A deep learning ensemble framework for robust classification of lung ultrasound patterns: covid-19, pneumonia, and normal

Shereen Morsy [a,1,*], Neveen Abd-Elsalam [a,2], Ahmed Kandil [a,3], Ahmed Elbialy [a,b,4], Abou-Bakr Youssef [a,5]

[a] Systems and Biomedical Engineering, Cairo University, Giza, Egypt
[b] Al-Shorouk Academy, Cairo, Egypt
[1] shereen.morsy.m@eng-st.cu.edu.eg; [2] neveen.mahmoud.n@eng-st.cu.edu.eg; [3] ahkandil@eng1.cu.edu.eg; [4] a.elbialy@sha.edu.eg; [5] aboubakryoussef51@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received December 1, 2024
Revised February 24, 2025
Accepted February 28, 2025
Available online February 28, 2025

Keywords
COVID-19
Pneumonia
Deep learning
Transfer learning
Ensemble method

ABSTRACT

To advance the automated interpretation of lung ultrasound (LUS) data, multiple deep learning (DL) models have been introduced to identify LUS patterns for differentiating COVID-19, Pneumonia, and Normal cases. While these models have generally yielded promising outcomes, they have encountered challenges in accurately classifying each pattern across diverse cases. Therefore, this study introduces an ensemble framework that leverages multiple classification models, optimizing their contributions to the final prediction through a majority voting mechanism. After training seven different classification models, the three models with the highest accuracies were selected. The ensemble incorporates these top-performing models: EfficientNetV2-B0, EfficientNetV2-B2, and EfficientNetV2-B3, and utilizes this framework to classify patterns in LUS images. Compared to individual model performance, the ensemble approach significantly enhances classification accuracy, achieving an accuracy of 99.25% and an F1-score of 99%. In contrast, the standalone models attained accuracies of 97.8%, 97.6%, and 98.1%, with F1-score of approximately 98%. This research highlights the potential of ensemble learning for improving the accuracy and robustness of automated LUS analysis, offering a practical and scalable solution for real-world medical diagnostics. By combining the strengths of multiple models, the proposed framework paves the way for more reliable and efficient tools to assist clinicians in diagnosing lung diseases.

1. Introduction

Ultrasound (US) is one of the most frequently employed medical imaging modalities, enabling real-time visualization of internal organs and providing immediate guidance for various interventional procedures. Compared to other imaging techniques, the US is increasingly recognized for its multifunctionality, portability, and cost-effectiveness [1]. Unlike X-rays or computed tomography (CT) scans, the US does not involve ionizing radiation. Therefore, it is safer for patients, including pregnant women. Lung sonography produces distinct results based on different pathologies, as ultrasound waves interact uniquely with various lung tissues. In COVID-19, the lung parenchyma is compromised by alveolar filling with fluid, resulting in specific artifacts. Clinicians utilize LUS images to evaluate patients for pleural wall thickening and areas of lung congestion, which are characteristic of pneumonia.

Pneumonia is an inflammatory condition affecting one or both lungs, often caused by bacterial infections, although it can also result from viral pathogens, such as the coronavirus responsible for COVID-19 [2]. Pneumonia symptoms can vary widely, from mild enough to go unnoticed to severe enough to require hospitalization [3], [4].

The World Health Organization (WHO) has reported that pneumonia is the primary infectious cause of death and child mortality worldwide. In 2019, it claimed the lives of 740,180 children under five years old, representing 14% of all fatalities in this category of age and a staggering 22% of deaths among children aged 1 to 5. In comparison, the pandemic of COVID-19 has had a significant impact, with an estimated 7,070,128 deaths attributed to the virus as of October 2024. The total number of confirmed COVID-19 cases has reached 776,546,006. Importantly, infections from various COVID-19 variants continue to pose a risk, highlighting ongoing public health challenges. These figures underscore the critical need for effective measures to combat pneumonia and COVID-19, particularly among vulnerable populations such as young children, to improve health outcomes worldwide [5]. Lung ultrasonography has become a significant bedside tool for diagnosing pneumonia, with greater sensitivity over chest radiography, particularly in detecting pleural effusion. Unlike chest CT, which remains the gold standard but is limited by high costs and significant radiation exposure, lung ultrasound provides a safer and more accessible alternative, especially for critically ill and unstable patients [6].

This research introduces a highly accurate, efficient, and scalable ensemble framework for LUS image classification, demonstrating the potential of ensemble learning to enhance automated medical diagnostics and support clinical decision-making. Seven deep learning models were evaluated based on accuracy, computational efficiency, and training time, with EfficientNetV2-B0, B2, and B3 emerging as the top-performing models. A majority voting ensemble framework was then developed, combining predictions from these three models. The ensemble significantly outperformed individual models, achieving an impressive 99.25% accuracy and a 99% F1-score, demonstrating the effectiveness of leveraging multiple models to enhance classification accuracy and robustness. The paper is structured as follows: Section I includes a detailed review of the relevant literature, emphasizing recent advancements and the performance of DL models on existing datasets. Section II describes the proposed methodology, including the research framework and architectural design. Section III discusses the results and performance evaluation of the proposed approach, accompanied by a critical analysis. Finally, Section IV concludes the study, summarizing key findings and implications.

2. Related Work

Born *et al.* [7] employed a convolutional neural network (CNN) model, POCOVID-NET, for automatically classifying lung conditions into three groups: Normal (healthy) lungs, COVID-19, and Pneumonia, achieving 89% accuracy using 5-fold cross-validation. Expanding on COVID-19 differentiation, Bahri *et al.* [8] applied texture analysis on LUS images to distinguish pneumonia from COVID-19. Similarly, Marco La Salvia *et al.* [9] leveraged DL models for COVID-19 and pneumonia classification, reporting a high F1-score of 98%. In another study, Elkhouly *et al.* [10] utilized the InceptionV1 model on LUS images, achieving an accuracy of approximately 84.3% for COVID-19 diagnosis. Shiyao *et al.* [11] proposed a system leveraging transfer learning with ResNet for detecting COVID-19 syndrome in LUS images. Their approach demonstrated significant performance improvements, with junior and senior radiologists achieving F1-scores of 91.33% and 95.79% on the balanced dataset and 94.20% and 96.43% on the unbalanced dataset, respectively. Similar methodologies have been explored in related studies [12]–[14] Additionally, Subramanyam *et al.* [15] utilized a pre-trained ResNet50 model specifically for the automated identification and localization of A-lines and B-lines in LUS images, further contributing to advancements in lung ultrasound analysis. Comparable research efforts have been reported in the literature [16]–[20].

