AI-Enhanced Pneumonia Detection with Visual Interpretability

Abrar Ahmed Shahok, Faizan Ali Memon, Kaleemullah Jalbani, M. Shoaib
Department of Computer Science Quaid-e-Awam University of Engineering, Science and Technology Nawabshah, Pakistan
***Correspondence**: abrarshahok@gmail.com, mahmoodfaizan033@gmail.com, kaleemullahjalbani1349@gmail.com, shaikhmuhammadshoaib37@gmail.com

Citation | Shahok. A. A, Memon. F. A, Jalbani. K, Shoaib. M, "AI-Enhanced Pneumonia Detection with Visual Interpretability", IJIST, Vol. 07 Special Issue. pp 118-126, May 2025
Received| April 15, 2025 **Revised|** May 13, 2025 **Accepted|** May 15, 2025 **Published|** May 17, 2025.

Pneumonia is a serious lung infection that can be life-threatening, particularly for young children, the elderly, and people with weakened immune systems. Early detection is crucial but difficult because pneumonia signs on X-rays can be subtle. Many AI tools can help diagnose pneumonia, but they often work like "black boxes," making it hard for doctors to trust their decisions. This study introduces a mobile app that uses Convolutional Neural Networks (CNNs) to detect pneumonia from X-rays. To improve transparency, we use Explainable AI (XAI) to highlight the areas of the X-ray that influenced the diagnosis. Additionally, we integrate a Large Language Model (LLM) to generate clear, structured medical reports. Our goal is to create a trustworthy and user-friendly tool for doctors in real-world settings.

Keywords: Deep Learning for Pneumonia Detection, Explainable AI in Medical Imaging, Pneumonia Classification with AI, Ethical AI in Healthcare, AI for Radiology and X-ray Analysis

Introduction:

Diagnosing pneumonia from chest X-rays is challenging, even for experienced radiologists, due to its subtle and variable indicators. While AI-powered tools exist, most function as "black boxes," providing results without explanations, which limits trust and adoption in clinical settings.

Our research found that current solutions are mostly web-based and offer little practical value for radiologists. A dedicated mobile app tailored to their workflow would be more effective. We aim to develop an AI-driven mobile app that not only detects pneumonia but also highlights key areas in X-rays using Explainable AI (XAI) techniques. This will help doctors interpret results with confidence.

To enhance usability, the app will include a Large Language Model (LLM) to generate structured, easy-to-understand medical reports. Designed for real-world healthcare environments, it will be rigorously tested for reliability and accuracy. By offering a transparent and accessible AI tool, we hope to support healthcare professionals in improving pneumonia detection and treatment outcomes.

Related Work & Literature Review:

AI-powered pneumonia detection from chest X-rays has become a major area of research, with the potential to make diagnoses faster and more accurate. Deep learning, especially convolutional neural networks (CNNs), has played a huge role in this progress.

One breakthrough was CheXNet, a 121-layer CNN that performed at a radiologist's level in detecting pneumonia. It even used heatmaps to highlight affected areas, making its decisions more transparent [1]. Researchers have also improved accuracy by adding attention mechanisms to models like EfficientNetB0 and DenseNet121, helping them focus on the most important parts of an X-ray [2].

Another major step forward is the use of anchor-free object detection, which simplifies pneumonia detection by cutting down on computational costs while maintaining accuracy [3]. More recently, Vision Transformers (ViTs) have been introduced, using self-attention to capture both local and global features in chest X-rays, further improving performance and interpretability [4][5].

Despite these advancements, AI models often work like "black boxes," making it difficult for doctors to understand why a model made a particular decision. That's where Explainable AI (XAI) comes in. Techniques like Grad-CAM create visual explanations by highlighting the regions of an image that influenced the AI's prediction [6][7]. Tools like LIME and SHAP go even further, breaking down AI decisions at the feature level, making them more trustworthy and easier to interpret [8].

