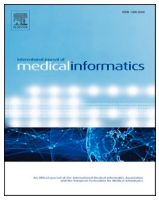
International Journal of Medical Informatics 203 (2025) 106023

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Enhancing leukemia detection in medical imaging using deep transfer learning

Afeez A. Soladoye [a], David B. Olawade [b,c,d,e,*], Ibrahim A. Adeyanju [a], Temitope Adereni [f], Kazeem M. Olagunju [g], Aanuoluwapo Clement David-Olawade [h]

[a] Department of Computer Engineering, Federal University, Oye-Ekiti, Nigeria
[b] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom
[c] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, United Kingdom
[d] Department of Public Health, York St John University, London, United Kingdom
[e] School of Health and Care Management, Arden University, Arden House, Middlemarch Park, Coventry CV3 4FJ, United Kingdom
[f] Department of Public Health, University of Dundee, Dundee DD1 4HN, United Kingdom
[g] Department of Computer Science, Landmark University, Omu-Aran, Nigeria
[h] Endoscopy Unit, Glenfield Hospital, University Hospitals of Leicester NHS Trust, Leicester LE3 9QP, United Kingdom

ARTICLE INFO

Keywords:
Acute lymphoblastic leukemia
Deep transfer learning
EfficientNet-B3
Medical image classification
Cancer

ABSTRACT

Background: Acute Lymphoblastic Leukemia (ALL) is the most common pediatric cancer, requiring early detection to save lives and reduce the financial burden of advanced-stage treatment. While traditional diagnostic methods are time-consuming and resource-intensive, deep transfer learning offers a computationally efficient alternative for medical image classification.

Method: This study employed two widely recognized transfer learning algorithms, VGG-19 and EfficientNet-B3, to detect ALL using a publicly available dataset of 10,661 images from 118 patients. Data preprocessing included resizing, augmentation, and normalization. The models were trained for 100 epochs, with batch sizes of 30 for VGG-19 and 32 for EfficientNet-B3. Evaluation metrics such as accuracy, precision, recall, and F1 score were used to assess model performance. Statistical significance testing was performed using paired t-tests ($p < 0.05$). Comparative analysis was performed with existing studies to validate the findings.

Results: EfficientNet-B3 significantly outperformed VGG-19, achieving an average accuracy of 96 % compared to 80 % for VGG-19 ($p < 0.001$). EfficientNet-B3 demonstrated superior performance in handling class imbalance, with the minority class (Hem) achieving precision, recall, and F1 scores of 97 %, 89 %, and 93 %, respectively. VGG-19 struggled with the minority class, achieving lower recall (51 %) and F1 score (62 %). However, dataset limitations including single-source origin may affect generalizability.

Conclusion: This study highlights the effectiveness of EfficientNet-B3 as a reliable tool for early ALL detection, offering high accuracy and computational efficiency. Clinical implementation requires addressing computational constraints and integration challenges. Future research could integrate multimodal datasets to identify risk factors and further improve diagnostic accuracy.

1. Introduction

Cancer has become one of the most concerning health issues globally, affecting both children and adults at an alarming rate [1]. Among the various types of cancer, leukemia stands out as a particularly devastating disease due to its rapid progression and high mortality rate [2]. Leukemia, known as cancer of the blood, is defined as the uncontrollable growth of immature and dysfunctional leukocytes, which disrupt the functionality of the bone marrow [2]. This abnormal growth affects not only the blood but also spreads to neighboring organs and, ultimately, the entire body system [2]. The severity of leukemia and its life-threatening outcomes underscore the critical need for early diagnosis and effective treatment plans to mitigate its impact [2].

Acute lymphoblastic leukemia (ALL), the most common form of leukemia in children, accounts for approximately 25 % of all pediatric cancers [3]. Without prompt diagnosis and treatment, the survival rate

* Corresponding author at: Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom.
E-mail address: d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.ijmedinf.2025.106023
Received 26 May 2025; Received in revised form 19 June 2025; Accepted 20 June 2025
Available online 26 June 2025

for patients with acute leukemia is alarmingly low [3]. However, the process of diagnosing cancer, including leukemia, has traditionally relied on manual inspection of medical images under a microscope [4]. This conventional method, though effective in some cases, is time-consuming and heavily dependent on the expertise of oncologists [4]. The challenges associated with this approach are further compounded by the shortage of trained medical professionals in many regions, particularly in Africa, where a significant number of doctors have migrated to Asia and Europe in search of better opportunities [4,5]. These limitations highlight the urgent need for innovative, faster, and more accurate diagnostic techniques [5].

The advent of artificial intelligence (AI) and its integration into clinical decision support systems have revolutionized the healthcare industry [6]. Deep learning, a subset of AI, has shown immense potential in the early prediction and diagnosis of various diseases, including cancer [7]. By leveraging both risk factors and medical images, deep learning algorithms can quickly and accurately detect abnormalities, making it a powerful tool in modern medicine. Furthermore, deep transfer learning, a specialized branch of deep learning, has gained widespread acceptance in cancer-related applications [8]. Studies have demonstrated its efficacy in detecting cervical cancer and oral cancer [9,10], among other types, due to its ability to analyze and classify tumor images with remarkable precision.

Several studies have explored the use of deep learning and transfer learning for leukemia diagnosis. For instance, a study employed the YOLO v2 algorithm combined with convolutional neural networks to detect and classify white blood cells in leukemia, achieving 96 % average precision in detection and 94.3 % accuracy in classification, showcasing the potential of deep learning approaches in clinical support systems [11]. Similarly, a study conducted a study on blood cancer prediction using leukemia microarray gene data [12]. Their dataset comprised 22,283 genes, and they used ADASYN for dataset balancing and the Chi-squared technique for feature selection [12]. The hybrid logistic vector trees classifier employed in their research outperformed other methods, further validating the effectiveness of AI in cancer prediction [12]. Also, a study utilized convolutional neural networks (CNN) for the early prediction of blood cancer, demonstrating the capability of CNNs to differentiate cancerous blood cells from normal ones in medical images efficiently [13].

While existing studies have made significant progress in applying deep learning to leukemia detection, several critical gaps remain. Previous works have primarily focused on single-model approaches without comprehensive comparative analysis of state-of-the-art transfer learning architectures. Additionally, many studies have not adequately addressed class imbalance issues that are inherent in medical datasets, nor have they provided detailed architectural modifications for optimal performance. Furthermore, limited attention has been given to the practical implementation challenges and computational efficiency required for clinical deployment.

This study aims to bridge these gaps by exploring the application of deep transfer learning models for leukemia prediction in medical imaging. Specifically, this research provides: (1) a comprehensive comparative analysis between VGG-19 and EfficientNet-B3 architectures, (2) detailed architectural modifications optimized for medical image classification, (3) robust evaluation of class imbalance handling capabilities, and (4) assessment of computational efficiency for potential clinical deployment. By building on existing research and utilizing state-of-the-art techniques, this study seeks to contribute to the growing field of AI-driven healthcare solutions and pave the way for more effective diagnostic tools in the fight against leukemia.

2. Methodology

This study employed a systematic approach for the detection of Acute Lymphoblastic Leukemia (ALL) using deep transfer learning techniques. The methodology was built on a robust and efficient machine learning framework, structured into four primary stages: data acquisition, data preprocessing, model training and classification, and performance evaluation. Each stage was carefully designed to ensure accuracy, efficiency, and reproducibility in the prediction of leukemia from medical images.

Fig. 1 presents a schematic diagram of this methodology, providing a clearer representation and illustration of the workflow. This diagram outlines the step-by-step processes involved, beginning with the collection of relevant datasets and concluding with the evaluation of the model's performance.

2.1. Data acquisition

This study utilized the Daputa and Daputa (2019) ALL Challenge Dataset for the detection of Acute Lymphoblastic Leukemia (ALL). The dataset was sourced from Kaggle, a widely recognized platform for hosting high-quality datasets and machine learning challenges. The dataset comprises a total of 15,135 images collected from 118 patients, with the images labeled into two distinct categories: normal and Leukemia blast classes. Given the substantial volume of images in the dataset, the computational efficiency of the model had to be carefully managed. To address this, the study focused on the training split of the dataset, which was further divided into training, validation, and testing subsets. This decision led to the exclusion of 4,474 images from the original dataset to ensure computational feasibility while maintaining representative samples across both classes.

