

LETTER OPEN ACCESS

Automated AI-Based Lung Disease Classification Using Point-of-Care Ultrasound

Nixson Okila¹  | Andrew Katumba^{2,3,4}  | Joyce Nakatumba-Nabende^{1,3,4}  | Sudi Murindanyi^{1,3,5}  | Jonathan Serugunda² | Cosmas Mwikirize² | Samuel Bugeza⁶ | Anthony Oriekot⁷ | Juliet Bosa⁸ | Eva Nabawanuka⁸

¹Department of Computer Science, Makerere University, Kampala, Uganda | ²Department of Electrical and Computer Engineering, Makerere University, Kampala, Uganda | ³Makerere Center for Artificial Intelligence and Data Science, Makerere University, Kampala, Uganda | ⁴Emergent AI, Kampala, Uganda | ⁵Center for Cognitive Interaction Technology (CITEC), Faculty of Technology, Bielefeld University, Bielefeld, Germany | ⁶Department of Radiology, Makerere University Hospital, Makerere University, Kampala, Uganda | ⁷Mulago Specialised Women and Neonatal Hospital, Kampala, Uganda | ⁸Mulago National Referral Hospital, Kampala, Uganda

Correspondence: Andrew Katumba (andrew.katumba@mak.ac.ug)

Received: 30 May 2025 | **Revised:** 29 August 2025 | **Accepted:** 17 October 2025

Funding: This work was supported by Makerere University, International Development Research Centre, and Swedish Foundation for International Cooperation in Research and Higher Education.

Keywords: computer vision | deep learning | explainable AI (XAI) | lung ultrasound

ABSTRACT

Timely and accurate diagnosis of lung diseases is critical for reducing related morbidity and mortality. Lung ultrasound (LUS) has emerged as a useful point-of-care tool for evaluating various lung conditions. However, interpreting LUS images remains challenging due to operator-dependent variability, low image quality, and limited availability of experts in many regions. In this study, we present a lightweight and efficient deep learning model, ParSE-CNN, alongside fine-tuned versions of VGG-16, InceptionV3, Xception, and Vision Transformer architectures, to classify LUS images into three categories: COVID-19, other lung pathology, and healthy lung. Models were trained using data from public sources and Ugandan healthcare facilities, and evaluated on a held-out Ugandan dataset. Fine-tuned VGG-16 achieved the highest classification performance with 98% accuracy, 97% precision, 98% recall, and a 97% F1-score. ParSE-CNN yielded a competitive accuracy of 95%, precision of 94%, recall of 95%, and F1-score of 97% while offering a 58.3% faster inference time (0.006 s vs. 0.014 s) and a lower parameter count (5.18 M vs. 10.30 M) than VGG-16. To enhance input quality, we developed a preprocessing pipeline, and to improve interpretability, we employed Grad-CAM heatmaps, which showed high alignment with radiologically relevant features. Finally, ParSE-CNN was integrated into a mobile LUS workflow with a PC backend, enabling real-time AI-assisted diagnosis at the point of care in low-resource settings.

1 | Introduction

Lung diseases are a diverse group of disorders that impair the function of the lungs, which are essential for gas exchange in the human body [1]. These conditions range from common infections like bacterial pneumonia, tuberculosis, and COVID-19 to chronic diseases such as asthma and COPD, affecting various

parts of the respiratory system and leading to symptoms like shortness of breath, coughing, and chest pain [2].

Globally, lung diseases represent a significant health burden, contributing to high morbidity and mortality rates [3]. Tuberculosis causes over 1.6 million deaths annually, with more than 10.5 million new cases reported each year [4].

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Applied AI Letters* published by John Wiley & Sons Ltd.

Bacterial pneumonia claims approximately 2.56 million lives annually [5], while by the end of 2024, the COVID-19 pandemic had resulted in more than 776 million confirmed cases and over 7 million deaths worldwide [6]. Early and accurate diagnosis is critical for improving clinical outcomes and reducing healthcare costs.

Conventional imaging modalities, such as chest x-ray and computed tomography (CT), remain the clinical reference standards but are associated with notable limitations. CT provides high sensitivity; however, its use entails substantial radiation exposure [7]. Chest x-ray, although widely accessible, often exhibits insufficient sensitivity for detecting early-stage pathologies [8]. Moreover, both modalities necessitate specialized infrastructure and incur considerable acquisition and maintenance costs, limiting their applicability in remote or resource-constrained settings [9].

Lung ultrasound (LUS) has emerged as a viable alternative, offering real-time, radiation-free, and low-cost imaging capabilities [10]. It is particularly effective in detecting interstitial syndromes, pleural effusions, and alveolar consolidations [11]. However, LUS interpretation is highly operator-dependent and susceptible to variability due to the presence of imaging artifacts [12], thereby limiting its scalability.

To mitigate these challenges, deep learning (DL) models, particularly convolutional neural networks (CNNs), have been increasingly applied for automated LUS image interpretation [13–17]. Classical CNN architectures such as VGG, Inception, and Xception have shown considerable promise in medical image analysis [18], while transformer-based models such as Vision Transformer (ViT) have recently demonstrated comparable or superior performance in image classification tasks [19, 20].

For instance, Born et al. [13] introduced POCOVID-Net, a VGG-16-based CNN designed for the automatic detection of COVID-19 from LUS images. Their model achieved a classification accuracy of 89%, with a sensitivity of 96% and specificity of 79%. However, performance limitations were noted due to training on an imbalanced dataset, which contributed to misclassification between COVID-19 and healthy cases.

In a related study, Diaz-Escobar et al. [14] evaluated four pre-trained CNN architectures—VGG19, InceptionV3, Xception, and ResNet50—for COVID-19 classification from LUS images. Among these, InceptionV3 demonstrated the highest performance, achieving an accuracy of 89.1%, a precision of 90.1%, a recall of 86.6%, and an F1-score of 88.0%. Similarly, Roberts and Tsiglikaridis [15] compared the performance of VGG-16 and ResNet18 for COVID-19 classification in LUS data, reporting superior results for VGG-16, with an accuracy of 85.99%.

Roy et al. [16] proposed a CNN model inspired by Spatial Transformer Networks to assess COVID-19 severity from LUS images. Their approach enabled both classification and localization of pathological features, highlighting the potential of spatially aware DL models in clinical imaging tasks.

Magrelli et al. [17] extended this work to three-class classification (healthy, bronchiolitis, and bacterial pneumonia) by

training four state-of-the-art DL models: VGG19, Xception, InceptionV3, and Inception-ResNet-v2 using pediatric LUS images. The InceptionV3 achieved the best performance, with an accuracy and sensitivity of 91.5%, and a specificity of 95.86%.

More recently, Fiorentino et al. [20] employed ViTs for COVID-19 pneumonia assessment from LUS images. Their implementation of the Swin Transformer Tiny architecture achieved an F1-score of 0.6513 on the International Consortium for Lung Ultrasound Study dataset, indicating the growing utility of transformer-based models in LUS analysis.

In parallel with model development, advancements in explainable AI (XAI) have played a vital role in enhancing the transparency and interpretability of DL models in medical imaging [21]. Among the most widely adopted XAI techniques is Gradient-weighted Class Activation Mapping (Grad-CAM), which generates visual explanations by highlighting salient regions in input images that most influence model predictions. This level of interpretability is particularly valuable in clinical settings, where understanding the rationale behind automated decisions fosters trust among healthcare professionals and facilitates the integration of AI tools into diagnostic workflows [21, 22].

The utility of Grad-CAM in LUS interpretation has been demonstrated in several studies. For example, Arntfield et al. [23] conducted a multicenter study employing Grad-CAM to distinguish between normal (A-lines) and abnormal (B-lines) lung parenchyma. Their model, trained on a large labeled LUS dataset, achieved high diagnostic accuracy while providing clinically meaningful visualizations. In related applications, Grad-CAM has been successfully applied to chest x-ray and CT images to localize COVID-19-related anomalies, supporting differential diagnosis between COVID-19 and other forms of pneumonia [24]. Furthermore, Duque et al. [25] explored the use of Grad-CAM in ultrasound image segmentation, demonstrating improved delineation of anatomical borders and strong alignment with expert annotations. These findings underscore Grad-CAM's potential to not only enhance model interpretability but also improve diagnostic performance in ultrasound-based clinical applications.

Despite recent progress, several limitations persist. Many existing models are computationally demanding, relying on architectures with millions of parameters, which limits their feasibility for real-time inference and deployment on mobile or edge devices. A further challenge lies in the scarcity of representative datasets from low- and middle-income countries (LMICs) such as Uganda, where both the prevalence and imaging manifestations of respiratory diseases may differ from those reported in high-income settings [26]. As a result, most existing approaches are trained and evaluated on data that lack geographical diversity, raising concerns about their generalizability and reliability when applied in LMIC contexts [27]. In addition, while XAI techniques such as Grad-CAM have gained traction for enhancing clinical trust by highlighting regions influencing model decisions, their adoption into clinical workflows is still minimal. Moreover, limited work has addressed the development of models specifically tailored to the LUS domain in terms of computational efficiency, robustness to noise, and suitability for mobile deployment. Most studies focus primarily on classification performance without evaluating practical deployment constraints.

