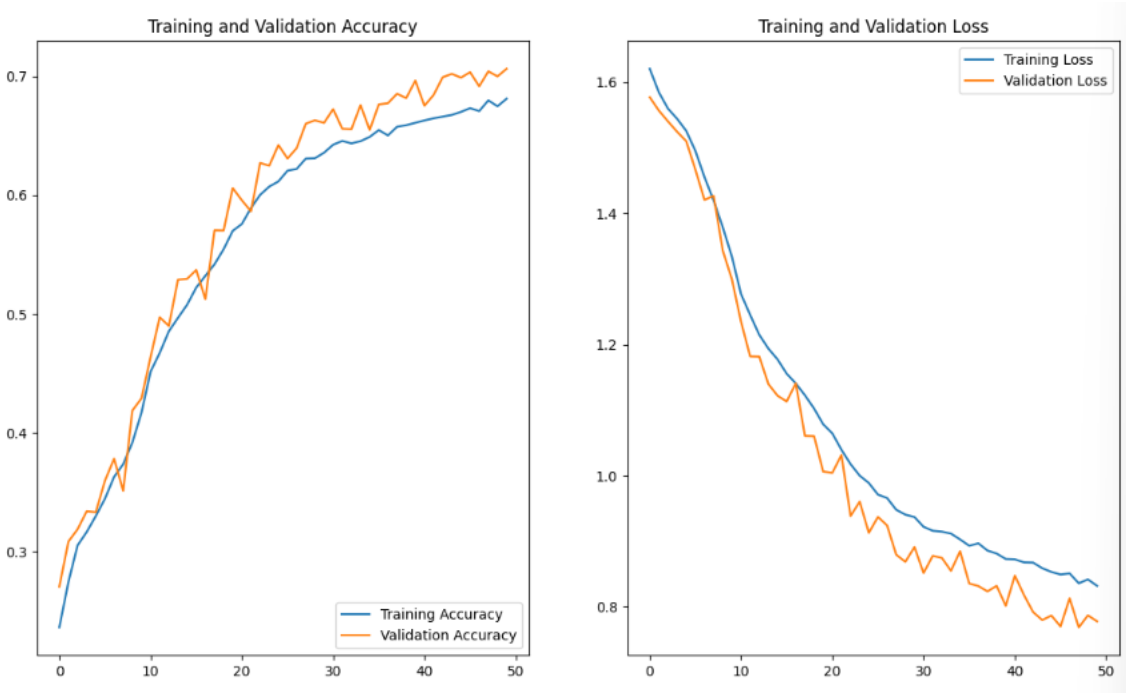

Figure 4.20: Training, Validation Accuracy and Loss of MaxViT model

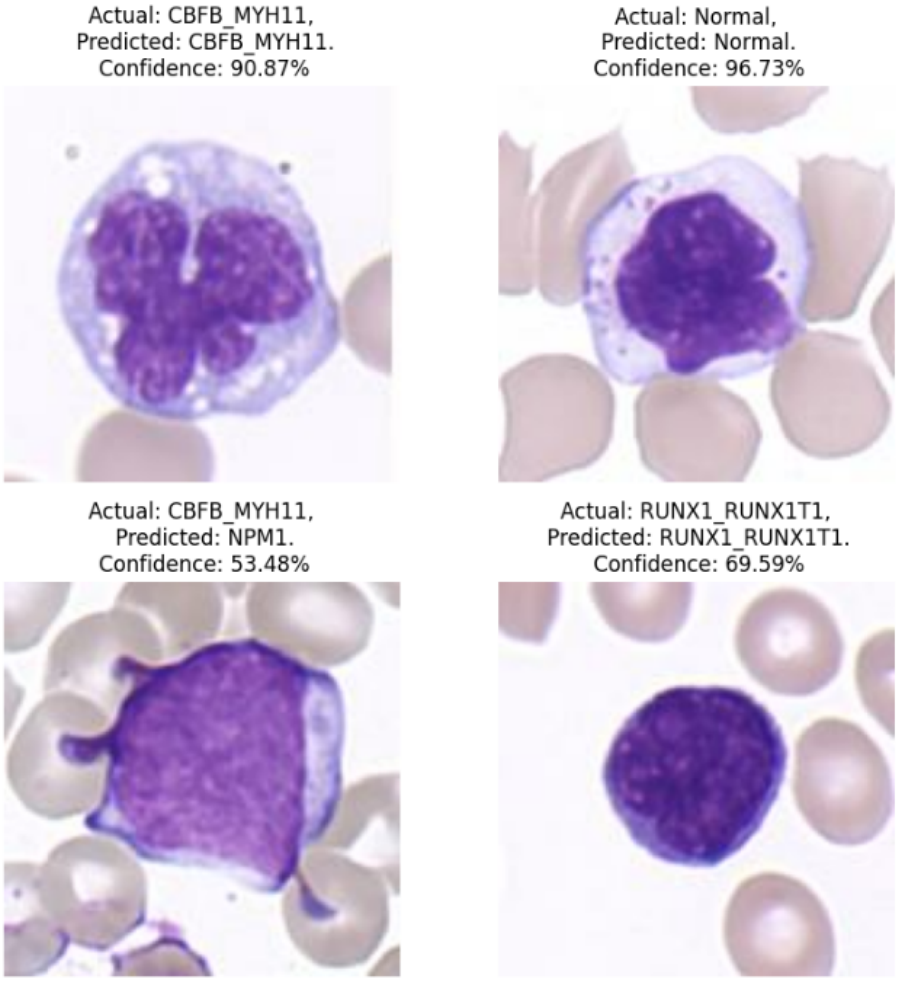

Figure 4.21: Sample Inference of MaxViT model with confidence levels

ViT

The ViT (Vision Transformer), a model bridging transformer structures with the adaptation of image classification tasks, documents the results of accuracy equal to 0.5189, F1 score equal to 0.5195, recall equal to 0.5189, and precision equal to 0.5269. Even though its attention hotspots enabled feature modeling reasonably well, it performed less effectively than other transformer-based models, such as MaxViT. Figures 4.22, 4.23, and 4.24, depict performance in terms of confusion metric, loss ratio, and delegated assistance diagrams.

