Kalpa Publications in Computing
Volume 19, 2024, Pages 475–487
Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

A Systematic Review on Pre-Trained Models on C-NMC Leukemia Using Deep Learning Techniques

Aniketh Vijesh[1], Kirti Sikka[2], and Remya S[3]

[1] Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amritapuri, Kerala, India
anikethvij464@gmail.com
[2] Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amritapuri, Kerala, India
kirtisikka972@gmail.com
[3] Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amritapuri, Kerala, India
remyas@am.amrita.edu

Abstract

The early detection of Acute Lymphoblastic Leukemia (ALL) poses a significant challenge in the medical field due to the subtle morphological features of ALL cells, which often resemble healthy cells. This necessitates the expertise of experienced hematologists, a reliance on human interpretation that introduces subjectivity and labor-intensive processes. Consequently, timely diagnosis and treatment initiation can be hindered. By leveraging the capabilities of machine learning,the paper aim to establish a system that can accurately distinguish between healthy and ALL cells, thereby reducing the reliance on subjective human interpretation and expediting the diagnostic process. This systematic review thoroughly examines the use of deep learning for classifying and detecting acute leukemia. This study discusses many stages such as preprocessing, augmentation, segmentation, and feature extraction that are taken before classification. It also addresses the issues faced by the authors in different datasets. This research study examined and compared several benchmark models VGG16, VGG19, Inception, Xception, Efficient NetB0, ResNet50, and ResNet101. Out of these models, ResNet101 came as the top performer with a Validation Accuracy of 76.36%, Validation Precision of 75.85%, and Validation Recall of 76.36%. This comparative analysis aims to elucidate the strengths and weaknesses of these models, contributing valuable insights.

1 Introduction

Blood primarily consists of three types of cells: Red Blood Cells (RBC), White Blood Cells (WBC), and platelets. RBC plays a crucial role in oxygen transportation from the heart to all tissues and in removing carbon dioxide. They constitute 50% of the entire blood volume.

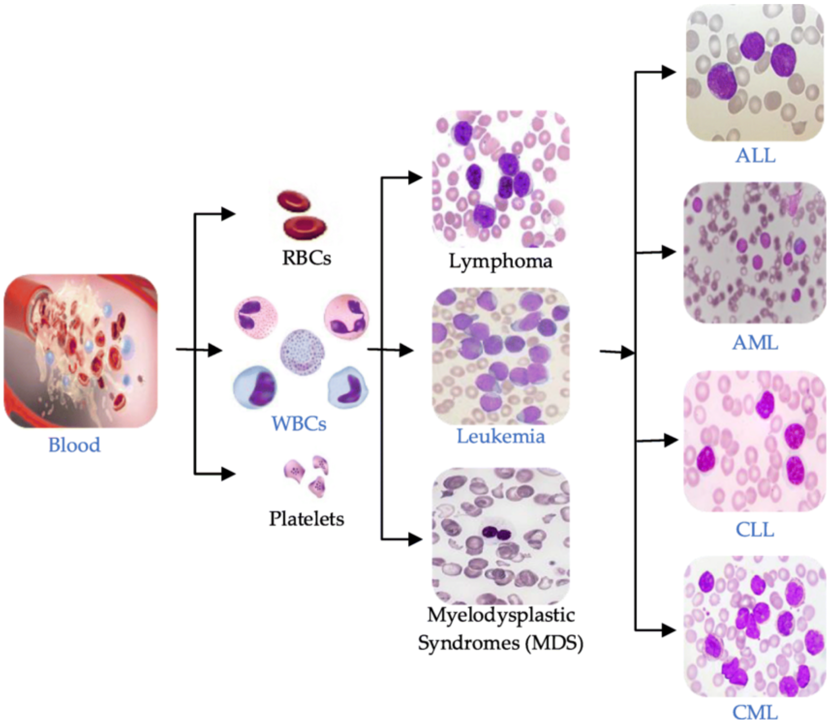

Figure 1: Types of Leukemia [?]

White blood cells have the crucial role of protecting the body from infections and illnesses. Leukaemia is a form of blood cancer that starts in the bone marrow. It is defined by an abnormal proliferation of blood cells known as blast or leukemia cells [?, ?, ?, ?]. These concerns include bleeding, bone discomfort, weariness, fever, and increased vulnerability to infections resulting from a lack of normal blood cells.