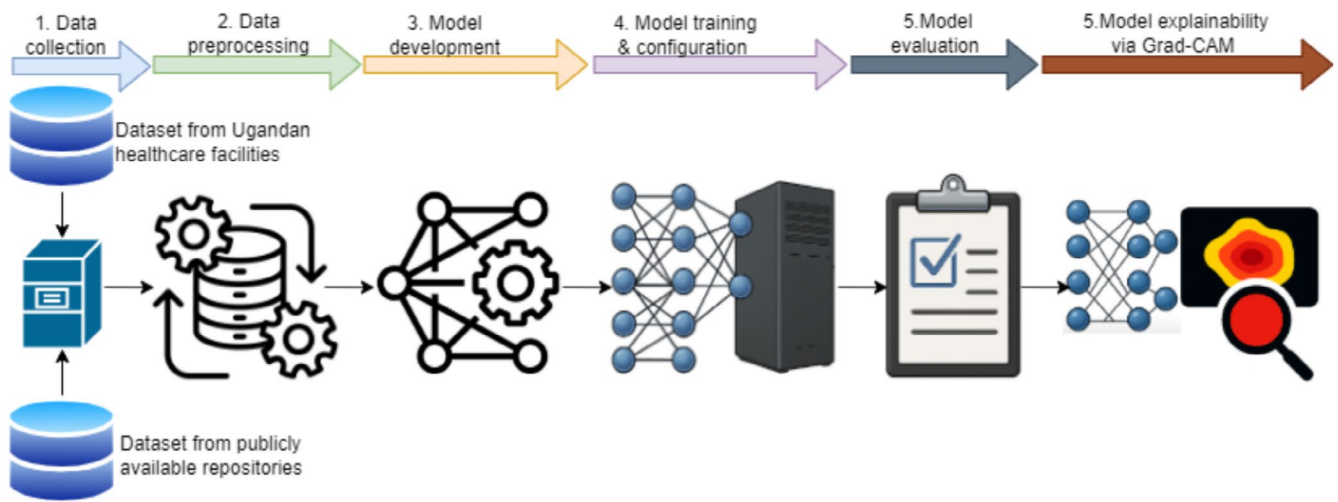


FIGURE 1 | A workflow diagram depicting the methodology used to develop the deep learning model for classifying lung diseases. It begins with data collection and progresses through stages of data preprocessing, model development, model training, model evaluation, and model explainability with Grad-CAM.

In this study, we address these challenges through the following key contributions:

1. We develop a custom lightweight CNN model, *ParSE-CNN*, specifically optimized for edge deployment, including mobile and portable devices. The model achieves competitive accuracy with substantially reduced parameters and inference time, making it well suited for point-of-care applications in low-resource settings.
2. We introduce a domain-specific preprocessing pipeline for LUS images, using OpenCV and Keras-OCR, to enhance image quality by isolating the region of interest (ROI) and removing distracting labels, an often overlooked step in prior works.
3. We fine-tune state-of-the-art models (*VGG-16*, *InceptionV3*, *Xception*, and *Vision Transformer*) through customized classification layers and selective layer unfreezing, and benchmark their performance against the proposed *ParSE-CNN*.
4. We integrate the proposed *ParSE-CNN* model into a mobile LUS screening workflow, running on a PC backend, to support real-time lung disease detection in remote areas lacking radiologists and high-end imaging equipment.

The remainder of this paper is organized as follows: Section 2 describes the methods employed in this study. Section 3 presents the experimental results, while Section 4 details the integration of the models into a mobile LUS screening tool. Section 5 reports the inference times before and after integration. Section 6 provides a discussion of the findings, and Section 7 concludes the paper.

2 | Methods

In this section, we present the methodology workflow used to develop the DL for classifying lung diseases as shown in Figure 1.

2.1 | Datasets

This study utilized datasets obtained from public repositories and from Ugandan healthcare facilities, specifically Mulago National Referral Hospital and Kiruddu National Referral Hospital, both located in Kampala.

2.1.1 | Dataset From Public Repository

We obtained a publicly available dataset from [28, 29]. The dataset included convex probe LUS videos categorized into COVID-19, bacterial pneumonia, viral pneumonia, and healthy lung classes. To harmonize the class structure with that of the dataset collected from Ugandan healthcare facilities, the bacterial and viral pneumonia categories were merged into a single class labeled “Other lung pathology.” In total, we acquired 113 LUS videos: 40 depicting COVID-19, 40 categorized as “Other lung pathology,” and 33 representing healthy lung conditions.

2.1.2 | Dataset From Ugandan Healthcare Facilities

The second dataset was acquired from patients diagnosed with COVID-19, as well as those presenting with pulmonary-related pathologies such as non-COVID-19-derived pneumonia. Data collection was conducted by three radiologists at clinical study sites situated in Mulago and Kiruddu National Referral Hospitals, Uganda. These sites were established following ethical approval from the Mulago Hospital Research Ethics Committee (No. MHREC 2021-45, dated August 24, 2021) and the Uganda National Council for Science and Technology (No. SIR74ES, dated October 22, 2021). Prior to participation, each patient provided informed consent, voluntarily agreeing to undergo LUS scanning with a clear understanding of the study's objectives and any associated risks.

During LUS scanning, radiologists divided each patient's thorax into 12 regions: six on the left to represent upper posterior, lower

posterior, upper anterior, lower anterior, upper lateral, and lower lateral and six on the right to represent upper posterior, lower posterior, upper anterior, lower anterior, upper lateral, and lower lateral. The radiologists conducted both longitudinal and transverse scans for each region. This systematic division allowed for a comprehensive evaluation of the lungs, facilitating the identification of localized abnormalities that might differ across regions [30]. Moreover, performing scans in both orientations provided multiple perspectives, improving the detection of hidden pathologies that could be missed with a single scanning approach [31]. All scans were performed using the Clarius C3 and C5 curvilinear handheld ultrasound probes [32], with a frame rate of 20 Hz, depth of 16.0 cm, gain set to 53, and power output of -0.3 dB.

For each patient, the radiologist acquired 24 LUS images, which were saved in either *png* or *jpeg* format within a uniquely identified folder. Each folder was labeled based on the patient's clinical diagnosis as "Probably COVID-19," "Diseased Lung but Probably Not COVID-19," or "Healthy Lung." For the purposes of this study, these categories were standardized and are hereafter referred to as COVID-19, Other lung pathology, and Healthy lung, respectively.

2.2 | Data Preprocessing

2.2.1 | Preprocessing of Publicly Available Dataset

We partitioned the videos into training and validation sets in an 80:20 ratio based on patient ID, ensuring that both sets did not contain videos from the same patient. We then extracted image frames from each video at a frame rate of 3 Hz using the OpenCV

TABLE 1 | Number of preprocessed lung ultrasound images for each lung condition category in the publicly available dataset, divided into training and validation subsets.

Lung condition	Training	Validation	Total
COVID-19	310	93	403
Other lung pathology	310	93	403
Healthy lung	310	93	403

Note: Each category includes an equal number of images in both subsets.

library [33]. In collaboration with radiologists, we selected 1209 high-resolution images with reliable diagnostic information to support DL model training. Table 1 presents the distribution of preprocessed images across the training and validation sets for each lung condition.

Figure 2 presents representative preprocessed image frames derived from the publicly available dataset.

2.2.2 | Preprocessing of Dataset From Ugandan Healthcare Facilities

We performed a quality assessment of each patient's images and identified the following challenges:

1. Certain images exhibited low resolution, which could potentially degrade the model's accuracy.
2. All image frames contained extensive regions of noninterest, which could increase the computational demands of the model.
3. Ultrasound labels were overlaid on the images, introducing artifacts that could interfere with feature extraction.

To address the first challenge, we removed all low-resolution images, resulting in a final dataset of 1062 images across the three lung condition classes. The dataset was subsequently partitioned into training, validation, and test sets using a 70:20:10 ratio at the patient level, ensuring that all images from a single patient were assigned to a single partition. Table 2 shows the distribution of images across the training, validation, and test sets used for model development and evaluation.

To overcome the second and third challenges, we designed a custom image preprocessing pipeline as shown in Figure 3 that utilizes the OpenCV library to crop the RoI from each image frame and employs Keras-OCR for inpainting to remove overlaid ultrasound labels.

The image preprocessing pipeline commenced by converting each image to grayscale using the OpenCV library [33] to reduce computational complexity while retaining relevant structural details.

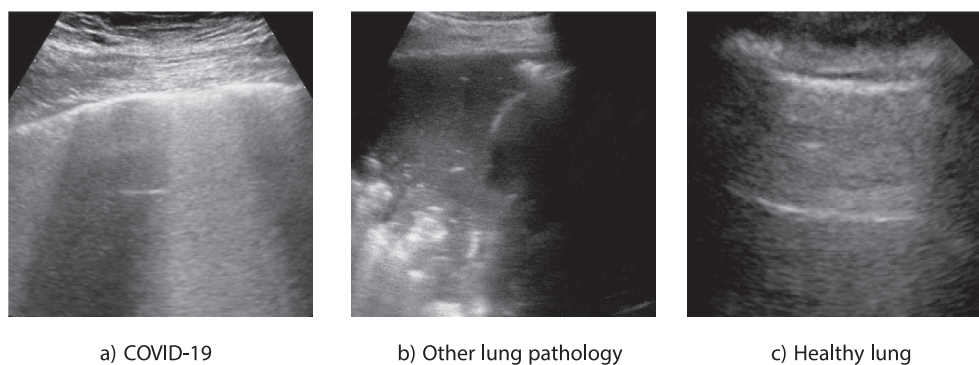


FIGURE 2 | Representative preprocessed lung ultrasound image frames from the publicly available dataset, illustrating the three lung condition categories used in model training. (a) COVID-19 case demonstrating vertical hyperechoic artifacts (B-lines). (b) Image from a patient with another pulmonary pathology showing consolidation. (c) Healthy lung image characterized by horizontal, repetitive, and equidistant reverberation artifacts (A-lines).

Otsu's thresholding [34] was applied to automatically determine an optimal binarization threshold, chosen for its robustness in segmenting images with bimodal histograms typical of LUS images containing high-contrast artifacts and background. To improve region continuity and suppress small holes or noise, morphological closing and dilation were performed. These operations were empirically selected based on visual inspection across a validation subset to ensure the enhancement of lung boundary structures without excessive merging of unrelated regions.

Contours were then extracted from the binary mask, and the largest contour was assumed to correspond to the lung field, based on prior anatomical knowledge that the pleural line and associated artifacts typically occupy the largest central area in

TABLE 2 | Number of lung ultrasound images for each lung condition category: COVID-19, Other lung pathology, and healthy lung collected from healthcare facilities in Uganda.

