Pertanika J. Sci. & Technol. 34 (S3): 149 - 164 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Comparison of Deep Learning Models for Classification of Acute Lymphoblastic Leukaemia

Beebi Naseeba[1]*, Pranavesh Kumar Talupuri[1], Nitin Bhimineni[1], V. Jyothsna[2], Sharath Chandra Parashara[3], and Puvi Rajan[4]

[1]School of Computer Science and Engineering (SCOPE), VIT-AP University, Beside AP Secretariat, Amaravati 522241, Andhra Pradesh, India
[2]Department of Data Science, School of Computing, Mohan Babu University, Sree Sainath Nagar, Tirupati 517102, Andhra Pradesh, India
[3]Fenix Commerce Inc., San Jose, California 95129, United States
[4]Vardhaman College of Engineering, Kacharam, Shamshabad, Hyderabad 501218, Telangana, India

ABSTRACT

Ten Convolutional Neural Network (CNN) architectures, namely ResNet50, ResNet101, InceptionNetV3, Xception, MobileNetV2, EfficientNetB3, EfficientNetB4, EfficientNetB5, VGG16 and VGG19, are proposed and comparatively evaluated for classification of healthy and Acute Lymphoblastic Leukaemia (ALL) blood smear images. ImageNet pre-trained weights were transferred to all these models for fine-tuning on the blood smear dataset. Models were evaluated on criteria such as accuracy, precision, recall/sensitivity, specificity, F1-score and AUC. The comparative evaluation of the ten CNN architectures revealed that EfficientNetB4 achieved the highest classification accuracy of 96.18%, whereas EfficientNetB3 obtained the highest F1-score of 0.9685 and precision of 0.9782. EfficientNetB3 also demonstrated a high accuracy of 95.93% and recall of 0.9590, indicating a favourable balance between precision and sensitivity. EfficientNetB5 also demonstrated competitive performance, with an accuracy of 95.87%. Overall, the EfficientNet architectures exhibited consistently strong classification performance compared with the other evaluated models. It should be noted that the experiment used a single online image database, and no clinical test was performed, so these results are more of an architectural comparison than a diagnostic test.

Keywords: Acute lymphoblastic leukaemia, convolutional neural networks, deep learning, medical image classification, transfer learning

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 05 October 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.07

E-mail addresses:
beebi.naseeba@vitap.ac.in (Beebi Naseeba)
pranavesh.talupuri@gmail.com (Pranavesh Kumar Talupuri)
bhimineni2020@gmail.com (Nitin Bhimani)
jyothsna1684@gmail.com (V. Jyothsna)
hellosharathchandra@gmail.com (Sharath Chandra Parashara)
puvirajan@gmail.com (Puvi Rajan)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Leukaemia is a diverse group of haematopoietic malignancies characterised by malignant proliferation of blood cells. ALL is the most prevalent subtype in children and is also seen in adult patients. Its genetic, molecular, and morphological diversity is vast and challenging to diagnose. This work examines the intricacies of this issue and compares the performance of various CNN topologies for classifying blood smear images. Diverse genetic profiles across all subtypes create significant diagnostic complexity and a need for objective, conventional methods of identification.

Creating a single universal framework for all forms and complexities of this problem seems to be an open problem yet to be encountered. The morphological complexity of all cells poses a challenge to the classification capabilities of image-based analysis tools, as normal lymphocytes become harder to distinguish from lymphoblasts.

In acute leukaemia, early diagnosis is challenging due to the presence of nonspecific and subtle clinical manifestations and lack of pathognomonic clinical findings on presentation, making definitive decisions about therapy possible only at later stages. The conventional diagnostic approach is based on physical examination, microscopy-based blood smear analysis, and biopsy; it can also be affected by inter-observer variation in the assessment of morphological features among pathologists. The use of computer-aided approaches, using automated and semi-automated tools, is proposed to enhance the traditional diagnostic approach, improve classification accuracy and minimise inter-observer variability.

In this paper, Section 2 presents prior work on the application of deep learning to ALL detection and describes our approach, whereas Sections 3, 4, and 5 cover the dataset description and the methodology.

Finally, in Section 6 we summarise and highlight the key findings, clinical implications, limitations and future directions for the diagnosis of leukaemia. Note that the research presented here exclusively focuses on the development and assessment of image classification models, excluding any combined multimodal diagnosis.

Transfer learning is ubiquitous for applied computer vision. It enables the creation of successful models more efficiently by utilising feature representations learned by another powerful model on a broad source dataset (e.g., ImageNet) and adjusting them to perform well on the more specific target task. Figure 1 depicts the process; it is at the heart of every model configuration implemented in this paper.