Leukemia is broadly classified into four main types based on the type of white blood cell affected and the rate of progression: Acute Lymphoblastic Leukemia(ALL), Acute Myeloid Leukemia(AML), Chronic Lymphoblastic Leukemia(CLL), and Chronic Myeloid Leukemia(CML) [?] as shown in Figure 1.

Acute Lymphoblastic Leukemia(ALL) is a common cancer disease that affects mostly children below the age of fifteen. The main reason is the rapid production of immature white blood cells, which decrease the number of good blood cells and platelets, making it harder for the body to fight against infections and diseases. As the name implies, this disease is ”acute”, that is it advances quickly and rapidly [?]. This may result in the disease spreading to different body parts like the lymph nodes, spleen, liver, brain, spinal cord, etc.

Due to the urgency and importance of early-stage detection, it is paramount that the detection is done in a precise and cost-efficient manner. This has proven to be a significant obstacle. Most procedures available in diagnostic laboratories are time-consuming and based on the experience of the hematologists [?, ?, ?]. To address these challenges, this research focuses on harnessing the capabilities of machine learning to develop a system capable of precise and efficient detection and classification of ALL cells.

This paper presents a comprehensive comparative analysis of seven distinct machine learning models, namely VGG16, VGG19, Inception, Xception, Efficient NetB0, ResNet101, and ResNet50. By examining the strengths and weaknesses of these models, it aims to contribute valuable insights into the development of an automated system for ALL detection. The objective is to reduce reliance on subjective human interpretation, expedite the diagnostic process, and ultimately improve outcomes for individuals affected by Acute Lymphoblastic Leukemia. The following sections provide an in-depth exploration of each model's performance and its potential applications in the context of leukemia detection.

The experiments and models were trained in a Python environment, with the Python version being 3.10.12. The Keras library and TensorFlow libraries were used to aid in training the models. Other libraries used were, numpy, scikit-learn, seaborn, etc.

The critical need for more effective and precise techniques for the early identification of acute lymphoblastic leukemia (ALL) is the driving force behind this study. The hematologist's knowledge is a major component of the current diagnostic processes, which can be labor-intensive and subjective, delaying diagnosis and treatment commencement. Furthermore, ALL cells and healthy cells have subtle physical similarities, making accurate identification even more difficult and emphasizing the need for cutting-edge technology solutions. This research intends to build a systematic approach that can reliably discriminate between healthy and ALL cells by utilizing the capabilities of machine learning, especially deep learning techniques. By facilitating prompt intervention, this technology would greatly decrease reliance on human interpretation, expedite the diagnostic procedure, and eventually enhance patient outcomes. Additionally, this study aims to thoroughly assess and maximize the efficacy of these methods in the context of acute leukemia detection by investigating different phases of deep learning, such as preprocessing, augmentation, segmentation, and feature extraction. The overall purpose of the research is to solve the significant obstacles related to ALL diagnoses, with the ultimate goal of improving patient care, diagnostic efficiency, and accuracy through creative technical solutions.

The subsequent sections of this paper will delve deeper into the existing literature related to leukemia detection and classification in Section 2, followed by a detailed exploration of the proposed methodology in Section 3. Section 4 will present the results obtained from the comparative analysis of machine learning models and draw conclusions based on their performance.

2 Literature Review

Extensive research efforts have been dedicated to developing efficient solutions in the field of leukemia cell classification. Various teams of researchers have proposed different algorithms and models, each aiming to enhance the accuracy and effectiveness of leukemia detection.

2.1 Particle Swarm Optimization and Neural Networks Approach

Agustin Arif Sukorini et al. [**?**] proposed a methodology employing Particle Swarm algorithms and neural networks for leukemia cell classification. Their approach involves image preprocessing followed by feature extraction using swarm algorithms. Classification occurs in two stages: firstly, distinguishing between lymphoid and non-lymphoid cells, and secondly, classifying lymphoid cells into malignant or non-malignant categories. Their method achieved an accuracy of 86.92

2.2 Data Augmentation and Convolutional Neural Networks

By utilising data augmentation approaches, Talaat Gamerl et al. [?] solved challenges triggered by a smaller dataset. Convolutional Neural Networks (CNNs) with ReLU activation function were used to extract features, and then an attention module was used to extract more features. Impressive results were obtained for the final classification of cells into benign and malignant categories: precision of 99.97%, recall of 100%, F1-Score of 99.98%, and accuracy of 99.98%.

2.3 Transfer Learning and Swarm Optimization

Using transfer learning, Sahlol et al. [?] developed a reliable leukaemia diagnosis model. Using a Salp Swarm Optimisation algorithm for feature selection and recursive feature elimination for refining is their method. While extracting less features, the model attained an accuracy of 83.2%, which is comparable to well-known convolutional neural networks like ResNet.