Lung condition	Training	Validation	Test	Total
COVID-19	241	66	31	338
Other lung pathology	241	66	55	362
Healthy lung	241	66	55	362

standard LUS frames. A bounding box around this RoI was computed, and the image was cropped accordingly.

To remove overlaid ultrasound labels, Keras-OCR [35] was chosen for its strong performance on low-resolution and varied-font text common in ultrasound annotations. Detected text regions were represented as quadrilaterals, and midpoints along their vertical edges were used to guide the drawing of thick black lines (thickness proportional to the Euclidean distance between text points) to effectively mask the text. This strategy allowed for adaptive coverage of text regardless of font size or rotation. Inpainting was then applied using OpenCV's Telea method, selected for its efficiency and ability to reconstruct smooth image backgrounds by interpolating surrounding pixel intensities. This approach preserved anatomical structures while eliminating textual noise. Figure 4 shows an example of an original and processed image.

2.3 | Model Development

To develop DL models for the classification of lung diseases from LUS images, we implemented three categories of models: state-of-the-art (SoTA) architectures, fine-tuned versions of these architectures, and a custom-designed CNN. These models were systematically trained and evaluated to assess their effectiveness in distinguishing between different lung pathologies.

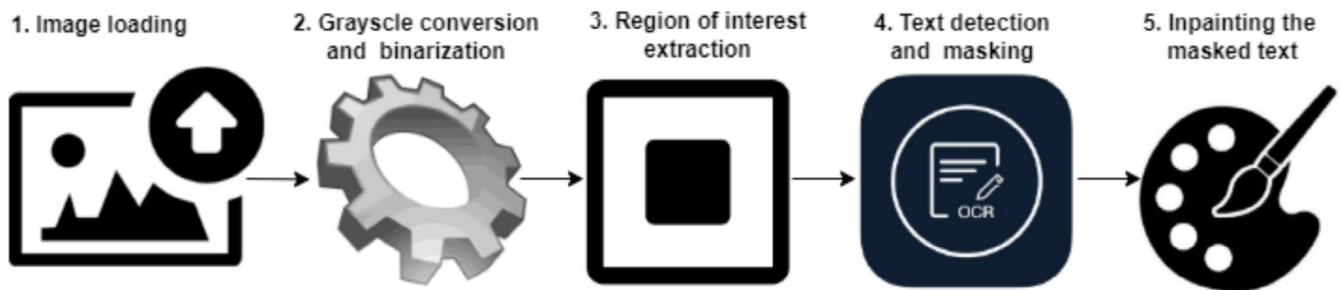


FIGURE 3 | Overview of the custom image preprocessing pipeline, which uses OpenCV to extract the region of interest from each lung ultrasound image and applies Keras-OCR with inpainting techniques to detect and remove overlaid ultrasound labels.

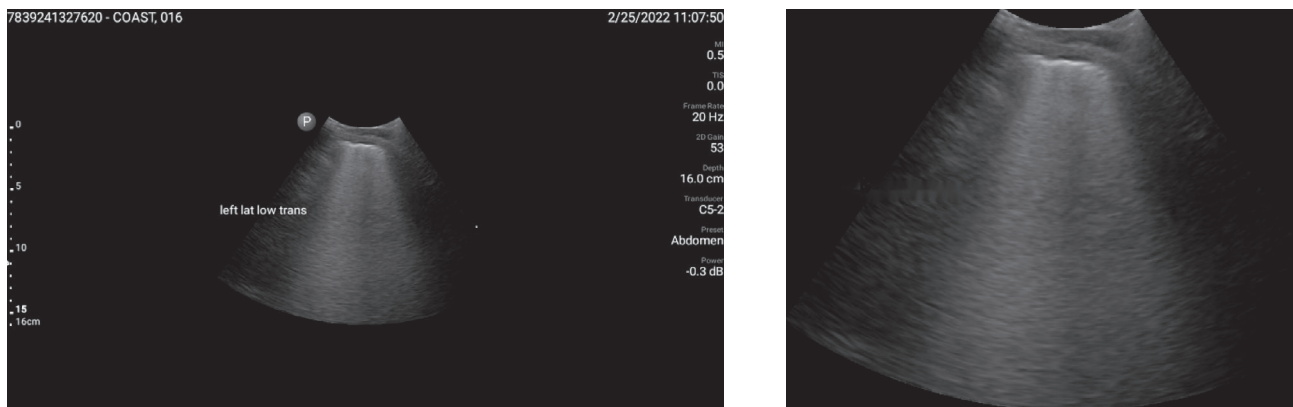


FIGURE 4 | Example of a raw lung ultrasound image and its corresponding preprocessed version. (a) Raw image containing irrelevant background and overlaid text labels. (b) Preprocessed image showing the cropped ROI with labels removed.

2.3.1 | DL Models

This study employed VGG-16, InceptionV3, Xception, and ViT as baseline models for lung disease classification in LUS images. VGG-16 is a deep CNN consisting of 16 weight (learnable parameter) layers and utilizes small 3×3 convolutional filters for hierarchical feature extraction [36]. Its architecture follows a uniform structure with max-pooling layers and fully connected (FC) layers for classification. InceptionV3 incorporates factorized convolutions and inception modules to capture multiscale features, thereby enhancing its ability to discriminate complex image patterns [37]. Xception, an extension of the Inception architecture, employs depthwise separable convolutions to improve parameter efficiency and enhance feature extraction [38]. ViT is a transformer-based model that partitions images into nonoverlapping patches and processes them using self-attention mechanisms, enabling it to capture long-range dependencies and contextual information [19].

The SoTA DL models, originally pretrained on the ImageNet dataset [39], were adapted for the LUS dataset by customizing their classification layers while keeping their convolutional layers frozen to establish baseline classification performance. The customized classification architecture consists of two FC layers. The first FC layer outputs 128 features, applies a rectified linear unit (ReLU) activation function, incorporates dropout with a probability of 0.5, and employs batch normalization. The second FC layer receives 128 input features and outputs 64 features while utilizing a ReLU activation function, dropout (0.5 probability), and batch normalization. Batch normalization was incorporated to stabilize the learning process, while dropout was applied to improve generalization and mitigate overfitting by randomly deactivating a fraction of input units during training. This strategy promotes the learning of more robust feature representations. Finally, a softmax activation function was applied to the output layer, which normalizes the network's predictions into a probability distribution across the three lung condition classes: COVID-19, other lung diseases, and a healthy condition.

2.3.2 | Fine-Tuned DL Models

In addition to modifying the classification layers of the SoTA DL models, we selectively unfroze specific convolutional layers to enable domain-specific feature learning. The decision to unfreeze layers was informed by the understanding that earlier layers in pretrained CNNs tend to capture general features such as edges and textures, while deeper layers encode more task-specific and abstract representations [40]. Therefore, we began unfreezing layers in the final convolutional block and progressively included earlier layers within the same block. This approach allowed us to adapt higher-level features to the LUS domain without significantly disrupting the general low-level representations. We avoided unfreezing all layers to reduce the risk of overfitting, especially given the limited size of our dataset, and to minimize training time. For each model, the configuration of unfrozen layers was selected based on the highest validation accuracy and smooth convergence of training and validation loss. Table 3 summarizes the convolutional or encoder layers unfrozen during fine-tuning, along with their associated memory utilization for trainable and frozen parameters.

TABLE 3 | Unfrozen convolutional or transformer encoder layers selected for fine-tuning each model, with corresponding memory usage (in megabytes) trainable and non-trainable parameters.

Model	Unfrozen layers	Trainable parameters (MB)	Non-trainable parameters (MB)
VGG-16	block5_conv1	39.29	29.13
	block5_conv2		
	block5_conv3		
InceptionV3	conv2d_85, conv2d_87	35.78	72.42
	conv2d_88, conv2d_91		
	conv2d_92, conv2d_93		
Xception	block14_sepconv1	67.12	61.49
	block14_sepconv2		
ViT	encoder_layer_10	54.49	273.22
	encoder_layer_11		

Note: This reflects layer-wise unfreezing to optimize domain adaptation on lung ultrasound data while balancing computational efficiency and model capacity.

2.3.3 | Custom-Designed CNN Model

To enable a comprehensive evaluation of classification accuracy, computational efficiency, and inference speed, we developed a custom CNN model to complement the SoTA architectures. The proposed model, referred to as ParSE-CNN, features a parallel CNN architecture integrated with Squeeze-and-Excitation (SE) blocks to enhance channel-wise feature recalibration and improve representational power, as depicted in Figure 5.

The network comprises two parallel convolutional branches, each designed to extract complementary spatial features from LUS images. The first branch utilizes smaller 3×3 convolutional kernels to capture fine-grained spatial details. It consists of a 32-filter convolutional layer followed by a 64-filter convolutional layer, both employing the ReLU activation function to introduce nonlinearity. These layers are followed by a 2×2 max pooling operation to reduce spatial dimensions while retaining essential feature information and reducing computational overhead [41]. An SE block is then applied to adaptively recalibrate the channel-wise feature responses, emphasizing informative features and suppressing less relevant ones, following the approach introduced by Hu et al. [42].

