Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

PERFORMANCE EVALUATION OF CNN MODELS USING TRANSFER LEARNING AND ENSEMBLE APPROACHES FOR AUTOMATED LEUKEMIA DIAGNOSIS

Nila Sultana, Sraboni Ghosh Joya, Mehedi Hasan

Department of Computer Science and Engineering,
City University, Birulia, Dhaka, Bangladesh

ABSTRACT

A significant portion of deaths related to cancer globally are caused by Leukemia, one of the most serious hematological malignancies. It is distinguished by the rapid growth of premature lymphocytes, which disrupts the regular operation of the bone marrow and blood. To increase survival rates and ensure prompt treatment, early and accurate diagnosis is essential. The majority of modern diagnostic techniques, however, rely on the laborious, ineffective, and error-prone manual interpretation of peripheral blood smear (PBS) images. Convolutional neural networks (CNNs), in particular, provide a dependable and automated substitute in deep learning. In this work, we utilize a carefully selected PBS dataset representing four main leukemia types collected in Bangladesh to evaluate six CNN architectures: VGG19, InceptionV3, MobileNetV2, Xception, DenseNet-201, and SecrensNet152. InceptionV3, MobileNetV2, DenseNet-201, VGG19, and SecrensNet152 were all subjected to transfer learning to enhance model generalization. To improve performance, we also developed an ensemble model called DEX, which combines DenseNet-121, EfficientNet-B7, and Xception. With an astounding accuracy of 99%, the trial findings show that DEX outperformed any of the separate CNN models. Accuracy improvements of up to 16% were observed with transfer learning compared to baseline models. These results open the door for the incorporation of CNN ensembles into real-time clinical decision

support systems and demonstrate their potential for providing extremely accurate leukemia diagnoses.

Keywords: *Leukemia*, *peripheral blood smear*, *CNN*, *Deep learning.*

Corresponding author: Nila Sultana can be contacted at nilasultanagodhulee@yahoo.com

1. INTRODUCTION

One of the deadliest hematological diseases, Leukemia, is characterized by excessive growth of white blood cells (WBCs). Red blood cells, white blood cells, and platelets are the three different lineages produced by the human blood-forming system. The primary oxygen transporters from the heart to the tissues are erythrocytes, also known as red blood cells, which comprise approximately 45% of the total blood volume. WBCs, on the other hand, serve as the immune system's first line of defense against infections and illnesses and are one of its primary components (Rehman et al., 2018). As a result, the early identification of aberrant WBCs sheds light on the type and progression of Leukemia. The four most common types of Leukemia can appear at any age.

Acute Myeloid Leukemia (AML): Rapid accumulation of WBCs with little practical use is caused by defective marrow cell growth. This damages the bone marrow and results in cancer. Its response to treatment improves with early detection. Symptoms of Tezal: Signs and symptoms may include weakness, unusual bleeding, or trouble breathing.

Acute Lymphocytic Leukemia (ALL): ALLis a disease that primarily affects children and is characterized by an excessive number of lymphoblasts. Radiation exposure, viral infections, and genetic

disorders, including Down syndrome, are risk factors. Among them are L1, L2, and L3. In fact, one of the intriguing aspects of the remission rate in children's ALL cases is that it usually performs far better than in adults.

Chronic Myeloid Leukemia (CML): When these myeloid cells have gene changes that impair their ability to fight off infections, CML develops. This subtype mainly affects adults and progresses slowly through three stages: the chronic period, the accelerated phase, and the blast phase. Early-stage cure is possible, but as the virus's ability to destroy blood cells intensifies, the disease's severity rises.

Chronic Lymphocytic Leukemia (CLL): B-lymphocyte-like cells proliferate in bone marrow and blood in CLL. Not only does the WBC count rise, but they are also not operating correctly. One of the most curable forms of Leukemia, CLL, is an adult condition. CLL can occasionally develop into ALL.

Among these, ALL is noteworthy since it accounts for 25% of juvenile cancer cases. One of its origins is bone marrow, and it develops in lymphoid tissue (Fujita et al., 2021). WBC growth is usually controlled, but in Leukemia, it increases uncontrollably. Although leukemic cells typically appear dark purple on stained PBS, their distinct appearance makes visual identification challenging. Smooth patterns, thin cytoplasmic rims, and nearly spherical nuclear outlines are characteristics of normal lymphocytes. Any patient with ALL will always have lymphoblasts with cytoplasmic vacuolization and an irregular nucleus, which are especially noticeable in more advanced stages of the illness. The EDVA's peak age (7–8) has a significant impact on patient outcomes; it declines until the twenties and

then increases again beyond 50. Therefore, it is crucial to identify and address the problem promptly. Morphological study of blood smears remains one of the most critical diagnostic methods for acute lymphoblastic Leukemia (ALL), even in contemporary times. It is difficult to distinguish between cancerous and normal lymphocytes when examining them under a microscope, though. Hematologists should allocate considerable effort and resources to the visual analysis of PBS images for the identification of disease subgroups.

Furthermore, early detection is particularly challenging in developing nations due to a lack of qualified specialists. Through end-to-end feature extraction and classification, deep learning approaches have made significant progress in medical image analysis over the past few years. CNN's versatility, capacity for self-learning, and robust generalization make it the most widely used model for various medical image and computer vision tasks (He al., 2016). Since conventional image classification methods and discriminative features may be learned from raw images, CNN does not rely on expert feature extraction. Despite being computationally expensive, CNNs may outperform databases and be more useful if sufficient learning databases are available. Transfer learning (Dosovitskiy, A., et al, 2020) is employed to enhance performance at a modest computational cost when training data are scarce.

In our study, the authors aim to identify the optimal deep learning model for leukemia identification using PBS images. Regarding this, we concentrate on the performance of the six well-known CNN architectures employing transfer learning: VGG19 (Huang, G.et al, 2017), InceptionV3 (Szegedy, C.et al, 2016), MobileNetV2, Xception (Chollet, F., 2017), DenseNet-

201, and SE-ResNet152 (Tan, M., & Le, Q.,2019). To compare baseline CNNs, transfer learning, and ensembles, we also investigate a single model ensemble (MDX) that comprises DenseNet-121 and MobileNetV2 as components. The fact that Leukemia is still diagnosed qualitatively under a microscope in many developing nations, despite growing manufacturing capacity, is a major driving force behind this work. This method is not only costly and time-consuming, but it is also prone to errors due to human error. Automated detection systems based on deep learning may offer reliable and scalable solutions that are more widely available for early diagnosis and treatment.

2. REVIEW OF LITERATURE

Leukemia is a blood and bone marrow cancer that ranks high among cancer-related deaths. It is difficult to identify early because it frequently has an abrupt attack and irreversible pathological alterations. Additionally, it is among the most thoroughly researched blood cancers. Even with a low mortality rate, early detection and appropriate therapy can reduce it, underscoring the importance of reliable diagnostics.