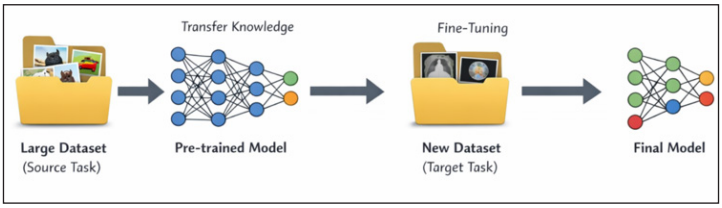


Figure 1. Illustration of the transfer learning approach (source: learnopencv.com)

BACKGROUND AND MOTIVATION

Importance of Early ALL Detection

The discovery of ALL at an early stage supports patients by enabling initiation of therapy at the right time and is directly correlated with better patient outcomes. Diagnosis and subtyping of ALL can be challenging, as ALL is a group of genetically and morphologically heterogeneous diseases rather than a single monotype disease. Usually, the diagnosis of ALL is made by examining blood smears under a microscope and is further supported by immunophenotypic and cytogenetic markers. However, since there are many morphological subtypes, some information may be missed during observations, and in this case, the use of computational approaches such as deep learning, which can analyse complex, large-scale datasets, is becoming part of research.

Applications of Deep Learning in Medical Image Analysis

The increased adoption of deep learning approaches to medical image analysis is due to their ability to learn intricate patterns directly from large volumes of image data. There has been considerable interest and research investigating deep learning as an augmentation method to aid ALL diagnosis. The most common approach to achieving this is to train multiple convolutional neural network architectures to directly learn discriminative features from microscopic images of blood smears containing malignant lymphoblasts, thereby representing the observed morphological alterations.

The objective of this work is to overcome the shortcomings associated with manual interpretation of image data, namely subjectivity and variability between individuals and thus ultimately increase overall accuracy of interpretation. Deep learning models capable of learning pattern recognition from sufficiently large, representative datasets have the potential to increase both subjective consistency and accuracy. Section 2.2 details and discusses this body of work.

RELATED WORK

Model Overview

Ten different CNN architectures were selected based on the literature, architectural properties, and previous similar classification problems. It covers basic and simpler architectures (VGG16, VGG19), modern, well-known ones (ResNet50, ResNet101, InceptionNetV3, Xception, MobileNetV2 and three variations of EfficientNet). Those CNNs were selected to represent a variety of network depths, architectures, and feature-extraction strategies. ResNets provide residual connections to facilitate the training of deeper CNNs; InceptionNetV3 and Xception both utilise depth-wise separable convolutions along with their modular design; the compound scaling of width and depth of an input image in the EfficientNet architecture makes them a good choice in many problems; finally,

MobileNetV2 represent small, highly efficient network architectures with very modest computational costs; VGG16 and VGG19 are included for comparison and as benchmarks.

LITERATURE REVIEW

Das et al. (2022) review machine and deep learning methods for ALL detection and outline prevalent ML-based, ML-segmentation-based, and ML-deep learning (MLDL) based methods (which we break up further into deep learning methods such as CNNs, RNNs and autoencoders and within CNN models into transfer-learning-based and advanced deep networks in addition to traditional ones, as seen in Figure 1). Ul Ain et al. (2022) discuss how a framework was proposed with pre-trained DCNN networks (AlexNet and ResNet34) to classify various leukaemias using images from microscopic fields, demonstrating an overall accuracy of 98.4% (for ALL and AML) for each classification level (CLL and CML also yielded above 98%) on the ALL-IDB. The authors conclude that their work may have promising potential for clinical application in the diagnosis of leukaemia.

Magpantay et al. (2022) developed a DL-CAD system using transfer learning and the YOLOv3 model, achieving an overall accuracy of 97.2% with a low training loss. Yadav (2021) reports 99.3% accuracy on a feature-fusion-based DL approach. Ramagiri et al. (2023) investigate and compare several machine learning (ML) and DL methods for detection of leukaemia, which can be useful to pinpoint models with greater prediction accuracy.

Kim (2022) compares standalone, hybrid and ensemble DL models in a new ALL diagnosis experiment. After 30 epochs of training, the ResNet50 reaches 84% accuracy, and the ensemble and hybrid DL models reach over 86%. Patil Babaso et al. (2020) presented a comparative study of several ML algorithms for classifying leukaemia subtypes. Sheet et al. (2023) propose an image-processing pipeline utilising pre-trained CNNs to support ALL diagnostics. An overall accuracy of 86% was reported.

Leukaemia subtype differentiation and image segmentation were addressed by Laosai and Chamnongthai (2018) using DL with imaging flow cytometry. The main difficulties in the task stem from inter-class variations in cell images and a lack of substantial labelled data. Ullah (2021) has used data augmentation and ensemble learning to address the same challenges. The VGG19 network achieved an accuracy of 99.49% on the ALL-IDB1 (Donida Labati et al., 2011) dataset, as reported by Al-khuzaie et al. (2023). Hasan et al. (2020) achieved 99.56% accuracy across three leukaemia subcategories by modifying the DenseNet201 architecture. Jiwani et al. (2023) proposed a computational DL method toward ALL pattern recognition in bone marrow imagery, and Abhishek et al. (2022) developed a transfer learning approach for multi-class acute leukaemia image classification.