In parallel, the second branch employs larger 5×5 convolutional kernels to capture broader contextual information. This branch

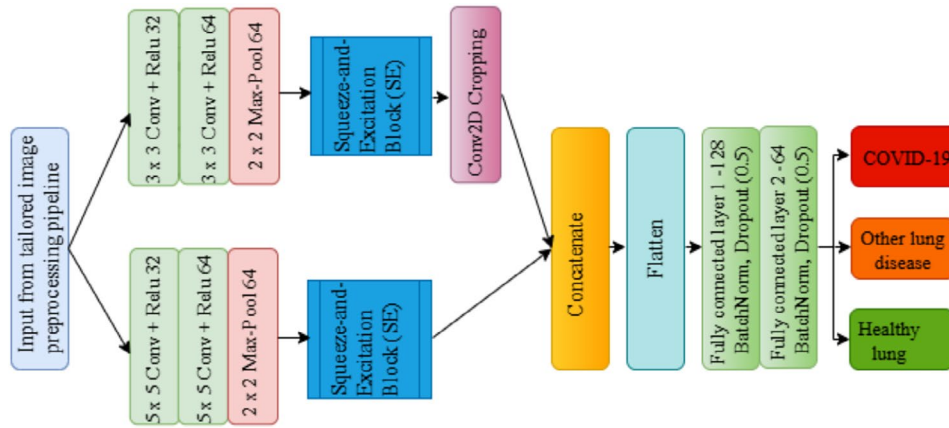


FIGURE 5 | Architecture of the proposed ParSE-CNN model, which consists of two parallel convolutional branches with different kernel sizes (3×3 and 5×5) to extract both fine-grained and coarse spatial features from lung ultrasound images. Each branch incorporates Squeeze-and-Excitation blocks for channel-wise attention, followed by feature concatenation and a fully connected classification head for three-class lung condition prediction.

mirrors the structure of the first, comprising 32- and 64-filter convolutional layers with ReLU activation, followed by 2×2 max pooling and an SE block to enhance feature sensitivity to relevant channels. Owing to the differing kernel sizes, the feature maps generated by the two branches may vary in spatial dimensions. To facilitate concatenation, the output of the first branch is spatially cropped to match the dimensions of the second.

The outputs of both branches are then concatenated to form a unified feature representation that combines both fine and coarse spatial details. This composite feature map is flattened into a one-dimensional vector and passed through a FC classification head. The head begins with a Dense layer of 128 units activated by ReLU, followed by Batch Normalization to promote stable learning and a Dropout layer (rate=0.5) to mitigate overfitting. A subsequent Dense layer with 64 units follows the same structure. The final output layer consists of three neurons with softmax activation, yielding class probabilities for the three target categories: COVID-19, other lung pathology, and healthy lung.

2.4 | Model Training and Parameter Configuration

We conducted our experiments on a Windows 10 64-bit operating system with an Intel Core i5-4200M CPU@2.50 GHz and 8GB of RAM. We implemented the VGG-16, InceptionV3, Xception, and ParSE-CNN models using the TensorFlow DL framework, while we used PyTorch for the ViT model. The models were trained and validated using three configurations: a publicly available dataset, a dataset collected from Ugandan healthcare facilities, and a combined dataset comprising both sources.

During training, we optimized the models using the Adam optimizer with a learning rate of 0.0001 and employed the categorical cross-entropy loss function. The input images were resized to 224×224 pixels, and data augmentation techniques, including horizontal flipping and zooming, were applied within the training pipeline. Horizontal flipping was implemented with a 50% probability, allowing each image an equal chance of being mirrored along the horizontal axis to enhance spatial invariance to left-right orientation changes. Zooming was applied with a

scaling factor of 0.2, enabling image magnification of up to 20%, thereby simulating variations in perspective and improving model generalization. We set the batch size to 32 and trained the models for 100 epochs, ensuring sufficient iterations for feature learning and convergence. Additionally, early stopping was employed to monitor validation loss and mitigate overfitting.

2.5 | Model Evaluation

To assess the effectiveness of the models on the LUS dataset, we employed key performance metrics, such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive evaluation of the models' classification performance, capturing different aspects of correctness and the balance between false positives and false negatives.

Accuracy is a fundamental metric that represents the proportion of correctly classified instances among all predictions, offering an overall measure of the model's performance [43]. Precision quantifies the accuracy of positive predictions, calculated as the ratio of true positive instances to all instances predicted as positive [43]. Recall, also referred to as sensitivity, measures the model's ability to correctly identify positive instances, computed as the ratio of true positives to all actual positive instances [43]. F1-Score provides a balanced assessment by incorporating both precision and recall, effectively capturing the trade-off between these two metrics [43].

Additionally, the confusion matrix offers a detailed breakdown of the model's predictions, displaying the number of true positives, true negatives, false positives, and false negatives for each class [44]. This information is particularly valuable for analyzing classification errors and identifying potential areas for model improvement.

2.6 | Model Explainability With Grad-CAM

Model explainability is essential in DL, particularly in medical imaging applications, where the interpretability and

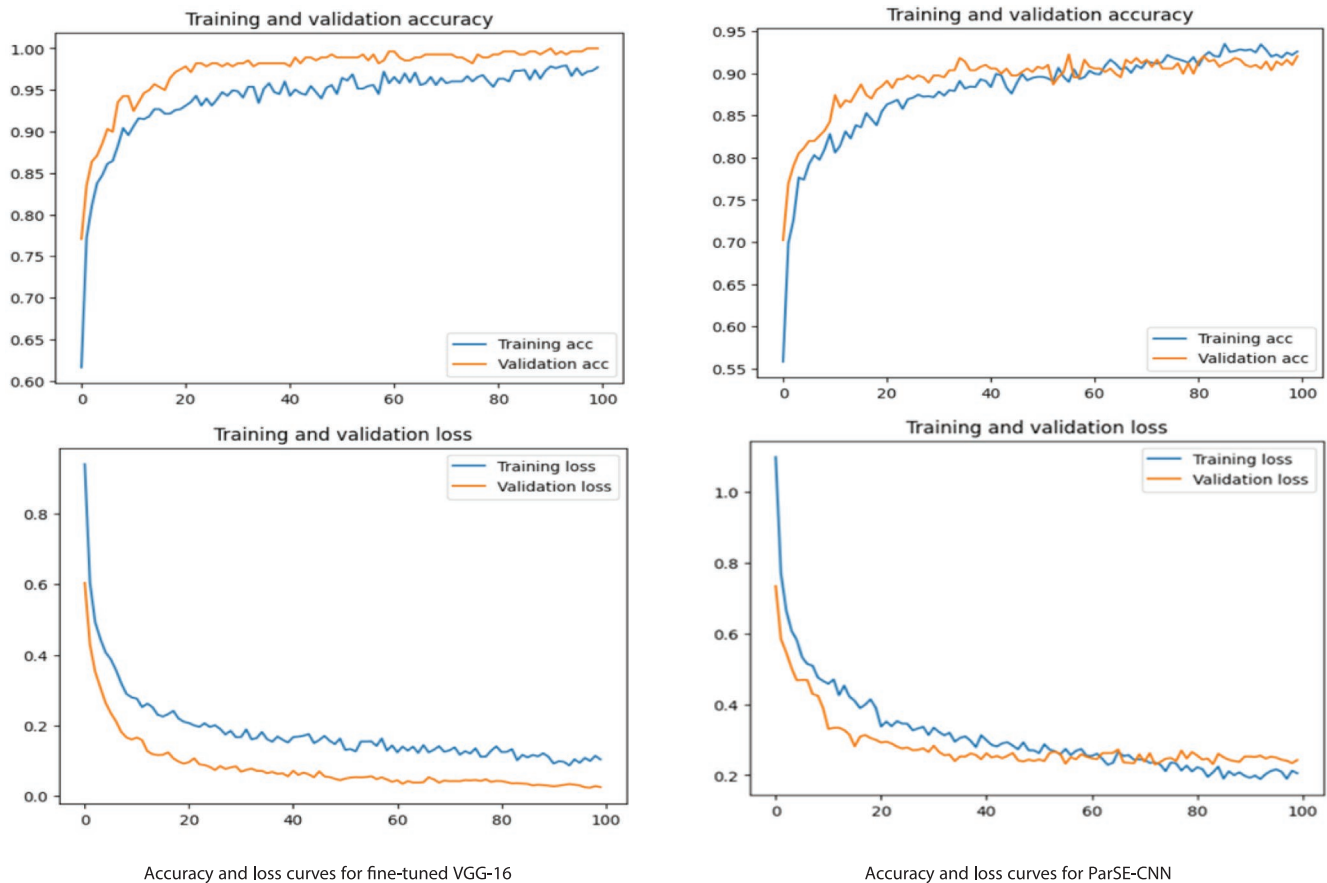


FIGURE 6 | Training and validation accuracy and loss curves for the fine-tuned VGG-16 and ParSE-CNN models. Top row: Training and validation accuracy. Bottom row: Training and validation loss.

reliability of decisions are critical [24]. In this study, we integrated Grad-CAM with the model to provide visual explanations of which regions in the input image contribute most to the model's decision. The Grad-CAM generates class-specific saliency maps [45], effectively identifying regions associated with COVID-19, other lung conditions, and healthy lung states. This approach facilitated the verification of model predictions by medical experts, ensuring that the classification of lung diseases from LUS images was based on clinically relevant features.

To generate the Grad-CAM visualizations, we first computed the gradients of the predicted class score concerning the feature maps of the final convolutional layer. These gradients were then averaged across all channels to obtain a weight vector representing each feature map's importance for the predicted class. Next, we applied a ReLU activation function to focus only on positive influences, resulting in a weighted combination of the feature maps highlighting significant areas in the input image. Finally, we upsampled this combined feature map to match the original image dimensions and overlaid it as a heatmap on the ultrasound images. This visualization enabled us to identify specific regions that contributed to the model's predictions, providing insights into its decision-making process and enhancing clinical interpretability.

3 | Results

3.1 | Accuracy and Loss Curves

The model performance was monitored over 100 epochs by recording accuracy and loss metrics for the fine-tuned VGG-16 and ParSE-CNN models using the combined dataset, which included both publicly available data and data collected from Ugandan healthcare facilities. The results are presented as training-validation accuracy and loss curves in Figure 6. Each model is represented by two plots: the accuracy curve (top row) and the loss curve (bottom row). The accuracy curves depict the training accuracy (blue line), which shows the model's performance on the training dataset over epochs, and the validation accuracy (orange line), which illustrates its performance on the validation dataset. Similarly, the loss curves display the training loss (blue line), indicating how the training error decreases over epochs, and the validation loss (orange line), representing the error on the validation set. It can be observed that both models improve over epochs, with training and validation accuracy increasing and loss decreasing. The fine-tuned VGG-16 outperforms ParSE-CNN with higher accuracy and better generalization due to transfer learning, while ParSE-CNN, trained from scratch with fewer parameters, converges more slowly and achieves slightly lower accuracy.