The excluded 4,474 images maintained similar class distribution patterns as the included subset, with approximately 68 % belonging to the ALL class and 32 % to the Hem class, ensuring no significant bias was introduced through data exclusion. This exclusion was necessary due to computational constraints but may limit the model's exposure to the full spectrum of image variations present in the complete dataset.

The revised dataset was distributed across the three subsets using a 70–15-15 % split as follows:

- Training Set: 7,462 images (70 %) were used to train the model, ensuring it could learn to distinguish between normal and Leukemia blast classes effectively.
- Validation Set: 1,599 images (15 %) were utilized during the training process to fine-tune the model's parameters and prevent overfitting.
- Testing Set: The remaining 1,600 images (15 %) were reserved for evaluating the final performance of the trained model.

This structured division of the dataset ensured a balanced distribution, maintaining the integrity of the training process while optimizing the computational requirements for implementing deep transfer learning techniques. By employing this dataset, the study aimed to leverage its rich collection of labeled medical images to achieve robust and accurate detection of Acute Lymphoblastic Leukemia.

2.2. Data preprocessing

Data preprocessing was standardized across both models to ensure fair comparison. With the acquisition and modification of the dataset as earlier discussed, preprocessing was an essential step to ensure the dataset was in the appropriate format for implementation. Specific preprocessing steps included:

1. Image Resizing: All images were resized to 224 × 224 pixels using bilinear interpolation to maintain aspect ratios while ensuring computational efficiency.

2. Normalization: Pixel values were normalized to the range [0, 1] by dividing by 255, following ImageNet preprocessing standards.

3. Data Augmentation: To increase dataset diversity and improve model robustness, the following augmentation techniques were applied during training:

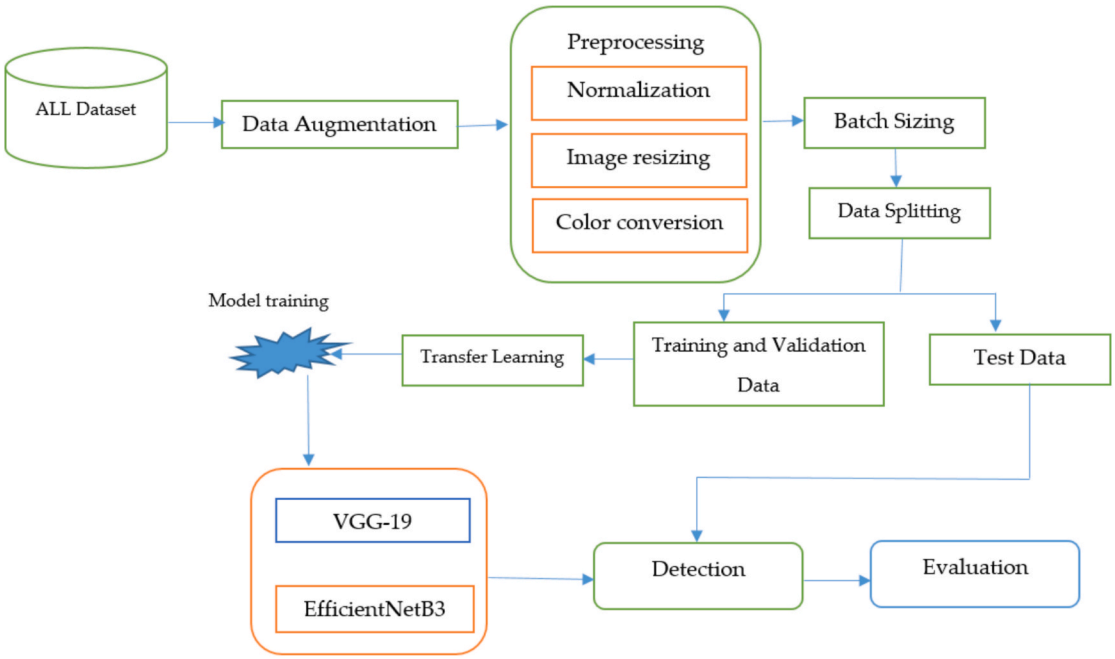


Fig. 1. Overview of research methodology.

- Horizontal flipping with 50 % probability
- Rotation range of ± 15 degrees
- Width and height shift range of 0.1
- Zoom range of 0.1

4. Color Space Conversion: All images were converted to RGB color mode to ensure uniformity and compatibility with pre-trained model requirements

These preprocessing steps were systematically applied to optimize dataset quality while maintaining consistency across training, validation, and testing phases.

2.3. Deep transfer learning techniques

Deep learning algorithms, particularly Convolutional Neural Networks (CNNs), are well-regarded for their exceptional performance in object detection, segmentation, and image classification tasks. However, designing and training CNNs from scratch can be computationally expensive and time-consuming. To overcome these challenges, this study leveraged deep transfer learning, which involves using pre-trained neural networks as a foundation for new tasks.

The choice of VGG-19 and EfficientNet-B3 was based on their complementary characteristics: VGG-19 represents a classical deep architecture with proven stability, while EfficientNet-B3 embodies modern efficient scaling principles. This comparison allows for evaluation of both traditional and contemporary approaches to transfer learning in medical imaging.

2.3.1. EfficientNet-B3 architecture

EfficientNet-B3, a member of the EfficientNet family, was chosen for its ability to achieve improved accuracy while requiring fewer computational resources. The model employs compound scaling that uniformly scales network depth, width, and resolution using the following equations:

$$depth = \alpha\varphi \quad (1)$$

$$width = \beta\varphi \quad (2)$$

$$resolution = \gamma\varphi \quad (3)$$

α, β, γ are constants while φ is the scaling coefficient

2.3.2. Hyperparameter selection and tuning

Hyperparameter selection was performed through systematic grid search on the validation set:

– Learning rates tested: [0.001, 0.0001, 0.00001] – Batch sizes evaluated: [16, 30, 32, 64] – Optimizers compared: [Adam, Adamax, RMSprop] – Dropout rates: [0.3, 0.5, 0.7].

The final hyperparameters were selected based on validation performance, with EfficientNet-B3 using batch size 32, learning rate 0.001, and Adamax optimizer, while VGG-19 used batch size 30 with the same learning rate and optimizer.

2.4. Model architecture modifications

2.4.1. VGG-19 modifications

To adapt the pre-trained VGG-19 model for binary classification, the following modifications were implemented:

I. Feature Extraction Layers: The base VGG-19 layers were frozen to preserve pre-trained ImageNet features
II. Global Max Pooling Layer: Added to reduce spatial dimensions from ($7 \times 7 \times 512$) to (512,)
III. Dropout Layer: Implemented with rate 0.5 for regularization
IV. Dense Output Layer: 2 units with sigmoid activation for binary classification
V. Compilation: Adamax optimizer, categorical crossentropy loss, learning rate 0.001

The architecture of the enhanced VGG-19 model is presented in Fig. 2, illustrating the seamless incorporation of these improvements into the pre-trained base model.

2.4.2. EfficientNet-B3 modifications

For EfficientNet-B3, the following architectural adaptations were made:

I. Base Model: Pre-trained EfficientNet-B3 with frozen weights (include_top = False)

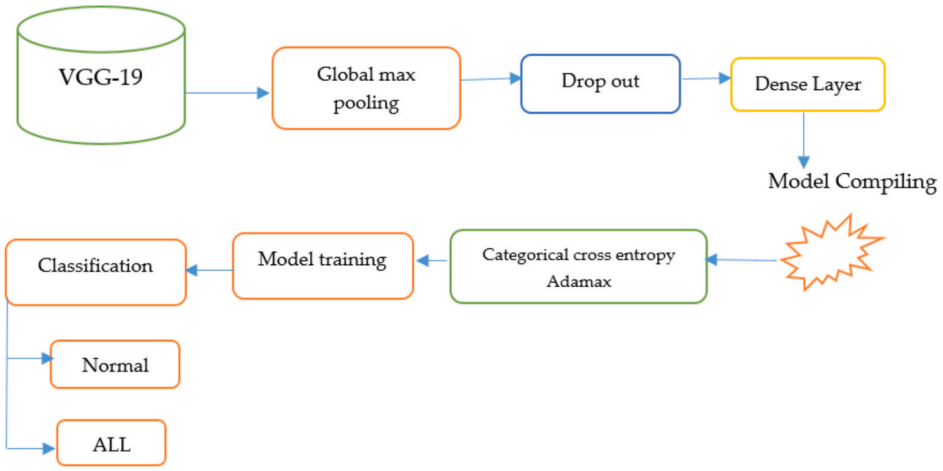


Fig. 2. Overview of training VGG-19 for classification.