2.4 Image Preprocessing and Machine Learning Classifiers

An automatic technique for identifying and categorizing white blood cells (WBCs) in microscopic images was presented by Putzu, et.al [?]. Acute lymphoblastic leukaemia (ALL), a form of blood cancer, can be found with this method in particular. Unlike other techniques, this one separates the WBC in its entirety at the outset, enabling a thorough examination of the cytoplasm and nucleus. The method first extracts properties like as texture, colour, and form, and then utilises machine learning models to identify the cells. High accuracy in categorising ALL cases (93%) and identifying WBCs (92%) is demonstrated by the data. With its potential to enhance speed, accuracy, and early illness diagnosis of disorders like ALL, this approach presents a promising solution for automated blood cell analysis.

2.5 Adaptive Image Processing and Deep Learning

Genovese Hosseini Piuri Plataniotis Scotti et al. [?] introduced a multi-step image processing and classification methodology. Their approach involves adaptive preprocessing techniques and classification using a pre-trained deep CNN. Qualitative and quantitative results demonstrate the effectiveness of their proposed algorithm, achieving a classification accuracy of 96.84%.

The literature presents various pre-processing methods and model architectures, each introducing innovative techniques to address leukemia cell classification. However, a comprehensive comparison of the performances of different pre-trained models is lacking. This paper aims to fill this gap by objectively evaluating popular pre-trained models, facilitating future research efforts to focus on the most effective models in leukemia detection.

3 Materials & Methods

To objectively evaluate the performances of popular pre-trained models for leukemia cell classification, a systematic approach was employed. The first step was compiling a large dataset that included a wide variety of leukemia cell images. Samples were taken from numerous sources, including the ALL-IDB1 and ALL-IDB2unsharp datasets. Preprocessing was done on the dataset to improve image quality and guarantee consistency for analysis. Then, certain well-known pre-trained models, such as VGG16, ResNet, and others, were selected and put to the test. Transfer learning methods were applied to the leukemia cell dataset to optimize each model [?]. Through extensive testing and cross-validation methods, performance parameters for each model were determined, including accuracy, precision, recall, and F1-score. Python programming and widely used deep learning frameworks were used to implement the entire procedure. This methodological approach ensures a robust and unbiased evaluation of pre-trained models, facilitating the identification of the most effective model for leukemia detection. The basic steps involved in the proposed methodology are described in Figure 2.

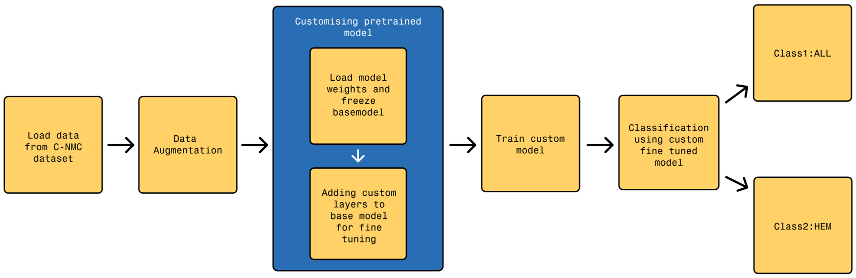
Load data from C-NMC dataset
Data Augmentation
Customising pretrained model
Load model weights and freeze basemodel
Adding custom layers to base model for fine tuning
Train custom model
Classification using custom fine tuned model
Class1:ALL
Class2:HEM

Figure 2: Proposed Methodology

3.1 Dataset

The initial step in the process is the collection of the images. The dataset used is the C-NMC dataset. The dataset used in this study is the publicly available C-NMC dataset. It comprises 15,135 images derived from 118 patients. These images have been meticulously segmented from microscopic samples, aiming to closely replicate real-world conditions by retaining staining noise and illumination errors. The Train Set comprises 73 subjects (47 ALL, 26 Normal) with 10,661 cell images (7,272 ALL, 3,389 Normal). The Preliminary Test Set includes 28 subjects (13 ALL, 15 Normal) with 1,867 cell images (1,219 ALL, 648 Normal). The Final Test Set has 17 subjects (9 ALL, 8 Normal) and 2,586 cell images. Figure 3 shows some samples of the images in the dataset which is being used to train the model.