TABLE 4 | Total number of trainable parameters for each deep learning model, reflecting differences in architectural complexity and their implications for computational cost and deployment feasibility.

Model	Number of trainable parameters
Fine-tuned VGG-16	10,299,523
Fine-tuned InceptionV3	9,380,611
Fine-tuned Xception	17,595,651
Fine-tuned ViT	14,283,011
ParSE-CNN	5,184,875

3.2 | Number of Trainable Parameters

To evaluate the computational resource requirements of the models, we analyzed the number of trainable parameters across their respective architectures. Table 4 summarizes the total number of trainable parameters for each model. Among them, ParSE-CNN exhibited the lowest parameter count (5,184,875), underscoring its lightweight design and potential for computational efficiency. In contrast, the fine-tuned Xception model had the highest number of trainable parameters (17,595,651), reflecting its more complex architecture.

3.3 | Classification Accuracy

We evaluated each model on the test data after training it separately on a publicly available dataset, a dataset collected from Ugandan healthcare facilities, and a combined dataset comprising both sources. Table 5 presents the classification accuracies, showing that all models achieved lower accuracy when trained on the publicly available dataset compared to the combined dataset.

3.4 | Precision, Recall, and F1 Score

In addition to evaluating classification accuracy, we assessed precision, recall, and F1-score for each model on the test set, across the three lung condition classes, following training on the combined dataset. The results are presented in Table 6.

For a more comprehensive evaluation of overall model performance beyond per-class metrics, we computed the average precision, average recall, and average F1-score for each model across the three lung condition classes. The results, summarized in Table 7, indicate that the fine-tuned VGG-16 achieved the highest average precision of 0.97, average recall of 0.98, and average F1-score of 0.97.

3.5 | Confusion Matrix

The confusion matrices for the fine-tuned VGG-16, InceptionV3, Xception, ViT, and ParSE-CNN models, trained on the mixed dataset and evaluated on the independent test set, are presented in Figure 7. Among these models, the fine-tuned VGG-16 achieved the lowest classification error,

TABLE 5 | Test set classification accuracy of baseline deep learning models, their fine-tuned counterparts, and the proposed ParSE-CNN model, evaluated after training on three datasets: publicly available, Ugandan healthcare facility, and a combined dataset.

Training dataset	Model	Accuracy (%)
Publicly available dataset	VGG-16	90.56
	InceptionV3	85.18
	Xception	90.07
	ViT	70.03
	Fine-tuned VGG-16	92.64
	Fine-tuned InceptionV3	89.25
	Fine-tuned Xception	91.04
	Fine-tuned ViT	78.72
	ParSE-CNN	91.71
Dataset from the Ugandan healthcare facilities	VGG-16	91.07
	InceptionV3	89.01
	Xception	90.49
	ViT	78.01
	Fine-tuned VGG-16	94.27
	Fine-tuned InceptionV3	90.78
	Fine-tuned Xception	92.91
	Fine-tuned ViT	82.98
	ParSE-CNN	92.95
Combined dataset	VGG-16	93.74
	InceptionV3	90.15
	Xception	91.33
	ViT	83.43
	Fine-tuned VGG-16	97.62
	Fine-tuned InceptionV3	92.71
	Fine-tuned Xception	94.33
	Fine-tuned ViT	90.02
	ParSE-CNN	94.68

Note: Bolded values indicate the highest-performing model for each dataset, highlighting the benefit of fine-tuning and dataset diversity.

misclassifying only three images, whereas the fine-tuned ViT exhibited the highest error with 12 misclassifications. These errors are likely attributable to the presence of coexisting

TABLE 6 | Precision, recall, and F1-score for each lung condition class (COVID-19, Other lung pathology, and Healthy lung) as predicted by baseline, fine-tuned, and ParSE-CNN models.

Model	Lung condition	Precision	Recall	F1 score
VGG-16	COVID-19	0.85	0.90	0.88
	Other lung pathology	0.93	0.93	0.93
	Healthy lung	1.00	0.96	0.98
InceptionV3	COVID-19	0.79	0.87	0.83
	Other lung pathology	0.91	0.91	0.91
	Healthy lung	0.96	0.91	0.93
Xception	COVID-19	0.82	0.87	0.84
	Other lung pathology	0.89	0.93	0.91
	Healthy lung	1.00	0.93	0.96
ViT	COVID-19	0.71	0.82	0.74
	Other lung pathology	0.91	0.83	0.88
	Healthy lung	0.91	0.93	0.91
Fine-tuned VGG-16	COVID-19	0.94	0.97	0.95
	Other lung pathology	0.98	0.96	0.97
	Healthy lung	1.00	1.00	1.00
Fine-tuned InceptionV3	COVID-19	0.80	0.90	0.85
	Other lung pathology	0.96	0.93	0.94
	Healthy lung	0.98	0.95	0.96
Fine-tuned Xception	COVID-19	0.81	0.97	0.88
	Other lung pathology	0.98	0.95	0.96
	Healthy lung	1.00	0.93	0.96
Fine-tuned ViT	COVID-19	0.77	0.87	0.78
	Other lung pathology	0.96	0.85	0.90
	Healthy lung	0.95	0.98	0.96
ParSE-CNN	COVID-19	0.83	0.97	0.90
	Other lung pathology	0.98	0.95	0.96
	Healthy lung	1.00	0.95	0.97

Note: All models were evaluated on the test set after training on the combined dataset. Bold figures indicate the top performing model in each category.

artifacts in LUS images. For instance, A-lines and B-lines may appear simultaneously as the lung transitions from a healthy to a mildly de-aerated state, with B-lines progressively emerging and eventually obscuring A-lines. Similarly, B-lines can coexist with consolidations in more advanced disease stages. Consequently, model performance could potentially be

TABLE 7 | Average precision, recall, and F1-score of each deep learning model, including the proposed ParSE-CNN, computed across all three lung condition classes (COVID-19, Other pathology, and Healthy lung).

Model	Avg. precision	Avg. recall	Avg. F1-score
VGG-16	0.93	0.93	0.93
InceptionV3	0.89	0.90	0.89
Xception	0.90	0.91	0.90
ViT	0.84	0.86	0.84
Fine-tuned VGG-16	0.97	0.98	0.97
Fine-tuned InceptionV3	0.91	0.93	0.92
Fine-tuned Xception	0.93	0.95	0.93
Fine-tuned ViT	0.89	0.90	0.88
ParSE-CNN	0.94	0.95	0.94

Note: Results are based on the test set following training on the combined dataset. Bold values denote the highest overall performance.

enhanced by incorporating additional training data that better represents such transitional cases.

3.6 | Explainability With Grad-CAM

To gain insights into the models' decision-making processes, we generated heatmaps that highlight the regions of the input image most influential to the models' predictions. Figure 8 presents the Grad-CAM outputs illustrating the salient regions identified by the fine-tuned VGG-16 and ParSE-CNN models during inference on LUS images representing healthy, COVID-19, and non-COVID-19 pneumonia conditions. Regions highlighted in red denote areas of highest relevance to the model's prediction, whereas blue regions indicate areas of minimal relevance.

Although the Grad-CAM visualizations in Figure 8 illustrate the capacity of the fine-tuned VGG-16 and ParSE-CNN models to localize clinically relevant regions across diverse lung conditions, it is important to note five instances: three from the fine-tuned VGG-16 and two from the ParSE-CNN model, where the generated attention maps failed to correspond with diagnostically meaningful features. Figure 9 presents representative examples of these cases, in which the models produced non-informative heatmaps, potentially reflecting a limitation in the models' learned representations within the input images.

4 | Model Integration into a Mobile LUS Screening Workflow

The mobile LUS screening system integrates three components: a Clarius handheld ultrasound scanner, a smartphone running the Clarius application (Clarius app), and a laptop executing the