Convolutional neural networks (CNNs) have recently shown remarkable proficiency in automatically diagnosing Leukemia from medical images. CNNs work well when it comes to detecting suspected and unknown abnormalities in peripheral blood smears. Leukemia can be detected using various technologies and approaches. However, many of these systems still have flaws (such as the ongoing inability to reliably achieve object discrimination) or overlap (in terms of accuracy), in addition to completely failing to meet practical performance standards.

To address these issues, the CNN model was used as our baseline architecture in the study, and other medical picture datasets (hematological images) were also suggested and used. Meanwhile, the efficacy of early leukemia identification has been further improved as the model structure is further developed and various technologies (such as ensemble learning and transfer learning are provided (Wu et.al., 2021).

Motivated by this procedure, we examine existing CNN-based techniques in this research, evaluate their effectiveness, and propose a new model to enhance accuracy.

2.1 CNNs for Leukemia Classification

Emerging artificial intelligence (AI) technology, deep learning (DL), has made a significant contribution to the research community in medical imaging. Among deep learning methods, the Convolutional Neural Network (CNN) is most successful in diagnosing hematological malignancies, such as Leukemia, due to its excellent feature extraction and classification abilities. In this section, we review current CNN methods aimed at improving leukemia diagnosis through the automatic analysis of blood smear images and other medical imaging modalities. CNNs are feedforward artificial neural networks that are biologically inspired and designed to process grid-like data, such as images, on mobile devices, including smartphones. They are also appealing for medical imaging applications, as they can infer hierarchical features directly from inputs at the image level without the need for manual feature crafting. For leukemia detection, CNN has been applied to WBC classification in microscopic images (Chollet, F., 2017), distinguishing between healthy and cancerous types of WBC (Fujita et al., 2021), or

between different categories of Leukemia, such as Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). An ordinary CNN has multiple convolutional layers that convolve the input images with learnable filters to capture spatial details, such as edges, textures, or shapes. These are followed by non-linear activation functions – such as ReLU, which is often used to introduce non-linearities in the network. Pooling techniques, such as max pooling and average pooling, aim to reduce the spatial size of feature maps, thereby improving computation efficiency and invariance to translation. In Chollet (2017), a CNN model is utilized on ALL-IDB datasets for public discriminative analysis of leukemic cells and results in 94% accuracy (I_ACCOUNTI) (a)(b). Figure 3: Across various feature extraction techniques, the constructed model was composed of (Figure 3c): three convolutional layers, subsequently pooling, and dense layers.

Similarly, a deep CNN with residual connections, as in ResNet-50, was proposed, which enables the training of deep networks without vanishing gradients. This technique proved effective in distinguishing types of TMP from both morphological and cytochemical standpoints. One of the significant advantages of using CNNs for leukemia diagnosis is that they are generalizable across different imaging conditions and staining protocols, provided that they have been trained on a sufficiently broad dataset. In addition, CNNs enable end-to-end learning, where the network learns to automatically extract meaningful features from raw images to produce classification labels. Moreover, transfer learning has been widely adopted to improve the model's performance, especially with small datasets. In our recent study, retraining pre-trained models (i.e., VGG16,

InceptionV3, and EfficientNet) on hematological datasets achieved exact classification results, and the models converged faster.

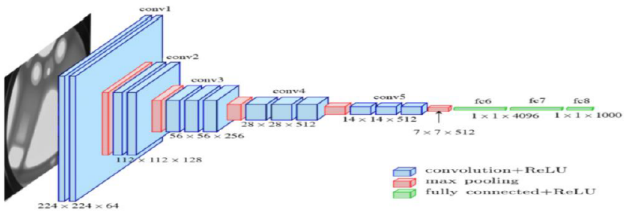


Figure 1. The Basic CNN Architecture. (Source-Google)

2.2 Comparison with State-of-the-Art CNN Architectures

During the past decade, a variety of CNN architectures have been proposed to mitigate different challenges in image recognition tasks, such as depth, representational efficiency (overfeat), computational cost (network-in-network), and feature generalization (Inception). Not only have these architectures improved classification accuracy, but they have also proposed novel approaches, such as residual connections, inception modules, depthwise separable convolutions, and feature recalibration. A comparison of some of the most popular CNN models, including their core contributions, parameter counts, classification errors, and the number of layers used, is presented in Table 1.

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

Table 1. A comparison of popular CNN architectures (M Million)

Architecture	Year	Main Contribution	No. of Parameters	Top-1 Accuracy on ImageNet	Depth / Layers
VGG-19	2015	Deep stack of 3×3 convolutions with uniform structure	138 M	71.5%	19
ResNet-152	2015	Introduced residual learning with identity mappings	60 M	78.3%	152
InceptionV3	2016	Factorized convolutions, auxiliary classifiers, and multi-scale filtering	23.6 M	78.8%	159
Xception	2017	Depthwise separable convolutions replacing Inception modules	22.8 M	79.0%	126
DenseNet-121	2018	Dense connectivity between layers for feature reuse and gradient flow	8 M	74.9%	121
ResNeXt-101	2018	Cardinality-based parallel paths improve performance over ResNet	84 M	79.6%	101
EfficientNet-B0	2019	Compound scaling across depth, width, and resolution	5.3 M	76.3%	~18
EfficientNet-B7	2019	Scaled-up version of EfficientNet-B0	66 M	84.3%	~40
EfficientNetV2-S	2021	Faster training, better accuracy, uses fused MBConv	24 M	84.6%	~24
ViT-B (Vision Transformer)	2020	Transformer architecture using patch embeddings for image classification	86 M	77.9%	12 Transformer blocks
ViT-H (Huge)	2020	Large-scale Transformer trained on JFT-300M	632 M	88.5%	32 Transformer blocks

DeiT-B	2021	Data-efficient training of ViT via knowledge distillation	86 M	81.8%	12 Transformer blocks
CvT-W24	2021	CNN-Transformer hybrid with convolutional token embeddings	~275 M	87.7%	Hierarchical, ~24 blocks
CSWin-Tiny	2022	Cross-shaped window attention with local global context	~23 M	85.4%	~12–24 attention blocks
TinyViT-21M	2023	Compact ViT model optimized via distillation and efficient training	21 M	84.8%	12–24 layers
Hybrid-Ms-S+	2024	Combines CNNs with Transformer/MLP blocks for balanced local-global modeling	63 M	83.9%	Multi-stage hybrid blocks

Note: The author(s) own work.