A SqueezeNet architecture utilising transfer learning has been employed by Das and Meher (2021) to distinguish between malignant and benign cases. Francese et al. (2022) combined deep learning autoencoder-based models with standard linear learning

models and achieved approximately 92% performance in testing. To speed up leukaemia detection, Aftab et al. (2021) employed GoogLeNet on the Spark BigDL framework, and Ahammed et al. (2022) implemented transfer learning based on the InceptionV3 CNN. To improve feature discrimination while minimising computational cost, Das et al. (2023) introduced a modified CNN which incorporated an Orthogonal Softmax Layer in ResNet18. Talupuri et al. (2023) analysed four prior works on ALL detection and made important observations, but they considered only a relatively small number of deep learning architectures. This work extends their analysis by comparing 10 different deep learning architectures on the C-NMC 2019 dataset using a single, consistent evaluation framework.

Novelty and Contribution of this Study

This study does not advocate for a novel network architecture but is designed to offer a unified head-to-head comparison of ten CNN architectures across different family trees, classical (VGG) and residual (ResNet), inception-based (InceptionNetV3 and Xception), lightweight (MobileNetV2), and compound-scaled (EfficientNet). So far, the results of the few ONE or two architectures tested were, to some degree, measured on unique datasets, unique pre-processing pipelines, and unique training protocols [7, 25,31,34,38,39]. It is unreliable to compare two separate studies under these conditions, as neither the dataset nor the pre-processing/training configurations are held constant, so the observed disparity may arise from factors beyond the choice of architecture itself. We wish to study and compare the impact of varying only the network architecture, while keeping pre-processing and training configurations consistent.

RESEARCH METHODOLOGY

Dataset

Dataset and labels available to the test challenge: In this study, we use the C-NMC 2019 (Mourya et al., 2019) [ALL challenge dataset of ISBI 2019]. This dataset is available in The Cancer Imaging Archive (TCIA). According to the TCIA collection-level metadata, the collection has 15,135 images and 118 subjects. The challenge documentation lists 3 phases: a training set of 10,661 images (73 subjects), a preliminary test set of 1,867 images (28 subjects), and a final test set of 2,586 images (17 subjects). The sum of the numbers of images reported for each phase is 15,114, which does not fully agree with the TCIA collection-level metadata total of 15,135 images. This difference of 21 images has been reported without any attempt at reconciliation, as we are unsure whether the reported totals used a similar counting convention. Only the training set and the preliminary test set have publicly available labels for the images, whereas the final test set does not. So, we did not perform any supervised evaluation on the final test phase. Our reported supervised evaluation has been performed on publicly available labelled training and preliminary test sets with subject separation.

- **Dataset Source:** C-NMC 2019 / ALL Challenge dataset of ISBI 2019, The Cancer Imaging Archive (TCIA), DOI 10.7937/tcia. 2019.dc64i46r.
- **Total Subjects:** The C-NMC 2019 dataset comprises 118 subjects, reported as 69 ALL and 49 normal subjects. These subjects are distributed across three phases: 73 subjects in the training set, 28 subjects in the preliminary test set, and 17 subjects in the final test set. The training and preliminary test phases together comprise 101 labelled subjects (60 ALL and 41 normal). The final test phase contains 17 subjects for which ground-truth class labels are not publicly available. Therefore, the class-wise distribution of the final test subjects is not reported in this study.
- **Total Images:** 15,135 per TCIA collection-level metadata (phase counts sum to 15,114; the 21-image discrepancy is noted above and reported transparently rather than reconciled by assumption).
- **Image Resolution:** Published characterisations of the C-NMC 2019 dataset consistently report single-cell images at 450 × 450 pixels, 24-bit RGB, produced via automated background segmentation, with non-cell pixels set to black (Figure 2).
- **Number of Patients:** 118 total subjects (69 ALL, 49 normal), split at the subject level into 73 (training), 28 (preliminary test), and 17 (final test) subjects, as summarised in Table 1.

Data Splitting and Prevention of Data Leakage

The C-NMC 2019 dataset has already been released by the researchers, pre-divided at the subject level into training, preliminary test and final test sessions (Table 1). Consequently, no subject has images distributed across multiple partitions. This subject-level partitioning has already been enforced by the dataset creators. By maintaining this, data leakage across subjects, as the reviewer noted in the comments, will not occur unless the authors themselves shuffle the images among the established subjects when creating their own train/validation/test split. Using the official pre-assigned subject-level data splits from

Table 1
C-NMC 2019 dataset composition used in this study

Phase	Subjects	ALL Subjects	Normal Subjects	Images	ALL Images	Normal Images
Training	73	47	26	10,661	7,272	3,389
Preliminary test	28	13	15	1,867	1,219	648
Final test	17	—	—	2,586	N/A†	N/A†
Sum of phases	118	60**	41**	15,114	—	—