For example, lightweight CNN models combined with Grad-CAM have been successfully used to detect pneumonia and even COVID-19, providing clear visual explanations [9]. The use of XAI in diagnosing lung diseases is proving to be essential, giving doctors deeper insights into how AI makes its decisions [10].

Gaps & Opportunities:

Despite these advancements, most existing solutions remain web-based and do not align with the practical needs of healthcare providers who require mobile-friendly intuitive applications. This research aims to bridge this gap by developing a mobile app that delivers accurate pneumonia detection with clear, visual explanations, ensuring that AI-powered diagnosis is both reliable and accessible for medical professionals.

Objectives:

- To develop a mobile-based AI tool capable of detecting pneumonia from chest X-rays using a custom Convolutional Neural Network (CNN).

- To enhance diagnostic transparency and trust by integrating Explainable AI (XAI) techniques such as Grad-CAM and LIME.
- To automate the generation of medical reports through a Large Language Model (LLM), improving efficiency and consistency.
- To provide an accessible, real-time solution tailored for healthcare professionals through a user-friendly mobile application.
- To ensure secure data handling and communication using a Flask backend and Firebase for privacy-compliant storage and authentication.

Novelty:

Unlike existing pneumonia detection solutions that are primarily web-based and offer limited interpretability, this study introduces an integrated mobile application combining a custom CNN model (PneumoniaNet), Explainable AI (XAI), and a Large Language Model (LLM) for real-time diagnosis and report generation. The proposed system bridges the gap between high-accuracy AI diagnostics and clinical usability by offering transparent predictions and structured outputs tailored to radiologists' workflows. This end-to-end pipeline—from image analysis to medical report delivery—is optimized for mobile environments, addressing key challenges in real-world deployment.

Implementation and Methodology:

Data Collection:

The dataset for this study was collected from Kaggle, combining two publicly available datasets:

- Chest X-ray Pneumonia Dataset
- COVID-19 Pneumonia Normal Chest X-ray Dataset

The dataset consists of two primary classes:

- NORMAL
- PNEUMONIA

Before pre-processing, the dataset contained:

- 3,896 NORMAL images
- 4,273 PNEUMONIA images

The datasets were merged by removing duplicate images, resolving naming conflicts through standardized renaming, and unifying class labels for consistency.

Data Preprocessing:

To enhance model generalization and improve performance, the following preprocessing techniques were applied:

Normalization:

- Images were rescaled to the range [0,1] by dividing pixel values by 255.

Resizing:

- Resized images to 224 × **224** pixels to keep them consistent with the model's requirements, make training faster, and use less memory.

Handle Class Imbalance:

- Handled class imbalance by applying the oversampling minority technique, duplicating NORMAL images until both classes had 4,273 samples each**.**
- This ensures the model learns from both classes equally, reducing bias and improving the accuracy of predictions.

Data Augmentation:

- To further improve model generalization and mitigate overfitting, augmentation techniques were applied.
- Random rotation, horizontal flipping, and zooming were applied during training.

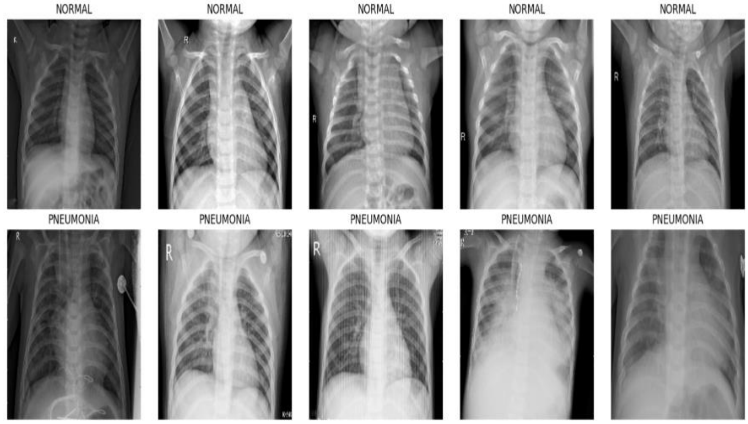

Figure 1. Dataset sample

Model Design & Training:

CNN Architecture: Implement, train, and fine-tune a custom CNN on the dataset and compare it with state-of-the-art models, including **ResNet50, DenseNet121, and EfficientNetB0.**

Training Strategy: Split the dataset into training, validation, and test sets (e.g., 80%, 10%, 10%). Use optimizers like Adam and loss functions such as categorical cross-entropy.