2.3 Transfer Learning

Transfer learning is an effective method that leverages a pre-trained CNN trained on large-scale datasets (e.g., ImageNet, with millions of images and thousands of categories). Rather than building a CNN from scratch, an expensive operation in terms of computational power and data requirements, transfer learning enables the repurposing of the learned features of a pre-existing model. This makes the load much lighter in terms of gathering a large annotated set and even lower when harvesting from scratch deep networks. There are two standard methods to perform transfer learning: Fine-tuning and Feature extraction. Fine tuning is an approach that partially freezes and re-trains the weights of a pre-trained CNN. Nonetheless, it's common practice to lock the first layers of the network, as they can extract low-level features (e.g., edges, textures, and shapes) that are general across various image

recognition tasks. The latter layers, on the other hand, retain more abstract and specialized features for a given task due to the specificity of the target domain. This selective retraining enhances the model's generalization, allowing it to perform well on new but related tasks. However, for a CNN as a feature extractor, the advantage of freezing all layers in the pre trained model is that one can only use intermediate layer outputs as fixed feature representatives. The features can then be used as inputs to an additional machine learning classifier (e.g., SVM or Random Forest), in which the deep network's weights remain unchanged. This approach is beneficial in cases where computational resources are limited and when the target dataset is small, because it alleviates the risk of overfitting due to retraining deep models. Transfer learning is widely used in medical image analysis, as it's expensive and challenging to annotate by domain experts in limited datasets. For identifying Leukemia, transfer learning is applied to leverage pre-trained natural image-based CNN models and generate meaningful features in microscopic blood smear images, thereby increasing classification accuracy even with small datasets. This flexibility and effectiveness have established transfer learning as the backbone of recent biomedical image processing.

2.4 Ensemble Technique

Ensemble learning is a technique of combining multiple weak base classifiers to yield a single strong prediction model. Ensemble techniques, which average outputs of multiple heterogeneous classifiers, usually outperform single models and achieve better accuracy, robustness, and generalization. The intuition behind ensembles is that if individual classifiers are all biased in a similar way or have various weaknesses, then the

bias or weaknesses cancel out and appear less pronounced, resulting in better overall predictive performance. Some of the widely used assembling techniques are bagging, boosting, and stacking. Bagging (Bootstrap Aggregating) fits several base models on various subsets of the training data and utilizes their averaged predictions to reduce variance and mitigate overfitting. Sequential training of base learners in boosting focuses on complex examples and learns from errors made by previous models, ultimately leading to their overall improvement in accuracy. Stacking (or stacked generalization) is a more advanced ensemble technique, in which the predictions made by base learners are fed into another learner at step N for modeling and predicting their collective performance. The meta-learner is designed to optimize the weights of base models by minimizing a loss function on a validation set (usually through cross-validation) (Kumar et al., 2020). This method leverages the strengths of various learners, overcomes their drawbacks, and ultimately generates a better prediction than any individual learner used in the input space.

Ensemble methods have also been employed in leukemia classification to enhance diagnostic accuracy. For instance, combining the predictions of CNNs, SVMs, and decision trees can result in robust classifications to counteract misclassifications occurring in individual models. Finally, ensemble particle swarm model selection (EPSMS) methods have been developed that perform simultaneous search and evaluation of classifiers to improve performance in medical imaging applications. Ensemble methods, in general, are beneficial for clinical applications where accurate prediction and trustworthiness are crucial.

2.5 Leukemia Blood Cancer Literature Review

Leukemia, also known as blood cancer, results from the rapid growth of white blood cells (WBCs) produced by an abnormal bone marrow, disrupting the production of normal and healthy blood cells. It is generally divided into two main types: acute Leukemia, which develops rapidly, and chronic Leukemia, which progresses more slowly. The mutated WBCs take over the blood and bone marrow, crowding out normal blood cells and thereby weakening the body's ability to fight infections and perform other essential blood tasks. Machine learning methods, particularly classifiers such as Support Vector Machines (SVMs), have been widely used for leukemia detection and classification from blood smear images. Hariprasath and Dharani employed an image analysis technique to distinguish between healthy blood and leukemic blood, demonstrating that an automated procedure can be effective in diagnosing blood cancer. Sinha et al. (2024) further investigated an image processing and machine learning method for enhancing the detection of Leukemia, again reflecting how computational tools can supplement such work. Among the AMLs with high-coverage whole-genome sequencing and mutational load information, acute myeloid Leukemia (AML) is the most studied subtype, as it is characterized by the clonal expansion of undifferentiated myeloid cells infiltrating the bone marrow, and blood, Originally considered untreatable, modernization has brought the survival rates to 35–40% in patients under 60 years of age and 5–15% in older patients (Rahman, et al., 2023). The accurate and early diagnosis of leukemia subtypes remains critical for successful therapy.

Present diagnostic methods, including morphological examination of blood smears by experienced hematologists, are labor-intensive, time-consuming, and subjective. Automated WBC identification and classification systems for microscopic images have been introduced to address these problems. Putzu et al. (2020) introduced an automated WBC classification method using image analysis to enhance accuracy and reduce processing time. Similarly, Ko et al. proposed a new segmentation method by integrating mean-shift clustering with gradient vector flow (GVF) snakes for accurate staining of WBC image segmentation.

Types of Leukemia include acute lymphoblastic Leukemia (ALL), as well as AML, chronic lymphocytic Leukemia (CLL), and chronic myeloid Leukemia (CML). Scientists have developed automated leukemia detection image-processing architectures to minimize dependence on manual examination (Sheng et al., 2020). Convolutional Neural Networks (CNNs) have become the dominant choice for such tasks due to their ability to learn hierarchical features from raw images, which is especially desirable when classifying different stages and subtypes of Leukemia (Khosla et al., 2018). New therapeutic approaches are emerging in AML, including targeted therapies and combinations of agents (e.g., venetoclax plus FLT3 or IDH inhibitors) that have demonstrated encouraging clinical activity in enhancing rates of remission and survival (Daver et al., 2020).

An accurate diagnosis remains important before applying treatment regimens. Standard diagnostic techniques, particularly manual microscopic examination, are challenging to use and prone to mistakes. Jagadev et al. (2017) also proposed an automated method for detecting and categorizing AML using

a CNN, which improved accuracy and efficiency. These findings indicate a trend toward the use of AI-based diagnosis for blood cancer.

Despite the advances in automatic leukemia detection, numerous issues remain to be addressed (Gonzalez & Woods, 1992), that are not exhaustive and require significant user intervention to verify their results (McNeill, 2008). Kumar et al. (2020) proposed a CNN-based efficient approach to classify Acute Lymphoblastic Leukemia (ALL) and Multiple Myeloma (MM) using the SN-AM dataset, achieving an accuracy of 97.2%, which outperforms traditional machine learning approaches such as SVM and Random Forest.