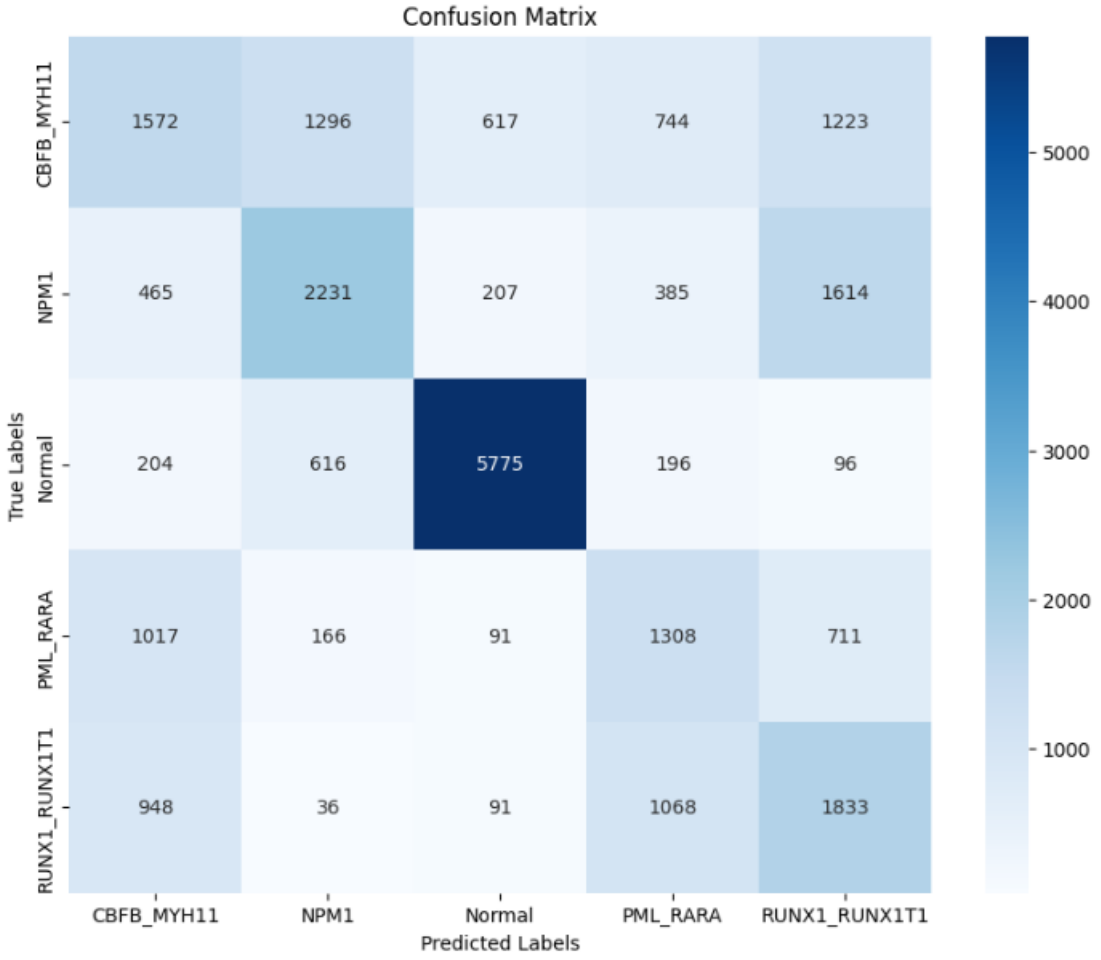

Figure 4.22: Confusion Matrix of ViT model

Figure 4.23: Training, Validation Accuracy and Loss of ViT model

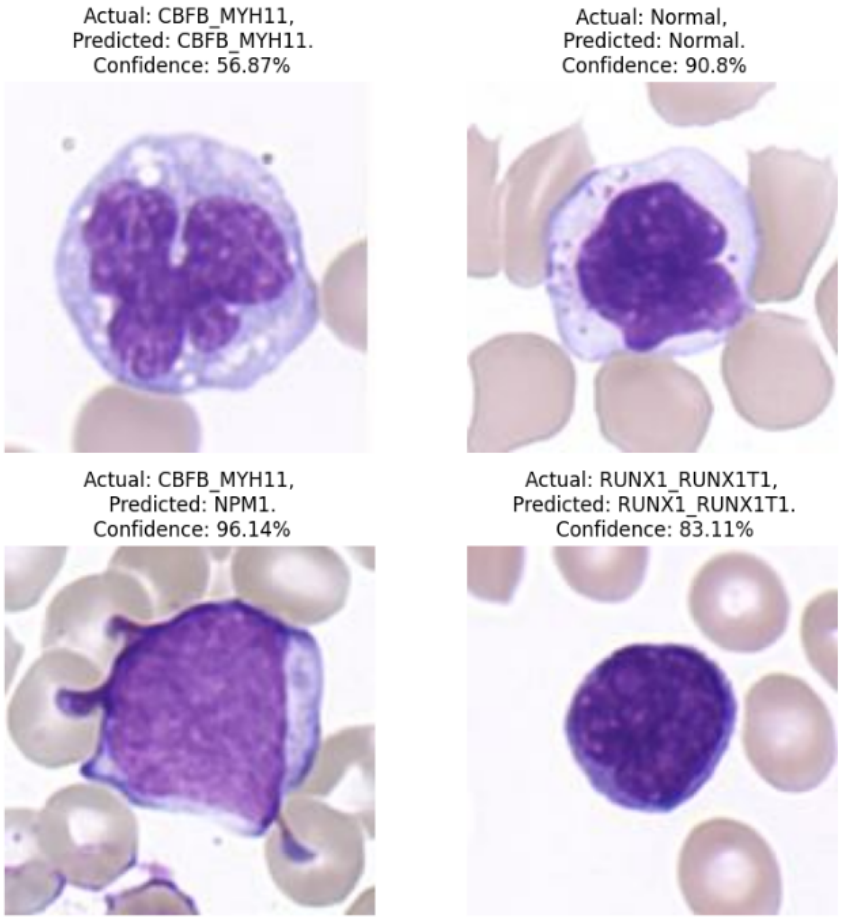

Figure 4.24: Sample Inference of ViT model with confidence levels

Hybrid Model(ResNet50+ViT)

An exception was the performance of the Hybrid model in which features of ResNet50's convolution were integrated with ViT's attention-based mechanisms. It showed an accuracy of 0.9287, an F1 score equal to 0.9285, a recall equal to 0.9287, and a precision equal to 0.9285 indicating its capacity to undertake feature integration. This unprecedented performance further accentuates its capability in terms of generalizing and learning complex data. The Confusion Matrix and the sample inference along with the confidence score have been given in Figures 4.25 and 4.26. A comparison has been shown in Table 4.1 for the combined performance metrics for different models.