Model Evaluation: Monitor performance using accuracy, precision, recall, and F1-score on the validation set.

Layer (type)	Output Shape	Param #
conv2d_14 (Conv2D)	(None, 224, 224, 32)	896
batch_normalization_14 (BatchNormalization)	(None, 224, 224, 32)	128
max_pooling2d_10 (MaxPooling2D)	(None, 112, 112, 32)	0
conv2d_15 (Conv2D)	(None, 112, 112, 64)	18,496
batch_normalization_15 (BatchNormalization)	(None, 112, 112, 64)	256
max_pooling2d_11 (MaxPooling2D)	(None, 56, 56, 64)	0
conv2d_16 (Conv2D)	(None, 56, 56, 128)	73,856
batch_normalization_16 (BatchNormalization)	(None, 56, 56, 128)	512
conv2d_17 (Conv2D)	(None, 56, 56, 128)	147,584
batch_normalization_17 (BatchNormalization)	(None, 56, 56, 128)	512
max_pooling2d_12 (MaxPooling2D)	(None, 28, 28, 128)	0
conv2d_18 (Conv2D)	(None, 28, 28, 256)	295,168
batch_normalization_18 (BatchNormalization)	(None, 28, 28, 256)	1,024
conv2d_19 (Conv2D)	(None, 28, 28, 256)	590,080
batch_normalization_19 (BatchNormalization)	(None, 28, 28, 256)	1,024
max_pooling2d_13 (MaxPooling2D)	(None, 14, 14, 256)	0
flatten_1 (Flatten)	(None, 50176)	0
dense_12 (Dense)	(None, 128)	6,422,656
dropout_6 (Dropout)	(None, 128)	0
dense_13 (Dense)	(None, 64)	8,256
dropout_7 (Dropout)	(None, 64)	0
dense_14 (Dense)	(None, 2)	130

Total params: 7,560,578 (28.84 MB)
Trainable params: 7,558,850 (28.83 MB)
Non-trainable params: 1,728 (6.75 KB)

Figure 2. CNN Architecture

Integration of Explainable AI:

- Implement Explainable AI (e.g., LIME and Grad-Cam) to make the model's decisions transparent.

- Grad-CAM and LIME both explain how a model makes decisions, but they do it in different ways.
- Grad-CAM highlights the most important areas of an image that influence the model's prediction, while LIME divides the image into sections and tests how removing parts affects the result.

LLM for Medical Report Generation:

- Use a Large Language Model (LLM) to automatically generate medical reports from X-ray images.
- This helps doctors by saving time, ensuring consistency, and providing detailed analysis.
- Train model (e.g., Llama) on medical data to create accurate and well-structured reports, making the diagnosis process more efficient.

Mobile Application and Back-end Development:

- Implement a Flask backend to serve the trained model's predictions via REST APIs.
- Ensure secure and efficient communication between the app and the server.
- Utilize Firebase for storing user's diagnostic results, logs, and images securely.
- Implement user authentication for data privacy.

Develop a user-friendly mobile app with Flutter that receives an X-ray image, sends it to the AI model deployed on the server, and retrieves the result