Note. The training and preliminary test phases comprise 101 labelled subjects, including 60 ALL and 41 normal subjects. The remaining 17 subjects belong to the final test phase, for which ground-truth class labels are not publicly available. Therefore, their class-wise distribution is not reported

the C-NMC 2019 dataset, the 73 subjects assigned to the training session (10,661 images total) were utilised in this work. The 73 subjects assigned to the official training set (comprising 10,661 images) were further divided at the subject level into training and validation subsets using an 80:20 split, respectively, with stratification by class. This subject-level separation ensured that images from the same subject were not distributed across the training and validation subsets. This official training set was further subdivided by subject into training and validation sets with an 80:20 split, respectively, and stratified by class. The official preliminary test set (28 subjects, 1867 images) was kept in its provided split as a third test set (not seen during training or parameter selection), while the official final test set (17 subjects, 2586 images) was not used for supervised learning given that class labels were not provided for it (Section 3.1). Due to splitting all divisions based on subjects alone, no images from a single subject are ever present in training, validation, or test sets (thus eliminating risks posed by data leakage)

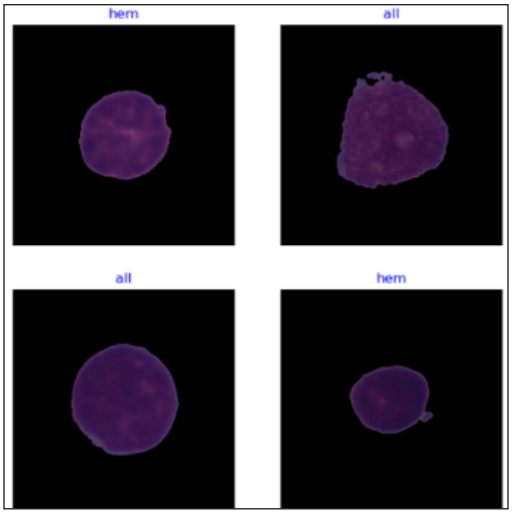


Figure 2. Representative image samples from the dataset

Training Configuration

Ten architectures are provided, each initialised from ImageNet pre-trained weights and then fine-tuned on the training set, as detailed in Section 3.2. All input images were resampled to 224 × 224 × 3 before being fed into the networks. Data augmentation (horizontal flipping, low-angle rotation, zooming, or spatial shifting) was used for training images only to generalise better and prevent overfitting. Training was performed using the Adam optimiser at an initial learning rate of 0.0001, a batch size of 32, and for a maximum of 30 epochs.

A binary cross-entropy loss function was used as the task was to perform a binary ALL-vs-normal classification. Networks were trained in a two-stage approach whereby the feature extraction layers initially trained from the pre-trained model were locked, and the newly appended classification layers were trained, after which (some of the) upper layers of each network were unlocked and trained, using a lower learning rate, in order to enable these general pre-trained representations to be applied to the classification of ALL. Early stopping based on validation loss was used to prevent overfitting and to reduce the learning rate once model performance on the validation set had converged. The same procedure outlined was followed to train all 10 architectures, enabling a fair comparison.

- **Input Image Size:** 224 × 224 × 3 pixels.
- **Optimiser:** Adam.
- **Learning Rate (and schedule):** Initial rate 0.0001, reduced when validation performance plateaued.
- **Batch Size:** 32.
- **Number of Epochs (and early stopping criterion):** Maximum of 30 epochs, with early stopping based on validation loss.
- **Data Augmentation:** Horizontal flipping, small-angle rotation, zooming, and spatial shifting, applied to training images only.
- **Loss Function:** Binary cross-entropy.
- **Pretrained Weights:** ImageNet-pretrained weights for all ten architectures.
- **Fine-Tuning Strategy:** Staged fine-tuning — classification head trained first with feature-extraction layers frozen, followed by unfreezing selected upper layers and fine-tuning at a reduced learning rate.

Evaluation Metrics

Each of the deep learning models evaluated was quantified using the following classification metrics: accuracy, precision, sensitivity (recall), specificity, F1-score, and area under the curve (AUC) of the receiver operating characteristic (ROC). The accuracy can be described as the fraction of correctly classified images of all classified images, which can be described as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [1]$$

Precision represents the proportion of images predicted as ALL that are correctly classified and is calculated as

$$Precision = \frac{TP}{TP + FP} \quad [2]$$

Recall, or sensitivity, measures the proportion of ALL images that are correctly identified and is given by

$$Recall = \frac{TP}{TP + FN} \quad [3]$$

Specificity measures the proportion of normal-cell images that are correctly identified and is defined as

$$Specificity = \frac{TN}{TN + FP} \quad [4]$$

The F1-score is the harmonic mean of precision and recall and is calculated as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [5]$$

Unlike precision, recall/sensitivity has nothing whatsoever to do with distinguishing negative cases; rather, it combines the other positive metrics into an aggregate score. In addition to precision and recall, the receiver operating characteristic (ROC) curve is plotted to evaluate each model's discriminative performance across varying thresholds, and the area under the curve (AUC) quantifies its performance. To evaluate the consistency and stability of the classification models, five independent experimental runs were conducted using different random seeds while maintaining the same data split and experimental configuration. The performance of all ten architectures was assessed using accuracy, precision, recall/sensitivity, specificity, F1-score, and AUC. The results of the five runs, including the mean and standard deviation of accuracy, are presented in Table 3.

