DEEP LEARNING FOR ACCURATE CLASSIFICATION OF PULMONARY EDEMA SEVERITY IN CHEST X-RAYS: A COMPARATIVE EVALUATION OF CNN MODELS WITH GRAD-CAM FEATURE VISUALIZATION

MANOSH SUR CHOUDHURY[1], RASHEDUR RAHMAN[1], SIAM TAHSIN BHUIYAN[1], SEFATUL WASI [2], SAADIA BINTE ALAM [1]

[1]Center for Computational & Data Sciences, Independent University, Bangladesh, Dhaka 1229, Bangladesh
[2]Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka 1229, Bangladesh
E-MAIL: manoshsc999@gmail.com, rashed.riyadh14@gmail.com, siamtbhuiyan@gmail.com, sefatulwasi@gmail.com, saadiabinte@iub.edu.bd

Abstract:

Accurate classification of pulmonary edema severity is essential for timely diagnosis and effective management, as each severity level requires distinct therapeutic interventions. Differentiating between edema severity classes is particularly challenging due to the overlapping radiographic features of Chest X-ray (CXR). Deep learning, especially convolutional neural networks (CNNs), presents a promising solution by automating classification and identifying subtle features often difficult to detect through conventional methods. This study evaluates the performance of six traditional deep learning models for classifying pulmonary edema severity. Among these, CheXNet achieved the best performance, with an accuracy of 91% and an overall AUC score of 0.88. The findings highlight the importance of using pretraining on CXR datasets, which significantly enhances model performance compared to general pretraining. Additionally, Grad-CAM visualization was employed to interpret model decisions, identifying key radiographic features that contribute to accurate classification across different severity levels.

Keywords:

CXR; Pulmonary Edema; Alveolar Edema; Vascular Congestion; Interstitial Edema; CNN; Grad-CAM.

1. Introduction

Pulmonary edema is the abnormal buildup of fluids in the lungs, causing breathing difficulties and potentially life-threatening complications. It can result from conditions like Congestive Heart Failure (CHF), cardiomyopathy, high altitude, or drug abuse. Severity ranges from moderate to severe, requiring timely treatment, especially in acute CHF cases. Accurate monitoring is essential for effective management, particularly in patients with other clinical conditions [1]. Classifying pulmonary edema severity is challenging and involves imaging, physical exams, and biomarkers to detect fluid buildup, symptoms, and underlying conditions. CXRs are the most common, affordable, and accessible method for diagnosing pulmonary edema, but require expert interpretation. Artificial Intelligence (AI), intense learning models like CNNs, can automate severity classification by analyzing medical images and patient data, detecting subtle patterns beyond human capability. These models enhance accuracy, consistency, and objectivity, and as they advance, they can integrate multi-modal data for a comprehensive edema assessment, aiding clinical decision-making.

1.1 Severity level and underlying conditions for Edema

Pulmonary Edema severity level is termed Vascular congestion, Interstitial Edema, and Alveolar Edema.

FIGURE 1. Edema Severity Level-wise Chest X-ray

Distinguishing between vascular congestion, interstitial edema, and alveolar edema is challenging, especially in mild cases, as vascular congestion leads to venous hypertension, causing interstitial edema and potentially progressing to alveolar flooding. Interstitial edema is primarily caused by increased pulmonary capillary hydrostatic pressure, leading to fluid leakage around the lungs, visible as peribronchial cuffing, septal thickening, and Kerley B lines on X-rays, which can progress to alveolar edema if severe [2]. Alveolar

edema occurs when fluid fills the alveoli due to high pressure, damaging the lung lining, often following interstitial edema and appearing centrally on imaging when pulmonary venous pressure exceeds 30 mmHg [3]. Radiologists diagnose edema levels using key features such as cephalization, Kerley lines, pleural effusion, bat wings, and infiltrates, which appear together rather than individually on CXRs. Analyzing these features and their anatomical significance can enhance diagnostic accuracy and deepen understanding.