New-generation sensing mechanisms have been achieved using novel technologies, such as photonic crystal fibers, for the early detection of blood cancer. This approach utilizes the refractive index difference between normal and Leukemia cells, resulting in increased visibility and a potential for rapid identification. In addition, CSCs (CSCs), especially leukemia stem cells (LSCs), are accountable for the progression and relapse of diseases, as they are resistant to conventional treatments. The selective elimination of LSCs, while preserving normal cells, represents a novel therapeutic approach. Biosensor technology utilizing coral nanostructures on ITO substrates has demonstrated potential in detecting genetic markers associated with Leukemia at a rate higher than that of currently used diagnostic approaches, resulting in improved diagnostic specificity and sensitivity (Mollah et al., 2020). Computerized diagnostic aids, such as CNNs applied to CT files, have achieved a maximum accuracy of 94.5% in cancer detection, further demonstrating

the importance of deep learning in early diagnosis (Rohaziat et al., 2020).

CNN models have been used for lymphoma cell detection with remarkable performance. In the blood cell classification problem, fused CNN architectures have demonstrated faster training and improved accuracy compared to hybrid RNN-CNN Models, highlighting the contribution of architectural innovations. Compared to spectral analysis alone, hyperspectral imaging combined with deep learning has enhanced both spectral and spatial characteristics, enabling more accurate classification of blood cells. It utilized the pre-trained AlexNet model, which has been fine-tuned on leukemia datasets, proving to be an effective strategy. This approach achieved high classification accuracy without requiring image segmentation at the microscopic level, demonstrating the effectiveness of transfer learning in medical image analysis.

However, several challenges remain, including small dataset size, class imbalance, and the need for a robust and interpretable model. It is essential to bridge these gaps to translate AI-based leukemia diagnostics into clinical applications.

3. RESEARCH METHODOLOGY

The experiments were executed using the Keras library and Google Colab. TensorFlow was chosen since it is regarded as one of the top Python deep learning libraries for machine learning implementation. Each model was hosted on Google's Collaboratory platform and trained in the cloud using a Tesla GPU. A 360 GB GPU cloud and 12 GB of RAM will be provided under the collaborative research framework.

3.1 Dataset Description

The centerpiece of this research is the Acute Lymphoblastic Leukemia (ALL) Image Database, an extensive collection of peripheral blood smear images. The dataset comprises 3,256 images from 89 subjects suspected of having ALL, as well as 25 subjects with benign hematogone/basophil cells in the bone marrow, and 64 subjects with confirmed ALL subtype. These subtypes are the Early Pre-B ALL (985 images), the Pre-B ALL (963 images), and the Pro-B ALL (804 images). The dataset is divided into benign and malignant clonal populations (hematogone and ALL, respectively), where the malignant subtype is further split into three subtypes, as described above. All pictures were taken with a Zeiss digital microscope at ×100 and saved in JPEG [24]. Medical professionals used flow cytometry to accurately corroborate subtypes.

Moreover, segmented images were generated by applying color thresholding in the HSV color space to segment target cellular parts. The training and testing datasets of the image distributions were classified in a 70:30 proportion from the dataset, as shown in Table 2.

Table 2. Images Used in Train, Test, and Validation Sets

	Total Images	Training Images	Validation Images
Original Dataset	3256	2277	979

Note: The author(s) own work.

3.2 Experimental Procedure

The proposed pipeline includes several significant steps, ranging from data collection to classification, as follows:

a. Image Acquisition

The first step was to collect crystal-clear images from the ALL-Image Database, which were then placed onto a white canvas. When the background was non-uniform or had color, the images were centered on a white background to maintain a similar format for input data.

b. Image Augmentation

Since the total size of medical imaging datasets is relatively small, we employed data augmentation to synthetically expand the range and quantity of training samples, thereby effectively preventing overfitting and enhancing model generalization. The data has been augmented through geometric transformations (scaling, cropping, horizontal and vertical flipping, random rotation between -15° and +15°, and 90° rotations) as well as photometric transformations (brightness, contrast, and saturation). Further augmentations included elastic distortions and changes in intensity, thereby reproducing the more realistic imaging conditions.

Ten augmented copies of each original image were created, which vastly increases the diversity of the training set. Pixel values for both original photos and augmented images were scaled to [0, 1] by dividing by 255 before training. All images were resized to a standard input size of 132×132 pixels, selected after manual testing for the effect of hardware limitations on model performance, and for the Xception architecture, which requires high computational power.

c. Model Training

Multiple CNNs, including ResNet-152, MobileNet, VGG-19, Xception, and Inception-V3, were trained and tested. Every model was trained using categorical cross-entropy, as is appropriate for the multi-class categorisation problem that we have. The ramp function was used as the activation in all hidden layers, and a softmax function was applied at the output layer, which promoted the use of probabilities for classification.

Training was performed using the Adam optimizer, an algorithm that achieves good results combining both RMSProp and Stochastic Gradient Descent with momentum, which we configured with the following hyperparameters: learning rate (α) of 0.0001, β1 equal to 0.9, and β2=0.999, ε = 10^{-7}. Early stopping was used with a patience of 10 epochs to prevent overfitting on the training set, so the training stopped after no

improvement was observed for 10 consecutive epochs. Both experiments used a batch size of 17 and were trained for no more than 25 epochs. We display the training time for each sub-model in Table 1, where training times were around 20 seconds per epoch for MobileNet and about 14 seconds per epoch for DenseNet-201 and InceptionV3. Model weights were saved in. h5 format for further examination and deployment.

d. Classification

The final stage involved applying the trained CNN models to distinguish between blood smear images: benign hematogone or malignant ALL subtype. Classification was done in an experiment-specific manner, by providing pre-processed images to the networks and reading out the softmax output

probabilities. The class with the highest predicted probability was considered the ultimate diagnosis. Therefore, neural networks were selected as an adequate model for capturing spatial relationships and features in biomedical images, especially for disease diagnosis from histopathological data.

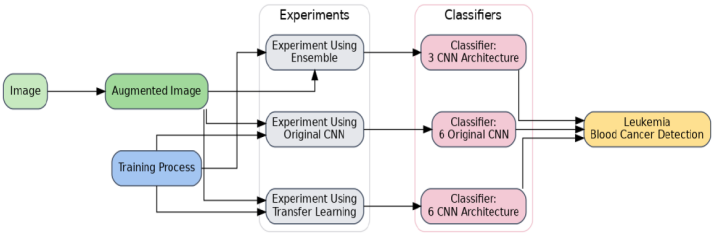


Figure 3. Process of Experiments

4. EXPERIMENTAL RESULTS

In this section, the experimental results are examined, followed by comparisons and additional commentary. Three types of experimental outcomes are distinguished: ensemble methods, transfer learning, and original individual network structures. Following a trial-and-error process, the ideal set of parameter configurations is presented for each constraint.