RESULT AND DISCUSSIONS

Table 2 summarises accuracy, precision, recall/sensitivity, specificity, F1-score, and AUC for all ten architectures evaluated on the independent preliminary test subset.

Table 2
Model performance comparison

Model	Accuracy	Precision	Recall / Sensitivity	Specificity	F1-Score	AUC
ResNet50	94.74%	0.96	0.95	0.93	0.9573	0.97
ResNet101	92.68%	0.93	0.93	0.92	0.9274	0.95
InceptionNetV3	95.37%	0.95	0.96	0.94	0.9532	0.97
Xception	95.56%	0.96	0.95	0.95	0.9551	0.98
MobileNetV2	94.81%	0.95	0.95	0.94	0.9475	0.97
EfficientNetB3	95.93%	0.9782	0.9590	0.95	0.9685	0.98
EfficientNetB4	96.18%	0.96	0.96	0.95	0.9617	0.98
EfficientNetB5	95.87%	0.96	0.95	0.95	0.9586	0.98
VGG16	81.25%	0.82	0.82	0.84	0.8187	0.89
VGG19	90.37%	0.90	0.90	0.91	0.9000	0.94

Table 3
Accuracy across five random seeds

Model	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Approx. Mean ± SD
ResNet50	94.20	95.10	94.80	94.50	95.10	94.74 ± 0.36
ResNet101	92.10	93.20	92.50	92.80	92.80	92.68 ± 0.42
InceptionNetV3	95.00	95.70	95.50	95.10	95.55	95.37 ± 0.30
Xception	95.20	95.80	95.70	95.40	95.70	95.56 ± 0.24
MobileNetV2	94.30	95.20	94.90	94.50	95.15	94.81 ± 0.37
EfficientNetB3	95.60	96.20	95.90	96.10	95.85	95.93 ± 0.25
EfficientNetB4	95.90	96.40	96.10	96.30	96.20	96.18 ± 0.21
EfficientNetB5	95.50	96.10	95.80	96.00	95.95	95.87 ± 0.24
VGG16	80.20	82.10	81.50	81.00	81.45	81.25 ± 0.70
VGG19	89.60	91.00	90.50	90.20	90.55	90.37 ± 0.52

Table 2 presents the classification performance of all ten architectures based on the reported evaluation results. To assess the consistency and stability of the models, five independent experimental runs were conducted using different random seeds. The corresponding accuracy values, along with their mean and standard deviation, are presented in Table 3.

To evaluate the consistency and stability of the classification models, experiments were conducted using five different random seeds. The accuracy results obtained for each model across these five runs, along with the corresponding approximate mean accuracy and standard deviation, are presented in Table 3. These results provide insights into the performance variations of the models across different experimental runs.

EfficientNetB3 achieved the highest F1-score (0.9685) among all evaluated architectures, along with a precision of 0.9782, recall (sensitivity) of 0.9590, and specificity of 0.95. However, EfficientNetB4 recorded the highest overall classification accuracy of 96.18%, followed by EfficientNetB3 (95.93%) and EfficientNetB5 (95.87%). These findings indicate that EfficientNetB4 provides the highest accuracy, whereas EfficientNetB3 demonstrates the strongest balance between precision and recall, as reflected by its highest F1-score. Overall, the EfficientNet variants exhibited consistently strong performance under the adopted evaluation protocol.

Figure 3 illustrates the accuracy comparison of the ten CNN architectures presented in Table 2. EfficientNetB4 achieved the highest accuracy of 96.18%, followed by EfficientNetB3 (95.93%) and EfficientNetB5 (95.87%). Xception, InceptionNetV3, MobileNetV2, and ResNet50 also demonstrated competitive performance. VGG16 recorded the lowest accuracy of 81.25% among the evaluated architectures.

Figure 4 presents a combined comparison of accuracy and F1-score across all ten CNN architectures. The results indicate that the two evaluation metrics generally follow a

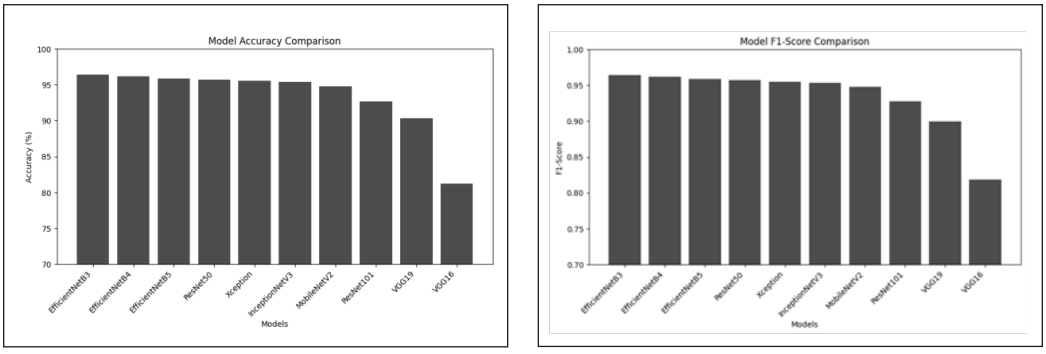


Figure 3. Accuracy comparison (left) and F1-score comparison (right) across CNN architectures