More recent studies employ ensemble mechanisms, such as the research by Dubey *et al.* [21] on the ICLUS-DB dataset that used ensemble models combining EfficientNet, ResNet, and other CNN architectures to classify lung patterns. The ensemble model performed better than individual models,

particularly in detecting pneumonia, COVID-19, and other lung pathologies. Another study by Shea *et al.* [22] explored ensemble learning using various CNN architectures such as ResNet and DenseNet for LUS data, enhancing prediction reliability by combining the strengths of each model in the ensemble. This method outperformed single models in LUS-based COVID-19 classification tasks (AUCs ≈0.9). The ensemble framework of Khan et al. [23] also combines RegNetX, ResNet50, and ResNet18 with spatial attention models, with each model's contribution to the final prediction adjusted via a weighted voting method. Each model's prediction is weighted according to its confidence level and overall performance in the predicted class. As a result, the final prediction is based on the majority class and the class having the most weight among the predictions. The ensemble approach outperforms individual models, achieving an F1-Score of 0.685.

These studies reflect the growing trend of using ensemble techniques to improve diagnostic accuracy in LUS analysis, particularly for complex tasks like COVID-19 detection and differentiation between lung conditions. Nevertheless, adaptive ensembles or approaches involving multiple models can lead to substantial computational demands, particularly during inference, which is the focus of our research.

3. Method

3.1. Dataset

The ultrasound dataset used in this study is the POCUS dataset [7], which is currently the most used in the literature for lung classification. It comprises a total of 13366 LUS images, divided as illustrated in Table 1, taken with a convex probe, then partitioned into 7130 images for training, 3565 for validation, and 2671 for testing. Preprocessing image data is a critical step when working with image classification tasks, as proper preprocessing can enhance the performance and convergence of machine learning models. Here, four steps were followed 1) To balance computational efficiency across multiple models with varying input size requirements, images were resized to a compromise resolution of 260x260. 2) Normalization (scaling pixel values): Neural networks perform better when the input values are within a small range, typically [0, 1]. 3) Standardization ensures that the pixel values have a mean of 0 and a standard deviation of 1, helping the model to converge faster. 4) Data Augmentation: Artificially expanding the training dataset size and introducing image variations to help the model generalize better and prevent overfitting.

Table 1. Dataset description with the number of images for each class

Class	No. of Images
COVID-19	5594
Pneumonia	4104
Normal	3668

3.2. Transfer Learning

Transfer learning (TL) is a cornerstone of modern AI, enabling faster deployment of sophisticated models in diverse applications. TL is a machine learning technique in which a model created for one task is used as an initial basis for another related task. It is especially beneficial when the dataset for the target job is limited or when training a DL model from scratch would be computationally expensive [24]. The key concepts of transfer learning include:

- Pre-trained model: Typically involves using a model pre-trained on a huge dataset (e.g., ImageNet for image classification) and adapting it to a new problem.
- Feature reuse: Early layers of deep networks learn generic features (e.g., edges, shapes), which can be useful across different tasks. Only the final layers, specific to the original task, may need modification or replacement for the new task.
- Fine tuning: Adapting a pre-trained model to the target task by continuing training with a lower learning rate. Depending on the similarity between tasks, it may involve training all layers or just the top few layers [25].

In this study, seven pre-trained models— EfficientNetV2B0, EfficientNetV2B1, EfficientNetV2B2, EfficientNetV2B3, InceptionV3, ResNet152V2, and InceptionResNetV2—were evaluated, and various hyperparameters were tested to identify the optimal configuration for classifying COVID-19, pneumonia and normal lungs. These specific models were chosen due to their strong performance in image classification tasks, architectural diversity, and proven capability to handle complex features in medical imaging datasets.

3.2.1. EfficientNet v2

EfficientNets are among the most advanced CNN architectures available. EfficientNetB0 [26] serves as the foundational model of the EfficientNet family, engineered to deliver high performance while utilizing fewer parameters and reducing computational costs compared to conventional deep learning models, as shown in Fig. 1.

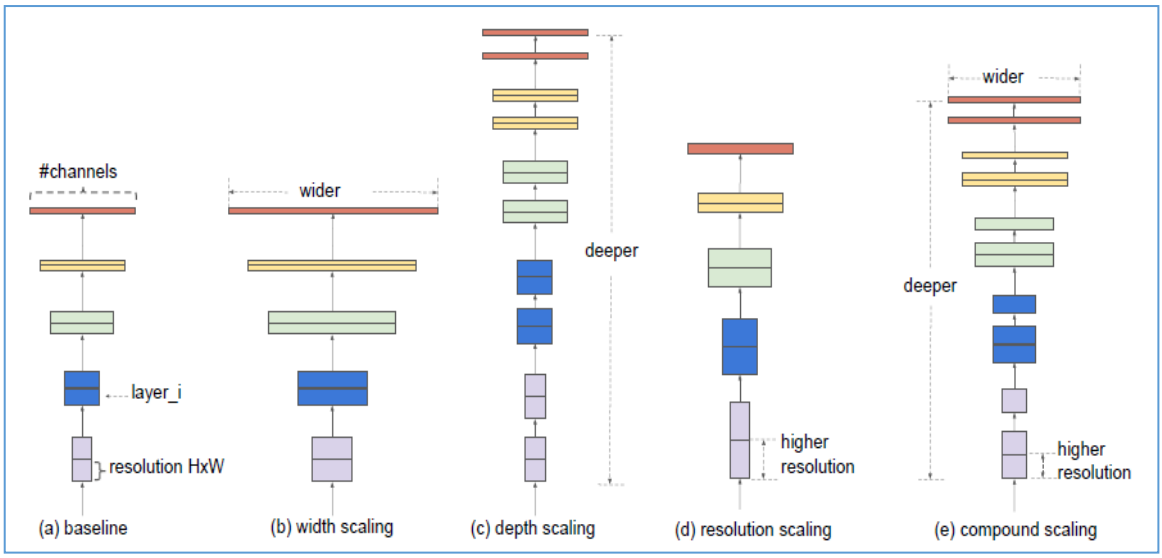

Fig. 1. Adjusting depth, width, and image resolution to generate various EfficientNet model variants [27]

The figure illustrates different strategies for scaling CNNs, focusing on EfficientNet's compound scaling approach. It consists of five subfigures.