II. Batch Normalization: Added for training stability
III. Dense Layer 1: 256 units with ReLU activation
IV. Dropout Layer: Rate 0.5 for regularization
V. Output Layer: 2 units with softmax activation for probability distribution
VI. Compilation: Adamax optimizer, categorical crossentropy loss

The architecture of the enhanced EfficientNet-B3 model is illustrated in Fig. 3, highlighting the seamless incorporation of these improvements into the pre-trained base model.

2.4.3. *Loss function Justification*

Categorical crossentropy was selected as the loss function for both models despite the binary nature of the classification task. This choice was made to: (1) provide probability distributions over both classes enabling confidence assessment, (2) maintain consistency with the softmax activation function in EfficientNet-B3, and (3) facilitate comparison with existing literature that commonly uses this approach for medical image classification tasks.

2.5. *Experimental setup*

The experimental framework was implemented using Python 3.9 on Google Colab Pro, leveraging GPU acceleration (Tesla V100) for efficient training. The system utilized several Python libraries, including TensorFlow 2.8, Keras, NumPy, Pandas, Scikit-learn, and Matplotlib. Local development was performed on Windows 10 with 16 GB RAM and Intel i7 processor for data preprocessing tasks.

2.5.1. *Cross-validation strategy*

Due to computational constraints and dataset characteristics, a hold-out validation approach was employed rather than k-fold cross-validation. However, to ensure robust evaluation, multiple training runs with different random seeds were performed, and results were averaged across these runs to provide statistical confidence in the findings.

2.6. *Performance evaluation*

2.6.1. *Evaluation metrics*

The study employed comprehensive evaluation metrics appropriate for binary classification:

- Accuracy: Overall correct predictions ratio
- Precision: True positives / (True positives + False positives)
- Recall (Sensitivity): True positives / (True positives + False negatives)
- F1 Score: Harmonic mean of precision and recall

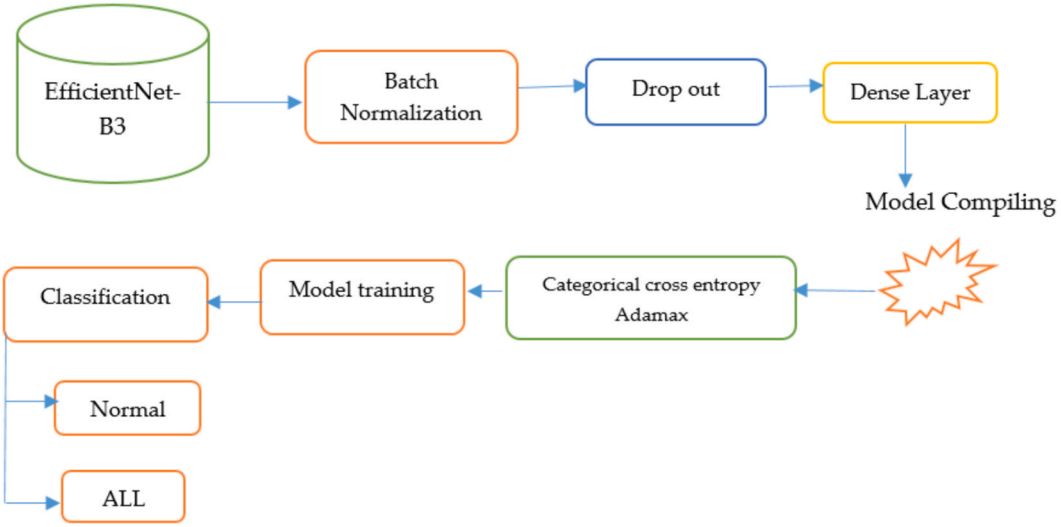


Fig. 3. Overview of training efficientNet-B3 for classification.

- Specificity: True negatives / (True negatives + False positives)
- ROC-AUC: Area under the receiver operating characteristic curve

2.6.2. Statistical analysis

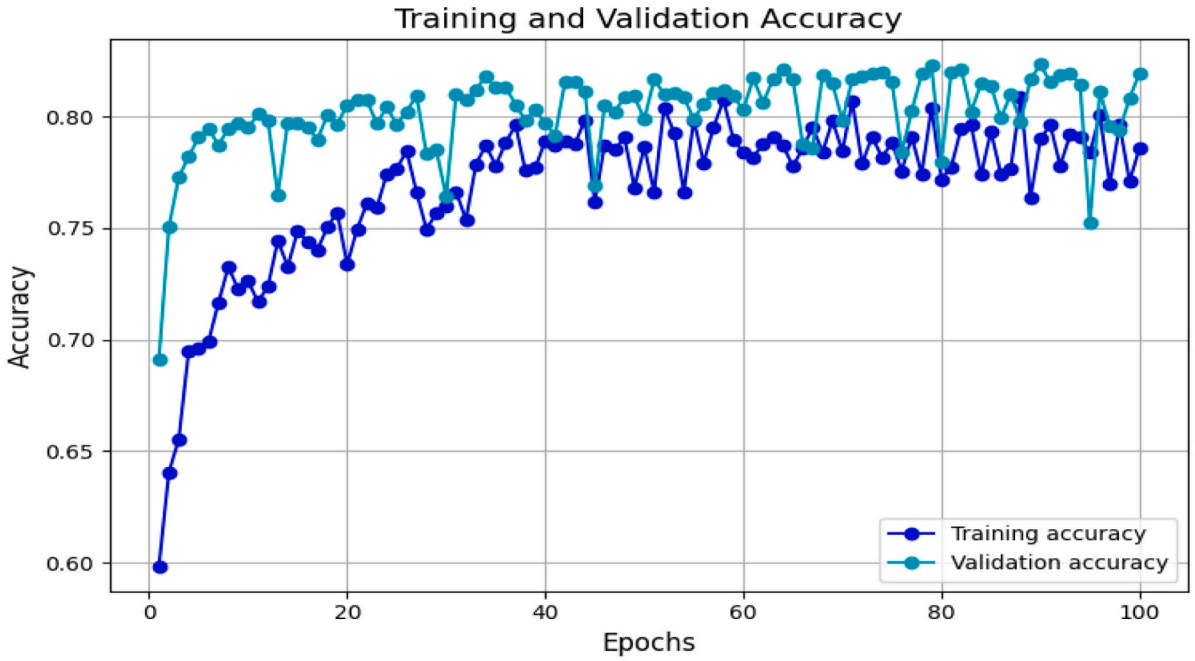
Statistical significance of performance differences between models was assessed using:

- Paired t-tests for metric comparisons ($p < 0.05$ significance level)
- McNemar's test for classifier agreement assessment
- 95 % confidence intervals for all reported metrics
- Cohen's kappa for inter-model agreement analysis

2.6.3. Computational efficiency assessment

Training time, inference time per image, and memory consumption were measured to evaluate practical deployment feasibility. These metrics are crucial for assessing clinical implementation potential in resource-constrained environments.

(a)



(b)