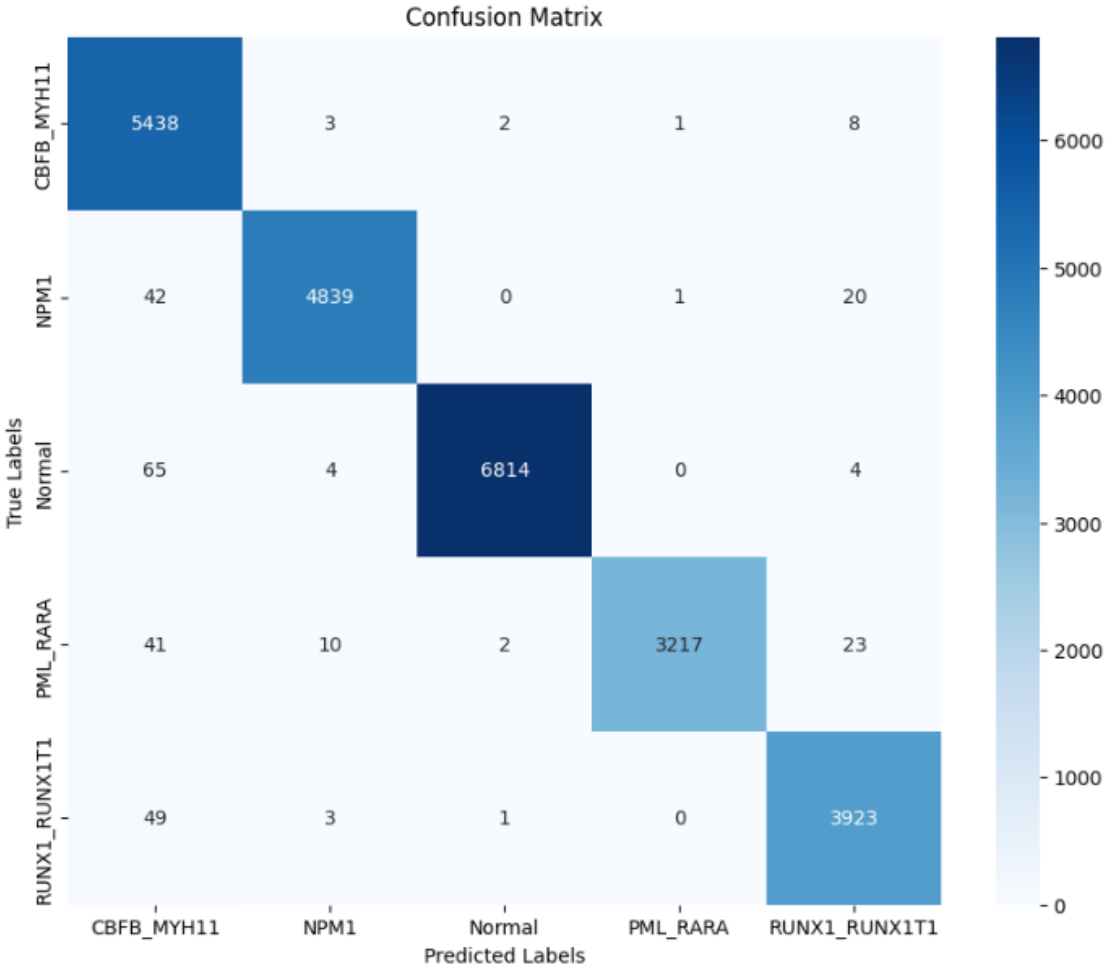

Figure 4.25: Confusion Matrix of Hybrid model

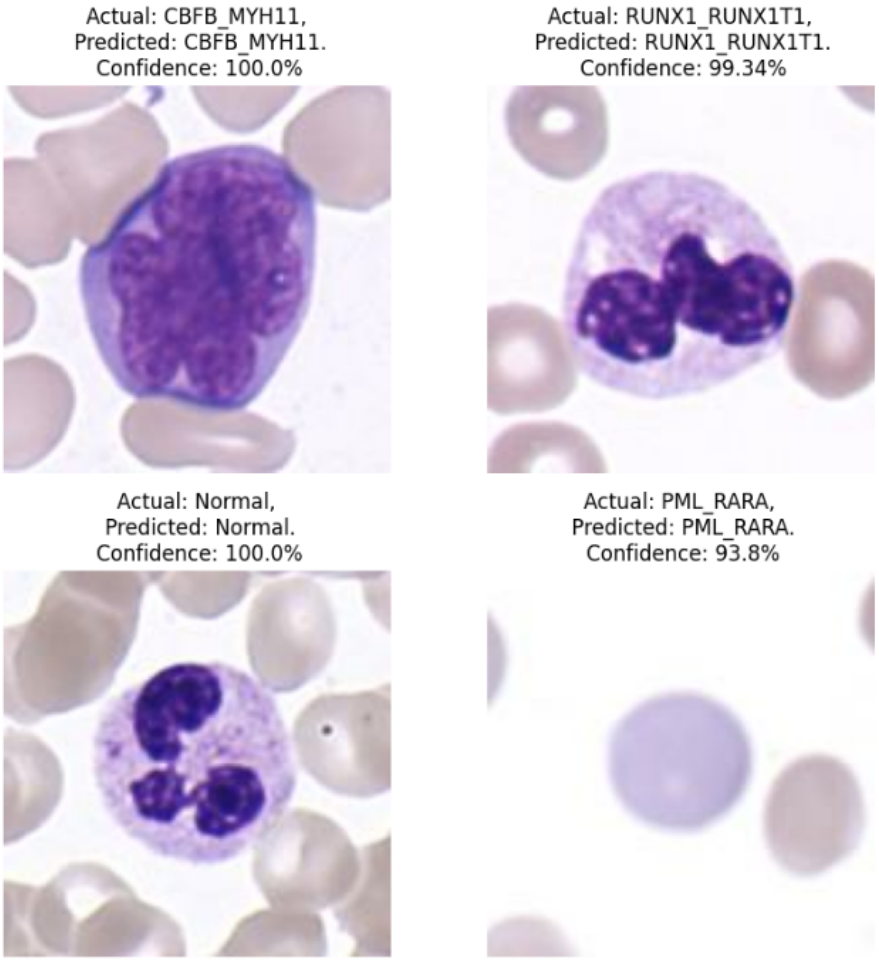

Figure 4.26: Sample Inference of Hybrid model with confidence levels

Model	Accuracy	F1 Score	Recall	Precision	Total Parameters	Space(MB)
Xception	0.4084	0.3949	0.4084	0.4061	21124397	80.58
VGG	0.4969	0.4893	0.4969	0.4867	14780997	56.39
ResNet	0.5461	0.5369	0.5461	0.5356	23850629	90.98
Inception	0.3485	0.2954	0.3485	0.3893	22065701	84.17
DenseNet	0.4778	0.4628	0.4778	0.4778	7169349	27.35
DaViT	0.5315	0.5320	0.5315	0.5335	14272517	54.45
MaxViT	0.5319	0.5294	0.5319	0.5285	217349	0.849
ViT	0.5189	0.5195	0.5189	0.5269	2739909	10.45
Hybrid	0.9287	0.9285	0.9287	0.9285	24376465	92.99

Table 4.1: Comparison table with combined performance metrics for different models.

4.2 Discussion and Key Findings

4.2.1 Performance Analysis of Deep Learning Models for Leukemia Detection

The analysis of the performance of deep learning models created for leukemia detection highlights the different efficacy and potential of such complex architectures on intricate healthcare datasets. This discussion further extends what was reported in the results section with emphasis on modeling strengths and patterns of the models across the hybrid model's performance that were indicated as the best.

As the results suggest, models of convoluted neural networks (CNN), for instance, Xception, VGG, ResNet, DenseNet, hybrid transformer models or DaViT, MaxViT, and ViT display varying capabilities in their ability to extract features and classify. CNN models focus on patterns or features in localized portions of an image which gradually leads to the formation of a deeper understanding of the whole image while transformers allow for a broader view of the picture in question. This is also the reason why this integration works, as the resulting structure was able to outperform all others: ResNet50-ViT Hybrid.