- Baseline Model – Represents the original architecture with a fixed number of layers, channels, and input resolution.
- Width Scaling – Increases the number of channels per layer, making the network wider. This allows the model to learn richer feature representations.
- Depth Scaling – Adds more layers, making the network deeper. This helps capture more complex hierarchical features but increases computation time.
- Resolution Scaling – Increases the input image resolution, enabling finer details to be captured and raising memory and processing demands.
- Compound Scaling – Simultaneously scales depth, width, and resolution in a balanced manner, optimizing performance while maintaining efficiency.

EfficientNetV2 models improve accuracy while reducing training time and inference latency by combining compound scaling (adjusting width, depth, and resolution) with neural architecture search. Incorporating Fused-MBConv blocks, they achieve up to 6.8× smaller sizes than prior models. Variants (B0–B3) differ in depth (layers), width (channels), and resolution, balancing complexity and accuracy. While deeper models capture richer features, they demand more computation. Compound scaling optimizes efficiency, but increasing model size raises resource requirements, including trainable parameters and FLOPs, as shown in Table 2 [28].

Table 2. EfficientNetV2 performance results on ImageNet

Model	Top-1 Acc.	Parameters (M)	FLOPs (B)
V2-B0	78.7%	7.4	0.7
V2-B1	79.8%	8.1	1.2
V2-B3	82.1%	14	3.0

3.2.2. InceptionV3

InceptionV3 [29], introduced by Szegedy *et al.* in 2016, brought significant advancements over previous architectures, incorporating approximately 24 million parameters. Key innovations included modifications to the optimizer, loss function, batch normalization, and auxiliary layers. A pioneering aspect was its use of batch normalization. Furthermore, InceptionV3 employed factorization techniques, replacing large convolutions (like 5x5 and 7x7) with sequences of smaller 3x3 convolutions and asymmetric 1xN and Nx1 convolutions.

This reduced computational cost and mitigated representational bottlenecks by lowering the dimensionality of layer inputs, leading to faster training and inference. Fig. 2 represents an Inception-based architecture for processing ultrasound images. The network comprises multiple convolutional layers (orange) interspersed with pooling layers (gray for max pooling, blue for mean pooling) to extract and refine features. The architecture follows a modular structure, with repeated inception-like blocks (3×, 4×, and 2×), each containing multiple small convolutions to enhance feature learning efficiency. Fully connected layers (black) are present at the end, followed by dropout layers (dark blue) to prevent overfitting and a Softmax layer (yellow) for final classification.

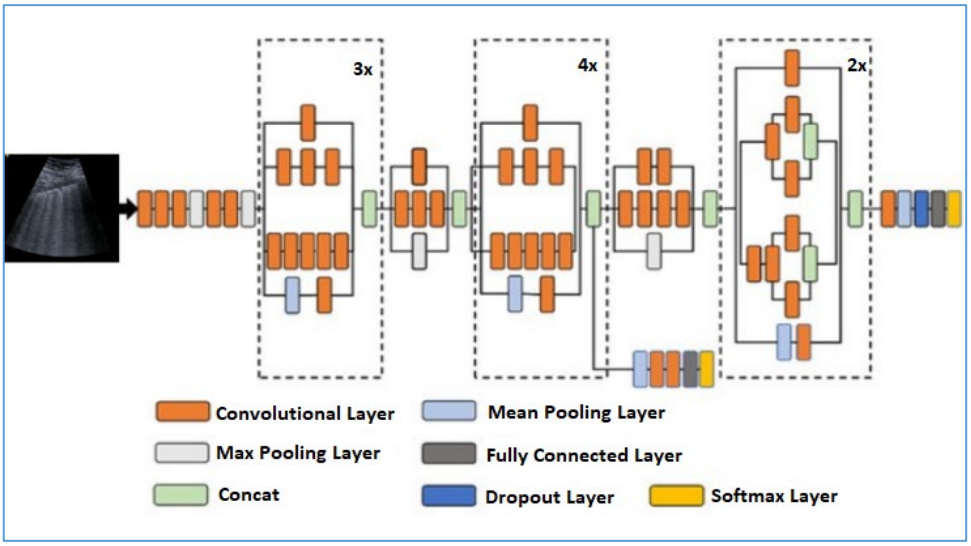

Fig. 2. Inception V3 architecture [30]

3.2.3. ResNet152V2

The Deep Residual Network (ResNet) architecture [31] was presented by He *et al.* in 2016. This groundbreaking work addressed the challenges of training very deep neural networks by introducing residual learning, which allows layers to learn residual functions with reference to the layer inputs, thereby facilitating the training of networks with substantially increased depth. The ResNet152V2 model is an advanced version of ResNet architecture and is part of the "ResNet v2" family, which improves upon the original ResNet by employing better training techniques and modifications to the residual blocks. The key feature of ResNet models is the use of residual blocks. These blocks incorporate skip connections (shortcuts) that facilitate the smooth flow of gradients through the network. ResNet152V2 is an advanced deep architecture consisting of 152 layers, designed for learning highly complex features in data. The modifications to the original design, include.

- Batch Normalization (BN) and ReLU activation are applied before the convolution layers instead of after, as in ResNet v1. That improves the optimization by allowing the skip connections to propagate the identity function more effectively.
- Each block is a bottleneck block, which reduces the computational cost. These blocks use 1x1, 3x3, and 1x1 convolutions [32].

Fig. 3 highlights the transition from the ResNet152V2 backbone (used for feature extraction) to the fully connected layers (used for task-specific prediction like LUS image classification into normal, pneumonia, or COVID-19). The architecture is adapted for a 3-class classification task, with dropout used to improve model generalization.