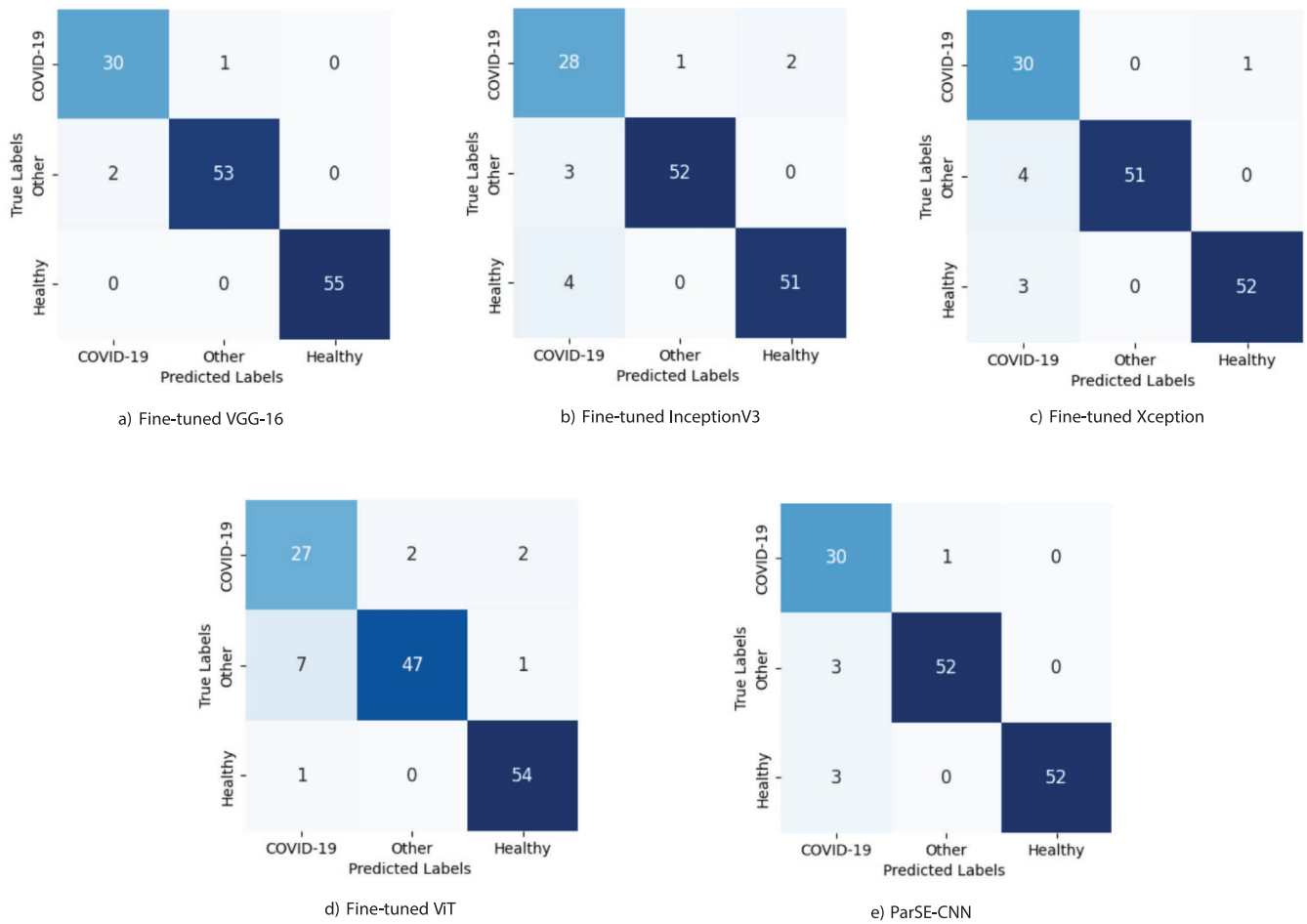


FIGURE 7 | Confusion matrices of the classification models. The label “Other” refers to the class of other lung pathology, while “Healthy” denotes the healthy lung class.

deployed DL model through the Clarius Cast API, as shown in Figure 10.

For integration, the Clarius application (v9.3) was installed on an Android smartphone from the Google Play Store, enabling acquisition and streaming of ultrasound images. The Clarius Cast API package (v9.1), including the required libraries (`cast.dll`, `cast.lib`, `pycaster.py`, and `pysidecaster.py`), was obtained from the Clarius developer repository¹ and deployed on the laptop. The `pysidecaster.py` script was modified to embed the ParSE-CNN model, allowing direct inference on incoming frames, which were converted to QuickTime format for compatibility with the visualization pipeline.

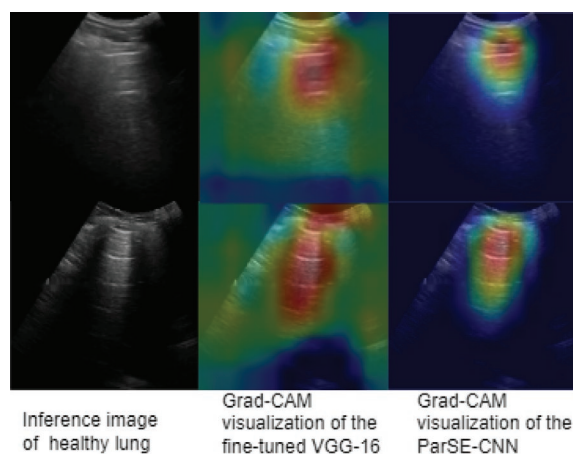
During operation, the scanner establishes a Wi-Fi Direct network that permits simultaneous connection of the smartphone and laptop without the need for an external router. The scanner streams ultrasound frames concurrently to the smartphone (for routine imaging) and to the laptop. The laptop-based model processes these frames in real time, generates classification outputs, and displays the diagnostic results, thereby facilitating immediate clinical interpretation at the point of care.

5 | Model Inferencing

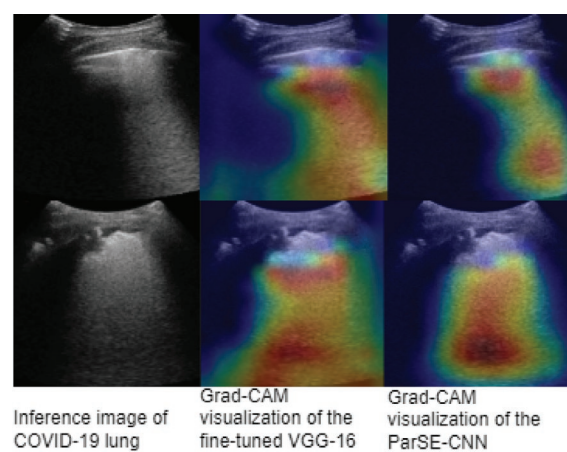
Model inference is essential for assessing the real-world applicability of DL systems, particularly in time-sensitive clinical workflows [46]. To evaluate responsiveness and quantify latency, inference times for the fine-tuned VGG-16 and ParSE-CNN models were measured both in the development environment (pre-deployment) and after deployment on a laptop (Intel Core i5 processor, 8GB RAM) within the mobile LUS screening workflow (post-deployment).

In the deployed setup, each model was independently hosted on the laptop, which functioned as the processing unit. Instead of live streaming from the scanner and smartphone, inference was simulated by sequentially looping through 20 test images stored locally, with each image forwarded to the model for classification. This approach enabled controlled measurement of latency while approximating real-time performance under resource-constrained point-of-care conditions. The average inference times are summarized in Table 8.

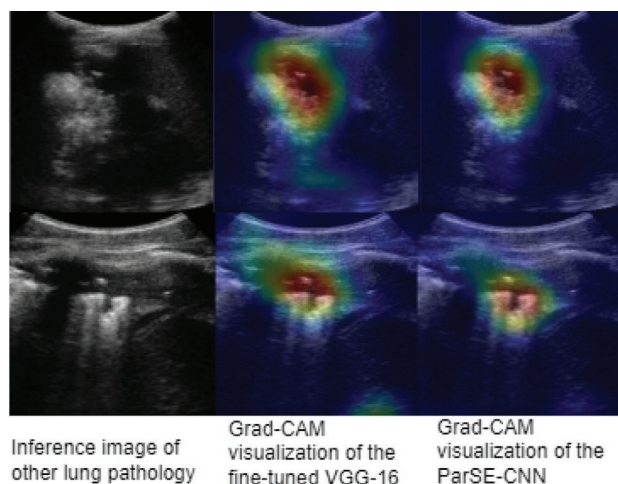
In the development environment, the average inference times were 0.012s for VGG-16 and 0.005s for ParSE-CNN, whereas deployment on the laptop yielded slightly higher values of



a) Heatmaps generated by VGG-16 and ParSE-CNN for healthy lung inference images



b) Heatmaps generated by VGG-16 and ParSE-CNN for COVID-19 inference images



c) Heatmaps generated by VGG-16 and ParSE-CNN for other lung pathology inference images

FIGURE 8 | Grad-CAM visualizations generated from the fine-tuned VGG-16 and ParSE-CNN models on lung ultrasound images. The highlighted regions indicate the discriminative features used for classification in (a) healthy lungs, (b) COVID-19 infection, and (c) other pathological lung conditions. These heatmaps enhance interpretability by revealing the areas most influential in the models' decision-making process.

0.014 s and 0.006 s, respectively. Both models maintained near real-time performance with only marginal latency increases in the deployed setup. Notably, ParSE-CNN, with substantially fewer parameters (5.18 M vs. 10.30 M), was over 2.3× faster than the fine-tuned VGG-16 across both conditions, highlighting its efficiency and suitability for real-time mobile clinical applications.

6 | Discussion

In this study, we evaluated state-of-the-art DL models: VGG-16, InceptionV3, Xception, ViT, and the custom ParSE-CNN, for classification of lung disease in LUS images. Models trained on a mixed dataset (publicly available dataset + dataset collected from Ugandan healthcare facilities) consistently outperformed those trained solely on public data, reinforcing prior findings [47] that dataset diversity improves generalizability. Exposure to a wider range of anatomical variations and imaging artifacts enhanced model robustness and reduced false classifications.

Among the evaluated architectures, the fine-tuned VGG-16 attained the highest diagnostic accuracy (Table 5). Its convolutional design, characterized by small receptive fields (3×3 filters), was well suited for capturing localized LUS features such as A-lines and B-lines. Fine-tuning further enabled adaptation to domain-specific patterns, consistent with prior observations [48]. The proposed ParSE-CNN delivered competitive accuracy with substantially fewer parameters and faster inference relative to VGG-16. Its parallel spatial extraction layers effectively captured localized features while reducing computational cost, corroborating recent findings on the efficiency of lightweight CNNs [49, 50]. This balance between accuracy and efficiency makes ParSE-CNN particularly attractive for real-time deployment in resource-constrained settings. InceptionV3 and Xception achieved moderate performance, benefiting from inception modules and depth-wise separable convolutions [37, 38], but did not exceed the fine-tuned VGG-16 or ParSE-CNN. Their reliance on multi-scale feature extraction may be less advantageous for LUS, where diagnostic cues are subtle and often artifact-driven. By