4.2.2 Performance of Convolutional Neural Networks (CNNs)

Xception

The Xception model uses depthwise separable convolutions to show the importance of computational resource savings. However, it is noted that the performance metrics report relatively low numbers, like an accuracy of only 40.84%, which raises concerns regarding adaptation to the complexity of the dataset. The limited gener-

alization may be a function of the design which is tailored to prioritize efficiency over the complexities of intricate pattern capturing. All the same, the Xception model sets the stage for a better appreciation of simpler models' trade-offs in performance and effectiveness.

VGG

Similar insights into the effects of increasing the depth of networks on feature representation extractive capacity are provided by deep networks like VGG. Though it was much better than Xception with its accuracy of 49.69%, its effectiveness was plateauing due to a lack of rudimentary residual connections or complex structures. Understandings gained from the evaluation of VGG model results emphasize that it is critical to address the potential of vanishing gradients and overfitting in deep networks by using more advanced solutions.

ResNet

With its new design of residual connections, ResNet was able to surpass VGG by a substantial margin with its accuracy being 54.61%. The ease of formulating and preserving stable gradients to enable multi-layer representation learning reveals some of the benefits of residual learning in traditional convolutional neural networks. This performance places ResNet as one of the most dependable architectures for feature extraction from medical datasets, which is consistent with its ability to manage sophisticated data.

Inception

The Inception model, despite convolving at multiple scales, achieved the lowest accuracy of 34.85% amongst the CNNs. Such results could be a result of the set construction not being in harmony with the building principles of the structure. The lower recall and F1 recall further highlight the concern that generalization was difficult, which suggests that there is a need for more specific optimization strategies during the deployment of such models in certain areas.

DenseNet

DenseNet which achieved moderate performance due to its feature reuse characteristic based on dense connectivity attained 47.78% accuracy. This model performed particularly well in general propagation, the behavior of the model's performance vis-a-vis the computational cost was pleasing. Nevertheless, their performance indicates that there is a need for more refinement to perform at par with more sophisticated architectures.

4.2.3 Performance of Transformer-Based Models

DaViT and MaxViT

The DaViT, MaxViT, and ViT, all transformer-based models have registered good results, with DaViT and MaxViT achieving comparable performance rates of 53.15% and 53.19%, respectively, following this trend. The merit of these models is that

they can focus on the entire data in a specific manner, making them superior to traditional CNNs. The results confirm with clear evidence the feasibility of using transformer models for medical imaging tasks, especially for datasets with diverse and intricate pattern formations.

Vision Transformer (ViT)

This was the case for the Vision Transformer (ViT) with an accuracy of 51.89% which succeeded to illustrate its potential in visual-spatial orientation modeling. On the other hand, the fact that it performed slightly lower than other transformers such as Max ViT, points out the benefits brought by some of the features of transformers architecture such as multi-axis attention in beefing up accuracy levels of classification.

4.2.4 Hybrid Model

ResNet50-ViT Hybrid

A Hybrid model that concatenates ResNet50 and ViT emerged as the highest-achieving model in this study, with an impressive performance rate of 92.87%. This combination of convolutional and attention-based mechanisms showcases the strength of the fusion of diverse architectures. The limitations of standalone architectures were solved by the Hybrid model, which made use of ResNet50's ability to extract local features with ViT's attention to a global perspective. It's impressive that it also has a high precision score and recall score on balance, this further enhances its trustworthiness in medical diagnosis as it demonstrates the ability to manage sensitivity and specificity.

The performance of the model is boosted even more with the incorporation of the Difference Enhancement-Random Sampling (DERS) technique. DERS is targeted at solving the general problem of class imbalance in a medical dataset by intensifying the contrast of a minority class of features and randomly sampling diverse augmented samples, thus ensuring exposure of the hybrid model to a more balanced and representative dataset, thus enhancing its generalizability and robustness to data variations. The results look pretty good for this hybrid approach. They have established an accuracy of 92.87%, which comfortably outperforms the other models tested in that study which are CNN- or transformer-based.

The hybrid ResNet50-ViT model was designed initially for leukemia classification. However, this architecture has its properties that can support segmentation, with appropriate modifications. ResNet50 extracts localized hierarchical feature maps from input images that mainly account for a spatial feature suitable for locating and isolating regions of interest, such as leukemia cells. These feature maps must be further exploited to perform pixel-level segmentation, where more layers such as upsampling or decoder modules can be appended to the existing network. On the other side, Vision Transformers (ViT) emphasize the global perspective, capturing long-range dependency and relation across the full image with self-attention mechanisms. This global context allows the model to learn spatial relationships in different areas, which in turn could help in feature-based separation between the cell bodies

and boundaries. In sum, the design of our hybrid architecture with local feature extractions from ResNet50 and the attention to global spatial relationships from ViT makes the architecture itself highly good for any segmentation task, though it would have to be adjusted with a segmentation head or decoder parts, say in a U-Net or an encoder-decoder framework.