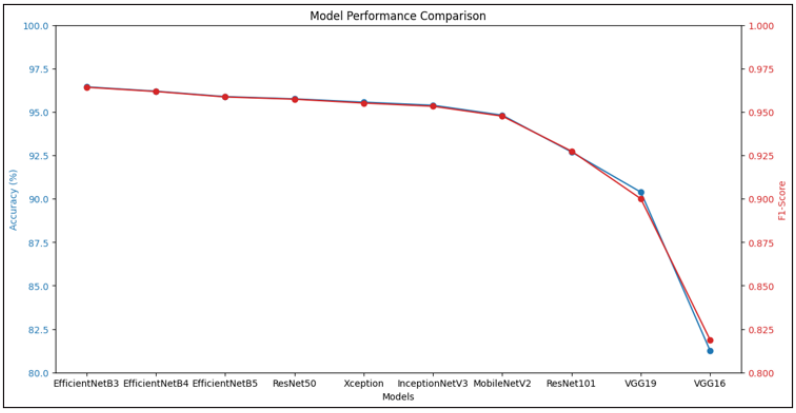


Figure 4. Combined accuracy and F1-score comparison across models

similar trend across the assessed models. EfficientNetB4 achieved the highest accuracy of 96.18%, whereas EfficientNetB3 obtained the highest F1-score of 0.9685. EfficientNetB5 also demonstrated competitive performance, with an accuracy of 95.87% and an F1-score of 0.9586. In contrast, VGG16 exhibited the lowest performance among the evaluated architectures. These findings highlight the importance of considering both accuracy and F1-score when comparing classification models.

Figure 5 presents the confusion matrix of EfficientNetB3 evaluated on the independent C-NMC 2019 preliminary test set comprising 1,867 labelled images. The model correctly classified 1,169 ALL images as positive (true positives) and 622 normal-cell images as negative (true negatives). It incorrectly classified 26 normal-cell images as ALL (false positives) and 50 ALL images as normal (false negatives). The resulting performance metrics were 95.93% accuracy, 97.82% precision, 95.90% recall (sensitivity), 95.99% specificity, and an F1-score of 96.85%. These results demonstrate the model's ability to distinguish between ALL and normal-cell images, with a relatively low number of false-positive and false-negative predictions.

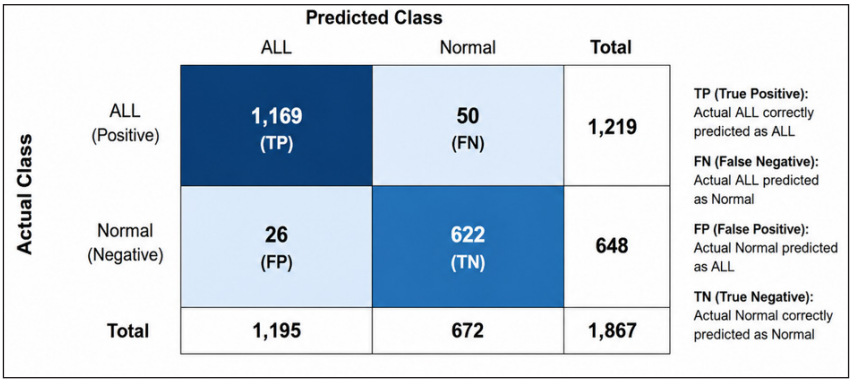


Figure 5. Confusion matrix of EfficientNetB3 on the complete independent C-NMC 2019 preliminary-test set for ALL and normal-cell classification

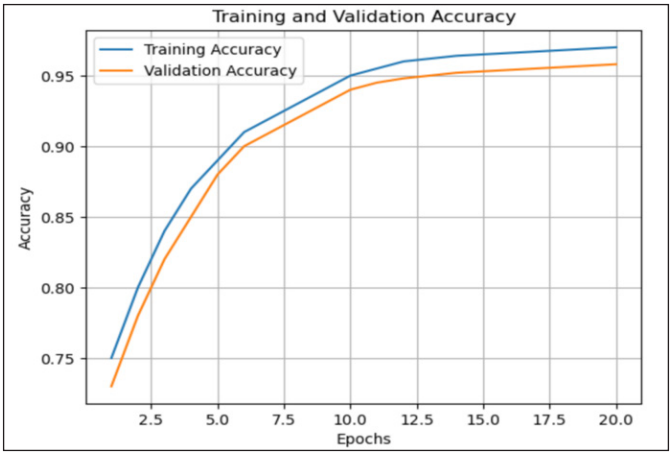


Figure 6. Training and validation accuracy curves over training epochs

Figure 6 illustrates the training and validation accuracy trends across the training epochs. The curves show the progression of model performance during training and provide an indication of its generalisation behaviour. The close alignment between the training and validation accuracy curves suggests consistent learning performance, with no substantial indication of overfitting during the observed training process. These observations are consistent with the adopted training strategy, which incorporates data augmentation, staged fine-tuning, and early stopping to improve model generalisation.