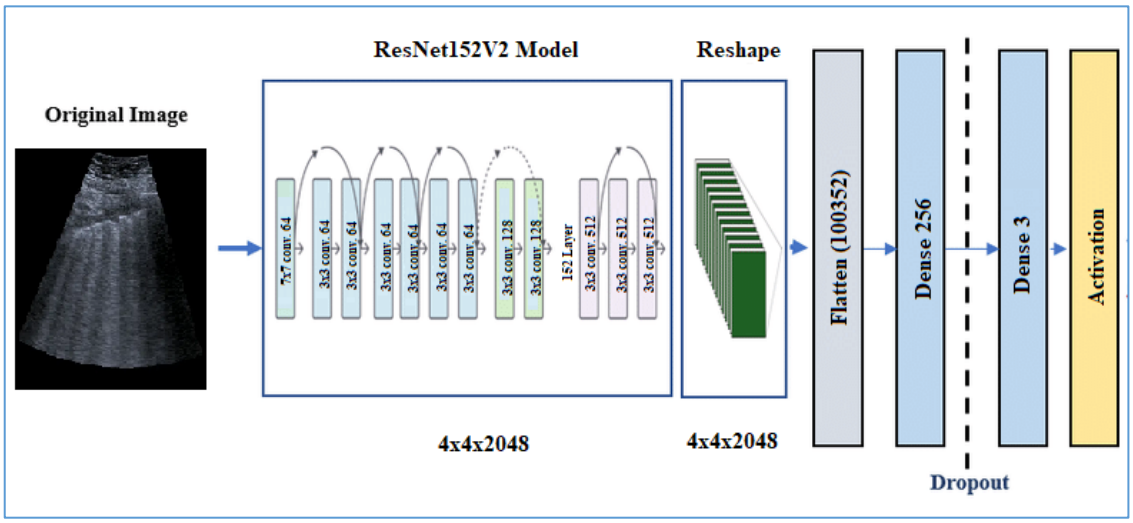

Fig. 3. ResNet152V2 architecture [32]

3.2.4. InceptionResNetV2

The InceptionResNetV2 model [33] is a powerful and efficient CNN architecture that benefits from the strengths of both Inception and ResNet, achieving high accuracy with manageable computational requirements. It represents a major advancement in DL model design and is significantly more complex than the InceptionV3 architecture. Unlike InceptionV3, this network has fewer parallel towers and features more streamlined Inception blocks. Each convolutional layer is paired with a ReLU activation function and is followed by batch normalization. The design of the InceptionResNetV2 layer is illustrated in Fig. 4.

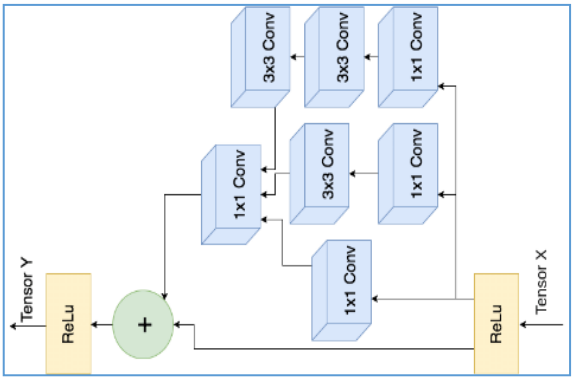

Fig. 4. InceptionResNetV2 layer [34]

3.3. The applied ensemble strategy

Ensemble learning is a technique that integrates multiple models or classifiers to enhance performance beyond what any individual model can achieve [35]. The predictions of each model are

merged to form an ensemble. This approach can be divided into four primary types: bagging, boosting, stacking, and voting. Bagging involves training multiple base models in parallel on different subsets of the original dataset to reduce variance and enhance stability. In contrast, boosting train base models sequentially, with each new model prioritizing correcting errors made by its predecessors, ultimately improving overall performance. Stacking, on the other hand, employs a meta-learner model that aggregates the predictions of various models to make the final decision [36]. In a voting ensemble, multiple models make predictions, and the final output is determined by majority voting for classification tasks or by averaging the predictions for regression tasks [37], [38]. This study employs the majority voting ensemble method to combine predictions from several pre-trained models, where the ensemble decision is formed by individually evaluating each model's prediction. By leveraging the diverse strengths of multiple models, majority voting enhances predictive accuracy while reducing overfitting through averaging individual biases. This approach not only improves robustness against noise and outliers, ensuring more stable predictions, but also offers a simple, interpretable, and easy-to-implement solution compared to more complex ensemble techniques like stacking or boosting [39], [40]. The implementation of the majority voting framework involved six key stages, as illustrated in Fig. 5.