4.1 Evaluation Criteria

The classification performance of Convolutional Neural Network (CNN)-based algorithms in the context of leukemia detection is evaluated using several performance metrics. These measures consist of the confusion matrix (CM), Recall, accuracy (AC), Precision, and F1-score. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are

the primary variables utilized in these measurements. Below is a description of the evaluation metrics used to assess the models' performance.

4.1.1 Accuracy

This metric considers the overall quantity of lessons correctly expected through the skilled version out of all feasible lessons. The ratio of efficaciously categorized images to the overall number of samples is described as accuracy.

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

4.1.2 Precision

This metric calculates the number of true positives among all positive cases. In the case of Leukemia, the model can correctly identify those patients who have Leukemia. When FP is more important than FN, Precision is a useful metric. The following equation mathematically defines it:

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

4.1.3 Recall

The recall metric measures how well the model highlights leukemia disease patients based on all relevant data. When FN triumphs over FP, Recall becomes a useful metric. The following equation is used to calculate it:

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

4.1.4 F-1 score

The F1 score is an additional measure of classification accuracy that combines Precision and Recall. Because the F1-score is a harmonic mean of Precision and Recall, it is a synthesis of these two metrics. When Precision equals Recall, it is at its peak. By combining Recall and precision values, this metric assesses the model's overall efficiency.

$$\mathbf{F-1\ Score} = 2 * \frac{\mathbf{Precision} * \mathbf{Recall}}{\mathbf{Precision} + \mathbf{Recall}}$$

Furthermore, the model's ability to generalize to new data is measured by the validation loss, whereas the training loss quantifies how well the model fits the training data. It means the fit is better if the loss is minor. However, an overemphasis on reducing training loss can lead to overfitting, which causes the model to perform poorly on unseen data.

A practical method for assessing classification performance is a confusion matrix (CM). It provides a table with numbers such as TP, FP, TN, and FN, which show the model's predictions in comparison to the actual results. These numbers are necessary to compute critical metrics, including F1-score, Recall, and Precision. The support is the quantity of real examples of each class in the dataset. To enhance model performance, imbalances in support may indicate the need for class rebalancing or stratified sampling.

4.2 Performance Of Original CNN Networks

In this section, the classification performance of seven original individual CNNs was estimated— SecrensNet152, MobileNetV2, VGG19, Xception, InceptionV3, and DenseNet-201. The training and model accuracy are provided, along with a brief analysis of the root cause of the low scores and suggestions on how to address them.

Table 3. Accuracy for Classification of Individual CNN Networks in Detecting Leukemia Blood Cancer (Original CNN Networks)

Architecture	Training Accuracy	Model Accuracy
VGG19	99%	89%
Inceptionv3	98%	90%
MobileNetv2	99%	89%
Xception	97%	90%
DenseNet-201	99%	91%
SecrensNet152	96%	90%

Note: The author(s) own work.

The accuracy of the SecrensNet152, MobileNet, VGG19, Xception, Inceptionv3, and DenseNet-201 is shown in Table III. The VGG19, DenseNet-201, and MobileNetv2 models had the highest accuracy of 99%, while the SecrensNet152 model had the lowest accuracy of 96%.

Table 4. Precision, Recall, F1, and Support (N) Result of Original CNN Networks (Based on the Number of Images, N Numbers)

	Vgg19			
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	100%	99%	97%	100%
Recall	97%	99%	100%	98%
F1-score	99%	99%	99%	99%
Support (N)	136	270	270	220

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

DenseNet-201				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	100%	99%	100%	100%
Recall	98%	100%	100%	100%
F1-score	99%	99%	100%	100%
Support (N)	137	271	267	221

Xception				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	100%	94%	100%	100%
Recall	88%	100%	100%	100%
F1-score	94%	97%	100%	100%
Support (N)	138	274	259	225

Seresnet152				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	100%	51%	88%	67%
Recall	29%	98%	95%	15%
F1-score	45%	67%	91%	25%
Support (N)	136	272	268	220

Mobilenetv2				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	98%	98%	100%	100%
Recall	96%	100%	100%	100%
F1-score	97%	99%	100%	100%
Support (N)	135	268	264	229

InceptionV3				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	96%	96%	100%	100%
Recall	93%	100%	98%	100%
F1-score	95%	98%	99%	100%

Support (N)	135	274	266	221

Note: The author(s) own work.

Table 4 shows the Precision, Recall, F1-score, and Specificity obtained by the SecrensNet152, MobileNet, VGG19, Xception, Inceptionv3, and DenseNet-201 models for each class. When the precision values for each architecture on the test dataset are considered, VGG19, DenseNet-201, and MobileNetV2 provide the best performance. According to the above table, the VGG19, DenseNet-201, SecrensNet152, and MobileNetv2 models correctly classified blood cancer as Leukemia. The InceptionV3 performed poorly, with the lowest identification.

4.3 Transfer Learning CNN Network Accuracy in Detecting Leukemia Blood Cancer

Table 5. Transfer Learning CNN Network Accuracy in Detecting Leukemia Blood Cancer

Architecture	Training Accuracy	Model Accuracy
VGG19	78%	68%
Inceptionv3	85%	78%
MobileNetv2	93%	89%
Xception	97%	90%
DenseNet-201	93%	85%
SecrensNet152	96%	88%

Note: The author(s) own work.

The performance of six transfer learning CNN architectures is presented in this section. SecrensNet152, MobileNet, VGG19, Xception, Inceptionv3, and DenseNet-201 models all had high accuracies in the test sets, as shown in Table 5. The number of properly-identified samples to the total number of samples was

used to compute the test accuracies displayed in Table 5. With a precision of 97 %, the Xception model was the most accurate. The accuracy improvement of the DenseNet-201 network from the initial network to transfer learning is notable.

Table 6. Precision, Recall, F1, and Specificity Results of CNN Networks with Transfer Learning (N= Numbers)

Vgg19				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	100%	65%	90%	82%
Recall	28%	86%	95%	95%
F1-score	50%	74%	93%	88%
Support (N)	140	266	269	221

DenseNet-201				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	93%	78%	99%	99%
Recall	52%	98%	97%	100%
F1-score	66%	87%	98%	99%
Support (N)	142	268	266	220

Mobilenetv2				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	97%	80%	97%	99%
Recall	61%	96%	97%	96%
F1-score	75%	87%	97%	98%
Support (N)	133	275	265	223

Seresnet152				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	98%	77%	98%	98%
Recall	45%	98%	97%	99%
F1-score	62%	86%	98%	98%
Support (N)	140	273	264	219

Xception				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	93%	79%	98%	97%
Recall	57%	97%	94%	97%
F1-score	70%	87%	96%	97%
Support (N)	141	272	259	224

InceptionV3				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	92%	73%	90%	92%
Recall	33%	90%	95%	95%
F1-score	49%	80%	92%	93%
Support (N)	137	271	265	223

Note: The author(s) own work.