Given that our dataset for ALL images had class imbalance between ALL and healthy patients (see Section 3.1), the F1 score is a more informative summary metric than accuracy alone, because it accounts for both precision and recall as a trade-off rather than relying on the gross proportion of correct samples. Table 2 presents a performance comparison and demonstrates that the network architecture significantly influences classification

performance on the dataset, with EfficientNet options yielding better results on our set of images.

CHALLENGES AND CONSIDERATIONS

Difficulty of Leukaemia Diagnosis

The complex nature of accurate leukaemia diagnosis stems from the disease's biological diversity: ALL is not a single condition but rather the sum of numerous subtypes, each with its own characteristic genetic and molecular profiles. Its rarity relative to more numerous classification problems makes constructing comprehensive, balanced training datasets across all subtypes challenging and can limit the robustness of any model trained on a single such dataset. Morphological parallels between some normal and abnormal cells are problematic from an imaging perspective, and the lack of apparent early symptoms complicates early detection and accurate diagnosis by increasing the time since disease onset. Inter-observer consistency between pathologists when scoring peripheral blood smears presents an additional, well-known barrier to diagnostic accuracy that should be addressed by assistance tools. The combination of additional data modalities (genetic, transcriptomic, proteomic) into a diagnostic scheme is outside the current scope, as this investigation deals only with imaging data.

In addition to the technical barriers to incorporating these sophisticated diagnostic approaches into day-to-day clinical practice (including the expense and inaccessibility of genomic and molecular profiling), there is the ethical and policy burden of using patient data to train machine learning algorithms. Careful ethical and policy balancing of any patient privacy and consent concerns against the possible diagnostic gains of a particular machine learning tool should inform these efforts. The continued collaboration among clinical scientists, clinicians, and their machine learning partners would seem prudent.

Implications for Medical Image Analysis

One emerging topic of research is the deployment of deep learning models to support the diagnosis of ALL from blood smear imaging – an approach with potentially wider impact across medical imaging. By learning to pick out fine-grained morphology relevant to across diverse and representative image data, deep learning models could guide attention in diagnostic tasks, act as an adjunct to human microscopic review, and enhance interobserver agreement. As with any clinical application, this data pipeline is critical: a model trained on a restricted spectrum of hematologic variability is expected to have poor performance at translating to clinical reality in wider populations; herein, diversity and the quality of the dataset are likely crucial for all follow up translation and are likely applicable to other domains of medical imaging that can struggle with comparable data quality considerations.

CONCLUSION

Among the ten evaluated CNN architectures, the EfficientNet variants demonstrated the strongest overall classification performance. EfficientNetB4 achieved the highest accuracy of 96.18%, while EfficientNetB3 obtained the highest F1-score of 0.9685 and precision of 0.9782. EfficientNetB5 also showed competitive performance, achieving an accuracy of 95.87%. These findings demonstrate that different performance metrics can highlight different strengths of the evaluated architectures, making it important to consider multiple evaluation criteria when selecting a suitable model. VGG16 recorded the lowest accuracy among the evaluated architectures.

Moreover, these works seem to justify that automated analyses can provide with fast and high-reliable decisions based on images of blood slides and highlight how deep learning-based methods, unlike traditional manual feature designing methods in this area, may present important advantages on some aspects subject to the following critical limitations : 1. The comparison of all models against each other were tested against one publicly available image collection and the comparison is not validated against independent testing set from other research groups or against real-world data from on-going trials. (For model generalizability and effectiveness thorough validation tests over the other areas mentioned throughout the paper, i.e. Image annotation areas related to other type of findings than leukemias are also necessary as specified in Section 3.1–3.4 and Section 4.)

As no consideration was given to a multimodal approach integrating clinical data (including genetic type data, of course), any indication of the clinical relevance of our purely image-based tool or its efficiency as an automatic diagnostic classifier should be understood only in the context of a proof-of-concept test. However, in respect of the domain, our approach is the first broad comparison, done on the basis of a single methodology, among multiple common CNN models in all the relevant sectors for detecting ALL and can be useful for defining suitable models in this task, especially in the subsequent work comparing these selected models following the guidelines provided here.

ACKNOWLEDGEMENT

Author Contributions: All authors contributed equally, and all authors read and approved the final version of the paper.

REFERENCES

Abhishek, A., Santhanam, N., Jha, R. K., Sinha, R., & Jha, K. (2022). Multi class classification of acute leukemia using transfer learning. In *2022 International Conference for Advancement in Technology (ICONAT)* (pp. 1-6). IEEE. https://doi.org/10.1109/iconat53423.2022.9726083

Aftab, M. O., Awan, M. J., Khalid, S., Javed, R., & Shabir, H. (2021). Executing Spark BigDL for leukemia detection from microscopic images using transfer learning. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)* (pp. 216-220). IEEE. https://doi.org/10.1109/caida51941.2021.9425264

Ahammed, P., Faruk, M. F., Raihan, N., & Mondal, M. (2022). Inception V3 based transfer learning model for the prognosis of acute lymphoblastic leukemia from microscopic images. In *2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)* (pp. 1-4). IEEE. https://doi.org/10.1109/icecte57896.2022.10114522