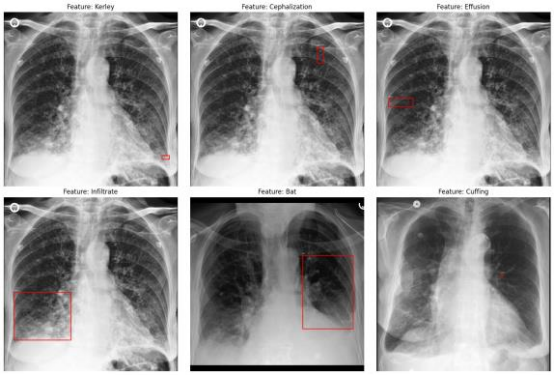

FIGURE 2. Different feature-wise Chest X-ray

Key features of pulmonary edema on CXRs include Kerley lines, indicating interstitial fluid; cephalization, an early sign of vascular congestion; pleural effusion, fluid in the pleural space seen in later stages; infiltrates, white opacities signaling alveolar fluid accumulation; bat-wing appearance, bilateral hilar opacities in severe cases; and cuffing, bronchial wall thickening due to peribronchial fluid.

1.2 Deep learning approaches for Edema classification

AI, particularly deep learning, has significantly improved pulmonary edema classification and diagnosis by enhancing accuracy and efficiency. CNNs excel in image-based classification by identifying subtle patterns. AI has shown strong potential in pulmonary edema detection, with a 2021 study achieving a 0.99 AUC for alveolar edema using MIMIC-CXR reports and another study reaching 92.3% accuracy in diagnosing canine cardiogenic pulmonary edema from CXRs [4]. A deep learning system, Non-local Channel Attention ResNet, was developed to assess pulmonary edema severity in COVID-19 pneumonia patients using 2,062 lung ultrasound images, showing strong performance and promising potential for monitoring [5]. The COVID-19 pandemic prompted significant research on lung disease classification using CXRs. This study presents a VGG-19-based CNN model that classifies pneumonia, lung cancer, TB, lung opacity, and COVID-19, achieving 96.48% accuracy and outperforming existing methods, highlighting its potential for improved diagnosis and treatment [6]. A multichannel deep learning approach using EfficientNet-B0, B1, and B2 with a stacked ensemble classifier achieved up to 99% accuracy in classifying pneumonia, TB, and COVID-19, outperforming existing methods for point-of-care diagnosis [7]. A key research trend is the development of interpretable CAD systems for CXR analysis, with the “CheXpert” dataset enabling CNNs to handle diagnostic uncertainty and outperform radiologists in detecting cardiomegaly, edema, and pleural effusion [8]. Beyond classification, some studies focus on detecting key edema features like cephalization and Kerley lines. This study used 1,000 annotated CXRs and a two-stage deep learning approach—lung segmentation and edema feature localization. The SABL network achieved the highest Average Precision (AP) (0.568), demonstrating effectiveness in identifying pulmonary edema features. [9].

This study compares six deep learning models (ResNet101, DenseNet201, DenseNet121, CheXNet [10], MobileNetV2, and EfficientNetB3) in diagnosing pulmonary edema severity from CXRs, evaluating the impact of general vs. medical-specific pretraining, and using Grad-CAM to align model explanations with radiologist-identified features for each edema class.

2. Materials and Methods

Methods can be described in 3 parts, first part dataset, which is also the Materials in this study, second part is the deep learning model, and third part is Grad-CAM for the explainability of the model.

2.1. Dataset

The dataset used in this study consists of CXRs from the Medical Information Mart for Intensive Care (MIMIC) database [9]. A thoracic radiologist with over 10 years of experience annotated 1,000 CXR studies from 741 patients, marking features related to pulmonary edema, such as cephalization, Kerley lines, pleural effusions, bat wings, and infiltrates. Cephalization and Kerley lines were annotated with polylines, while other features used binary masks. The dataset's size balances practicality with robust model training needs. Additionally, it includes the SLY dataset, with CXR AP (anterior to posterior) and lateral view images, enabling the connection between the bounding box and Grad-CAM features. The severity levels annotated include “No edema,” “Vascular Congestion,” “Interstitial Edema,” and “Alveolar Edema.” The dataset consists of 741 CXRs with the following class frequencies: 21 for No Edema, 74 for Vascular Congestion, 51 for Interstitial Edema, and 595 for

Alveolar Edema. The dataset exhibits significant class imbalance, with Alveolar Edema (595 X-rays) being the most common and Interstitial Edema (51 X-rays) the least. To compensate for limited normal X-rays, 681 external samples were added. High augmentation was applied to the training set, including horizontal flipping, random rotations (−10° to 10°), brightness and contrast adjustments (0.8 to 1.2), and zooming (90% to 120%) to enhance model generalization by introducing spatial and lighting variations. After the augmentation and oversampling, the dataset was,

TABLE 1. Final data distribution after augmentation

Classes	Train	Validation	Test
Alveolar Edema	1000	174	150
Interstitial Edema	945	100	100
Normal	1000	19	19
Vascular Congestion	1260	13	19

The dataset was processed by resizing images to 224x224 pixels with 3 channels, and normalizing pixel values to the range [0,1] for consistent intensity levels and improved model stability. Besides, Up sampling also explored focal loss with alpha scores [0.45, 0.25, 0.20, 0.10] to solve class imbalance. But Up sampling performed better than the Focal loss technique.