The Precision, Recall, F1-score, and Specificity findings from CNN networks incorporating transfer learning are shown in Table 6. In general, a model with high Precision, Recall, and support is a superior model. With a 92% accuracy, the trial results show that Inceptionv3 has low Precision in leukemia blood cancer.

Table 7. Precision, Recall, F1, and Specificity Results of CNN Networks with Ensemble Techniques (N= Numbers)

Ensemble DEX Model (DenseNet121, EfficientNetB7, and Xception)				
	Benign	Early	Pre-B ALL	Pro-B ALL
Precision	99%	89%	99%	100%
Recall	78%	99%	99%	99%
F1-score	87%	94%	99%	99%
Support (N)	141	275	263	217

Note: The author(s) own work.

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

4.4 Confusion Matrix (CM) After Original CNN Networks (Based on The Number of Images)

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	132	0	0	0
Early	4	268	0	0
Pre-B ALL	0	2	270	5
Pro-B ALL	0	0	0	215

Figure 4 (A): CM after Original Vgg19

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	134	0	0	0
Early	3	271	0	0
Pre-B ALL	0	0	267	0
Pro-B ALL	0	0	0	221

Figure 4 (B): CM after Original DenseNet201

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	132	0	0	0
Early	4	268	0	0
Pre-B ALL	0	2	270	5
Pro-B ALL	0	0	2	215

Figure 4 (C): CM after Original Xception

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	40	0	0	0
Early	75	266	14	162
Pre-B ALL	7	3	254	24
Pro-B ALL	14	3	0	34

Figure 4 (D): CM after Original Seresnet152

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	126	0	4	0
Early	9	274	2	0
Pre-B ALL	0	0	260	0
Pro-B ALL	0	0	0	220

Figure 4 (E): CM after Original InceptionV3

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	129	1	0	1
Early	6	267	0	0
Pre-B ALL	0	0	264	0
Pro-B ALL	0	0	0	228

Figure 4 (F): CM after Original MobileNetV2

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

4.5 Confusion matrix (CM) after Transfer Learning (TL) (Based on the number of images)

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	4	0	0	0
Early	105	229	8	9
Pre-B ALL	5	21	257	0
Pro-B ALL	26	16	4	212

Figure 5 (A): CM after TL Vgg19

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	74	4	1	0
Early	67	263	3	0
Pre-B ALL	1	1	260	0
Pro-B ALL	0	0	2	220

Figure 5 (B): CM after TL of DenseNet-201

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	81	5	2	0
Early	55	266	10	4
Pre-B ALL	1	1	245	1
Pro-B ALL	4	0	2	219

Figure 5 (C): CM after TL of Xception

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	64	1	0	0
Early	73	268	5	2
Pre-B ALL	0	4	258	0
Pro-B ALL	3	0	1	217

Figure 5 (D): CM after TL of Seresnet152

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	46	2	2	0
Early	78	245	7	5
Pre-B ALL	4	18	253	6
Pro-B ALL	9	6	3	212

Figure 5 (E): CM after TL of InceptionV3

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	82	2	0	0
Early	51	265	6	7
Pre-B ALL	0	7	259	0
Pro-B ALL	0	1	0	216

Figure 5 (F): CM after TL of MobileNetV2

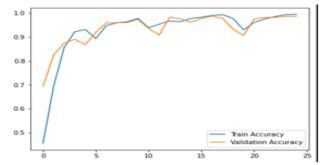
Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

4.6 Confusion matrix (CM) after Ensemble Technique (Based on the number of images)

	Benign	Early	Pre-B ALL	Pro-B ALL
Benign	111	0	0	1
Early	30	274	1	0
Pre-B ALL	0	1	262	0
Pro-B ALL	0	0	0	216

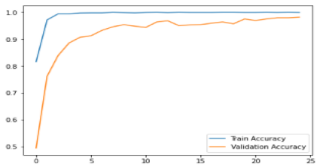
Figure 6: CM after Ensemble Technique of DEX.

4.7 Training and Validation Accuracy and Loss of Original CNN Networks

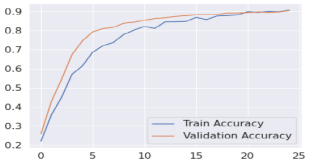


(a) Vgg19

(b) DenseNet-201

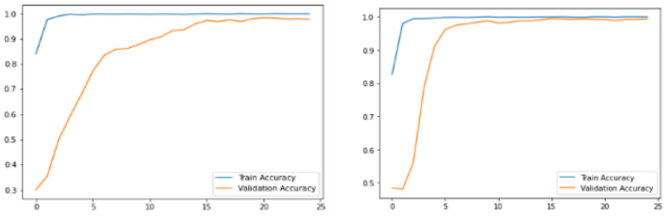


(c) Xception



(d) Seresnet152

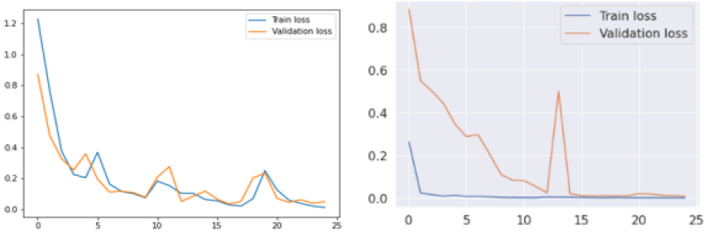
Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801



(e) InceptionV3 (f) MobileNetV2

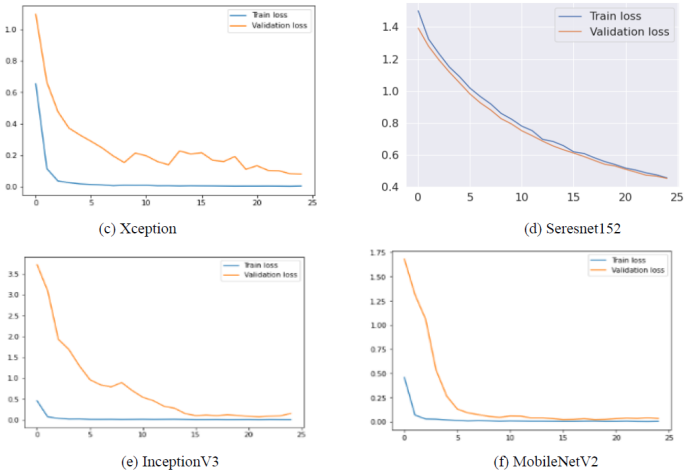
Figure 7 (a), (b), (c), (d), (e), (f): Training and validation accuracy over the epochs (Original CNN Networks).