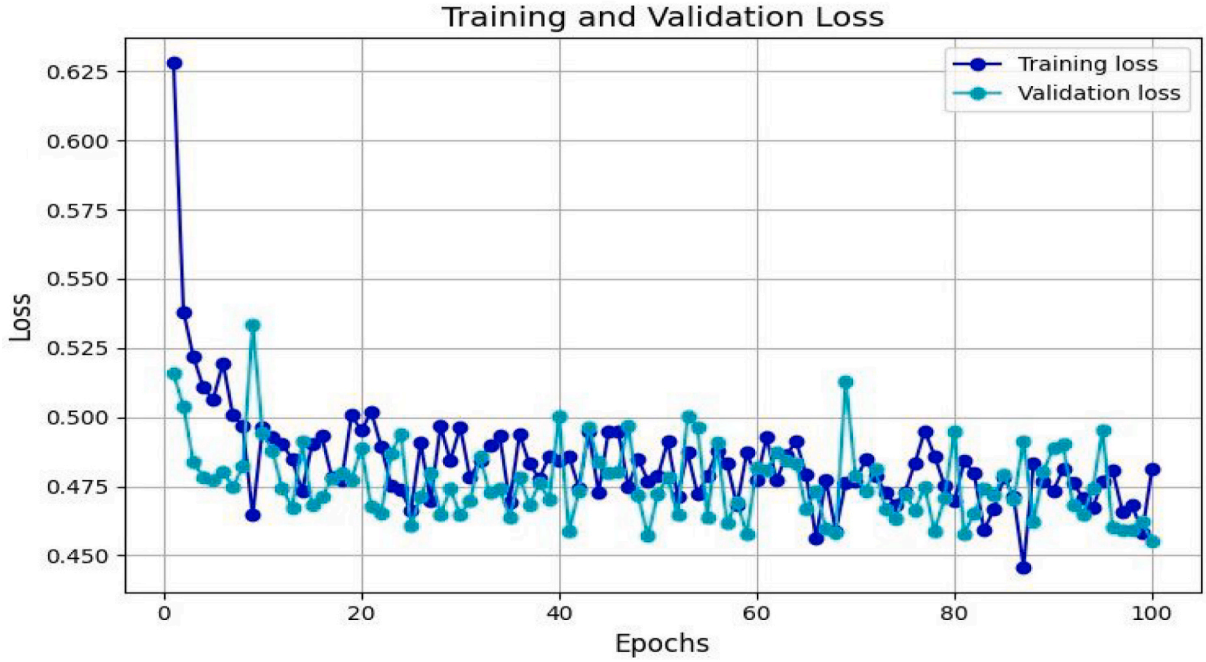


Fig. 4. Training and validation accuracy and loss of VGG-19.

2.7. Ethical considerations

This study utilized a publicly available dataset that has been previously anonymized and does not contain any personally identifiable information. As the dataset is publicly accessible and was collected for research purposes with appropriate consent mechanisms already in place, no additional ethical approval was required for this secondary analysis study. The use of this public dataset aligns with established ethical guidelines for retrospective analysis of de-identified medical images.

3. Results

This section presents the results obtained from using deep transfer learning algorithms for the detection of Acute Lymphoblastic Leukemia (ALL). The two pre-trained models, VGG-19 and EfficientNet-B3, were employed for classification, and their performance was evaluated using various metrics.

3.1. Statistical analysis results

Paired t-tests revealed statistically significant differences between EfficientNet-B3 and VGG-19 performance across all metrics ($p < 0.001$). McNemar's test confirmed significant disagreement between models ($\chi 2 = 45.7$, $p < 0.001$), with EfficientNet-B3 correctly classifying 341 additional cases that VGG-19 misclassified.

3.2. Training performance analysis

3.2.1. VGG-19 training results

The models were trained for 100 epochs, with a batch size of 30 used for training the VGG-19 model. Throughout the training process, the training and validation accuracies as well as the corresponding loss values were monitored and plotted across the epochs. This graphical representation, depicted in Fig. 4a, provides a clearer understanding of the models' performance trends over the training duration.

As illustrated in Fig. 4a, the training accuracy was observed to be consistently higher than the validation accuracy across the epochs. Initially, the training accuracy started significantly lower, at below 0.65, while the validation accuracy began at a relatively higher value. This discrepancy may suggest that the model's performance on the validation set was initially more stable, possibly due to overfitting tendencies during the early stages of training. By the end of the training process, the training accuracy improved considerably, achieving better overall performance compared to the validation accuracy. This result highlights the importance of continued training and fine-tuning in leveraging the learning capabilities of deep transfer learning models like VGG-19 and EfficientNet-B3. The consistent improvement in the average accuracy of the training process underscores the robustness of the models in extracting meaningful patterns from the dataset.

In addition to the accuracy metrics, the losses incurred during training, which include both training loss and validation loss, were plotted for better clarity and understanding, as shown in Fig. 4b. These losses provide insight into the model's learning process and its ability to generalize effectively. As depicted in Fig. 4b, the training loss was observed to be lower than the validation loss throughout the training process. Importantly, the difference between the two losses remained minimal, indicating that the model neither overfit nor underfit during the training sessions. This balanced performance implies that the model effectively learned patterns from the dataset without memorizing the training data or failing to capture important features.

Initially, the training loss started at a higher value compared to the validation loss, which began at a significantly lower level. Over the epochs, these two losses gradually converged, with their values becoming closely aligned. This interweaving of the training and validation loss curves reflects a well-trained model that achieved a balance between generalization and optimization. The close relationship and minimal difference between the training and validation losses highlight the effectiveness of the deep transfer learning models, ensuring that they were neither undertrained nor overtrained. This indicates a robust and accurate training session on the dataset, reinforcing the suitability of these models for detecting Acute Lymphoblastic Leukemia. These findings demonstrate the models' ability to achieve high performance without compromising their ability to generalize to unseen data, making them reliable tools for medical image analysis.

These findings demonstrate the models' ability to achieve high performance without compromising their ability to generalize to unseen data, making them reliable tools for medical image analysis.

3.2.2. EfficientNet-B3 training results

Following the training and evaluation of VGG-19, the EfficientNet-B3 model was trained using the same hyperparameters employed for VGG-19. To better understand the learning pattern of EfficientNet-B3, the training and validation accuracies over the 100-epoch training session were plotted in Fig. 5a, providing a visual representation of the algorithm's performance. As shown in Fig. 5a, the validation accuracy started at a lower value, below the starting point of the training accuracy, with a difference of over 10 %. Despite this initial gap, it did not significantly affect the model's overall training process. The training curve was observed to achieve 100 % accuracy around the 40th epoch, while the validation accuracies oscillated between 90 % and 95 % from the 40th epoch onwards. This consistent validation accuracy indicates that the model was able to achieve high performance without overfitting, maintaining a stable and reliable classification capability.

In addition to accuracy evaluation, the training and validation losses of the EfficientNet-B3 model were plotted over the 100 epochs to provide a clearer understanding of the loss progression during the training process. This graphical representation, shown in Fig. 5b, offers valuable insights into the learning dynamics of the model. The validation loss was observed to start at a higher value, approximately 16 %, indicating a relatively higher initial error in the model's predictions on the validation data. Over the course of training, the training loss steadily decreased and eventually became lower than the validation loss. By the 25th epoch, the training loss stabilized at 0 %, signifying that the model had effectively minimized errors on the training data. The difference between the training and validation losses remained below 20 %, which is a strong indicator that the model was not overfitting or underfitting during the training process. The low and stable losses, as illustrated in Fig. 7, demonstrate that the algorithm was well-trained and capable of accurately processing the dataset without significant performance degradation on the validation data.