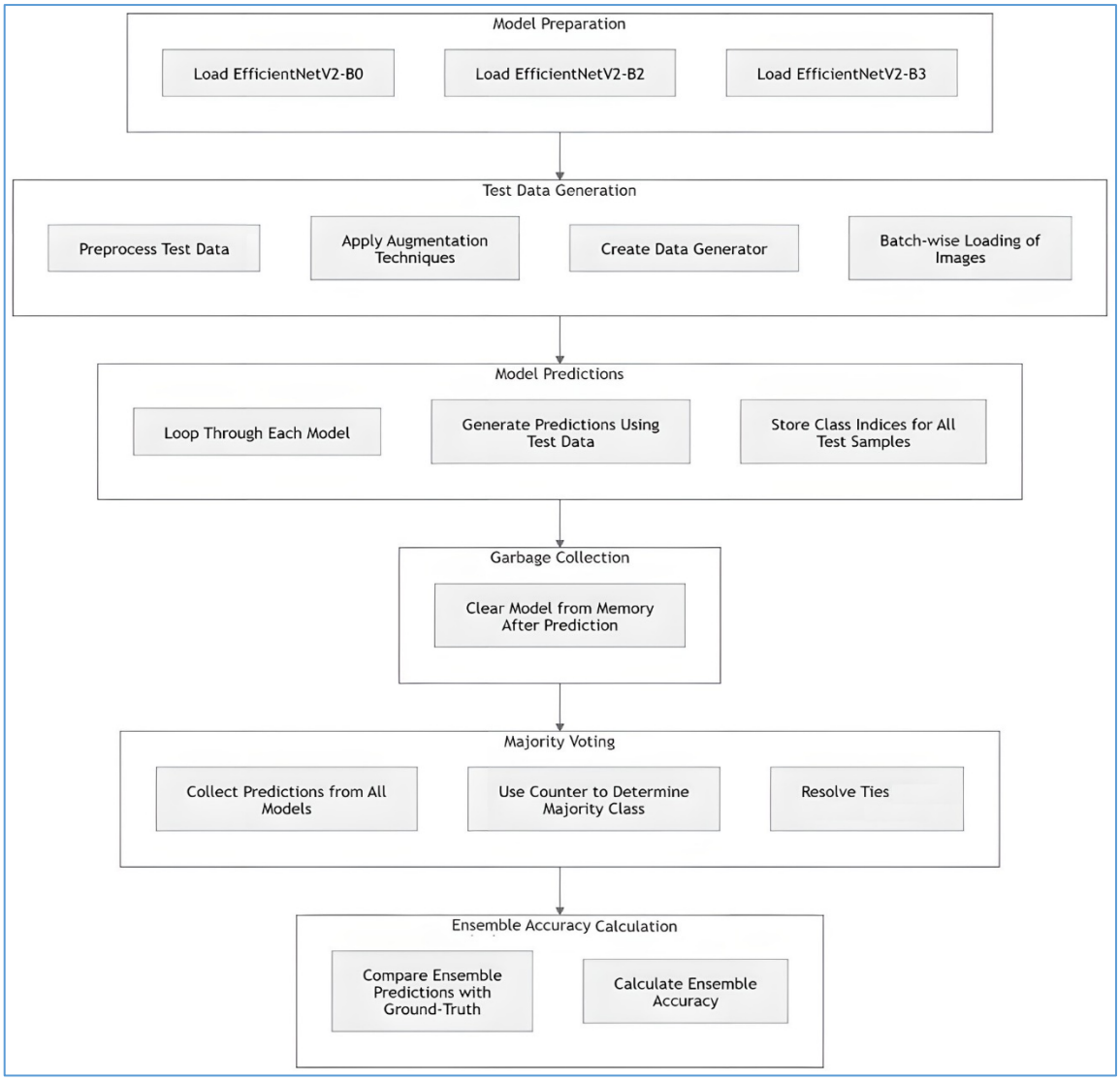

Fig. 5. Workflow for ensemble classification using majority voting

- Model Preparation: The process begins with loading the three best-performing pre-trained models, utilizing their saved versions for predictions.

- Test Data Generation: This step preprocesses the test data by applying various augmentation techniques such as zoom, rotation, shift, shear, and flip. A data generator is then created for efficient batch-wise image loading.
- Model Predictions: Each model is iterated over, using the test data to generate predictions. The predicted class indices for all test samples are stored for subsequent steps.
- Garbage Collection: After processing each model, memory is cleared to optimize RAM usage and prevent unnecessary resource consumption.
- Majority Voting: Predictions from all models are aggregated for each test sample. Using the Counter method, the majority class is determined.
- Ensemble Accuracy Calculation: Finally, the ensemble's predictions are compared with the ground truth labels to compute the overall accuracy of the framework. This stepwise approach ensures computational efficiency, robustness, and improved classification accuracy while leveraging the strengths of multiple DL models.

3.4. Evaluation

The dataset was partitioned into training, validation, and test sets to evaluate model performance rigorously. Training data, while useful for model development, cannot reliably assess predictive capability. Therefore, validation set performance provides a more accurate estimate of generalization to unseen data. The final model was tested on a held-out test set following validation to determine its classification accuracy on entirely novel data. The metrics commonly used in classification tasks are.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision+Recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{3}$$

where: TP – True Positive, TN – True Negative, FP – False Positive, and FN – False Negative

4. Results and Discussion

Practically, a Python-based implementation was employed, leveraging Nvidia RTX 3060 GPU capabilities. In the first phase, a lung classification model was developed using a pre-trained model followed by a custom classifier (average pooling, three fully connected layers, and a Softmax output for three classes). The training utilized the Adam optimizer (learning rate 1e-4), a batch size of 64, and a maximum of 100 epochs, with early stopping implemented after 10 epochs of no validation loss improvement. A preprocessing step was done by means of input normalization and augmentations, including geometric transformations (zoom, rotation, shifts, shear, flips). Each of the seven models was evaluated independently, and the results of their accuracies are demonstrated in Fig. 6. The evaluation metrics include accuracies of training, validation, and testing. A callback function was employed to save the best model during training, using validation accuracy as the selection criterion, which is useful in preventing overfitting and retaining the most effective model.

The EfficientNetV2 models displayed robust and consistent performance across the training, validation, and test datasets, with validation and test accuracies consistently exceeding 97%. Among these, EfficientNetV2-B1 achieved the highest test accuracy (98.32%), underscoring its superior generalization ability within this model family. EfficientNetV2-B3 followed closely, with a slightly lower test accuracy of 98.09%, maintaining strong overall performance and a balanced relationship between training and validation metrics. In comparison, ResNet152V2 also delivered impressive results, achieving one of the highest validation and test accuracies (98.50% and 97.98%, respectively), further highlighting its excellent generalization capabilities. Meanwhile, InceptionResNetV2 demonstrated moderate

effectiveness with a test accuracy of 96.67%, which was lower than both the EfficientNetV2 models and ResNet152V2. Lastly, InceptionV3 performed the worst out of all the models, with a test accuracy of 96.37%.

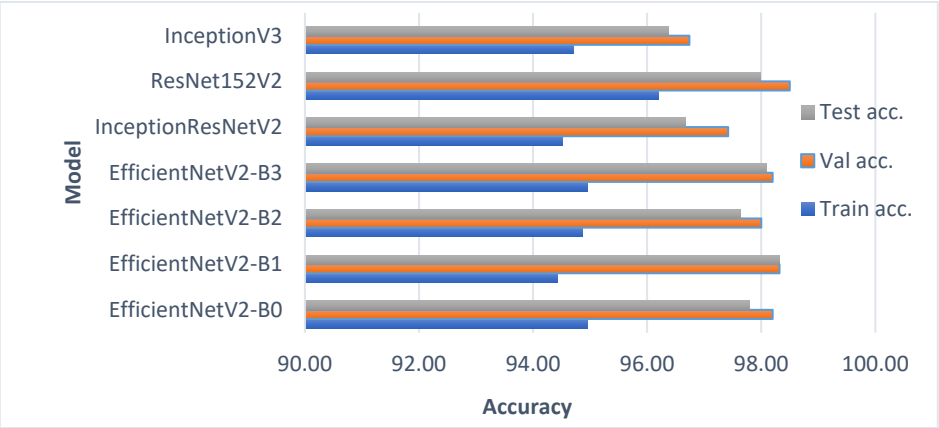

Fig. 6. Training, validation and test accuracies of the seven pre-trained models

To select the top three models, additional criteria, such as model complexity and efficiency were considered. Table 3 provides a detailed comparison of the seven models, including their performance with regard to a number of model parameters (measuring complexity), FLOPs (computational cost), and training time (efficiency). These metrics were evaluated to ensure the best possible balance between computational efficiency and performance.