The use of these models in the diagnosis of leukemia has great prospects for application in clinical settings. The ability to conduct automated diagnosis can help enhance the precision and effectiveness of act hematology measures in the sense that it would allow for proactive and individualized interventions. The Hybrid model has shown great results and such results compel integration of such models into diagnostic systems which is likely to change the face of care in hematology.

In addition, the different results for the performance of the evaluated models serve to emphasize the need for particular model choices and modifications in the tasks of medical imaging. As effective CNNs are adequate, combining advanced attention mechanisms and hybrid networks is beneficial for sophisticated data. This work presents the positive effects that could be achieved by coupling various advanced techniques provided that the end clinical requirements for the solution are well defined.

In summary, the findings illustrate the evolving capabilities of deep learning models in medical diagnostics. The integration of advanced architectures like ResNet and ViT into hybrid models represents a significant step forward in leveraging AI for healthcare applications. By enabling accurate and efficient leukemia diagnosis, these models pave the way for enhanced patient outcomes and the broader adoption of AI-driven solutions in medicine.

4.2.5 Limitations of the Hybrid Model

Despite the promising results, certain limitations must be acknowledged.

- The study's dataset size and diversity were constrained, potentially limiting the generalizability of the findings across broader clinical applications.
- The computational complexity of transformer-based and hybrid models may pose challenges for deployment in resource-limited settings.
- The efficiency of the model depends on high quality and well-illustrated images therefore, the presence of inconsistency or noise in diagnostic images may affect the accuracy.
- Real-world testing in clinical environments is necessary for validation of its stability and consistency.

Chapter 5

Conclusion

5.1 Future Work

Future research should address these issues by exploring the following :

- Refining the components of the CNN-transformer model, such as building an architecture using more models with different features could improve optimization techniques and perform better
- Working with a larger and more diverse dataset with the inclusion of rare leukemia variants and hematologic abnormalities could enhance the model to work better.
- Addition of patient information, genetic markers, and clinical data can improve diagnosis with better accuracy.
- Real-world testing should be done to validate and enhance the applicability of these models in clinical practice.
- Optimization of the model for working efficiently in low-resource places can help to diagnose leukemia in underdeveloped and remote regions.

Incorporating these practices can help the hybrid model to work better with enhanced performance and higher accuracy for detection of leukemia on blood cells.

5.2 Conclusion

The most accurate model for leukemia diagnosis was sought; at the same time, the abilities and the limitations of each of the tested architectures including CNN were analyzed. As the results showed, typical CNN architectures comprising Xception, VGG, ResNet, Inception, and DenseNet have different levels of success in feature extraction and classification of medical images. Out of the CNNs, ResNet got the best results among the three groups of participants because of its residual connections that enhance gradient flow and facilitate deeper feature learning.

Models based on transformers like DaViT, MaxViT, and ViT managed to greatly improve the performance of medical image processing due to their ability to identify complex structures in the data quickly. These models stress the contribution of

attention mechanisms to model performance and MaxViT was arguably the best solo transformer model.

The hybrid model ResNet50-ViT turned out to be the best, as it was able to balance its precision and recall and obtain a very high accuracy. This is a result of the hybrid models combining local and global attention as one of the features of ViT is attention. The hybrid model did not only perform well but it was one of the best components for leukemia diagnostics. These results again demonstrate the effectiveness of hybrid models in extending the field of medical diagnostics including complex diseases such as leukemia.

The clinical consequences of this work are significant. The inclusion of such models in diagnostic algorithms can completely change the workflows in hematology, as the assessment of leukemia will be more accurate, prompt, and cost-effective. Reducing reliance on human interpretation and facilitating faster decision intervention can improve patient prognoses significantly. In addition, the work paves the way for further work on the implementation of AI technologies into medicine, spurring initiatives on effective and efficient diagnostic solutions.

Bibliography

[1] H. R. Gralnick, D. A. G. Galton, D. Catovsky, C. Sultan, and J. M. Bennett, "Classification of acute leukemia," *Ann. Intern. Med.*, vol. 87, no. 6, pp. 740–753, Dec. 1977. DOI: 10.7326/0003-4819-87-6-740.

[2] P. Melville, N. Shah, L. Mihalkova, and R. J. Mooney, "Experiments on ensembles with missing and noisy data," in *Multiple Classifier Systems*, F. Roli, J. Kittler, and T. Windeatt, Eds., Berlin, Heidelberg: Springer, 2004, pp. 293–302. DOI: 10.1007/978-3-540-25966-4_29.

[3] W. Shitong and W. Min, "A new detection algorithm (nda) based on fuzzy cellular neural networks for white blood cell detection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 5–10, Jan. 2006. DOI: 10.1109/TITB.2005.855545.

[4] S.-H. Hsieh, Z. Wang, P.-H. Cheng, I.-S. Lee, S.-L. Hsieh, and F. Lai, "Leukemia cancer classification based on support vector machine," in *2010 8th IEEE International Conference on Industrial Informatics*, Jul. 2010, pp. 819–824. DOI: 10.1109/INDIN.2010.5549638.

[5] M. Roy, G. Jin, D. Seo, M.-H. Nam, and S. Seo, "A simple and low-cost device performing blood cell counting based on lens-free shadow imaging technique," *Sens. Actuators B Chem.*, vol. 201, pp. 321–328, Oct. 2014. DOI: 10.1016/j.snb.2014.05.011.