3.2 Data Augmentation

When it comes to improving the resilience and generalization capacities of deep learning models used for acute leukaemia classification and detection, data augmentation is essential. Data augmentation is carried out to increase the diversity of the dataset and provide better and more comprehensive results by introducing controlled variations to the existing data points. This tactic thus expands the size and complexity of the dataset without requiring the collection of new data [**?**,**?**]. This enhances the capacity of the model to generalize to new data. By employing methods like rotation, translation, scaling, and flipping the enhanced dataset encompasses a wider range of morphological changes found in leukemia cells, allowing the model to acquire more thorough representations of the underlying characteristics. In the end, this augmentation strategy helps to develop more reliable and accurate classification models for the early detection of acute lymphoblastic leukemia (ALL). It does this by increasing the size and complexity of the dataset and making training easier by offering a richer set of training examples.

Utilizing Tensorflow's ImageDetector function, we specifically introduced horizontal flips to the training images. This augmentation technique enriches the training data by creating

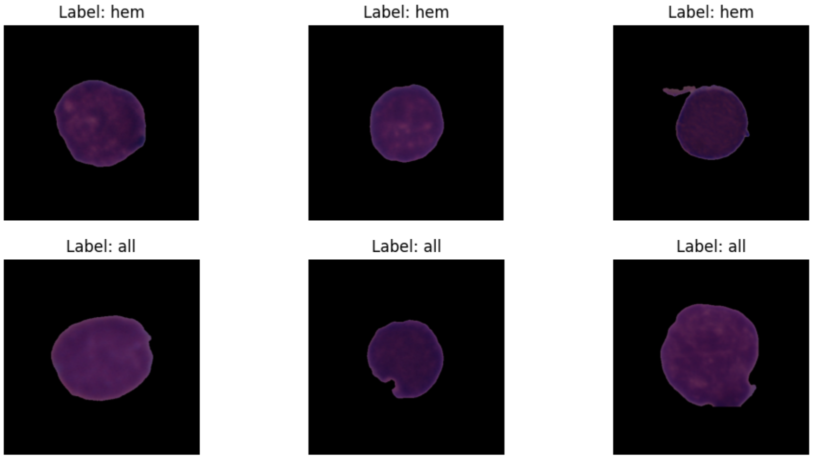

Figure 3: Samples from the C-NMC dataset

mirror versions of the original. This augmentation was chosen to ensure that no misleading information was introduced into the training set hence the more simplistic approach. The testing data remained unaltered. This decision was made to ensure an unbiased evaluation of the model's performance on unseen data, as it reflects the model's ability to generalize to data beyond the training set's specific augmentation to a better extent.

Furthermore, a batch size of 40 was selected for optimal training efficiency and stability. After careful experimentation with various values, 40 emerged as the sweet spot given in the dataset, computational resources, and setup. This configuration led to significantly faster convergence compared to the previous trails with batch sizes of 32 and 28. This selection ultimately resulted in accelerating the learning process without compromising performance.

3.3 Segmentation

One of the most important steps in the analytical pipeline is the segmentation of leukemia cells from surrounding tissues. The most recent image processing techniques were used to obtain reliable segmentation. Before that, the visual quality of the photographs was enhanced through the use of techniques like contrast stretching and histogram equalization. Afterward, thresholding methods were applied to isolate the foreground leukemia cells from the background, such as the triangle method or the Zack thresholding approach. Furthermore, the accuracy of cell segmentation was examined using sophisticated segmentation algorithms based on deep learning techniques, such as U-Net and Mask R-CNN. The dependability of the segmented regions for additional analysis was ensured by carefully verifying the segmentation process by manual annotation and comparison with ground truth labels. Strengthening the leukemia detection system overall, strong segmentation approaches allow classification and assessment of pre-trained models to be carried out on well-defined leukemia cells.

3.4 Feature Extraction

Leukaemia cell classification is facilitated by feature extraction, which derives discriminative information from segmented cell pictures [?, ?, ?]. With the use of the insights gleaned from the segmentation procedure, several feature extraction methodologies were explored to illustrate the intrinsic characteristics of leukemia cells. Both manually developed qualities and those based on deep learning were incorporated into these techniques. Statistical descriptors such as mean intensity and standard deviation were combined to create a texture that included Haralick texture features extracted from the segmented regions. Additionally, deep learning-based feature extraction methods like convolutional neural networks (CNNs) were used to extract automatically generated hierarchical representations straight from the segmented cell images. Models like ResNet and VGG16 were refined on the segmented leukemia cell dataset in order to extract high-level features that capture complex patterns and structures. By combining multiple feature extraction techniques, the classification system can differentiate between malignant and non-malignant cells, providing an accurate diagnosis and prognosis for leukemia.