2.2. Methodology

This study employed various CNN architectures as backbone models for edema severity classification from CXR images, using transfer learning to fine-tune pre-trained weights for better adaptation to the dataset. Six CNN models are used here, and their comparison is provided below.

TABLE 2. Model summary

Model	Pretrained dataset	Depth	Parameters
ResNet101	ImageNet	209	44.7
DenseNet121	ImageNet	402	20.2
DenseNet201	ImageNet	242	8.1
CheXNet	ImageNet & Chest X-ray 14	242	8.1
MobileNetV2	ImageNet	105	3.5
EfficientNetB3	ImageNet	210	12.3

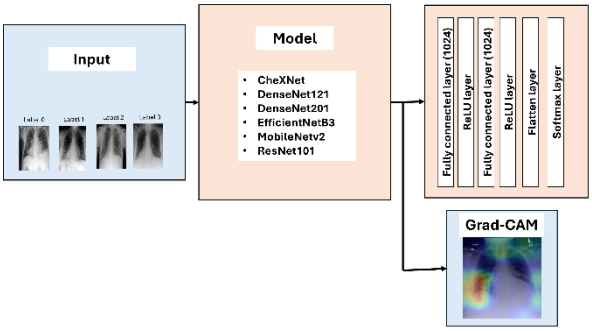

FIGURE 3. Model Training

In the case of model heads, it has a linear sequential layer followed by a ReLU (Rectified Linear Unit) activation function. Another Sequential layer followed by a ReLU activation function. Also, flatten layer is used in neural networks to convert multi-dimensional feature maps (usually from convolutional layers) into a 1D vector before passing it to fully connected (dense) layers. Finally, a classification layer for 4 heads with the SOFTMAX activation function. To train the models, we used the same parameters. Every model is trained for 100 epochs. The batch size for the Train, Validation, and Test sets was 32. Also used Adam optimizer with $1x10e^{-3}$ learning rate. Also, adaptive learning rate adjustment was used, which monitored validation loss for 10 epochs; if it did not decrease, it decreased the learning rate at a fixed rate.

2.3. Metrics and Grad CAM

Model performance is evaluated using metrics like accuracy, precision, recall, F1 score, and AUC. Accuracy measures overall correctness. Precision minimizes false positives, while recall focuses on capturing actual positives. F1 score balances precision and recall, and AUC assesses how well the model distinguishes between classes across thresholds.

Grad-CAM is a visualization technique for interpreting CNNs by highlighting image regions that influence the model's predictions. It uses gradients from the final convolutional layer to create a heatmap, showing key areas contributing to the decision. Grad-CAM is widely used in medical imaging and explainable AI to improve model transparency and trust.

3. Results and Discussion

The performance of six CNN models—CheXNet, EfficientNetB3, ResNet101, DenseNet201, DenseNet121,

and MobileNetV2 was evaluated using precision, recall, F1 score, accuracy, and AUROC, as presented in the table below. Among these models, only CheXNet is pretrained on CXRs, giving it a distinct advantage in feature extraction over ImageNet-pretrained models. DenseNet121 shares the same architecture as CheXNet, while DenseNet201 offers deeper layers for more feature extraction. MobileNetV2 is the most lightweight model, providing efficiency at the cost of depth. EfficientNetB3, selected from the EfficientNet family (B1–B7), offers a balanced trade-off between accuracy and computational efficiency. Detailed evaluation metrics for all models are summarized in the following table.

TABLE 3. Result summary of 5 models

Test set	Dense Net 121	Dense Net 201	Chexnet	Mobile net v2	Resne t 101	Efficient net B3
Accuracy	0.89	0.90	**0.91**	0.88	0.86	0.88
Precision	0.74	0.78	**0.79**	0.75	0.71	0.74
Recall	0.72	**0.77**	0.75	0.72	0.69	0.75
F1 score	0.73	**0.77**	**0.77**	0.73	0.70	0.75
AUC score	0.86	0.87	0.88	**0.89**	0.87	0.84

CheXNet outperforms other models in accuracy, precision, and F1-score, while DenseNet201 excels in recall, which is crucial for medical classification as it minimizes false negatives. MobileNetV2 achieves the highest AUC, indicating strong class discrimination despite its lightweight design. Overall, CheXNet, optimized for CXR tasks, balances accuracy and feature extraction well, whereas DenseNet201 is more reliable for capturing positive cases, and MobileNetV2 excels in computational efficiency with robust AUC performance.