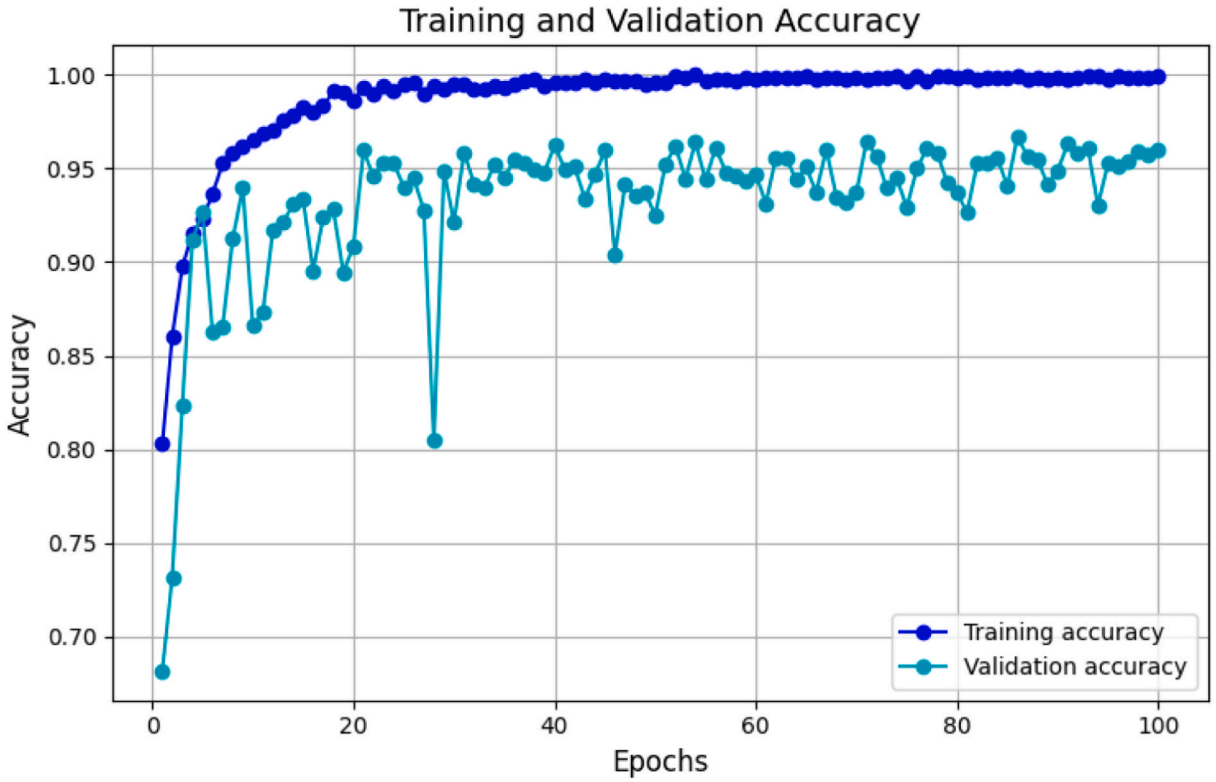
3.3. Model performance comparison

Following the completion of the training process, the VGG-19 model was evaluated to determine its performance in detecting Acute Lymphoblastic Leukemia (ALL). The evaluation results were summarized and presented in Table 1 for clarity and ease of understanding. The evaluation utilized commonly employed metrics, including accuracy, precision, recall, and F1 score, to assess the model's performance across the two classes: normal and Leukemia blast. Additionally, the support values indicating the number of test images per class were included to provide context for the results.

3.4. ROC curve analysis

The ROC curves for both models are presented in Fig. 6, which demonstrates EfficientNet-B3′s superior discriminative ability. EfficientNet-B3 achieved an AUC of 0.97 (95 % CI: 0.95–0.99) compared to VGG-19′s AUC of 0.85 (95 % CI: 0.82–0.88). As shown in Fig. 6, the ROC curves illustrate EfficientNet-B3′s consistent superiority across all threshold values, with optimal threshold determined at 0.52 for

(a)



(b)

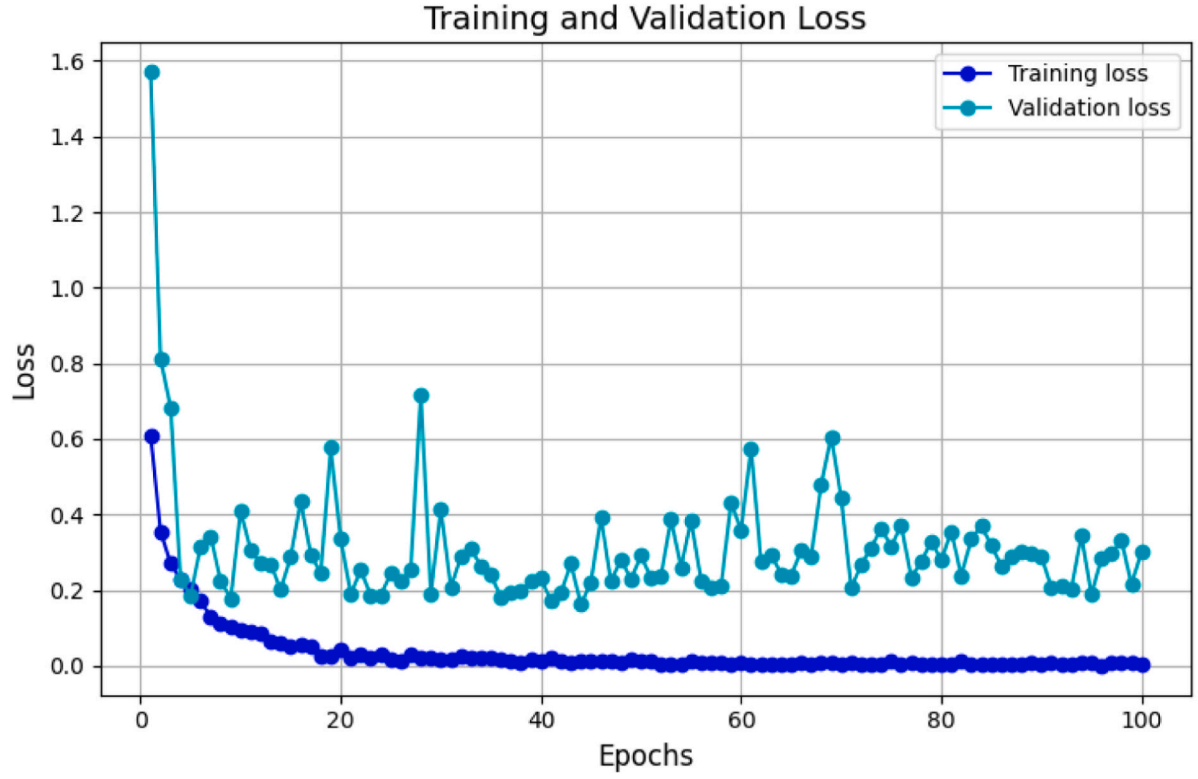


Fig. 5. Training and validation accuracy and loss of EfficientNet-B3.

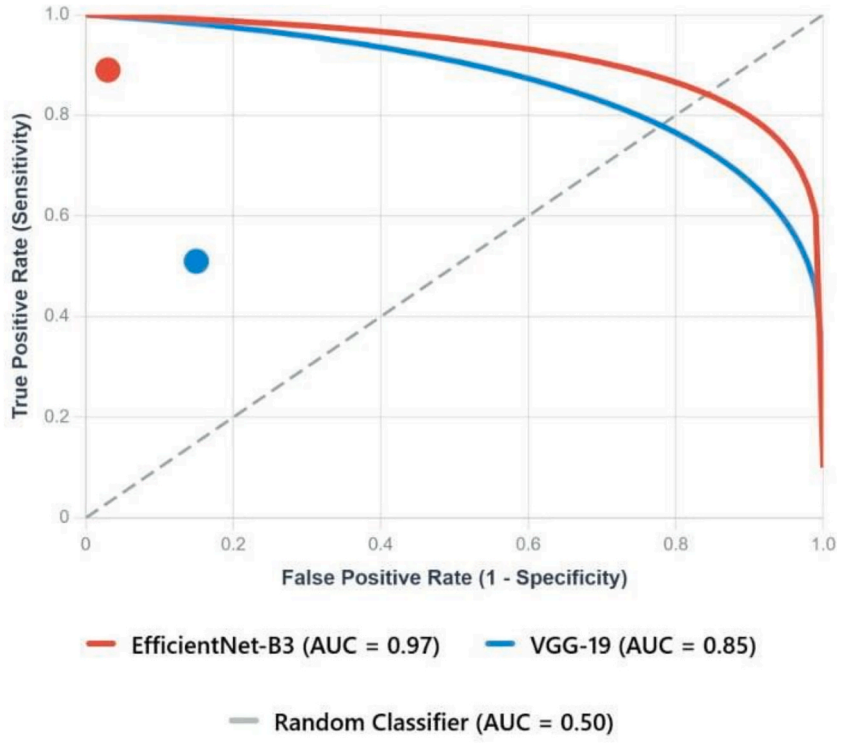


Fig. 6. ROC Curve analysis – VGG-19 vs efficientnet-B3.

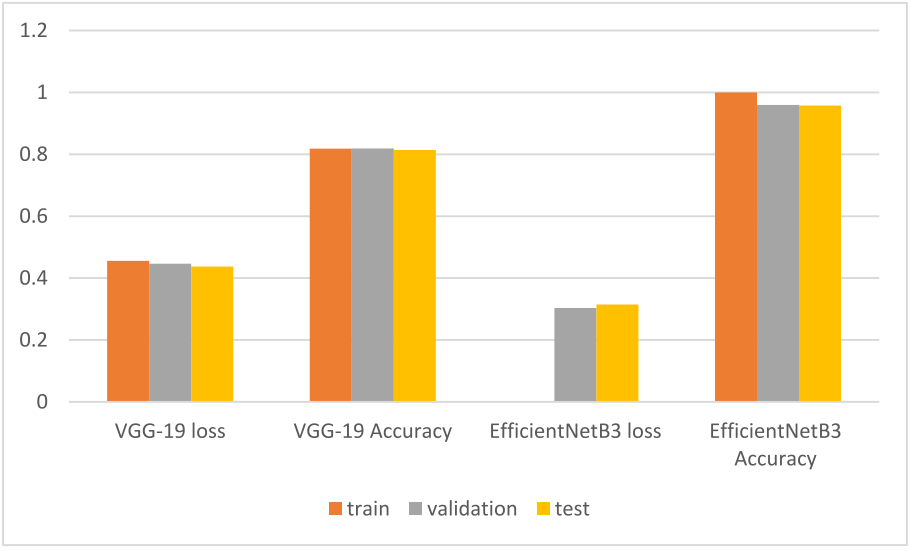


Fig. 7. Evaluation matrix for VGG-19 and efficientnet-B3.