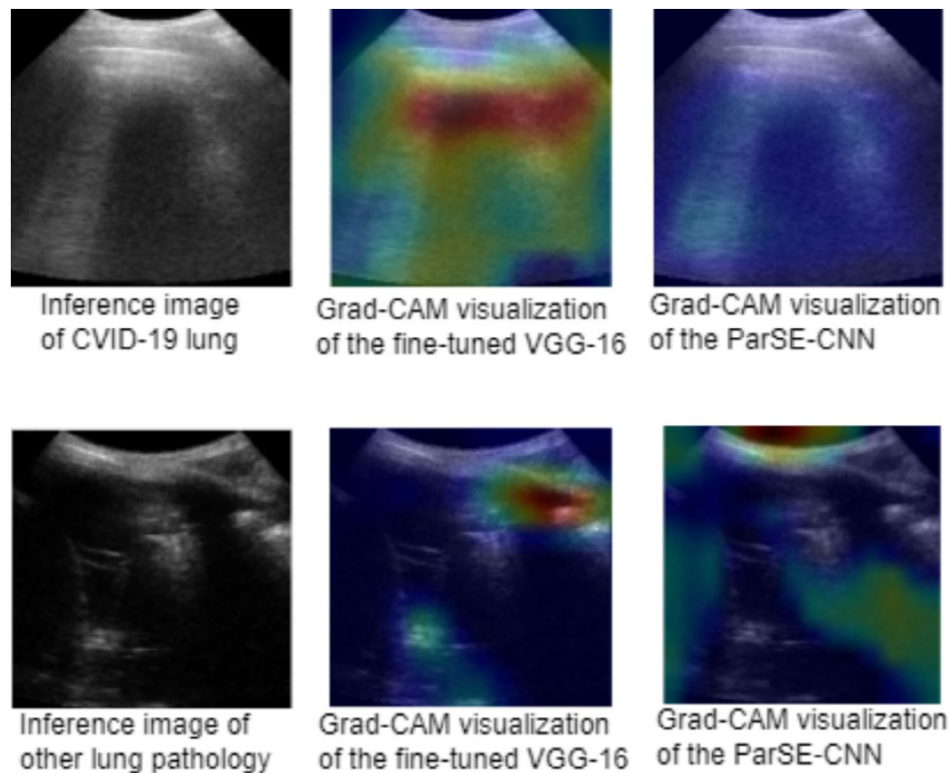


FIGURE 9 | Representative failure cases of Grad-CAM where the fine-tuned VGG-16 and ParSE-CNN produced attention maps misaligned with diagnostically relevant regions in lung ultrasound images. The first row corresponds to COVID-19 infection and the second row to other pathological lung conditions, highlighting limitations in the models' ability to consistently capture clinically meaningful features.

contrast, ViT exhibited the lowest performance, underscoring the difficulty of training transformer-based models on limited LUS datasets [19]. Despite data augmentation, its reliance on patch-based global attention limited its ability to capture fine-grained features. This underscores the importance of aligning model architecture with dataset scale and imaging modality.

Grad-CAM heatmaps (Figure 8) showed strong concordance with radiologists' assessments, supporting the interpretability and clinical relevance of the trained models. The highlighted regions typically overlapped with diagnostically salient LUS features, including pleural lines, A-lines, B-lines, and consolidations, which are established biomarkers for pulmonary pathologies. These results are consistent with previous studies [29, 51] demonstrating the value of Grad-CAM in enhancing transparency and trust in DL-based medical diagnosis. However, a subset of cases (from both the fine-tuned VGG-16 and the ParSE-CNN) yielded heatmaps that did not correspond to clinically meaningful regions (Figure 9). Such outcomes highlight the limitations of gradient-based attribution, particularly in instances where subtle features influence classification. Potential contributing factors include variability in artifact visibility and sensitivity to spurious correlations. Furthermore, prior work has shown that saliency methods may be inconsistent and dependent on the internal representations of the model [52, 53]. These findings indicate that visual explanations should be interpreted cautiously and considered complementary rather than definitive tools in clinical decision support.

The integration of ParSE-CNN into a mobile LUS screening workflow: comprising a Clarius handheld ultrasound scanner,

an Android smartphone, and a laptop hosting the deployed model, highlights its translational potential. Moving toward edge deployment on mobile and portable devices further supports global health equity by expanding access to diagnostic imaging in underserved regions.

Overall, the findings highlight the importance of model architecture, dataset diversity, and computational efficiency in optimizing LUS classification. Classical convolutional models, when fine-tuned, remain competitive for domain-specific tasks, while lightweight custom models like ParSE-CNN provide a practical solution for real-time mobile deployment.

7 | Conclusion

This study presents a DL framework for automated classification of lung diseases in LUS images, emphasizing computational efficiency, interpretability, and suitability for edge deployment. A customized preprocessing pipeline was introduced to isolate ROIs and remove distracting labels, thereby improving data quality for robust model training. Among the benchmarked state-of-the-art models, the fine-tuned VGG-16 achieved the highest classification accuracy, while the proposed ParSE-CNN delivered comparable performance with substantially reduced parameters and faster inference, demonstrating its suitability for point-of-care applications in resource-limited settings. Grad-CAM further enhanced interpretability by localizing diagnostically relevant regions, supporting clinical decision-making despite occasional inconsistencies. Finally, the integration of the ParSE-CNN into a mobile LUS screening workflow—comprising

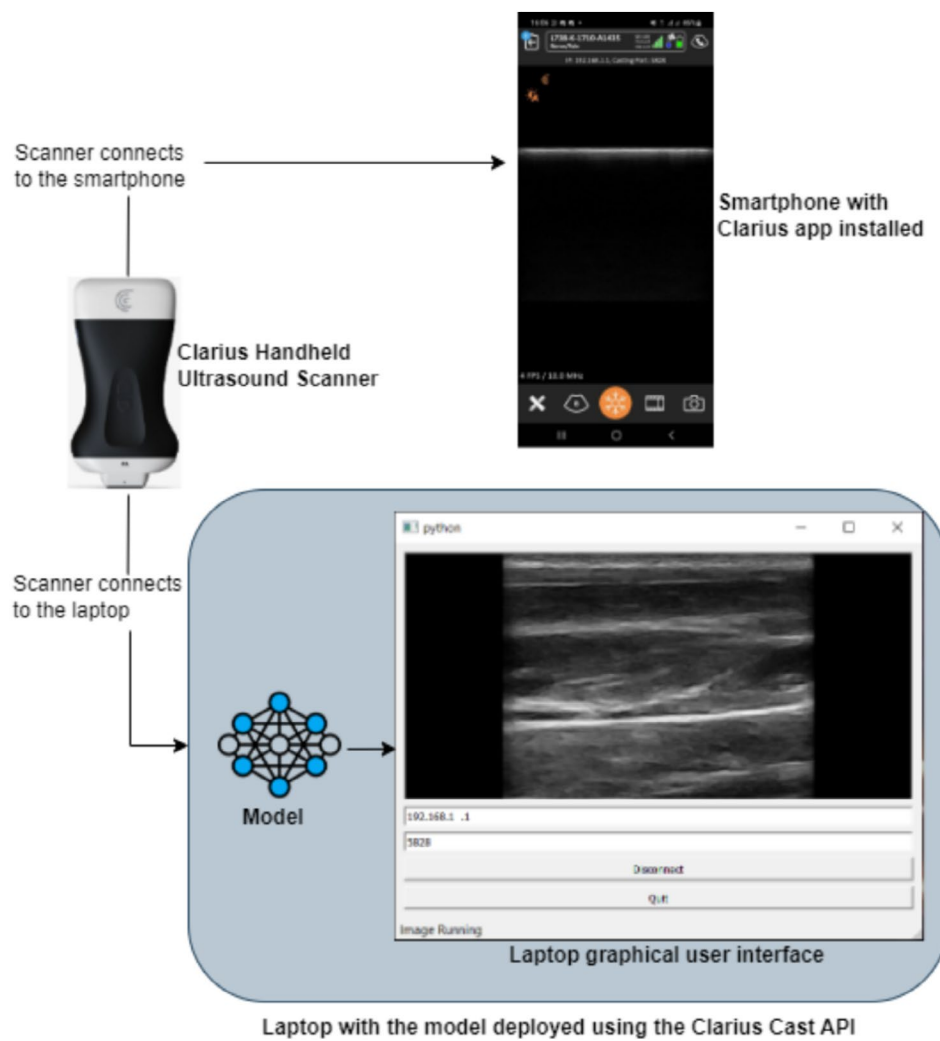


FIGURE 10 | Mobile lung ultrasound screening workflow. The setup comprises a Clarius handheld ultrasound scanner, a smartphone with the Clarius application for image acquisition, and a laptop running a deep learning model deployed via the Clarius Cast API for real-time diagnostic inference and visualization.

TABLE 8 | Average inference times (in seconds) for the fine-tuned VGG-16 and ParSE-CNN models, measured both in the development environment (pre-deployment) and after deployment (post-deployment) on a laptop within the mobile lung ultrasound screening workflow.

Model	Average pre-deployment inference time (s)	Average post-deployment inference time (s)
Fine-tuned VGG-16	0.012	0.014
ParSE-CNN	0.005	0.006

Note: Average inference times were computed over 20 test images. Bold values highlight the model with the shortest average inference time in each setting.

a handheld ultrasound scanner, smartphone, and laptop back-end—illustrates its translational potential to expand diagnostic access and advance health equity.

Nonetheless, this work has certain limitations. The use of static LUS images does not capture the dynamic characteristics present in video sequences, which are integral to clinical interpretation. In addition, comparative evaluation against traditional machine learning classifiers was not conducted, as the study primarily focused on end-to-end DL approaches for clinical deployment.

The current evaluation was restricted to a laboratory-based prototype; future research will extend testing across diverse mobile and edge devices, including lower-end hardware, to assess scalability and real-world performance. Further directions include incorporating video-based analysis to exploit temporal diagnostic cues, broadening the range of disease classes considered, and applying advanced explainability techniques to enhance transparency and clinical trust. Benchmarking against traditional classifiers will also be pursued to provide a more comprehensive performance comparison.

Acknowledgments

The authors would like to thank the Government of Uganda through Makerere University Research and Innovation Funds, International Development Research Centre, and Swedish International Development Cooperation Agency for funding the study and Mulago National Referral Hospital, Department of Radiology as a clinical collaborator.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Mendeley Data at <https://data.mendeley.com/datasets/hb3p34ytxv/1>, reference number 10.17632/hb3p34ytxv.1.

Endnotes

¹<https://github.com/clariusdev/cast>.