4 Results and Discussion

4.0.1 Experimental Setup

Within a meticulously chosen computing environment, the training and experimentation procedures were carried out with precision. For experimentation, the Python 3 version is used. Model construction and training made heavy use of TensorFlow version 2.14.0, a state-of-the-art deep learning framework known for its effectiveness and scalability. The computational power required for training the intricate models was furnished by an NVIDIA RTX 3050 GPU, boasting 8GB of dedicated RAM. This high-performance GPU architecture expedited the training process, allowing for rapid iteration and optimization of model parameters. By harnessing this advanced computational infrastructure, the experimental procedures were conducted with utmost precision and efficiency, culminating in robust and reliable results.

4.1 Customizing the pre-trained models

4.1.1 Loading the model weights and freezing the model

The main objective of this study is to leverage state-of-the-art pre-trained models, which are established and trained for expansive datasets for general tasks, as the foundational framework for the classification of leukemia cells. This provides us with a robust starting point for this task.

As mentioned earlier, the main point of using a pre-trained model is the ability to use the knowledge acquired from the training. The process of freezing a model is done to preserve the already learned weights and the architecture used in the beginning.

4.1.2 Adding custom layers to the model

After freezing the model the main step is to add custom layers to facilitate more efficient binary classification. Three additional layers are added to the frozen model, A GloabalMaxPooling layer, a Dropout layer, and finally a Dense layer with the sigmoid activation function and two nodes, which act as the output layer. The pooling layer is used to flatten the spatial dimensions while retaining the most important features. The dropout layer is used to introduce regularisation of the data and 50% of the input units are dropped out. This helps in preventing overfitting. The output layer is where the final classification happens. The sigmoid function is used as it is specifically optimal for binary classification [?, ?, ?]. The sigmoid function also called the logistic function is a sigmoidal function that maps real values to values between 0 and 1. Mathematically, the formula is defined as follows:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

The graph for the function is shown in Figure 4.

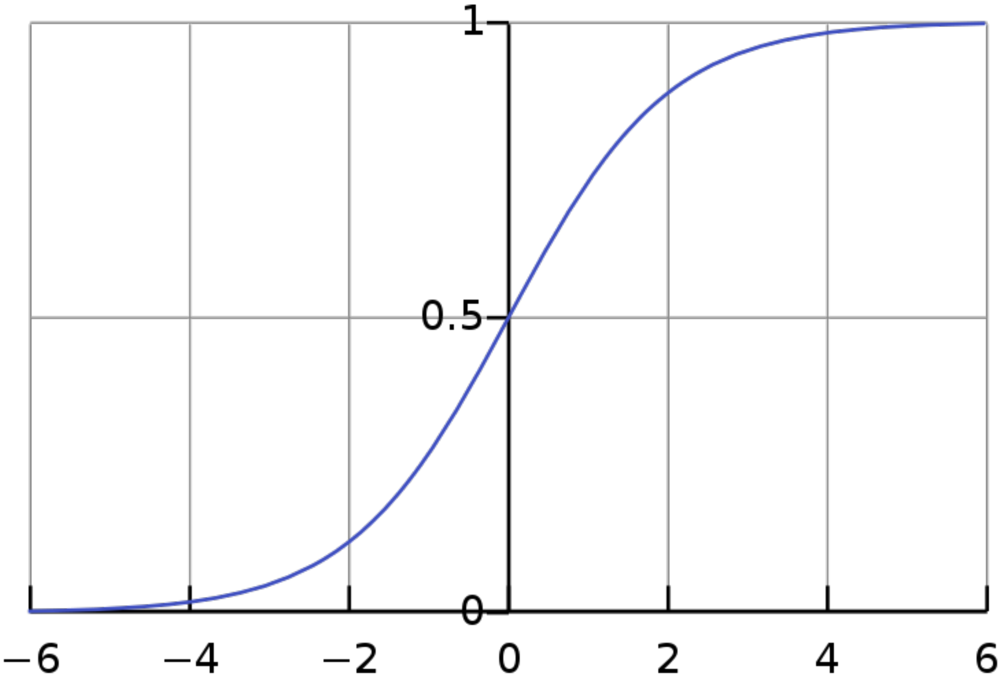

Figure 4: The Sigmoid function

As the real values get mapped to the range between 0 and 1, it is one of the best activation functions to be used in the output layer for a binary classification problem such as this.