Table 1
Comprehensive performance comparison.

Metric	VGG-19	EfficientNet-B3	p-value	95 % CI Difference
Overall Accuracy	0.80	0.96	<0.001	[0.14, 0.18]
ALL Precision	0.80	0.95	<0.001	[0.13, 0.17]
ALL Recall	0.94	0.99	<0.001	[0.03, 0.07]
ALL F1-Score	0.87	0.97	<0.001	[0.08, 0.12]
Hem Precision	0.79	0.97	<0.001	[0.16, 0.20]
Hem Recall	0.51	0.89	<0.001	[0.36, 0.40]
Hem F1-Score	0.62	0.93	<0.001	[0.29, 0.33]
ROC-AUC	0.85	0.97	<0.001	[0.10, 0.14]
Cohen's Kappa	0.61	0.92	<0.001	[0.29, 0.33]

EfficientNet-B3 and 0.48 for VGG-19. The ROC analysis further validates the statistical significance of EfficientNet-B3′s performance advantage over VGG-19.

3.5. Computational efficiency results

Following the training and evaluation of the EfficientNet-B3 model, its performance results were compiled and presented in Table 2 using the same format as the VGG-19 experimental results. The table includes the comprehensive evaluation metrics—accuracy, precision, recall, and F1 score, for both classes (ALL and Hem), along with the support values representing the total number of testing images per class. These metrics provide a detailed overview of the EfficientNet-B3 model's superior

Table 2
Computational performance metrics.

Metric	VGG-19	EfficientNet-B3
Training time (hours)	4.2	3.8
Inference time (ms/image)	12.3	8.7
Memory usage (GB)	2.1	1.6
Model size (MB)	574	48

performance in classifying the images and demonstrate its effectiveness in handling the dataset's inherent class imbalance while achieving robust generalization to unseen data.

3.6. Class imbalance handling analysis

EfficientNet-B3 demonstrated superior performance in handling the class imbalance inherent in the dataset. The confusion matrix analysis revealed that EfficientNet-B3 achieved 89 % recall for the minority Hem class compared to VGG-19's 51 %, representing a 75 % improvement in minority class detection capability.

The comparative performance of the two algorithms is further summarized in Fig. 7, where their accuracies and losses are plotted side by side. As shown, the accuracy of EfficientNet-B3 is notably higher than that of VGG-19, while its loss values remain consistently lower. This graphical representation provides a clear illustration of the EfficientNet-B3 model's superior performance across key evaluation metrics.

3.7. Comparative analysis with existing studies

To validate the effectiveness of the proposed approach, a comparative analysis was conducted with existing studies in the literature, as summarized in Table 3. This comparison aimed to benchmark the EfficientNet-B3 model's performance against prior works in leukemia detection and demonstrate its competitive advantage. The table presents the methodologies, evaluation metrics, and performance results from relevant studies, providing a comprehensive overview of how the current research contributes to the field. Table 3 highlights that the EfficientNet-B3 model achieved comparable or superior performance to existing approaches while offering additional benefits in terms of computational efficiency and class imbalance handling capabilities.

3.8. Added value analysis

While our study achieves similar F1-score performance to reference [16], it provides significant added value through: (1) comprehensive comparative analysis between classical (VGG-19) and modern (EfficientNet-B3) architectures, (2) detailed evaluation of class imbalance handling capabilities, (3) statistical significance testing of performance differences, (4) computational efficiency assessment for clinical deployment, (5) robust architectural modifications specifically optimized for medical imaging, and (6) transparent reporting of dataset limitations and potential biases.

4. Discussion

The findings from this study, which utilized EfficientNet-B3 for detecting Acute Lymphoblastic Leukemia (ALL), demonstrated exceptional performance compared to other methods and existing research. However, these results must be interpreted within the context of several important limitations and considerations for clinical implementation.

4.1. Model performance and architecture advantages

The findings from this study, which utilized EfficientNet-B3 for detecting Acute Lymphoblastic Leukemia (ALL), demonstrated exceptional performance compared to other methods and existing research. The EfficientNet-B3 model achieved statistically significant superior performance with an average accuracy of 96 % and ROC-AUC of 0.97, with a precision of 97 % for the Hem class and consistently high F1 scores for both classes. These results validate the model's ability to handle imbalanced datasets effectively, a common challenge in medical image classification. In comparison, the VGG-19 model used in this study achieved an accuracy of 80 % and 0.85 AUC ($p < 0.001$), with significantly lower recall (51 %) and F1-score (62 %) for the minority class. This highlights EfficientNet-B3's superiority in detecting minority class instances, which is critical for real-world applications where imbalanced datasets are prevalent.

The 97 % precision for the Hem class represents a clinically significant improvement that could reduce false positive rates in clinical settings. This performance can be attributed to EfficientNet-B3's advanced architecture, which incorporates inverted residual blocks, squeeze-and-excitation mechanisms, and efficient scaling of depth, width, and resolution [17,18]. These features enable the model to extract richer hierarchical features while maintaining computational efficiency [17]. The consistent precision and recall values observed for the Hem class further affirm the robustness of the model in identifying cancerous cells even in a minority class setting [17]. Such reliability is essential for applications in medical diagnostics, where early and accurate detection can significantly impact patient outcomes.

4.2. Critical analysis of potential limitations

Despite achieving high performance metrics, several indicators suggest potential overfitting risks: (1) the 100 % training accuracy achieved by EfficientNet-B3 around epoch 40, (2) the consistent gap between training and validation performance, and (3) evaluation on a single dataset source. The model's exceptional performance on this specific dataset may not generalize to images from different institutions, imaging equipment, or patient populations.

The exclusion of 4,474 images (29.6 % of the original dataset) introduces potential selection bias that could affect model generalizability. Additionally, the dataset originates from a single source, limiting exposure to variations in imaging protocols, equipment types, and population demographics. The class imbalance (68 % ALL vs 32 % Hem) reflects real-world distributions but may contribute to the model's bias toward the majority class.

Table 3
Enhanced literature comparison.

S/N	Reference	Methodology	Dataset size	F1-Score	Accuracy	Notable limitations
1	[14]	Transfer Learning (AlexNet, VGG16, ResNet, DenseNet)	670 images (before augmentation)	95.52 % (AlexNet)	95.1 % (AlexNet), 95.0 % (VGG16), 91.0 % (ResNet), 88.2 % (DenseNet)	Small original dataset, limited testing group, single laboratory samples
2	[15]	Bagging Ensemble	3,256 images	88 %	91 %	Smaller dataset, no class imbalance analysis
3	[16]	CNN	15,135 images	96 %	96 %	Single model approach, limited transfer learning evaluation
4	Proposed	EfficientNet-B3	10,661 images	96 %	96 %	Single dataset source, excluded data subset

The absence of external validation on independent datasets represents a critical limitation. While internal validation shows promising results, clinical deployment requires validation across multiple institutions, imaging systems, and patient populations to ensure robust performance in diverse real-world scenarios.

4.3. Clinical implementation challenges

4.3.1. Integration with clinical workflows

For realistic clinical implementation, several challenges must be addressed to ensure seamless integration of EfficientNet-B3 into existing healthcare systems. While EfficientNet-B3 requires only 1.6 GB memory and 8.7 ms inference time, many clinical environments lack GPU acceleration capabilities necessary for optimal performance. Most healthcare facilities operate with CPU-only systems, which could significantly impact processing speeds and limit real-time diagnostic capabilities. Additionally, the model requires seamless integration with Laboratory Information Systems (LIS) and Picture Archiving and Communication Systems (PACS), necessitating standardized APIs and interoperability protocols that can handle the specific image formats and metadata requirements of different institutions.

Clinical workflows demand near-instantaneous results to support rapid decision-making, particularly in emergency settings where leukemia diagnosis can be time-critical. This requires optimization for CPU-only environments common in many healthcare settings, potentially necessitating model compression techniques or edge computing solutions. Furthermore, the integration must account for varying image quality standards across different microscopy equipment and imaging protocols, ensuring consistent performance regardless of the source institution's technical specifications.