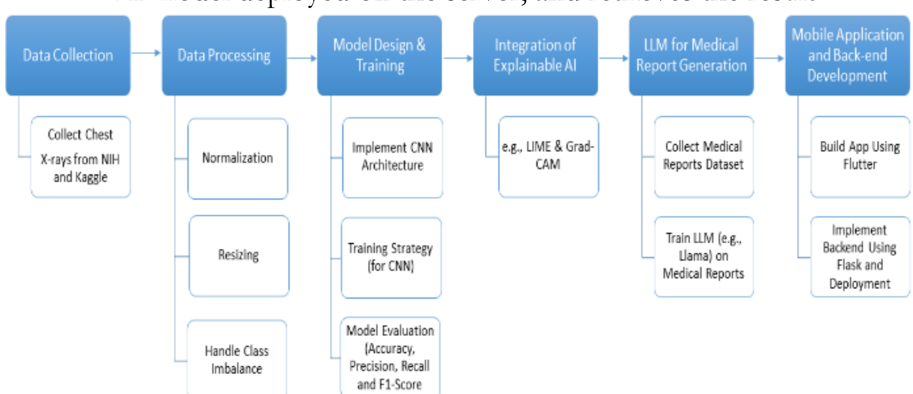

Figure 3. Methodology Flowchart

Results:

Model Performance:

- PneumoniaNet (our custom CNN model) achieved 96.14% accuracy, with balanced performance across both classes.
- It had 96.03% precision, 96.25% recall, and a 96.14% F1-score for normal cases, while pneumonia cases showed 96.25% precision, 96.03% recall, and a 96.14% F1-score.

	precision	recall	f1-score	support
NORMAL	0.960280	0.962529	0.961404	427.000000
PNEUMONIA	0.962529	0.960280	0.961404	428.000000
accuracy	0.961404	0.961404	0.961404	0.961404
macro avg	0.961405	0.961405	0.961404	855.000000
weighted avg	0.961406	0.961404	0.961404	855.000000

Figure 4. Classification report of PneumoniaNet

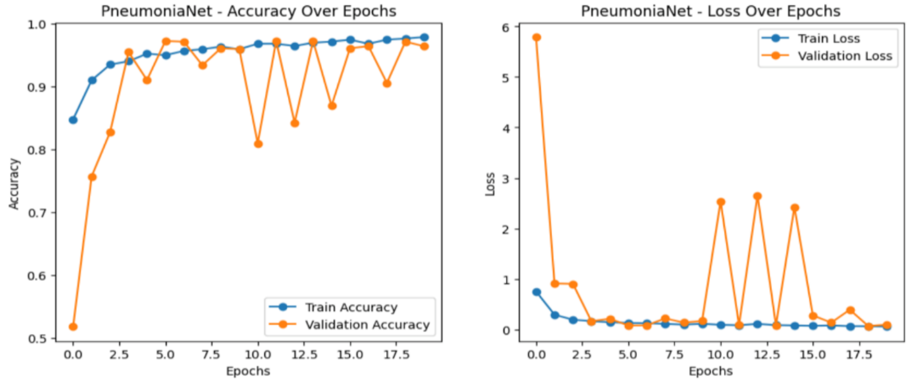

Figure 5. Loss and Accuracy plot over epochs

- The confusion matrix confirms minimal misclassification, with 411 correct predictions per class and only 33 errors out of 855 samples.
- The 99% ROC curve demonstrates strong model performance, with a high area under the curve (AUC), indicating excellent discrimination between normal and pneumonia cases.