As a primary, it can be seen that Models with dense block like CheXNet, Densenet201, DenseNet121 are the best performing models. By class-wise metrics result more information can be retrieved.

TABLE 4. CheXNet result for individual classes

Model	Alveolar Edema	Interstitial Edema	Normal	Vascular Congestion
Precision	0.92	0.56	0.98	0.71
Recall	0.95	0.53	1.00	0.53
Specificity	0.94	0.97	1.00	0.97
F1-score	0.93	0.54	0.99	0.61

TABLE 5. DenseNet 201 result for individual class

Model	Alveolar Edema	Interstitial Edema	Normal	Vascular Congestion
Precision	0.93	0.58	0.96	0.65
Recall	0.92	0.58	1.00	0.58
Specificity	0.91	0.97	1.00	0.97
F1-score	0.93	0.58	0.98	0.61

DenseNet201 and CheXNet perform equally well for the Normal class, achieving perfect recall and specificity (1.0). For Alveolar Edema, CheXNet has a slightly higher recall (0.95 vs. 0.92), while DenseNet201 has a marginally better precision (0.93 vs. 0.92), making their F1-scores nearly identical (0.93). DenseNet201 outperforms CheXNet in Interstitial Edema, achieving a higher F1-score (0.58 vs. 0.54) due to better recall (0.58 vs. 0.53). Both models struggle with Vascular Congestion, but DenseNet201 shows slightly improved recall (0.58 vs. 0.53) while maintaining the same F1-score (0.61). Overall, DenseNet201 demonstrates more consistency, particularly in difficult cases, whereas CheXNet maximizes recall for Alveolar Edema. However, CheXNet struggles with Interstitial Edema and Vascular Congestion, leading to low recall, precision, and F1-scores, highlighting its difficulty in balancing detection performance across all edema types.

This study also evaluates the impact of pretrained weights on model performance. Despite having the same architecture, CheXNet, pretrained on CXRs, outperforms DenseNet121 in all metrics. CheXNet even rivals DenseNet201, demonstrating that domain-specific pretrained weights enable better feature extraction with fewer layers.

Grad-CAM is a powerful tool for identifying which image regions a model focuses on when making predictions. In diagnosing pulmonary edema, it helps to reveal whether the model prioritizes lung opacities or other relevant features. This dataset includes bounding boxes for interim features, allowing comparison between model-focused areas and radiologist-identified regions. Among the six models used, CheXNet, DenseNet201, and DenseNet121 performed best, so their Grad-CAM visualizations were compared with the provided CXRs with bounding boxes. For Alveolar Edema, which is the most severe form of Edema. Anatomically, Cephalization, infiltration, and bat wings are the predominant features of this class. Cephalization is mostly dominant in the early stages of Alveolar Edema. Infiltrates are mostly seen in central and prehilar areas. Bat Wing is a unique feature of this stage. Best Performing mode CheXNet focuses on the costophrenic angles. Mostly, Effusion can be seen from blunting the costophrenic angles.

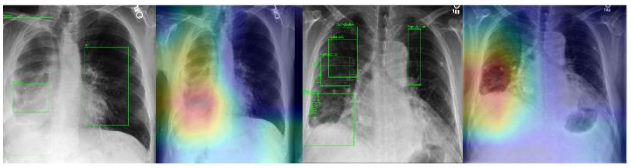

FIGURE 4. Ground Truth (1st and 3rd), CheXNet Grad-CAM (2nd and 4th) for Alveolar Edema. CheXNet prioritizes Effusion features.

So, from Figure 4, CheXNet prioritizes Effusion features for Alveolar Edema classification. As CheXNet has a bias towards Costophrenic angles. It misses features like bat wings and cephalization. DenseNet201 also focuses on the costophrenic angle or the lower half of the lung region. Due to the extra layers than CheXNet, DenseNet201 can find subtle features like cephalization.