[6] E. Suryani, W. Wiharto, and N. Polvonov, "Identification and counting white blood cells and red blood cells using image processing case study of leukemia," *arXiv*, vol. arXiv:1511.04934, Nov. 2015. DOI: 10.48550/arXiv.1511.04934.

[7] K. A. S. A. Daqqa, A. Y. A. Maghari, and W. F. M. A. Sarraj, "Prediction and diagnosis of leukemia using classification algorithms," in *2017 8th International Conference on Information Technology (ICIT)*, May 2017, pp. 638–643. DOI: 10.1109/ICITECH.2017.8079919.

[8] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, S36–S40, Apr. 2017. DOI: 10.1016/j.metabol.2017.01.011.

[9] A. S. Negm, O. A. Hassan, and A. H. Kandil, "A decision support system for acute leukemia classification based on digital microscopic images," *Alexandria Engineering Journal*, vol. 57, pp. 2319–2332, 2018. DOI: 10.1016/j.aej.2017.08.025.

[10] S. Rajpurohit, S. Patil, N. Choudhary, S. Gavasane, and P. Kosamkar, "Identification of acute lymphoblastic leukemia in microscopic blood image using image processing and machine learning algorithms," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2018, pp. 2359–2363. DOI: 10.1109/ICACCI.2018.8554576.

[11] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technol. Cancer Res. Treat.*, vol. 17, p. 1 533 033 818 802 789, Jan. 2018. DOI: 10.1177/1533033818802789.

[12] T. Tran, J.-H. Park, O.-H. Kwon, K.-S. Moon, S.-H. Lee, and K.-R. Kwon, "Classification of leukemia disease in peripheral blood cell images using convolutional neural network," *Journal of Korea Multimedia Society*, vol. 21, no. 10, pp. 1150–1161, 2018. DOI: 10.9717/kmms.2018.21.10.1150.

[13] M. I. Daoud, S. Abdel-Rahman, and R. Alazrai, "Breast ultrasound image classification using a pre-trained convolutional neural network," in *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Dec. 2019, pp. 167–171. DOI: 10.1109/SITIS.2019.00037.

[14] S. Hariprasath, T. Dharani, and B. N. S. Mohammad, *Automated detection of acute lymphocytic leukemia using blast cell morphological features*, Social Science Research Network, Apr. 2019. DOI: 10.2139/ssrn.3372907.

[15] I. Ahmad, Y. Liu, D. Javeed, and S. Ahmad, "A decision-making technique for solving order allocation problem using a genetic algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 853, no. 1, p. 012 054, May 2020. DOI: 10.1088/1757-899X/853/1/012054.

[16] Z. G. et al., "Learning physical properties in complex visual scenes: An intelligent machine for perceiving blood flow dynamics from static ct angiography imaging," *Neural Netw.*, vol. 123, pp. 82–93, Mar. 2020. DOI: 10.1016/j.neunet.2019.11.017.

[17] Z. G. et al., "Privileged modality distillation for vessel border detection in intracoronary imaging," *IEEE Trans. Med. Imaging*, vol. 39, no. 5, pp. 1524–1534, May 2020. DOI: 10.1109/TMI.2019.2952939.

[18] A. Bodzas, P. Kodytek, and J. Zidek, "Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception," *Front. Bioeng. Biotechnol.*, vol. 8, Aug. 2020. DOI: 10.3389/fbioe.2020.01005.

[19] P. H. Kasani, S.-W. Park, and J.-W. Jang, "An aggregated-based deep learning method for leukemic b-lymphoblast classification," *Diagnostics*, vol. 10, no. 12, p. 12, Dec. 2020. DOI: 10.3390/diagnostics10121064.

[20] M. Loey, M. Naman, and H. Zayed, "Deep transfer learning in diagnosing leukemia in blood cells," *Computers*, vol. 9, no. 2, p. 29, Apr. 2020. DOI: 10.3390/computers9020029.

[21] A.-A. Nahid, N. Sikder, A. K. Bairagi, *et al.*, "A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neural network," *Sensors*, vol. 20, no. 12, p. 3482, 2020. DOI: 10.3390/s20123482.

[22] A. T. Sahlol, P. Kollmannsberger, and A. A. Ewees, "Efficient classification of white blood cell leukemia with improved swarm optimization of deep features," *Sci. Rep.*, vol. 10, no. 1, p. 2536, Feb. 2020. DOI: 10.1038/s41598-020-59215-9.

[23] C. Xu, L. Xu, P. Ohorodnyk, M. Roth, B. Chen, and S. Li, "Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal gans," *Med. Image Anal.*, vol. 62, p. 101 668, May 2020. DOI: 10.1016/j.media.2020.101668.

[24] M. Z. U. et al., "An attention-based convolutional neural network for acute lymphoblastic leukemia classification," *Appl. Sci.*, vol. 11, no. 22, p. 22, Jan. 2021. DOI: 10.3390/app112210662.

[25] S. Alagu, A. P. N., G. Kavitha, and B. B. K. Bhoopathy, "Automatic detection of acute lymphoblastic leukemia using unet based segmentation and statistical analysis of fused deep features," *Applied Artificial Intelligence*, vol. 35, no. 15, pp. 1952–1969, 2021. DOI: 10.1080/08839514.2021.1995974.