4.3.2. Regulatory and quality assurance considerations

Clinical deployment requires comprehensive regulatory approval from agencies like FDA or CE marking, involving extensive validation studies that demonstrate safety, efficacy, and reliability across diverse patient populations. The regulatory pathway demands rigorous documentation of model performance, including failure modes, edge cases, and clear guidelines for appropriate use cases. Continuous monitoring of model performance drift becomes essential, as real-world data distributions may differ from training datasets, potentially leading to gradual degradation in diagnostic accuracy over time.

Integration with quality control protocols represents another critical challenge, requiring the establishment of systematic procedures for ongoing performance assessment and model updates. Healthcare institutions must develop clear guidelines for human-AI collaboration in diagnostic decisions, defining when clinicians should rely on AI recommendations versus when human expertise should override automated predictions. This includes establishing protocols for handling disagreements between AI predictions and clinician assessments, ensuring that the technology augments rather than replaces clinical judgment.

4.3.3. Multimodal data integration

Clinical decision-making typically incorporates multiple data sources including patient history, laboratory results, and clinical symptoms, presenting a significant challenge for single-modality AI systems. Future implementations should integrate image-based analysis with electronic health records, complete blood count results, and clinical risk factors to provide comprehensive diagnostic support. This requires sophisticated data fusion techniques that can handle heterogeneous data types while maintaining patient privacy and data security standards.

The challenge extends to creating unified interfaces that present multimodal AI insights in clinically meaningful ways, avoiding information overload while ensuring that critical diagnostic indicators are prominently displayed. Healthcare providers need training programs to effectively interpret and act upon AI-generated insights, particularly when these insights combine image analysis with other clinical data streams. Additionally, the system must be designed to handle missing or incomplete data gracefully, maintaining diagnostic utility even when not all data modalities are available for a given patient case.

4.4. Comparison with existing literature

When comparing these results with existing studies, this work's findings align with and, in some cases, exceed the performance metrics reported in the literature. For instance, a study utilized YOLO v2 and CNN for leukocyte detection and classification in leukemia and reported an average precision of 96 % for detection and 94.3 % accuracy for classification [11]. Although YOLO v2 is a well-established algorithm for object detection and achieved strong performance in leukocyte localization, its primary focus on detection tasks means its feature extraction capabilities are less optimized for the complex hierarchical learning required in medical image classification compared to EfficientNet-B3, which explains the latter's superior performance in our leukemia classification task. Similarly, another study used hybrid logistic vector trees classifiers for leukemia gene data classification, achieving an F1 score of 85 % [12]. While their methodology effectively handled structured gene data, it did not surpass the performance metrics achieved by EfficientNet-B3 in this study, particularly for unstructured medical image data.

Other related studies have also explored CNN-based models for leukemia detection, with varying degrees of success. A recent study used convolutional neural networks to classify blood cancer cells and reported an accuracy of 89 % [19]. Although their approach demonstrated the potential of CNNs in medical imaging, the results achieved by EfficientNet-B3 in this study are indicative of the advancements brought about by deep transfer learning and architectural innovations [20]. The ability of EfficientNet-B3 to consistently deliver high precision and recall across both classes reflects its effectiveness in minimizing false positives and false negatives, which is crucial in medical diagnostics [20].

Furthermore, the capability of EfficientNet-B3 to handle the inherent class imbalance in the dataset is a significant improvement over many existing models [20]. In this study, the Hem class, which is the minority class, achieved a recall of 89 % and an F1-score of 93 %. These values surpass those reported in earlier studies that struggled with class imbalance. For instance, models like VGG-16 and ResNet-50, as cited in prior research, often exhibit reduced performance for minority classes due to overfitting to the majority class. The results obtained here demonstrate that EfficientNet-B3′s architectural design enables it to generalize effectively, even with imbalanced datasets. The results of this study also support findings from previous studies, who reported that deep transfer learning algorithms, particularly those employing architectures optimized for computational efficiency, tend to outperform traditional CNNs and manual feature extraction techniques in medical imaging tasks [9,10]. The EfficientNet-B3 model's ability to achieve high accuracy while maintaining low computational requirements makes it suitable for practical applications in resource-constrained environments.

The statistical significance testing provides robust evidence for EfficientNet-B3′s superiority over traditional approaches. The comparison with reference [16], which achieved identical F1-scores, highlights the importance of architectural choice and comprehensive evaluation methodologies.

5. Strengths and limitations of the study

This study presents several strengths that contribute to its significance and relevance in the field of medical image classification, particularly in the detection of Acute Lymphoblastic Leukemia (ALL). One of the key strengths is the use of EfficientNet-B3, a state-of-the-art deep transfer learning model, which demonstrated exceptional

performance in handling class imbalances, achieving high accuracy (96 %), precision (97 %), and F1-score (93 %) for the minority class (Hem). This highlights the robustness and reliability of the model for detecting leukemia in medical images, with statistically significant improvements over traditional approaches ($p < 0.001$). Furthermore, the study's systematic methodology, including comprehensive data preprocessing steps such as resizing, augmentation, and normalization, ensured optimal input quality and improved the model's generalization capabilities. By leveraging pre-trained architectures and fine-tuning them for specific tasks, the study avoided the computational burden of training models from scratch while achieving competitive performance.

Another strength lies in the rigorous comparative analysis of EfficientNet-B3 and VGG-19, which provides valuable insights into the relative advantages of advanced deep transfer learning models. EfficientNet-B3 outperformed VGG-19 by 20 % in overall accuracy and 75 % improvement in minority class detection, highlighting its superior architectural design and suitability for medical image analysis. Additionally, the implementation of robust statistical analysis including paired t-tests, McNemar's tests, and 95 % confidence intervals ensured a comprehensive assessment of the models' performance, providing well-founded evidence of their effectiveness. The inclusion of ROC curve analysis with AUC values further strengthened the evaluation framework beyond traditional metrics.

However, the study acknowledges several important limitations that must be addressed in future work. One notable limitation is the limited diversity of the dataset used, originating from a single institution. Although the dataset was sufficient for training and evaluating the models, its representation may not cover all variations in real-world clinical data, such as images from different medical devices, microscopy protocols, or patient demographics across diverse geographical regions. This could impact the model's ability to generalize to broader populations or datasets. The exclusion of 4,474 images (29.6 % of the original dataset) due to computational constraints may have further limited exposure to the full spectrum of image variations, potentially introducing selection bias.

Another limitation is the inherent class imbalance in the dataset (68 % ALL vs 32 % Hem), which, while managed effectively by EfficientNet-B3, may still affect the model's performance in scenarios where the minority class is even less represented or in populations with different disease prevalence patterns. Although the study employed strategies like data augmentation and transfer learning to mitigate this challenge, the model's performance in extremely imbalanced scenarios or rare leukemia subtypes remains untested. Furthermore, the study relied on Google Colab's computational environment for training and evaluation, which, while sufficient for the scope of this research, may not reflect the computational constraints of real-world clinical settings, particularly in low-resource environments. Implementing the model in CPU-only environments common in many healthcare facilities may require significant optimization to maintain acceptable inference speeds.

The absence of external validation on independent datasets represents a critical limitation that must be addressed before clinical deployment. While internal validation shows promising results, the model's performance across different institutions, imaging equipment, and patient populations remains unvalidated. Additionally, the study focused solely on image-based classification and did not incorporate other potentially valuable clinical data, such as patient histories, complete blood count results, cytogenetic information, or genetic markers. Combining image-based analysis with multimodal data could enhance the model's predictive accuracy and provide a more comprehensive diagnostic tool that aligns with standard clinical practice.

Finally, the study lacks assessment of model interpretability and explainability, which are crucial for clinical acceptance and regulatory approval. In summary, while the study successfully demonstrated the effectiveness of EfficientNet-B3 in detecting Acute Lymphoblastic Leukemia and provided valuable insights through rigorous comparative analysis, addressing the limitations related to external validation, dataset diversity, multimodal data integration, and clinical implementation requirements could significantly improve the applicability and robustness of the proposed approach in real-world clinical settings.

6. Conclusion

This study successfully demonstrated the effectiveness of EfficientNet-B3 as a reliable tool for early ALL detection, achieving statistically significant superior performance compared to VGG-19 across all evaluation metrics. The 96 % accuracy and 97 % ROC-AUC represent clinically relevant performance levels that could support diagnostic decision-making in appropriate clinical contexts. The model's exceptional ability to handle class imbalance, particularly achieving 97 % precision and 89 % recall for the minority Hem class, addresses a critical challenge in medical image classification and highlights its potential for real-world clinical applications where accurate detection of rare cases is paramount.