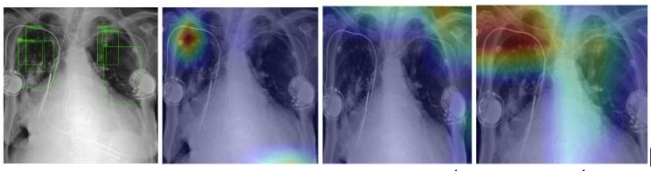

FIGURE 5. Ground Truth (1st), DenseNet201(2nd), CheXNet(3rd), DenseNet121(4th) for Alveolar Edema. Here, CheXNet missed Cephalization, but DenseNet was successful.

Here at figure 5, where CheXNet failed to focus on cephalization features, both Dense Net architectures were successful at focusing on cephalization. DenseNet architecture can find subtle features like cephalization and bat wings. Interstitial Edema is the second stage of Edema. It is very hard to detect because of subtle features. Still, anatomically, Kerley B Lines are the most common features for Interstitial Edema. All models performed badly in the Interstitial edema classification. In Figure 6, CheXNet is mainly focusing on the Hilar area of the CXR, which mainly hosts the bat structures. It focuses on the bat wing area for identifying Interstitial Edema.

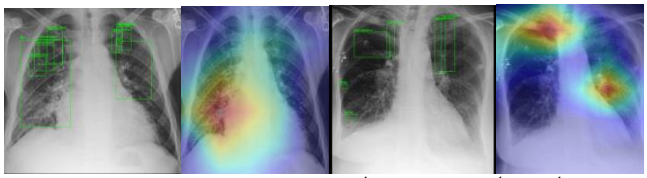

FIGURE 6. Ground Truth (1st and 3rd), CheXNet (2nd and 4th) for Interstitial Edema. CheXNet is focusing on the Hilar structure.

Besides Bat wing's structure, CheXNet is prioritizing Cephalization for Interstitial Edema classification in figure 6 (4th). DenseNet201 performed better for Interstitial Edema classification. DenseNet201 also focuses on Cephalization and Bat wings, while Interstitial Edema classification is shown in Figure 7. DenseNet201 is giving Cephalization the most favor than Bat wings, unlike CheXNet.

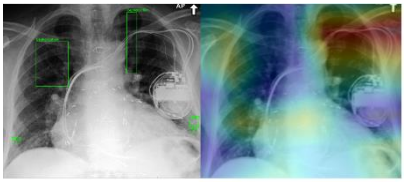

FIGURE 7. Ground Truth (left), DenseNet201 Grad-CAM (right) for Interstitial Edema. DenseNet201 was giving priority to Cephalization.

DenseNet201 can detect underlying subtle structures better than CheXNet.

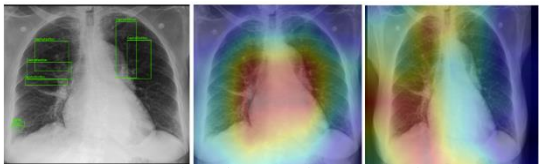

FIGURE 8. Ground Truth (1st), CheXNet Grad-CAM (2nd), DenseNet201 Grad-CAM (3rd) for Interstitial Edema. CheXNet failed, but DenseNet201 was successful in detecting subtle features.

Here at Figure 8, CheXNet failed to classify this as Interstitial Edema, but DenseNet 201 was successful in classifying it. Because DenseNet 201 was successful in detecting subtle features like cephalization that's why it decided to classify this as Interstitial Edema. DenseNet121 performs worse than the other 2 models, but it shows a pattern of finding features on the border of the lungs, and Kerley B lines are also present on the pleural border of the lungs as shown in figure 9. So the model does detect some Kerley lines features, but the model does not show any reliability like the other 2 models on any particular feature.

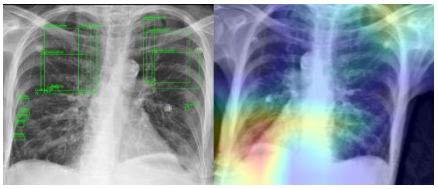

FIGURE 9. Ground Truth (left), DenseNet121 Grad-CAM (right) for Interstitial Edema. DenseNet121 does not show any reliability to any features.

Vascular Congestion is the initial Edema level. According to the Radiologists, Cephalization is the most dominant feature. Cephalization is also a dominant feature of Alveolar Edema. It creates confusion between Vascular Congestion and Alveolar Edema. CheXNet is the most suitable model for Vascular Congestion classification. In Figure 10, it is shown that, typically for Vascular Congestion, CheXNet is focusing on the upper part of the chest.