Figure 7 illustrates the training and validation accuracy of the original model, where the number of epochs is plotted on the x-axis and the accuracy and loss percentages are represented on the y-axis. The training and validation data are appropriately separated in the figure, and there is no over-fitting.



(a) Vgg19 (b) DenseNet-201

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801



(c) Xception (d) Seresnet152

(e) InceptionV3 (f) MobileNetV2

Figure 8 (a), (b), (c), (d), (e), (f): Training and validation loss over the iteration(Original CNN Networks).

Figure 8 depicts the training and validation losses of the original over epochs. To optimize an architecture, CNN employs a loss function. The loss is calculated using training and validation data, and its significance is determined by the model's performance in these two sets. It is the total number of errors committed for each example in each training or validation set. The loss value represents how well or poorly a model performs after each iteration of optimization.

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

4.8 Training and Validation Accuracy and Loss of TL

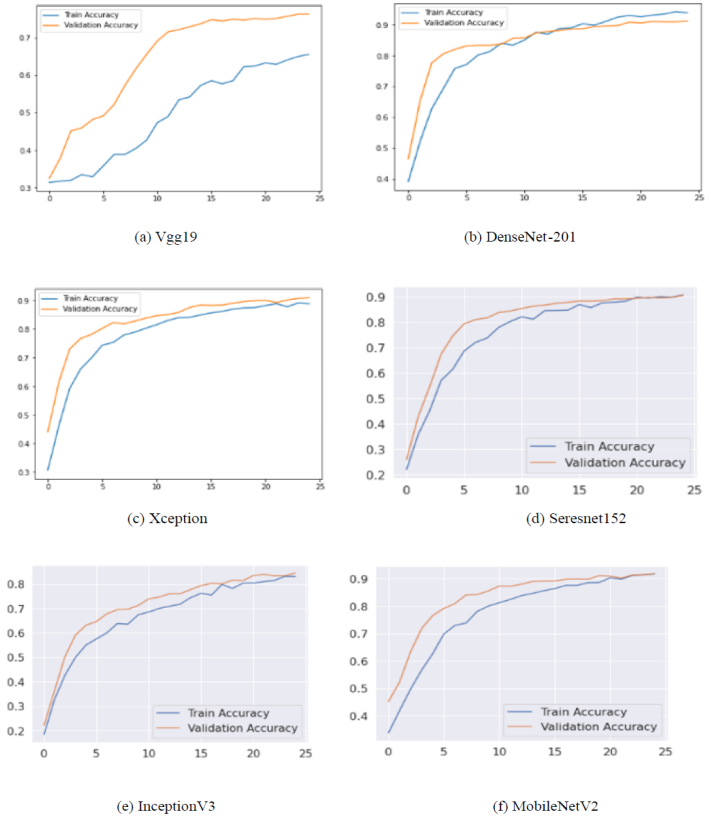


Figure 9 (a), (b), (c), (d), (e), (f): Training and validation accuracy over the epochs (TL CNN Networks).

Figure 9 illustrates the training and validation accuracy of the TL version, where the x-axis represents the number of epochs and the y-axis represents the accuracy and loss values. Within

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

the figure, the training and validation data areas should be segregated, and there should be no over-fitting.

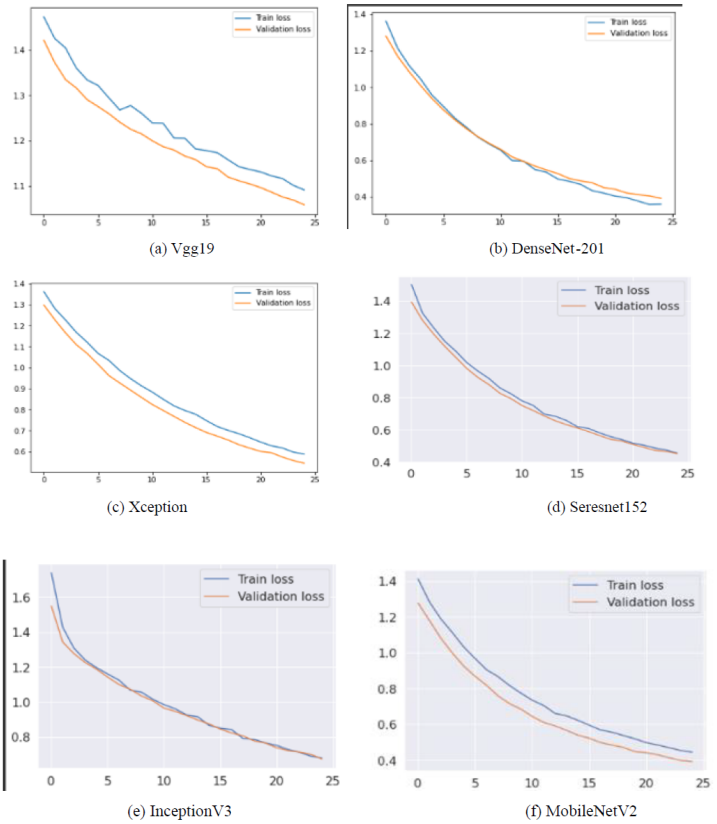


Figure 10 (a), (b), (c), (d), (e), (f): Training and validation loss over the iteration (TL CNN Networks).

The training and validation losses of the TL are shown in Figure 10 over the course of epochs. CNN uses a loss function to

optimize an architecture. The loss was determined using both training and validation data, and the model's overall performance in these sets was used to assess its significance. It is the total number of errors assigned to each instance in each training or validation collection. After each optimization iteration, the loss value represents how well or poorly a model performs.

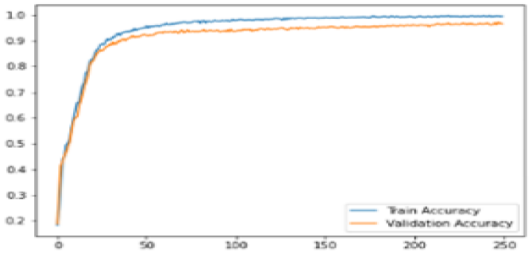


Figure 11 (a): Training and validation accuracy over the epochs (Ensemble Technique).

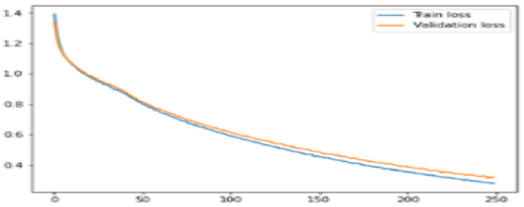


Figure 11 (b): Training and validation loss over the iteration (Ensemble Technique).

