Interpretable Deep Learning Framework for Chest X-Ray Classification of Pneumonia and Lung Abnormalities

[1]Anjul Singh [2]Akanksha Kapoor [3]Shilpa Gupta [4] Kumud Dixit
[5] *Sujeet Kumar
[1,4]D.S. College, Aligarh,U.P.,India- 202001
[2,3]Jagannath International Management School, New Delhi, India-110085
[5]Greater Noida Institute of technology,Greater Noida, U.P.,India-201310

[1]anujsingh1212@gmail.com [2]akankshakapoor2292@gmail.com
[3]shilpa.gupta039@gmail.com [4]Kumuddixit30@gmail.com
[5v]Sujeetsingh142000@gmail.com

Abstract: Pneumonia and other lung abnormalities continue to be significant health issues of global concern and any diagnostic support to aid clinical decision-making must be quick and precise. The Chest X-ray imaging has been extensively utilized in respiratory assessment, but the manual interpretation is time-consuming and is likely to be intra-clinician variability. This research paper suggests a deep learning model that can be interpreted to classify pneumonia and related pulmonary abnormalities using chest radiographs. The model combines a convolutional neural network framework that is optimized to extract features with explainable AI algorithms like Grad-CAM, SHAP, and Integrated Gradients to visualize and confirm the behavior of the model. The interpretability tools underscore meaningful lung areas that lead to predictions and, therefore, make sure that such decisions are based on clinical significance, and not on arbitrary trends. High accuracy, sensitivity, and specificity have been shown by experimental evaluation in the ability to differentiate between normal and pathological images and explainability results indicate consistency with radiologically significant appearances like opacities and consolidation patterns. Model reliability is enhanced by the inclusion of transparent decision mechanisms and thus its use as a reliable diagnostic tool. All in all, the framework facilitates the gap between performance and interpretability, which leads to safer and more responsible AI-based medical imaging systems to detect respiratory diseases.

Keywords: *Chest X-ray classification, pneumonia detection, deep learning, explainable AI, Grad-CAM, SHAP, Integrated Gradients, model interpretability, medical image analysis, pulmonary abnormalities*

1. Introduction

Pneumonia, bronchopneumonia, and other lung abnormalities are respiratory diseases that remain number one causes of morbidity and mortality in most parts of the world particularly in children, old adults and immunocompromised people [7]. Chest X-ray imaging is one of the most available, affordable, and fast modalities of identifying pulmonary infections and structural aberrations in the clinical setting [3]. Radiographic interpretation proved to be a variable issue, even though it is extensively used, because of the variations in the knowledge of clinicians, workload pressure, and subtlety of early signs of pathology [14]. The

P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_12

challenges have shown that automated and reliable diagnostic tools can be helpful in assisting radiologists in their decision-making and lessening the burden of manual interpretation [5].

Deep learning innovations have introduced many changes to the analysis of medical imaging, especially using convolutional neural networks capable of learning high-level hierarchical patterns using raw pixel data [8]. Instead, CNN-based systems have demonstrated high accuracy in detecting characteristics like opacities, consolidation, ground-glass and textural irregularities of pneumonia and other lung conditions [12]. Despite this development, a large number of high-performing models are black boxes in which they generate predictions without insight into the manner in which such decisions were arrived at. The limitation also creates some worries about clinical reliability, the ethical adoption, and the fact that it may fall back to the existence of spurious correlations like image borders, device marks, or dataset artifacts instead of medically significant areas [1].

Explainable Artificial Intelligence (XAI) provides solutions to these issues because of introducing mechanisms that give insights into the inner-working of a model. Grad-Cam, Integrated Gradients, occlusion mapping, and SHAP are techniques that produce either visual or quantitative explanations that indicate the areas of the X-ray image that lead to the classification output [10]. The tools are used to ensure that the model focuses on the appropriate lung regions, including areas of consolidation or infiltrate, instead of the external noise [4]. XAI can enable a more open line of passage to the incorporation of deep learning systems in the actual diagnostics processes and enhance their acceptability by radiologists and healthcare regulatory bodies [9].

A variety of available literature has investigated the use of deep learning to classify pneumonia; most of these works aim to obtain high accuracy, but not interpretability, and clinical acceptability of model interpretations [6]. It has been shown that, according to some studies, models can make misclassifications because of biases due to imbalanced distribution of datasets, image acquisition conditions or annotation bias [13]. The other factor that should be taken into account is that the clinical imaging setting is highly dynamic, with regard to patient population, disease distribution and imaging modalities. A powerful classification system should hence be shown to be strong in generalizing various X-ray samples and at the same time be interpretable in meaningful subgroups [11]. Without any obscurity, clear-cut models assist the radiologists to distinguish between the grey cases and build confidence in the diagnosis. Interpretable deep learning can be used in validation as well as other aspects. It also contributes to the understanding of education among clinical practitioners and medical trainees in terms of radiographic indicators of various lung diseases [2]. In addition, interpretability is also in line with international initiatives of producing safe, accountable and conscientious AI systems which meet upcoming regulations regarding explainability of medical devices [15]. Although there has been the current development in XAI, there are still difficulties. The reliability of many methods of explanation may depend on the model architecture, perturbation noise or sensitivity to hyperparameters [5]. To illustrate, Grad-CAM heatmaps can occasionally identify areas of large, diffuse regions and it is challenging to identify particular pathological structures. SHAP provides pixel- or region-wise contribution information, but the cost of the computations increases with model complexity [7]. Gradient approaches indicate spatial attention, and SHAP shows the contribution of features by reasoning through perturbation. When these approaches are combined, the framework yields a more interpretable profile, which is more stable and can be trusted by the radiologists [4]. This study is a part of the trend of transparent AI in healthcare, where the aim is not only to surpass human readers, but to supplement human knowledge in a safe and trustworthy way [9]. Comprehensively, the introduction sets the basis of a deep learning system that can be interpreted, and is both accurate and transparent.