References

1. G. S. Gould, J. R. Hurst, A. Trofor, et al., "Recognising the Importance of Chronic Lung Disease: A Consensus Statement From the Global Alliance for Chronic Diseases (Lung Diseases Group)," *Respiratory Research* 24, no. 1 (2023): 15.
2. A. Wheeldon, "The Respiratory System and Associated Disorders," *British Journal of Nursing* 32, no. 13 (2023): 613–619.
3. S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid Deep Learning for Detecting Lung Diseases From X-Ray Images," *Informatics in Medicine Unlocked* 20 (2020): 100391, <https://doi.org/10.1016/j.imu.2020.100391>.
4. M. J. Reid and E. Goosby, "Improving Quality is Necessary to Building a TB-Free World: Lancet Commission on Tuberculosis," *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases* 19 (2020): 100156.
5. B. Dadonaite and M. Roser, "Burden of Pneumonia," Tech. Rep., Our World in Data (2019).
6. B. Qu and D. Zhang, "Evaluation of COVID-19 Booster Vaccine Effectiveness," *Viruses* 17, no. 2 (2025): 179.
7. C. F. Cao, K. L. Ma, H. Shan, et al., "CT Scans and Cancer Risks: A Systematic Review and Dose-Response Meta-Analysis," *BMC Cancer* 22, no. 1 (2022): 1238.
8. Y. Yang, Y. Wu, and W. Zhao, "Comparison of Lung Ultrasound and Chest Radiography for Detecting Pneumonia in Children: A Systematic Review and Meta-Analysis," *Italian Journal of Pediatrics* 50, no. 1 (2024): 12.
9. Z. F. Al Nufaiei and K. M. Alshamrani, "Comparing Ultrasound, Chest X-Ray, and CT Scan for Pneumonia Detection," *Medical Devices: Evidence and Research* 18 (2025): 149–159.
10. R. Bassiouny, A. Mohamed, K. Umapathy, and N. Khan, "An Interpretable Object Detection-Based Model for the Diagnosis of Neonatal Lung Diseases Using Ultrasound Images," in *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2021), 3029–3034.
11. L. Demi, F. Wolfram, C. Klersy, et al., "New International Guidelines and Consensus on the Use of Lung Ultrasound," *Journal of Ultrasound in Medicine* 42, no. 2 (2023): 309–344, <https://doi.org/10.1002/jum.16088>.
12. T. Yang, O. Karakuş, N. Anantrasirichai, and A. Achim, "Current Advances in Computational Lung Ultrasound Imaging: A Review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 70, no. 1 (2022): 2–15.
13. J. Born, G. Brändle, M. Cossio, et al., "POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS)," Preprint, arXiv, arXiv:2004.12084 (2020).
14. J. Diaz-Escobar, N. E. Ordóñez-Guillén, S. Villarreal-Reyes, et al., "Deep-Learning Based Detection of COVID-19 Using Lung Ultrasound Imagery," *PLoS One* 16 (2021): e0255886.
15. J. Roberts and T. Tsiligkaridis, "Ultrasound Diagnosis of COVID-19: Robustness and Explainability," Preprint, arXiv, arXiv:2012.01145 (2020).
16. S. Roy, W. Menapace, S. Oei, et al., "Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound," *IEEE Transactions on Medical Imaging* 39, no. 8 (2020): 2676–2687.
17. S. Magrelli, P. Valentini, C. De Rose, R. Morello, and D. Buonsenso, "Classification of Lung Disease in Children by Using Lung Ultrasound Images and Deep Convolutional Neural Network," *Frontiers in Physiology* 12 (2021): 1–22.
18. M. J. Lim, D. E. Kim, D. K. Chung, H. Lim, and Y. M. Kwon, "Deep Convolution Neural Networks for Medical Image Analysis," *International Journal of Engineering & Technology* 7 (2018): 115–119.
19. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," Preprint, arXiv, arXiv:2010.11929 (2020).
20. M. C. Fiorentino, R. Rosati, A. Melnic, E. Conti, and P. Zingaretti, "Vision Transformer Approaches for COVID-19 Pneumonia Assessment in Lung Ultrasound Images," in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)* (2024), 83–88.
21. S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-Cam Explainable in Medical Images?," in *Proceedings of the International Conference on Computer Vision and Image Processing* (Springer, 2023), 124–135.
22. H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A Deep Learning and Grad-CAM Based Color Visualization Approach for Fast Detection of COVID-19 Cases Using Chest X-Ray and CT-Scan Images," *Chaos, Solitons & Fractals* 140 (2020): 110190.
23. R. Arntfield, D. Wu, J. Tschirhart, et al., "Automation of Lung Ultrasound Interpretation via Deep Learning for the Classification of Normal Versus Abnormal Lung Parenchyma: A Multicenter Study," *Diagnostics* 11, no. 11 (2021): 2049.
24. M. F. Rahman, T. L. Tseng, M. Pokojovy, et al., "Machine-Learning-Enabled Diagnostics With Improved Visualization of Disease Lesions in Chest X-Ray Images," *Diagnostics* 14, no. 16 (2024): 1699.
25. V. G. Duque, A. Marquardt, Y. Velikova, et al., "Ultrasound Segmentation Analysis via Distinct and Completed Anatomical Borders," *International Journal of Computer Assisted Radiology and Surgery* 19 (2024): 1419–1427.
26. J. Yang, N. T. Dung, P. N. Thach, et al., "Generalizability Assessment of AI Models Across Hospitals in a Low-Middle and High Income Country," *Nature Communications* 15, no. 1 (2024): 8270.
27. J. Yang, L. Clifton, N. T. Dung, et al., "Mitigating Machine Learning Bias Between High Income and Low-Middle Income Countries for Enhanced Model Fairness and Generalizability," *Scientific Reports* 14, no. 1 (2024): 13318.
28. A. Ebadi, P. Xi, A. MacLean, S. Tremblay, S. Kohli, and A. Wong, "COVIDx-US: An Open-Access Benchmark Dataset of Ultrasound Imaging Data for AI-Driven COVID-19 Analytics," *Frontiers in Bioscience (Landmark Edition)* 27, no. 7 (2022): 198.
29. J. Born, N. Wiedemann, M. Cossio, et al., "Accelerating Detection of Lung Pathologies With Explainable Ultrasound Image Analysis,"

- Applied Sciences* 11, no. 2 (2021): 672, <https://doi.org/10.3390/app11020672>.
30. T. J. Marini, D. J. Rubens, Y. T. Zhao, et al., "Lung Ultrasound: The Essentials," *Radiology: Cardiothoracic Imaging* 3, no. 2 (2021): e200564.
 31. N. M. Duggan, H. Shokoohi, A. S. Liteplo, C. Huang, and A. J. Goldsmith, "Best Practice Recommendations for Point-of-Care Lung Ultrasound in Patients With Suspected COVID-19," *Journal of Emergency Medicine* 59, no. 4 (2020): 515–520.
 32. Clarius Mobile Health, "Clarius – Portable Ultrasound Scanner," accessed May 6, 2022 (2022).
 33. G. Bradski and A. Kaehler, "OpenCV," *Dr. Dobb's Journal of Software Tools* 3, no. 2 (2000): pp. 120 and 122–125.
 34. J. Yousefi, *Image Binarization Using Otsu Thresholding Algorithm*, vol. 10 (University of Guelph, 2011), 1–4.
 35. M. Fauser, "Keras-OCR" (2020), accessed May 6, 2025, <https://keras-ocr.readthedocs.io/en/latest/>.
 36. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Preprint, arXiv, arXiv:1409.1556 (2014).
 37. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 2818–2826.
 38. F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 1251–1258.
 39. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–255.
 40. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How Transferable Are Features in Deep Neural Networks?," in *Advances in Neural Information Processing Systems* (Springer, 2014), 3320–3328.
 41. A. Zafar, M. Aamir, N. Mohd Nawi, et al., "A Comparison of Pooling Methods for Convolutional Neural Networks," *Applied Sciences* 12, no. 17 (2022): 8643.
 42. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018): 7132–7141.
 43. R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (2020): 79–91.
 44. A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Woźniak, and M. F. Ijaz, "Implementing Vision Transformer for Classifying 2D Biomedical Images," *Scientific Reports* 14, no. 1 (2024): 12567.
 45. S. Murindanyi, C. Kirabo, N. P. Kirabo, N. Hellen, and G. Marvin, "Trustworthy Machine Emotion Intelligence Using Facial Micro-Expressions," in *Proceedings of International Conference on Intelligent Vision and Computing (ICIVC 2023)* (Springer, 2023), 46–62.
 46. M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical Image Analysis Using Deep Learning Algorithms," *Frontiers in Public Health* 11 (2023): 1273253.
 47. D. E. Shea, S. Kulhare, R. Millin, et al., "Deep Learning Video Classification of Lung Ultrasound Features Associated With Pneumonia," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023* (2023) 3103–3112.
 48. T. Fatima and H. Soliman, "Application of VGG16 Transfer Learning for Breast Cancer Detection," *Information* 16, no. 3 (2025): 227.
 49. X. Hu, L. Chu, J. Pei, W. Liu, and J. Bian, "Model Complexity of Deep Learning: A Survey," *Knowledge and Information Systems* 63 (2021): 2585–2619.
 50. M. Zawish, S. Davy, and L. Abraham, "Complexity-Driven Model Compression for Resource-Constrained Deep Learning on Edge," *IEEE Transactions on Artificial Intelligence* 5, no. 8 (2024): 3886–3901.
 51. S. Tengjongdee, C. Chaitan, S. Hanmanop, T. Jongsiri, M. Khongtan, and S. Pechprasarn, "Comparative Analysis of Deep Learning Networks for COVID-19 and Pneumonia Identification: Grad-CAM Visualization of Chest X-Ray Images," in *2024 16th Biomedical Engineering International Conference (BMEiCON)* (2024), 1–5.
 52. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Inc, 2018).
 53. N. Arun, N. Gaw, P. Singh, et al., "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging," *Radiology: Artificial Intelligence* 3, no. 6 (2021): e200267.