Transfer learning, the pivotal strategy employed in this paper involves adapting the pre-trained models to the specific classification task of leukemia cell classification. Instead of starting the training process from scratch, transfer learning allows us to capitalize on the knowledge learned from a different but related domain. As an extension of the transfer learning process, fine-tuning is used to make the pre-trained model into a more specialized model for the new task [?].

4.2 Training the custom model

The custom model which is the pre-trained model to which the new layers were added, is trained on the augmented images. This process is called fine-tuning as the model is tuned to perform the new classification task. The training is done over 15 epochs. This decision was made as most models reached convergence by the 7th or 10th epoch. The Early Stopping method was employed to come to this conclusion. The patience value was set to 5. Early stopping is a regularization method that is used to reduce overfitting and decrease the training time [?, ?]. The optimization function used is Adamax with a learning rate of 0.001. Adamax was used for its adaptability to varying gradient magnitudes. Also studies related to classification problems have shown that the Adamax function is one of the best optimizer functions [?]. The learning rate was set to 0.001 through trial and error. The loss function used is binary cross-entropy, as it is the optimal and standard function to use for binary classification problems [?, ?],. The binary cross-entropy function can be defined as follows:

$$L(y, \hat{y}) = -\left(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right)$$

The model classifies the data into classes namely, class 1, ALL; which are cells inflicted to leukemia, and class 2, HEM, which are normal healthy blood cells.

4.2.1 Performance Evaluation

The experimental results presented in Table 1 demonstrate the performance evaluation of various deep learning models. The models considered for evaluation include VGG16, VGG19, Inception, Xception, Efficient Net B0, ResNet101, and ResNet50. The evaluation metrics employed for assessing the models' performance are Validation Accuracy, Validation Precision, and Validation Recall.

4.2.2 Evalutation Metrics

Precision and recall, fundamental metrics in classification tasks, provide deeper insights into the models' abilities. Precision, denoted as

$$\text{Precision} = \frac{\text{TrP}}{\text{TrP + FaP}}$$

measures the accuracy of positive predictions. Recall denoted as

$$\text{Recall} = \frac{\text{TrP}}{\text{TrP + FaN}}$$

gauges the ability of the model to capture all positive instances.

A confusion matrix is a table that summarizes the performance of a classification model [?]. It consists of four components:

- **TrP:** Instances that are actually positive and were correctly classified as positive.
- **FaP:** Instances that are actually negative but were incorrectly classified as positive.
- **TrN:** Instances that are actually negative and were correctly classified as negative.
- **FaN** Instances that are actually positive but were incorrectly classified as negative.

The confusion matrix associated with each of the models are presented in Figure 5.

The first two models, VGG16 and VGG19 exhibit comparable Accuracy(70.46% and 72.27% respectively) and Precision(50.35% and 54.97% respectively), but a significant difference arises in the Recall values with VGG16 outperforming VGG19 by almost 8% (98.86% and 91.82% respectively). This is a very interesting and valuable insight as this means that the model was able to identify True Positive values almost 99% of the time that is, correctly labeling ALL cells. But is also important to acknowledge that the models have relatively low Precision scores. This