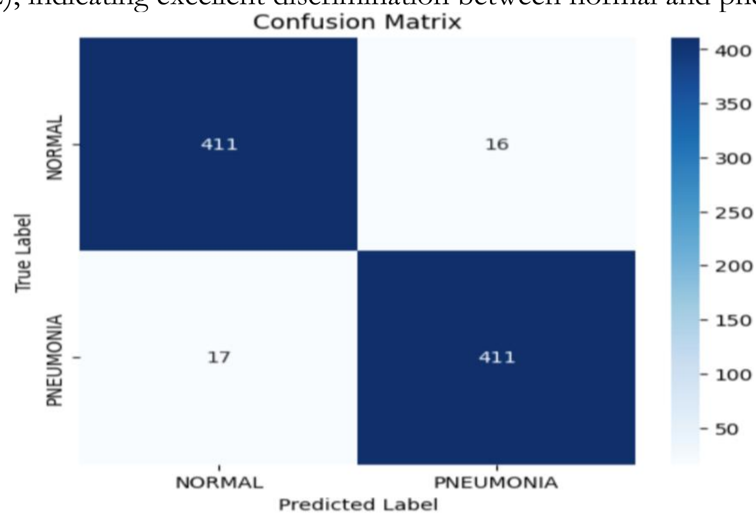

Figure 6. Confusion Matrix of Pneumonia Net

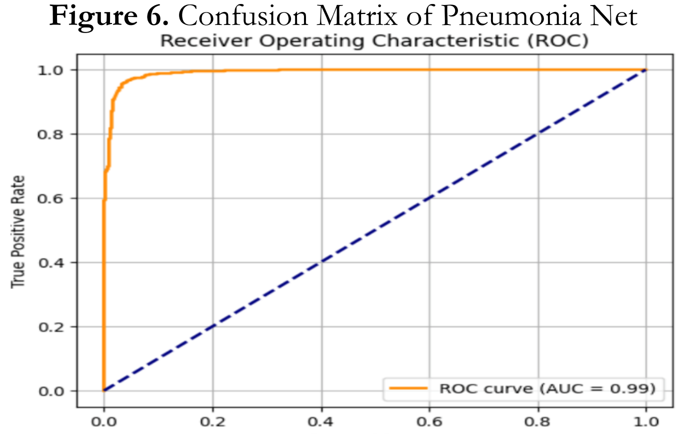

Figure 7. ROC of Pneumonia Net

Model Comparison:

- **EfficientNetB0** demonstrated superior classification performance.
- **However,** Pneumonia Net (our custom CNN model) offers the best trade-off between accuracy and speed (3.6 times faster), making it a strong candidate for real-time medical imaging applications.

Model	Num Parameters	Accuracy	Precision	Recall	F1-score	Avg Inference Time (s)
PneumoniaNet	7.6M	0.961404	0.961406	0.961404	0.961404	0.003497
ResNet50	23.6M	0.925146	0.928909	0.925146	0.924986	0.009809
DenseNet121	7.0M	0.955556	0.955916	0.955556	0.955548	0.016529
EfficientNetB0	4.1M	0.982456	0.982522	0.982456	0.982455	0.012843

Figure 8. Pneumonia Net vs other state-of-the-art Models

XAI (LIME vs Grad-CAM):

- In the comparison image, the red areas (Grad-CAM) focus on specific key regions, while the blue areas (LIME) are more spread out and sometimes less precise.
- Grad-CAM is better for medical use because it's much faster and more focused. Unlike LIME, which must test many different parts of the image multiple times, Grad-CAM quickly finds the important regions in one step.
- This makes it more efficient, especially when working with large numbers of medical images. Plus, its heatmaps clearly show where the model is looking, which helps doctors trust and understand AI predictions more easily.

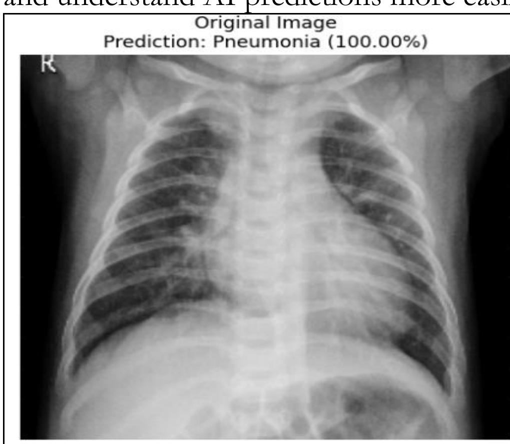

Figure 9. Original Image with Prediction

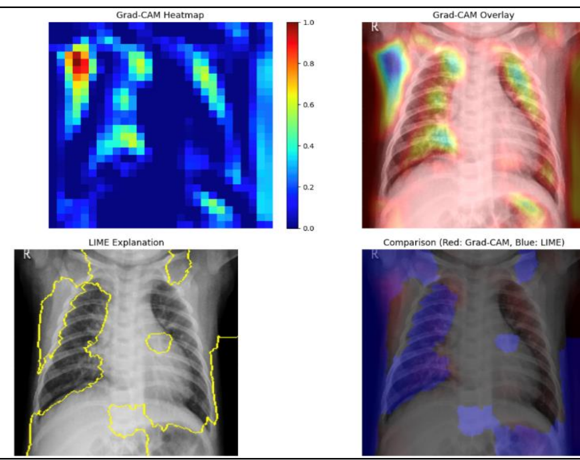

Figure 10. LIME vs Grad-CAM

Potential Challenges and Mitigation Strategies:

Challenge	Mitigation Strategy
Data Limitations	Use data from different sources to make the model more accurate and less biased.
Class Similarity in Medical Imaging	Utilize advanced CNN architectures (e.g., ResNet, EfficientNet) to capture subtle differences.
High Computational Requirements	Leverage cloud computing platforms (e.g., Google Colab, Kaggle Notebook) to utilize GPUs/TPUs for faster training and optimization.
LLM Training for Medical Reports	Use verified data, train the model on medical terms, and have experts review the results.

Model Calibration and Uncertainty Estimation for Medical Usage:

In our research, model calibration, and uncertainty estimation were critical to ensure the clinical reliability of Pneumonia Net, our custom CNN model. Despite achieving a high accuracy of 96.14% and balanced precision, recall, and F1-scores for both normal and pneumonia cases, medical deployment requires more than strong performance metrics. Calibration analysis was used to assess whether the model's predicted probabilities aligned with actual outcomes—vital for determining how much trust clinicians can place in the AI's confidence scores. Additionally, we considered uncertainty estimation by evaluating prediction variability in borderline cases. This helps flag uncertain predictions that may require further human review. Given the high-stakes nature of medical diagnosis, such as identifying pneumonia, incorporating well-calibrated outputs and uncertainty-aware mechanisms enhances the safety, transparency, and acceptance of AI systems in clinical settings.

Limitations:

While our AI-enhanced pneumonia detection model demonstrates promising results, several limitations must be acknowledged:

1. **Data Limitations**: Our dataset comes from public sources and may not fully reflect diverse patient groups. Variations in imaging methods, equipment, and demographics could affect the model's accuracy across different populations.
2. **Model Interpretability:** We have used Explainable AI (XAI) techniques like Grad-CAM and LIME, but they have limitations. Grad-CAM highlights important features but may miss clinically relevant areas, while LIME is computationally heavy and less precise.
3. **Real-World Implementation Challenges:** Bringing AI into healthcare requires thorough validation, regulatory approval, and clinician trust. Moving from research to real-world use is complex, involving legal, ethical, and operational challenges.
4. **Computational Constraints:** Deep learning models need high processing power, making them hard to run on low-end mobile devices. Cloud-based solutions help but raise concerns about latency, cost, and privacy.
5. **Dependency on High-Quality X-rays:** AI accuracy depends on high-quality chest X-rays. Poor scans with noise or artifacts can impact performance.
6. **Ethical and Privacy Concerns:** Handling medical data requires strict security and compliance with regulations like HIPAA and GDPR to protect patient privacy.

Future Work:

To improve and expand this research, we plan to:

1. **Enhance Model Accuracy and Generalization:** Fine-tune Pneumonia Net with more diverse datasets to further improve accuracy and ensure it works well across different hospitals and imaging conditions.
2. **Develop a Mobile Application:** Build a user-friendly mobile app that allows doctors to upload X-rays, receive AI-powered diagnoses, and view clear visual explanations.
3. **AI-generated Medical Reports:** Train the Large Language Model (LLM) on more medical data to generate detailed and reliable reports, enhancing diagnostic support and efficiency.
4. **Validate with Clinical Trials:** Conduct real-world testing with medical professionals to assess the system's **reliability, usability, and effectiveness** in hospital environments.
5. **Optimize for Performance and Accessibility:** Explore cloud-based AI services **and** edge computing to make the model run faster, ensuring smooth performance even on low-end mobile devices.