Table 3. Comparison of Model Complexity, Computational Cost, and Training Time for the Seven Deep Learning Architectures

Model	Parameters (M)	FLOPs (B)	Training Time (Min.)
EfficientNetV2-B0	2.11	5.9	29.35
EfficientNetV2-B1	2.9	6.9	53.87
EfficientNetV2-B2	3.42	8.77	40.26
EfficientNetV2-B3	4.72	12.9	29.35
InceptionResNetV2	17.9	54.3	58.25
ResNet152V2	32	58.3	61.43
InceptionV3	7.88	21.8	72.25

EfficientNetV2-B0 is highly efficient, featuring a lightweight architecture (2.11M parameters), low computational cost (5.9B FLOPs), and a remarkably short training time of 29.35 minutes, making it an excellent option for resource-constrained tasks. EfficientNetV2-B1 delivers slightly better performance than B0 but comes at the expense of increased training time (53.87 minutes), reducing its efficiency for rapid deployment scenarios. EfficientNetV2-B2 offers a balanced trade-off between computational cost (8.77B FLOPs) and training time (40.26 minutes), presenting a middle ground within the EfficientNetV2 family. EfficientNetV2-B3, despite being more computationally complex with 12.9B FLOPs, retains a relatively short training time of 29.35 minutes, highlighting an effective balance between efficiency and performance. InceptionResNetV2, with 17.9M parameters and 54.3B FLOPs, is significantly heavier and more resource-intensive compared to the EfficientNetV2 models, making it less practical for efficiency-focused applications. Similarly, ResNet152V2, while demonstrating strong performance, is highly resource-demanding with 32M parameters and 58.3B FLOPs, which diminishes its suitability for real-world deployments. Lastly, InceptionV3 stands out as the least efficient in terms of training time (72.25 minutes) and computational cost (21.8B FLOPs), making it less competitive in scenarios requiring both performance and efficiency. Based on these results, the top three models selected were.

- EfficientNetV2-B0: Best choice for its exceptional computational efficiency and minimal training time while maintaining strong performance.
- EfficientNetV2-B3: A well-rounded model with a good trade-off between computational cost and training time, offering strong results.
- EfficientNetV2-B2: Offers a balance between model complexity and computational cost, slightly heavier than B0 but still efficient.

These models are chosen for their high accuracy and efficiency, making them ideal for scalable, resource-constrained applications. In the second phase, an ensemble of the three models was developed using majority voting, combining their predictions to enhance overall accuracy and robustness. It evaluates each model individually on the test dataset and uses the predictions of each model to form an ensemble decision. Fig. 7 presents the ensemble framework. In cases of a tie, EfficientNetB3 (the model having the best test accuracy) is utilized as the tie-breaker.

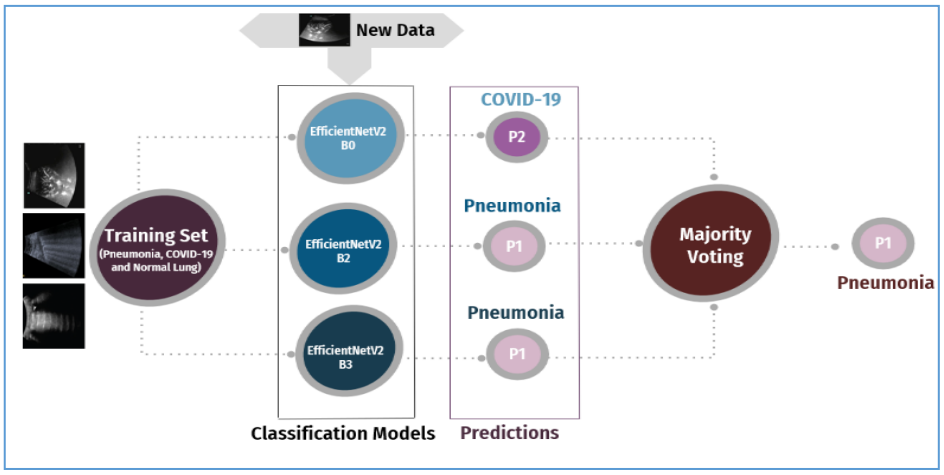

Fig. 7. Majority voting ensemble for lung ultrasound image classification

Fig. 8. A compares test dataset performance (accuracy, precision, F1-score) for three EfficientNetV2 models (B0, B2, B3) and the ensemble. Individually, EfficientNetV2-B3 exhibited the highest accuracy (98.09%), while B0 and B2 showed slightly lower, but still strong, performance (97.79% and 97.64%, respectively). All three models achieved near-perfect F1-score and precision (around 98%). Notably, the ensemble method significantly outperformed each model, achieving 99.25% accuracy and a 99% F1-score and precision. This improvement highlights the effectiveness of the majority voting ensemble in mitigating individual model limitations and improving overall classification accuracy and robustness. Although EfficientNetV2-B3 was the best-performing single model, the substantial performance gain from the ensemble demonstrates its value in enhancing the reliability of the classification system. The confusion matrix, shown in Fig. 8.b, shows excellent performance of the ensemble model on the three-class classification problem (Normal, Pneumonia, Covid). It strongly suggests that the model exhibits high accuracy and precision in classifying ultrasound images. Also, the very low number of misclassifications in each category indicates a robust and reliable model.

A further comparison with individual models in terms of model complexity and efficiency appears that the ensemble combines models but remains computationally manageable, as its parameter count (10.25M) is modest compared to larger standalone models like ResNet152V2 (-32M parameters). This indicates that the ensemble uses only the most efficient models without unnecessary complexity. Also, at 25.7B FLOPs, the ensemble is computationally heavier than EfficientNetV2-B0 (5.9B) or B3 (12.9B) but achieves significantly higher accuracy (99.25% compared to B3's 98.09%). The increased computation is justified by the substantial improvement in results. With an execution time of 1.20 minutes on a test dataset, the ensemble offers a practical balance of performance and speed, making it suitable for scalable applications such as medical diagnostics, where time efficiency is important.

Nevertheless, the encouraging outcomes of this study have some limitations. First, the comprehensive dataset might not adequately represent the diversity of LUS images encountered in real-world clinical practice. Variations in image quality, acquisition techniques, and patient demographics could affect model generalizability. Second, the ensemble framework, while highly accurate, requires the deployment of multiple models, which may increase computational overhead in some scenarios. Finally, the study focused on three specific lung conditions (Normal, Pneumonia, and COVID-19), and the framework's performance on other lung diseases or multi-label classification tasks remains unexplored.