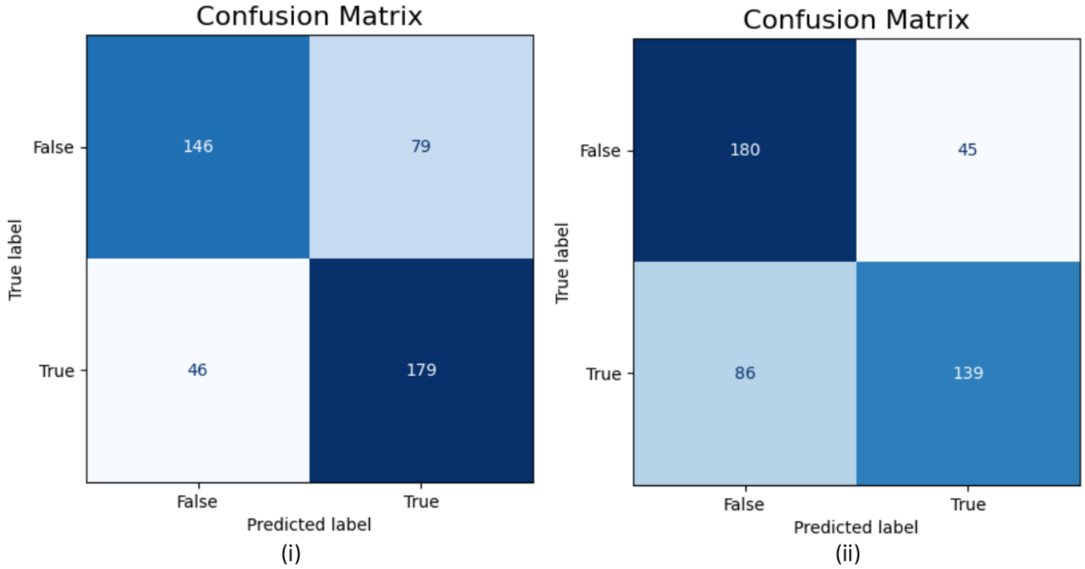

Figure 5: Confusion matrix (i)Vgg16,(ii)vgg19

indicated that healthy cells may get flagged as ALL, which would require further confirmation and investigation.

Inception and Xception well known for their efficient architecture design, exhibit similar and comparable results as shown in Figure 6. Both models achieve very similar Accuracy and Recall scores. It is to be noted that the Xception showed a marginally higher Precision score over Inception. However, as the differences are relatively small, it indicates that both models perform competitively in this task. These performances suggest that both models are adept at correctly identifying both healthy and ALL cells. This is a good improvement when compared to the previous models.

EfficientNet B0 [**?**] showed a striking behavior in which it only classified the images into one class, making it by far the worst model. This was first considered to be some error in the pipeline that was being employed, but after meticulous testing, the output remained the same. Even though it achieved a Recall score of 100%, both the Precision and Validation Accuracy are 50%. The most likely explanation for this is that the model might be overfitted and the smaller, less complex architecture of the EfficientNet B0 model might also have cont02201859-202003030-00041ributed to this result.

The ResNet models, ResNet101 and ResNet50 have showcased some of the best results as shown in Figure 7, emerging as the strongest contenders in the analysis. The models have a Validation Accuracy of 76.36% and 75.91%, Precision and Recall scores of 75.85% and 76.36% respectively. These models have shown the best all-round results and this has demonstrated their ability to accurately differentiate healthy and ALL cells with considerable consistency.

ResNet101 emerges as the top-performing model with a Validation Accuracy of 76.36%, Validation Precision of 75.85%, and Validation Recall of 76.36%. Similarly, ResNet50 closely follows, demonstrating a Validation Accuracy of 76.36%, Validation Precision of 75.85%, and Validation Recall of 76.36%. Notably, VGG16, VGG19, Inception, and Xception exhibit competitive performances, with Validation Accuracy values ranging from 70.46% to 73.86%.