Cephalizations are mostly seen upper part of the lung.

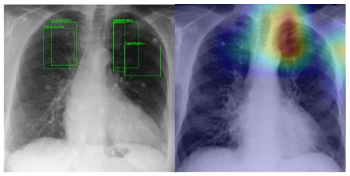

Figure 10. Ground Truth (left), CheXNet Grad-CAM (right) for Vascular Congestion. CheXNet focuses on the upper part of the lung.

Dense Net 201 and Dense Net 121 both performed badly in Vascular congestion classification. They don't follow any consistent features. But DenseNet121 follows a pattern of following the feature upper lung like Cephalization, as shown in Figure 11 (2^{nd}).

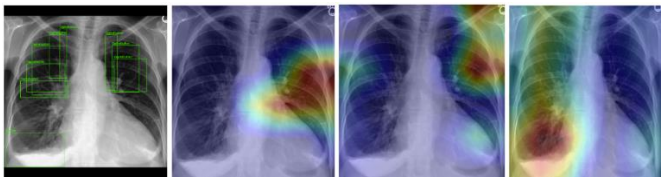

Figure 11. Ground Truth (1st), CheXNet (2nd), DenseNet121(3rd), DenseNet201(4th) for Vascular Congestion. Only DenseNet121 gives importance to the upper chest for Vascular Congestion classification.

This study explored the correlation between radiologist-selected and model-identified features using Grad-CAM, revealing key regions influencing the model's decisions, but due to the small dataset and class imbalance, further research with larger datasets and attention-based models is needed for more reliable conclusions.

4. Conclusions

This study aimed to evaluate the performance of traditional CNN models in classifying edema severity levels and identifying key factors influencing model performance. Six models were tested, with CheXNet achieving the highest accuracy, precision, and F1 score. DenseNet201 excelled in recall, while MobileNetV2 outperformed others in AUC. Class-wise analysis showed CheXNet performed best in classifying vascular congestion, while DenseNet201 was superior for alveolar and interstitial edema. Both models showed nearly perfect performance in distinguishing normal cases from edema classes.

Pretrained CheXNet outperformed DenseNet121 in edema detection by leveraging X-ray learned features. Grad-CAM showed CheXNet effectively identified key edema patterns (pleural effusion, bat-wing signs, cephalization), while DenseNet201 detected subtler features but struggled with vascular congestion. Dense blocks proved most effective for severity classification, though DenseNet models showed inconsistent feature focus.

Acknowledgements

This paper is partially supported by a grant from the Independent University, Bangladesh (IUB).

References

[1] Horng S, Liao R, et al., "Deep Learning to Quantify Pulmonary Edema in Chest," *Radiology: Artificial Intelligence*, vol. 3, no. 2, 2021.

[2] Murray JF., "Pulmonary edema: pathophysiology and diagnosis." *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease*, vol. 15, 2011.

[3] Khan AN, Al-Jahdali H, Al-Ghanem S, Gouda A., "Reading chest radiographs in the critically ill (Part II): Radiography of lung pathologies common in the ICU patient," *Annals of thoracic medicine,* vol. 4, 2009.

[4] Kim E, Fischetti AJ, Sreetharan P, Weltman JG, Fox PR. "Comparison of artificial intelligence to the veterinary radiologist's diagnosis of canine cardiogenic pulmonary edema," *Veterinary Radiology & Ultrasound,* vol. 63, no. 2, 2022.

[5] Q. Huang, Y. Lei, et al., "Evaluation of Pulmonary Edema Using Ultrasound Imaging in Patients With COVID-19 Pneumonia Based on a Non-local Channel Attention ResNet," *Ultrasound in Medicine & Biology,* vol. 5, p. 48, 2022.

[6] M. A. Mufarah, Q. Ni, R. Jiang, et al., "A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images," *Alexandria Engineering Journal*, vol. 64, 2023.

[7] Ravi, V., Acharya, V. & Alazab, M., "A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images." *Cluster Computing*, vol. 26, no. 2, 2023.

[8] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.. 33, no. 01, pp. 590-597, 2019.

[9] Viacheslav V Danilov, Anton O Makoveev, Alex Proutski, Irina Ryndova, Alex Karpovsky, Yuriy Gankin, "Explainable AI to identify radiographic features of pulmonary edema," *Radiology Advances,* vol. 1, no. 1, 2024.

[10] Rajpurkar P, et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, 2018.