These advancements will further strengthen the model's real-world applicability in clinical settings.

Conclusion:

In this study, we gathered and combined datasets from multiple sources to make our model more reliable and adaptable. Using this data, we built PneumoniaNet, an AI-powered system designed to detect pneumonia from chest X-rays. To ensure accuracy, we used Convolutional Neural Networks (CNNs), and to improve trust, we incorporated Explainable AI (XAI) techniques so doctors could see why the model made a particular diagnosis.

Our results show that PneumoniaNet achieved an accuracy of 96.14%, making it a strong candidate for real-world medical use. However, there are still some challenges to tackle before it can be widely adopted in clinical settings. These include making sure the model works well across different datasets, optimizing it for mobile devices, and ensuring it meets medical regulations.

By addressing these challenges, we move closer to creating an AI tool that can truly support healthcare professionals in diagnosing pneumonia more effectively.

Acknowledgments:

We acknowledge Quaid-e-Awam University of Engineering, Science and Technology Nawabshah for supporting this research and providing resources for development.

References:

[1] A. Y. N. Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv:1711.05225*, 2017, doi: https://doi.org/10.48550/arXiv.1711.05225.

[2] W. S. Qiuyu An, Wei Chen, "A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble," *Diagnostics*, vol. 14, no. 4, p. 390, 2024, doi: https://doi.org/10.3390/diagnostics14040390.

[3] W. L. & F. G. Linghua Wu, Jing Zhang, Yilin Wang, Rong Ding, Yueqin Cao, Guiqin Liu, Changsheng Liufu, Baowei Xie, Shanping Kang, Rui Liu, "Pneumonia detection based on RSNA dataset and anchor-free deep learning detector," *Sci. Rep.*, vol. 14, no. 1929, 2024, doi: https://doi.org/10.1038/s41598-024-52156-7.

[4] T. T. Ifty, S. A. Shafin, S. M. Shahriar, and T. Towhid, "Explainable Lung Disease Classification from Chest X-Ray Images Utilizing Deep Learning and XAI," *2024 IEEE 3rd Int. Conf. Comput. Mach. Intell. ICMI 2024 - Proc.*, 2024, doi: 10.1109/ICMI60790.2024.10586202.

[5] Dikai Li, "Attention-enhanced architecture for improved pneumonia detection in chest X-ray images," *BMC Med. Imaging*, vol. 24, no. 6, 2024, doi: https://doi.org/10.1186/s12880-023-01177-1.

[6] K. A. & S. S. Sukhendra Singh, Manoj Kumar, Abhay Kumar, Birendra Kumar Verma, "Efficient pneumonia detection using Vision Transformers on chest X-rays," *Sci. Rep.*, vol. 14, no. 2487, 2024, doi: https://doi.org/10.1038/s41598-024-52703-2.

[7] C.-L. W. and W.-C. C. R. -K. Sheu, M. S. Pardeshi, K. -C. Pai, L. -C. Chen, "Interpretable Classification of Pneumonia Infection Using eXplainable AI (XAI-ICP)," *IEEE Access*, vol. 11, pp. 28896–28919, 2023, doi: 10.1109/ACCESS.2023.3255403.

[8] Anthony Novokshanov, "Using Explainable Artificial Intelligence to Locate Pneumonia," *Res. Arch. Rising Sch.*, 2024, doi: Using Explainable Artificial Intelligence to Locate Pneumonia.

[9] S. A. B. Zubaira Naz, Muhammad Usman Ghani Khan, Tanzila Saba,Amjad Rehman, Haitham Nobanee, "An Explainable AI-Enabled Framework for Interpreting Pulmonary Diseases from Chest Radiographs," *Cancers (Basel).*, vol. 15, no. 1, p. 314, 2023, doi: https://doi.org/10.3390/cancers15010314.

[10] B. B. Izegbua E. Ihongbe ,Shereen Fouad ,Taha F. Mahmoud,Arvind Rajasekaran, "Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens," *PLoS One*, 2024, doi: https://doi.org/10.1371/journal.pone.0308758.