5. DISCUSSION

In this work, we compared the performance of original single CNNs with transfer learning ones. Healthcare The SecrensNet152, MobileNet, VGG19, Xception, InceptionV3, and DenseNet-201 architectures were used to categorize four

groups of PBS images. The dataset comprised 3,256 PBS images, of which 2,277 were used for training and 979 were tested after rotation. Our experimental results demonstrate that VGG19, DenseNet-201, and MobileNet-92 performed well in classifying leukemia blood cancer. On small datasets, our results indicate that transfer learning performs slightly better than individual original CNN networks. Transfer learning outperformed the original CNN networks to a certain extent; however, this superiority was only evident after these networks had been trained for extended periods. The transfer learning adopted in this study was pre-training and then feature extraction. Notably, VGG19 was even reduced to 21% in accuracy after transfer learning. Unsupervised training using ImageNet (which includes generic photos of animals) or MURA (which contains X-ray images for different body regions, except the chest) also performed better than not using transfer learning. Not surprisingly, the deep learning ensemble outperformed the individual CNN architectures. This implies that an ensemble model can provide better performance.

6. CONCLUSION AND FUTURE WORK

Leukemia, the blood cancer, continues to be among the significant causes of death from cancer. In recent studies, deep-learning-based systems and transfer learning approaches have achieved high accuracy in diagnosing Leukemia. Yet, research is still being conducted to improve deep learning models. The experiments in this research provide valuable insights for modeling in cases involving small datasets. The proposed model was evaluated and compared to transfer learning and six state-of-the-art CNN architectures. Tests were performed on the train and augmented datasets.

In summary, across all models, the DEX Model (DenseNet121, EfficientNetB7, and Xception) achieved the highest average accuracy and Precision. However, this study has several limitations. We conducted experiments with limited computational resources available at Google Colab. Therefore, no experiments on hyperparameter searching or training with base models that differ from ImageNet are performed, and we do not train such an image model with other optimizers (such as Adadelta, FTRL, NAdam, and others). We also utilized available secondary data from the public domain, rather than primary field data, which may limit the generalizability of our findings.

In the next step, it requires more computational resources, and larger datasets are necessary to achieve better model performance. In this regard, we anticipate that with high-performance computing machinery and increased access to larger-scale data, our model will progressively enhance its overall generalization power and prediction accuracy in the future, primarily by increasing the number of frames per second (fps) and fine-tuning our models using more robust optimization algorithms.

REFERENCES

Rehman, A., Abbas, N., Saba, T., Rahman, S. I. U., Mehmood, Z., & Kolivand, H. (2018). Classification of Acute Lymphoblastic Leukemia using deep learning. Microscopy Research and Technique, 81(11),1310-1317.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern (pp. 2818-2826).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Tan, M., & Le, Q. (2019, May). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on machine learning (pp. 6105-6114). PMLR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16 times 16 words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data- Efficient image transformers & distillation through attention. In International Conference on machine learning (pp. 10347-10357). PMLR.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions To vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22-31).

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., ... & Guo, B. (2022). Cswin transformer: A general Vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12124-12134).

Tran, T., Kwon, O. H., Kwon, K. R., Lee, S. H., & Kang, K. W. (2018, December). Blood cell image Segmentation using deep learning semantic segmentation. In 2018, the IEEE International Conference on Electronics and Communication Engineering (ICECE) (pp. 13-16). IEEE.

Fujita, T. C., Sousa-Pereira, N., Amarante, M. K., & Watanabe, M. A. E. (2021). Acute Lymphoid Leukemia Etiopathogenesis. Molecular Biology Reports, 48(1), 817-822.

Li, L., & Wang, Y. (2020). Recent updates for antibody therapy for acute lymphoblastic Leukemia. hematology & oncology, 9(1), 1-11.

Dharani, T., & Hariprasath, S. (2018, October). Diagnosis of Leukemia and its types using digital image Processing techniques. In 2018, the 3rd International Conference on

Communication and Electronics Systems (ICCES) (pp. 275-279). IEEE.

Ratley, A., Minj, J., & Patre, P. (2020, January). Leukemia disease detection and classification using a Machine learning approach: a review. In 2020, the First International Conference on Power, Control, and Computing Technologies (ICPC2T) (pp. 161-165). IEEE.

Jagadev, P., & Virani, H. G. (2017, May). Detection of Leukemia and its types using image processing and Machine learning. In the 2017 International Conference on Trends in Electronics and Informatics (ICEI) (pp. 522-526). IEEE.

Khosla, E., & Ramesh, D. (2018, March). Phase classification of chronic myeloid Leukemia using. Convolution neural networks. In 2018, the 4th International Conference on Recent Advances in Information Technology (RAIT) (pp. 1-6). IEEE.

Daver, N., Wei, A. H., Pollyea, D. A., Fathi, A. T., Vyas, P., & DiNardo, C. D. (2020). New directions for Emerging therapies in acute myeloid Leukemia: The next chapter. Blood Cancer Journal, 10(10), 1-12.

Kumar, D., Jain, N., Khurana, A., Mittal, S., Satapathy, S. C., Senkerik, R., & Hemanth, J. D. (2020). Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks. IEEE Access, 8, 142521-142531.

Mollah, M. A., Yousufali, M., Ankan, I. M., Rahman, M. M., Sarker, H., & Chakrabarti, K. (2020). Twin-core photonic crystal fiber refractive index sensor for early detection of blood cancer. Sensing and Bio-Sensing Research, 29, 100344.

Sheng, B., Zhou, M., Hu, M., Li, Q., Sun, L., & Wen, Y. (2020). A blood cell dataset for lymphoma Classification using faster R-CNN. Biotechnology & Biotechnological Equipment, 34(1), 413-420.

Rohaziat, N., Tomari, M. R. M., Zakaria, W. N. W., & Othman, N. (2020). White Blood Cells Detection Using YOLOv3 with CNN Feature Extraction Models. International Journal of Advanced Computer Science and Applications, 11(10).

Mehrad Aria, Mustafa Ghaderzadeh, Davood Bashash, Hassan Abolghasemi, Farkhondeh Asadi, and Azamossadat Hosseini, "Acute Lymphoblastic Leukemia (ALL) image dataset." Kaggle (2021).

Wang, S. H., & Zhang, Y. D. (2020). DenseNet-201-based deep neural network with composite Learning Factor and Precomputation for Multiple Sclerosis Classification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(2s), 1-19.

Rahman, W., Faruque, M. G. G., Roksana, K., Sadi, A. S., Rahman, M. M., & Azad, M. M. (2023). Multiclass blood cancer classification using a deep CNN with optimized features. Array, 18, 100292.

Sinha, R., Sinha, K. K., Patel, M., Gupta, S., & Priya, S. (2024, July). Detection of leukemia disease Using a convolutional neural network. In 2024, the 5th International Conference on Image Processing and Capsule Networks (ICIPCN) (pp. 451-456). IEEE.