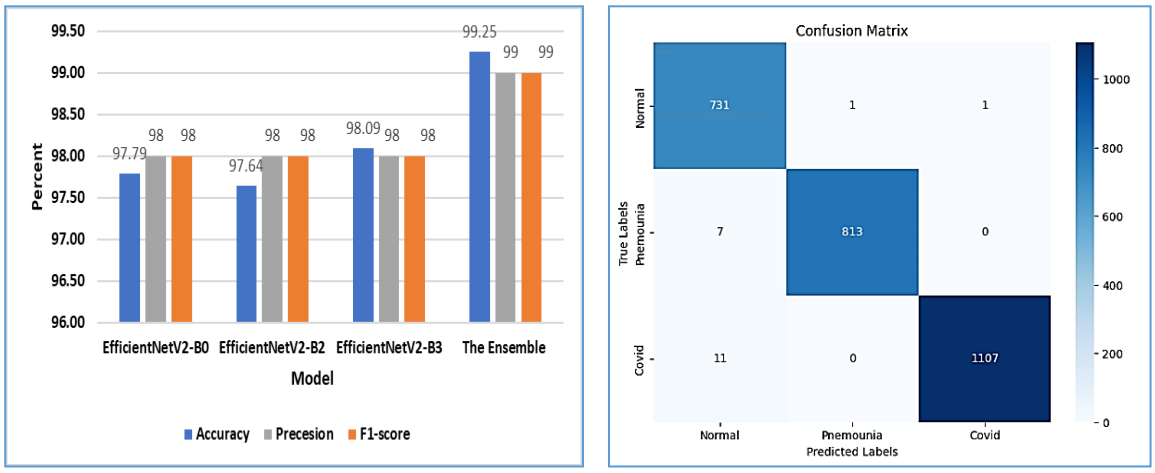

Fig. 8. Performance metrics: a) Comparison of individual EfficientNetV2 models and the ensemble model based on Accuracy, Precision, and F1-score, b) The confusion matrix for the ensemble model

5. Conclusion

This study investigated the application of an ensemble learning approach, specifically majority voting, to improve the accuracy of lung disease classification using LUS images. Three EfficientNetV2 models (B0, B2, and B3) were selected for their balance of performance and efficiency. The individual model evaluation revealed strong performance, with EfficientNetV2-B3 demonstrating the highest accuracy (98.09%). However, the ensemble method significantly outperformed all individual models, achieving a remarkable 99.25% accuracy, as shown by the confusion matrix and detailed evaluation metrics. This substantial improvement underscores the power of integrating multiple models to leverage their strengths and minimize weaknesses. The ensemble's computational efficiency, with a parameter count significantly lower than some larger models and a reasonable inference time, makes it a practical and scalable solution for real-world applications in medical diagnostics, where fast and accurate classification is crucial. Future work could explore expanding the ensemble to include a more diverse set of models or investigating alternative ensemble methods to enhance performance further. The findings suggest that ensemble learning offers a valuable strategy for enhancing the robustness and accuracy of automated lung disease diagnosis using ultrasound imaging.

Acknowledgment

The author(s) would like to express sincere gratitude to the Information Technology Industry Development Agency (ITIDA) – Information Technology Academia Collaboration (ITAC) for their financial support of this research. Their contributions have been invaluable in facilitating the successful completion of this study.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. This work is financially supported by the Information Technology Industry Development Agency (ITIDA) – 351 Information Technology Academia Collaboration (ITAC) program under grant number CFP 223.

Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] J. L. Díaz-Gómez, P. H. Mayo, and S. J. Koenig, "Point-of-Care Ultrasonography," *N. Engl. J. Med.*, vol. 385, no. 17, pp. 1593–1602, Oct. 2021, doi: 10.1056/NEJMra1916062.

[2] H. Scott, A. Zahra, R. Fernandes, B. C. Fries, H. C. Thode, and A. J. Singer, "Bacterial infections and death among patients with Covid-19 versus non Covid-19 patients with pneumonia," *Am. J. Emerg. Med.*, vol. 51, pp. 1–5, Jan. 2022, doi: 10.1016/j.ajem.2021.09.040.

[3] P. R. Joshi, "Pulmonary Diseases in Older Patients: Understanding and Addressing the Challenges," *Geriatrics*, vol. 9, no. 2, p. 34, Mar. 2024, doi: 10.3390/geriatrics9020034.

[4] D. Davis *et al.*, "Advancements in the Management of Severe Community-Acquired Pneumonia: A Comprehensive Narrative Review," *Cureus*, vol. 15, no. 10, Oct, pp. 1-11, 2023, doi: 10.7759/cureus.46893.

[5] World Health Organization "Coronavirus disease (COVID-19)." [Online]. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[6] F. J. Rodríguez-Contreras *et al.*, "Lung Ultrasound Performed by Primary Care Physicians for Clinically Suspected Community-Acquired Pneumonia: A Multicenter Prospective Study," *Ann. Fam. Med.*, vol. 20, no. 3, pp. 227–236, May 2022, doi: 10.1370/afm.2796.

[7] J. Born *et al.*, "POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS)," *Eur. PMC*, no. April, pp. 1–14, 2020. [Online]. Available at: https://europepmc.org/article/PPR/PPR268507.

[8] S. Bahri, Suprijanto, and E. Juliastuti, "Texture Analysis of Ultrasound Images to Differentiate Pneumonia and Covid-19," in *IEEE International Biomedical Instrumentation and Technology Conference (IBITeC), Yogyakarta, Indonesia*, 2021, pp. 24–28, doi: 10.1109/IBITeC53045.2021.9649067.

[9] M. La Salvia *et al.*, "Deep learning and lung ultrasound for Covid-19 pneumonia detection and severity classification," *Comput. Biol. Med.*, vol. 136, p. 104742, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104742.

[10] O. Elkhuoly, M. Malhat, A. Keshk, and M. Elsabaawy, "A Comparative Analysis of COVID-19 Diagnosis Using Lung Ultrasound Based on Convolutional Neural Networks," *IJCI. Int. J. Comput. Inf.*, vol. 10, no. 1, pp. 0–0, Sep. 2022, doi: 10.21608/ijci.2022.151629.1079.

[11] S. Shang *et al.*, "Performance of a computer aided diagnosis system for SARS-CoV-2 pneumonia based on ultrasound images," *Eur. J. Radiol.,* vol. 146, p. 110066, Jan, doi: 10.1016/j.ejrad.2021.110066.

[12] S. Morsy *et al.*, "Impact of Dataset Size on the Performance of Deep Learning Models: A Case Study on Lung Ultrasound and X-Ray Image Classification," *SSRN Electron. J.*, vol. 8, no. 4, pp. 21–30, Aug. 2024, doi: 10.2139/ssrn.5044890.