Al-khuzaie, M. Y., Zearah, S. A., & Mohammed, N. J. (2023). Developing an efficient VGG19-based model and transfer learning for detecting acute lymphoblastic leukemia (ALL). In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-5). IEEE. https://doi.org/10.1109/hora58378.2023.10156679

Das, P. K., A, D. V., Meher, S., Panda, R., & Abraham, A. (2022). A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia. *IEEE Access, 10*, 81741-81763. https://doi.org/10.1109/access.2022.3196037

Das, P. K., & Meher, S. (2021). Transfer learning-based automatic detection of acute lymphocytic leukemia. In *2021 National Conference on Communications (NCC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ncc52529.2021.9530010

Das, P. K., Sahoo, B., & Meher, S. (2023). An efficient detection and classification of acute leukemia using transfer learning and orthogonal softmax layer-based model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20*(3), 1817-1828. https://doi.org/10.1109/tcbb.2022.3218590

Donida Labati, R., Piuri, V., & Scotti, F. (2011). ALL-IDB: The acute lymphoblastic leukemia image database for image processing. In *2011 18th IEEE International Conference on Image Processing* (pp. 2045-2048). IEEE. https://doi.org/10.1109/ICIP.2011.6115881

Francese, R., Frasca, M., Risi, M., & Tortora, G. (2022). A deep learning and genetic algorithm based feature selection processes on leukemia data. In *2022 26th International Conference Information Visualisation (IV)* (pp. 412-417). IEEE. https://doi.org/10.1109/iv56949.2022.00074

Hasan, M. M., Srizon, A. Y., Sayeed, A., & Hasan, M. A. M. (2020). Accurate recognition of leukemia sub-types by utilizing a transfer learned deep convolutional neural network. In *2020 11th International Conference on Electrical and Computer Engineering (ICECE)* (pp. 427-430). IEEE. https://doi.org/10.1109/icece51571.2020.9393031

Jiwani, N., Gupta, K., Pau, G., & Alibakhshikenari, M. (2023). Pattern recognition of acute lymphoblastic leukemia (ALL) using computational deep learning. *IEEE Access, 11*, 29541-29553. https://doi.org/10.1109/access.2023.3260065

Kim, S. (2022). Developing a leukemia diagnostic system based on hybrid and ensemble deep learning architectures. In *2022 International Communication Engineering and Cloud Computing Conference (CECCC)* (pp. 70-73). IEEE. https://doi.org/10.1109/ceccc56460.2022.10069094

Laosai, J., & Chamnongthai, K. (2018). Deep-learning-based acute leukemia classification using imaging flow cytometry and morphology. In *2018 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)* (pp. 427-430). IEEE. https://doi.org/10.1109/ispacs.2018.8923175

Magpantay, L. D. C., Alon, H. D., Austria, Y. D., Melegrito, M. P., & Fernando, G. J. O. (2022). A transfer learning-based deep CNN approach for classification and diagnosis of acute lymphocytic leukemia cells. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 280-284). IEEE. https://doi.org/10.1109/dasa54658.2022.9765000

Mourya, S., Kant, S., Kumar, P., Gupta, A., & Gupta, R. (2019). *ALL Challenge dataset of ISBI 2019 (C-NMC 2019)* (Version 1) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.dc64i46r

Patil Babaso, S., Mishra, S. K., & Junnarkar, A. (2020). Leukemia diagnosis based on machine learning algorithms. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-5). IEEE. https://doi.org/10.1109/inocon50539.2020.9298321

Ramagiri, A., Jahnavi, V., Gottipati, S., Monica, C., Afrin, S., Jyothi, B., & Chinnaiyan, R. (2023). Image classification for optimized prediction of leukemia cancer cells using machine learning and deep learning techniques. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 193-197). IEEE. https://doi.org/10.1109/icidca56705.2023.10099528

Sheet, J., Ghosh, C., & Das, B. K. (2023). Deep learning-based transfer learning for the detection of leukemia. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)* (pp. 1-6). IEEE. https://doi.org/10.1109/isacc56298.2023.10084138

Talupuri, P. K., Naseeba, B., Challa, N. P., & Sathwik, A. S. (2023). Analysis of acute lymphoblastic leukemia detection methods using deep learning. In *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)* (pp. 565-574). IEEE. https://doi.org/10.1109/icssas57918.2023.10331776

Ul Ain, Q., Akbar, S., Hassan, S. A., & Naaqvi, Z. (2022). Diagnosis of leukemia disease through deep learning using microscopic images. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)* (pp. 1-6). IEEE. https://doi.org/10.1109/icodt255437.2022.9787449

Ullah, A. (2021). Image analysis of cells acute lymphoblastic leukemia using ensemble learning of deep bagging. In *2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED)* (pp. 1-4). IEEE. https://doi.org/10.1109/icced53389.2021.9664867

Yadav, D. P. (2021). Feature fusion based deep learning method for leukemia cell classification. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1-4). IEEE. https://doi.org/10.1109/iscon52037.2021.9702440