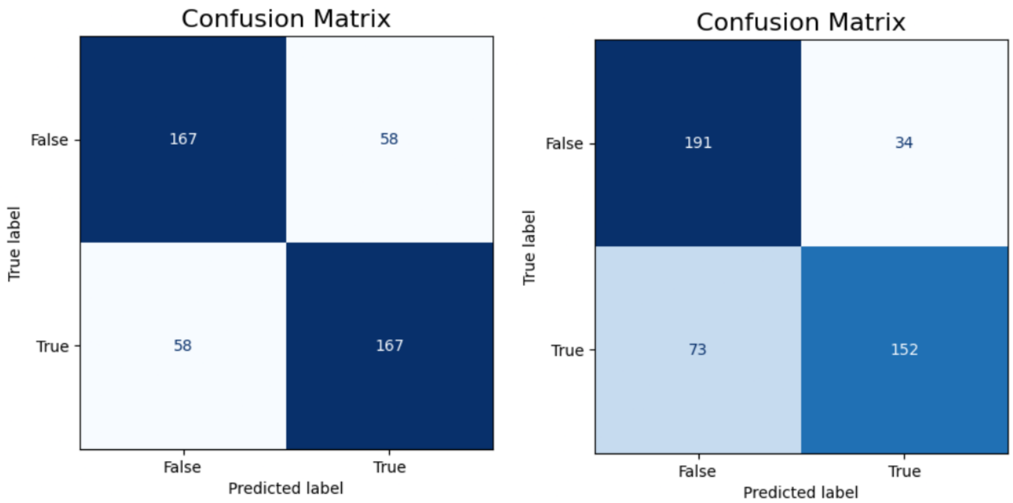

Figure 6: Confusion matrix (iii)Xception,(iv)Inception

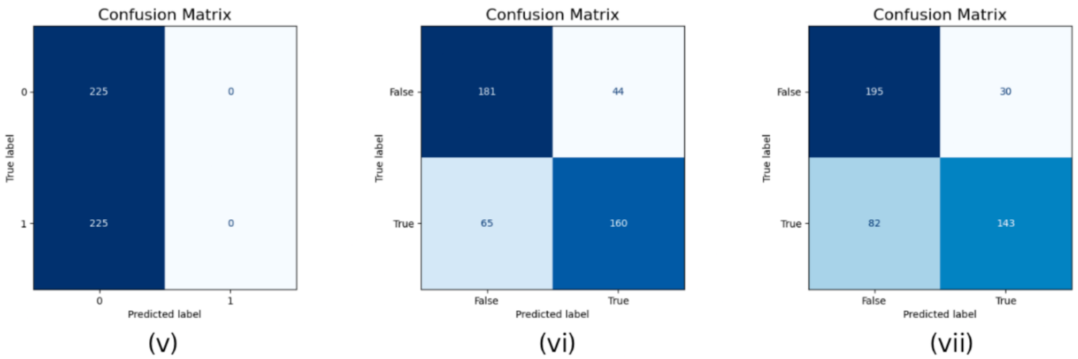

Figure 7: Confusion matrix (v)EfficientNet B0,(vi)ResNet101 (vii)ResNet50

Model	**Validation Accuracy**	**Validation Precision**	**Validation Recall**
VGG16	0.70455	0.5035	0.9886
VGG19	0.7227	0.5497	0.9182
Inception	0.7341	0.7101	0.7682
Xception	0.7386	0.7310	0.7659
Efficient Net B0	0.5000	0.5000	1.0000
ResNet101	0.7636	0.7585	0.7636
ResNet50	0.7591	0.7585	0.7636

Table 1: Validation Metric

5 Challenges in the Classification of Acute Leukemia

Efforts towards accurate classification of acute leukemia face several challenges, ranging from data availability to computational constraints.

5.1 Datasets are not available in the public domain

One significant hurdle encountered in acute leukemia classification pertains to the scarcity of publicly available datasets. Robust classification model creation and evaluation are hindered by limited access to diverse and well-curated datasets. Improper performance and restricted real-world applicability could result from trained predictors' inability to generalise from suitably big and representative datasets.

5.2 Generalization capabilities

Guaranteeing that trained predictors can generalize to other demographic and clinical contexts is another crucial difficulty. Because leukemia cells have diverse morphologies and genetic traits, models that are capable of generalizing to a variety of patient demographics, disease subtypes, and experimental circumstances are required. If this problem remains unsolved, models may be used in clinical settings with limited effectiveness.

5.3 Computational time

Because they demand a lot of processing power, complex classification models are especially difficult to train and assess. Deep learning models require a lot of processing power due to their complex structures, resulting in the need for strong GPUs and prolonged training periods. Moreover, the computational expense associated with iterative experimentation and hyperparameter tuning extends the development cycle and complicates the rapid prototyping and deployment of classification systems.

To solve these problems, the scientific community has to collaborate to develop algorithms, enhance computational procedures, and initiate data-sharing initiatives. Ultimately, overcoming these challenges will improve patient outcomes and diagnostic accuracy while also speeding up the classification of acute leukemia.

6 Conclusion and Future Scope

In summary, this study explored complexities related to the classification of acute leukemia, addressing significant issues and utilizing creative approaches to resolve them. Many methods and strategies used by researchers were compiled after an extensive analysis of the literature, demonstrating the variety of methodologies used for leukemia cell classification. To train and assess classification models, a state-of-the-art suite of tools and technologies was utilized, as detailed in the materials and methods section. There is a need for cooperative efforts and breakthroughs in data sharing, algorithmic development, and computational infrastructure because of challenges including restricted data availability, generalization of predictors, and computational limits.

As move forward, there are several interesting directions that this research will likely go in the future. The creation of more reliable and broadly applicable categorization models is made possible, first and foremost, by efforts to improve data accessibility and sharing among the research community. The accuracy and scalability of classification systems could also be enhanced by developments in algorithmic techniques, such as the incorporation of multimodal data sources, transfer learning, and domain adaptation strategies. It is also possible to expedite the development cycle and simplify experimentation by optimizing computational procedures and leveraging cloud-based resources. Lastly, interdisciplinary collaborations between researchers, clinicians, and industry stakeholders can foster innovation and facilitate the translation of research findings into clinical practice, ultimately improving diagnostic accuracy and patient outcomes in the realm of acute leukemia classification.