[26] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis, and F. Scotti, "Acute lymphoblastic leukemia detection based on adaptive unsharpening and deep learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 1205–1209. DOI: 10.1109/ICASSP39728.2021.9414362.

[27] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis, and F. Scotti, "Histopathological transfer learning for acute lymphoblastic leukemia detection," in *2021 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Jun. 2021, pp. 1–6. DOI: 10.1109/CIVEMSA52099.2021.9493677.

[28] D. Javeed, T. Gao, and M. T. Khan, "Sdn-enabled hybrid dl-driven framework for the detection of emerging cyber threats in iot," *Electronics*, vol. 10, no. 8, p. 918, Apr. 2021. DOI: 10.3390/electronics10080918.

[29] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for diagnosis of acute lymphoblastic leukemia based on vit-cnn ensemble model," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, p. 7 529 893, 2021. DOI: 10.1155/2021/7529893.

[30] S. Rezayi, N. Mohammadzadeh, H. Bouraghi, S. Saeedi, and A. Mohammadpour, "Timely diagnosis of acute lymphoblastic leukemia using artificial intelligence-oriented deep learning methods," *Computational Intelligence and Neuroscience*, vol. 2021, p. 5 478 157, 2021. DOI: 10.1155/2021/5478157.

[31] A. S. et al., "A deep learning-based approach for the diagnosis of acute lymphoblastic leukemia," *Electronics*, vol. 11, no. 19, Jan. 2022. DOI: 10.3390/electronics11193168.

[32] Y.-M. Chen, F.-I. Chou, W.-H. Ho, and J.-T. Tsai, "Classifying microscopic images as acute lymphoblastic leukemia by resnet ensemble model and taguchi method," *BMC Bioinformatics*, vol. 22, no. 5, p. 615, Jan. 2022. DOI: 10.1186/s12859-022-04558-5.

[33] M. Ghaderzadeh, A. Hosseini, F. Asadi, H. Abolghasemi, D. Bashash, and A. Roshanpoor, "Automated detection model in classification of b-lymphoblast cells from normal b-lymphoid precursors in blood smear microscopic images based on the majority voting technique," *Sci. Program.*, vol. 2022, no. 1, p. 4 801 671, 2022. DOI: 10.1155/2022/4801671.

[34] D. Javeed, T. Gao, M. T. Khan, and D. Shoukat, "A hybrid intelligent framework to combat sophisticated threats in secure industries," *Sensors*, vol. 22, no. 4, p. 1582, Feb. 2022. DOI: 10.3390/s22041582.

[35] M. A. Razib, D. Javeed, M. T. Khan, R. Alkanhel, and M. S. A. Muthanna, "Cyber threats detection in smart environments using sdn-enabled dnn-lstm hybrid framework," *IEEE Access*, vol. 10, pp. 53 015–53 026, 2022. DOI: 10.1109/ACCESS.2022.3172304.

[36] A. R. Revanda, C. Fatichah, and N. Suciati, "Classification of acute lymphoblastic leukemia on white blood cell microscopy images based on instance segmentation using mask r-cnn," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 5, pp. 625–634, 2022. DOI: 10.22266/ijies2022.1031.54.

[37] M. T. H. K. Tusar and R. K. Anik, *Automated detection of acute lymphoblastic leukemia subtypes from microscopic blood smear images using deep neural networks*, Undergraduate Project, City University, Dhaka, Bangladesh, 2022. DOI: https://doi.org/10.48550/arXiv.2208.08992.

[38] P. K. Mallick, S. K. Mohapatra, G.-S. Chae, and M. N. Mohanty, "Convergent learning–based model for leukemia classification from gene expression," *Personal and Ubiquitous Computing*, vol. 27, pp. 1103–1110, 2023, Published online 16 October 2020. DOI: 10.1007/s00779-020-01467-3.

[39] R. F. Mulya, E. Utami, and D. Ariatmanto, "Classification of acute lymphoblastic leukemia based on white blood cell images using inceptionv3 model," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 4, pp. 947–952, 2023. DOI: 10.29207/resti.v7i4.5182.

[40] Z. Wang, T. Li, J.-Q. Zheng, and B. Huang, "When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation," in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., Springer Nature Switzerland, 2023, pp. 424–441. DOI: 10.1007/978-3-031-25082-8_28.

[41] A. Badruzzaman and A. M. Arymurthy, "A comparative study of convolutional neural network in detecting blast cells for diagnose acute myeloid leukemia," *J. Electron. Electromed. Eng. Med. Inform.*, vol. 6, no. 1, pp. 84–91, Jan. 2024. DOI: 10.35882/jeemi.v6i1.354.

[42] R. U. Khan, S. Almakdi, M. Alshehri, A. U. Haq, A. Ullah, and R. Kumar, "An intelligent neural network model to detect red blood cells for various blood structure classification in microscopic medical images," *Heliyon*, vol. 10, e26149, 2024, Available online 13 February 2024. DOI: 10.1016/j.heliyon.2024.e26149.

[43] T. C. I. A. (TCIA), *Aml cytomorphology dataset - munich leukemia laboratory (mll)*, Accessed: 2025-01-29, 2025. [Online]. Available: https://www.cancerimagingarchive.net/collection/aml-cytomorphology_mll_helmholtz/.