[13] E. A. Nehary, S. Rajan, and C. Rossa, "Lung Ultrasound Image Classification Using Deep Learning and Histogram of Oriented Gradients Features for COVID-19 Detection," in *2023 IEEE Sensors Applications Symposium (SAS)*, 2023, pp. 1–6, doi: 10.1109/SAS58821.2023.10254002.

[14] J. Himasree, K. Aravindhan, K. P. Keerthana, and C. Gobinath, "Design of an Efficient Deep Learning Framework for Covid-19 Image Classification," in *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, 2023, vol. 1, pp. 1–4, doi: 10.1109/ICCAMS60113.2023.10525806.

[15] A. Subramanyam and M. Sucharitha, "Multi-classification of Lung Diseases Using Lung Ultrasound Imaging BT - AI Technologies for Information Systems and Management Science," 2024, pp. 510–521, doi: 10.1007/978-3-031-66410-6_40.

[16] G. Madhu, S. Kautish, Y. Gupta, G. Nagachandrika, S. M. Biju, and M. Kumar, "XCovNet: An optimized xception convolutional neural network for classification of COVID-19 from point-of-care lung ultrasound images," *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 33653–33674, 2024, doi: 10.1007/s11042-023-16944-z.

[17] P. T. H. Nhat *et al.*, "Clinical benefit of AI-assisted lung ultrasound in a resource-limited intensive care unit.," *Crit. Care*, vol. 27, no. 1, p. 257, Jul. 2023, doi: 10.1186/s13054-023-04548-w.

[18] L. Howell, N. Ingram, R. Lapham, A. Morrell, and J. R. McLaughlan, "Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound.," *Ultrasonics*, vol. 140, p. 107251, May 2024, doi: 10.1016/j.ultras.2024.107251.

[19] W. Xing *et al.*, "Automated lung ultrasound scoring for evaluation of coronavirus disease 2019 pneumonia using two-stage cascaded deep learning model," *Biomed. Signal Process. Control*, vol. 75, p. 103561, May 2022, doi: 10.1016/j.bspc.2022.103561.

[20] R. Chaudhary *et al.*, "Diagnostic accuracy of an automated classifier for the detection of pleural effusions in patients undergoing lung ultrasound," *Am. J. Emerg. Med.*, vol. 90, pp. 142–150, Apr. 2025, doi: 10.1016/j.ajem.2025.01.041.

[21] A. K. Dubey *et al.*, "Ensemble Deep Learning Derived from Transfer Learning for Classification of COVID-19 Patients on Hybrid Deep-Learning-Based Lung Segmentation: A Data Augmentation and Balancing Framework," *Diagnostics*, vol. 13, no. 11, p. 1954, Jun. 2023, doi: 10.3390/diagnostics13111954.

[22] D. E. Shea *et al.*, "Deep Learning Video Classification of Lung Ultrasound Features Associated with Pneumonia," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3103–3112, doi: 10.1109/CVPRW59228.2023.00312.

[23] U. Khan, A. Smargiassi, R. Inchingolo, and L. Demi, "A Novel Weighted Majority Voting-Based Ensemble Framework for Lung Ultrasound Pattern Classification in Pneumonia Patients," in *2023 IEEE International Ultrasonics Symposium (IUS)*, 2023, pp. 1–4, doi: 10.1109/IUS51837.2023.10308194.

[24] M. Iman, K. Rasheed, and H. R. Arabnia, "A Review of Deep Transfer Learning and Recent Advancements," pp. 1-14, Jan. 2022, doi: 10.3390/technologies11020040.

[25] X. Li *et al.*, "Principled and efficient transfer learning of deep models via neural collapse," *arXiv Prepr. arXiv2212.12206*, pp. 1-30, 2022. [Online]. Available at: https://arxiv.org/abs/2212.12206.

[26] M. W. Ahdi, Khalid, A. Kunaefi, B. A. Nugroho, and A. Yusuf, "Convolutional Neural Network (CNN) EfficientNet-B0 Model Architecture for Paddy Diseases Classification," in *2023 14th International Conference on Information & Communication Technology and System (ICTS)*, Oct. 2023, pp. 105–110, doi: 10.1109/ICTS58770.2023.10330828.

[27] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114. [Online]. Available at: https://arxiv.org/abs/1905.11946v5.

[28] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," *in Proceedings of Machine Learning Research*, 2021, vol. 139, pp. 10096–10106, [Online]. Available at: https://arxiv.org/abs/1905.11946.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[30] A. Arini, M. Azhari, I. I. A. Fitri, and F. Fahrianto, "Performance Analysis of Transfer Learning Models for Identifying AI-Generated and Real Images," *J. Tek. Inform.*, vol. 17, no. 2, pp. 139–152, Oct. 2024, doi: 10.15408/jti.v17i2.40453.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[32] N. Parimala and G. Muneeswari, "Modified ResNet152v2: Binary Classification and Hybrid Segmentation of Brain Stroke Using Transfer Learning-Based Approach," *Polish J. Med. Phys. Eng.*, vol. 30, no. 1, pp. 24–35, Mar. 2024, doi: 10.2478/pjmpe-2024-0004.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, pp. 4278–4284, Feb. 2017, doi: 10.1609/aaai.v31i1.11231.

[34] R. P. Tripathi, S. K. Khatri, and D. V. G. Baxodirovna, "A transfer learning approach to implementation of pretrained CNN models for Breast cancer diagnosis," *J. Posit. Sch. Psychol.*, pp. 5816–5830, 2022. [Online]. Available at: https://journalppw.com/index.php/jpsp/article/view/4358.

[35] P. Schneider and F. Xhafa, *Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to EHealth and Patient Data Monitoring*. Elsevier Science, pp. 1-406, 2022, doi: 10.1016/B978-0-12-823818-9.00013-4.

[36] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[37] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, 2022, doi: 10.1016/j.engappai.2022.105151.

[38] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren, "StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1547–1557, Feb. 2021, doi: 10.1109/TVCG.2020.3030352.

[39] A. Desiani *et al.*, "Majority Voting as Ensemble Classifier for Cervical Cancer Classification," *Sci. Technol. Indones.*, vol. 8, no. 1 SE-Articles, pp. 84–92, Jan. 2023, doi: 10.26554/sti.2023.8.1.84-92.

[40] M. Azad, T. H. Nehal, and M. Moshkov, "A novel ensemble learning method using majority based voting of multiple selective decision trees," *Computing*, vol. 107, no. 1, p. 42, 2024, doi: 10.1007/s00607-024-01394-8.