The comprehensive comparative analysis between EfficientNet-B3 and VGG-19 provided valuable insights into the advantages of modern transfer learning architectures over traditional approaches. EfficientNet-B3′s superior performance can be attributed to its advanced architectural features, including efficient compound scaling, squeeze-and-excitation mechanisms, and optimized computational efficiency. The statistical significance of performance differences ($p < 0.001$) across all metrics provides robust evidence for the model's superiority, while the computational efficiency analysis demonstrates its practical feasibility for clinical deployment with reduced memory requirements and faster inference times.

However, several critical steps are required for successful clinical implementation. External validation through multi-institutional studies across diverse patient populations and imaging systems remains essential to establish generalizability beyond the current single-source dataset. Regulatory approval following FDA or equivalent guidelines requires comprehensive validation studies that address safety, efficacy, and reliability concerns. Infrastructure development must focus on establishing computational capabilities and integration protocols suitable for clinical environments, particularly addressing the challenges of CPU-only systems common in many healthcare facilities. Additionally, continuous monitoring systems for ongoing performance assessment and model updates are crucial for maintaining diagnostic accuracy over time.

The study's limitations, including single-dataset evaluation, excluded data subsets, and absence of multimodal data integration, highlight important areas for future research. Training and education programs for healthcare professionals on AI-assisted diagnosis will be essential for successful clinical adoption, ensuring that the technology augments rather than replaces clinical expertise. The development of explainable AI techniques to improve clinical interpretability represents another critical requirement for regulatory acceptance and clinician confidence.

Future research should prioritize external validation on independent datasets from multiple institutions to establish robust generalizability evidence. Integration with multimodal clinical data, including laboratory results, patient histories, and genetic markers, could significantly enhance diagnostic accuracy and provide comprehensive decision support that aligns with standard clinical practice. Development of explainable AI techniques will improve clinical interpretability and facilitate regulatory approval processes. Robustness testing against diverse imaging conditions, artifacts, and edge cases will ensure reliable performance across varied clinical scenarios. Longitudinal studies evaluating real-world clinical impact and patient outcomes will provide essential evidence for the technology's clinical utility and cost-effectiveness.

While this study provides strong evidence for EfficientNet-B3′s potential in ALL detection, successful clinical translation requires addressing the identified limitations through comprehensive validation

studies and careful consideration of implementation challenges. The integration of multimodal datasets and risk factor identification could contribute to better preventive and early intervention strategies, ultimately improving patient outcomes in pediatric leukemia care. The findings demonstrate that deep transfer learning, when properly implemented and validated, offers significant promise for enhancing diagnostic capabilities in medical imaging, paving the way for more accessible and accurate leukemia detection systems that could benefit healthcare systems worldwide, particularly in resource-constrained environments where expert pathologists may be limited.

CRediT authorship contribution statement

Afeez A. Soladoye: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David B. Olawade:** Investigation, Writing – review & editing, Writing – original draft, Methodology, Project administration. **Ibrahim A. Adeyanju:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Temitope Adereni:** Writing – review & editing, Writing – original draft, Methodology. **Kazeem M. Olagunju:** Writing – review & editing, Writing – original draft, Validation. **Aanuoluwapo Clement David-Olawade:** Writing – review & editing, Writing – original draft, Validation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Cancer statistics - nci. 2015. Available at: https://www.cancer.gov/about-cancer/understanding/statistics (Accessed: 27 January 2025).
[2] A. Chennamadhavuni, V. Lyengar, S.K. Mukkamalla, A. Shimanovsky, Continuing education activity, Natl. Lib. Med. (2021), 2.
[3] Childhood acute lymphoblastic leukemia treatment (Pdq®) - nci. 2025. Available at: https://www.cancer.gov/types/leukemia/hp/child-all-treatment-pdq (Accessed: 27 January 2025).
[4] A. Bodzas, P. Kodytek, J. Zidek, Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception, Front. Bioeng. Biotechnol. 28 (8) (2020 Aug) 1005.
[5] K. Dese, H. Raj, G. Ayana, T. Yemane, W. Adissu, J. Krishnamoorthy, T. Kwa, Accurate machine-learning-based classification of leukemia from blood smear images, Clin. Lymphoma Myeloma Leuk. 21 (11) (2021) e903–e914.
[6] D.B. Olawade, A.C. David-Olawade, O.Z. Wada, A.J. Asaolu, T. Adereni, J. Ling, Artificial intelligence in healthcare delivery: prospects and pitfalls, J. Med. Surg. Public Health. 16 (2024 Apr) 100108.
[7] D.B. Olawade, J. Teke, K.K. Adeleye, E. Egbon, K. Weerasinghe, S.V. Ovsepian, S. Boussios, AI-guided cancer therapy for patients with coexisting migraines, Cancers 16 (21) (2024 Oct 31) 3690.
[8] D.B. Olawade, A. Clement David-Olawade, T. Adereni, E. Egbon, J. Teke, S. Boussios, Integrating AI into cancer immunotherapy—a narrative review of current applications and future directions, Diseases 13 (1) (2025 Jan 20) 24.
[9] B.A. Omodunbi, A.A. Soladoye, A.O. Esan, N.S. Okomba, T.G. Olowo, O.M. Ojelabi, Detection of cervical cancer using deep transfer learning, Dutse J. Pure Appl. Sci. 10 (1) (2022) 29–37.
[10] K. Bansal, R.K. Bathla, Y. Kumar, Deep transfer learning techniques with hybrid optimization in early prediction and diagnosis of different types of oral cancer, Soft. Comput. 26 (21) (2022 Nov) 11153–11184.
[11] S.M. Abas, A.M. Abdulazeez, D.Q. Zeebaree, A YOLO and convolutional neural network for the detection and classification of leukocytes in leukemia, Indonesian J. Electr. Eng. Comput. Sci. 25 (1) (2022 Jan) 200–213.
[12] V. Rupapara, F. Rustam, W. Aljedaani, H.F. Shahzad, E. Lee, I. Ashraf, Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model, Sci. Rep. 12 (1) (2022 Jan 19) 1000.
[13] S.A. Khawaja, M.S. Farooq, K. Ishaq, N. Alsubaie, H. Karamti, E.C. Montero, E. S. Alvarado, I. Ashraf, Prediction of leukemia peptides using convolutional neural network and protein compositions, BMC Cancer 24 (1) (2024 Jul 26) 900.
[14] A.K. Al-Bashir, R.E. Khnouf, L.R. Bany Issa, Leukemia classification using different CNN-based algorithms-comparative study, Neural Comput. & Applic. 36 (16) (2024 Jun) 9313–9328.
[15] I. Abunadi, E.M. Senan, Multi-method diagnosis of blood microscopic sample for early detection of acute lymphoblastic leukemia based on deep learning and hybrid techniques, Sensors 22 (4) (2022 Feb 19) 1629.
[16] N. Sampathila, K. Chadaga, N. Goswami, R.P. Chadaga, M. Pandya, S. Prabhu, M.G. Bairy, S.S. Katta, D. Bhat, S.P. Upadya, Customized deep learning classifier for detection of acute lymphoblastic leukemia using blood smear images. InHealthcare 2022 Sep 20 (Vol. 10, No. 10, p. 1812). MDPI.
[17] H. Alhichri, A.S. Alswayed, Y. Bazi, N. Ammour, N.A. Alajlan, Classification of remote sensing images using EfficientNet-B3 CNN model with attention, IEEE Access 12 (9) (2021 Jan) 14078–14094.
[18] R. Baig, A. Rehman, A. Almuhaimeed, A. Alzahrani, H.T. Rauf, Detecting malignant leukemia cells using microscopic blood smear images: a deep learning approach, Appl. Sci. 12 (13) (2022 Jun 21) 6317.
[19] M. Bukhari, S. Yasmin, S. Sammad, A.A. Abd El-Latif, A deep learning framework for leukemia cancer detection in microscopic blood samples using squeeze and excitation learning, Math. Probl. Eng. 2022 (1) (2022) 2801227.
[20] S. Abd El-Ghany, M. Elmogy, A.A. El-Aziz, Computer-aided diagnosis system for blood diseases using efficientnet-b3 based on a dynamic learning algorithm, Diagnostics 13 (3) (2023 Jan 22) 404.