2. Related Works

The study of deep learning-based analysis of chest X-rays has quickly developed, particularly alongside the concomitant emergence of explainable artificial intelligence (XAI). The initial investigations in this field were more focused on maximizing precision in the detection of pneumonia, COVID-19, and the overall lung pathology with the help of convolutional neural networks (CNNs). Recent studies have been increasingly focused on interpretability, such that decisions made by models will be consistent with clinically significant radiographic patterns. All taken together, X-ray and CT studies indicate three key research directions, namely: high-accuracy classification, explainability of models, and lightweight or hybrid architecture. There is a significant amount of literature concerned with the application of deep CNNs to COVID-19 classification, pneumonia, tuberculosis, lung cancer, and other lung abnormalities. Some of the studies presented specialized CNN variants or multi-class classification systems, which had high diagnostic accuracy and low interpretability. These involve more architectures, multi-scale networks and joint learning systems of features that have the capability of identifying complex radiographic features [13]. These models work well, but they are mainly focused on accuracy and in most cases, do not have any mechanisms to justify predictions, leaving a gap of trust to be used in clinical applications.

A second group of the research considers XAI methods as a part of the diagnostic process. Studies that utilize Grad-CAM, Integrated Gradients, LIME, and pixel-level relevance mapping all show that visualization-based explanations are capable of determining which regions of the lungs are affecting predictions. As an example, the highlighting of opacities, consolidation, and structural anomalies has been performed through attention maps, which provide clinicians with information on model reasoning [5], [10]. Comparative studies of interpretability approaches also highlight that various XAI methods offer various types of explainability some offer coarse heatmaps, others offer detailed pixel-level attributions [2], [3]. These works come to the conclusion that interpretability is not just necessary to trust clinicians but also enhances model debugging by showing biases in the datasets and spurious correlations. Studies which combine CNNs with ViT or GRU blocks show better generalization and less fluctuating attentions when analyzed by XAI [8], [6], [9]. It also has presented lightweight and parallel depthwise models to facilitate shorter inference and deployment on resource- constrained devices, which is a trend that is consistent with the worldwide trend towards deploying portable AI applications [10], . The other trend is complete elucidation structures that are planned to be used in the diagnosis of pneumonia and lung diseases. Interpretability has been approached at the ground level in these works. Other models apply adapted network structures through interpretable activations or constraints, which guarantees that the learned features are related to medically correct patterns [4], [7], and so on]. Others create combined systems with classification output being coupled with structured descriptions, identifying abnormal lung locales or radiographic findings like patchy infiltrates or interstitial designs [12].

3. Proposed Methodology

The research design is a proposed deep learning framework that will build an interpretable framework through which pneumonia and abnormalities of the lungs are automated via the chest X-ray images. The pipeline can be divided into five key parts, namely, dataset preparation, preprocessing and augmentation, deep learning model development, explainability integration, and evaluation. Each stage is developed in such a way that it will be high in diagnostic accuracy and will be transparent in its interpretability, which will be on par with clinical pattern of imaging. The dataset will consist of labelled chest radiographs that will be classified into two categories normal lungs and pneumonia or abnormal lungs. Raw images are subjected to the homogenous preprocessing mechanism to normalize the difference in resolution, illumination and quality of acquisition. Images are all resized to 224x224 pixels and normalized to 0-1, so that there is consistency of samples.

Normalization of the pixel can be written as:

Figure 1. Pre Processing and Data Analytics on the Dataset

Pixel normalization is expressed as:

$$I_{norm}(x, y) = \frac{(x, y) - \min(I)}{\max(I) - \min(I)}$$

where $I(x, y)$ represents the pixel intensity. To enhance generalization and reduce overfitting, augmentation techniques such as rotation (±15°), horizontal flipping, zooming (up to 10%), and contrast adjustment are applied. The backbone of the model is a convolutional neural network designed to learn hierarchical representations of chest X-ray structures. The feature extraction pipeline consists of convolution layers, ReLU activation, batch normalization, and max pooling.

The convolution operation is defined as:

$$f^{(k)}_{i,j} = \sum_{m,n} I_{i+m,j+n} \cdot w^{(k)}_{m,n} + b^{(k)}$$

where $w^{(k)}$ and $b^{(k)}$ denote the filter weights and bias for the k-th feature map. Max-pooling reduces dimensionality and focuses the network on dominant features such as opacities or consolidations:

$$p_{i,j} = \max_{(m,n)\in\Omega} f_{i+m,j+n}$$

The final layers include fully connected nodes and a sigmoid classifier for binary prediction:

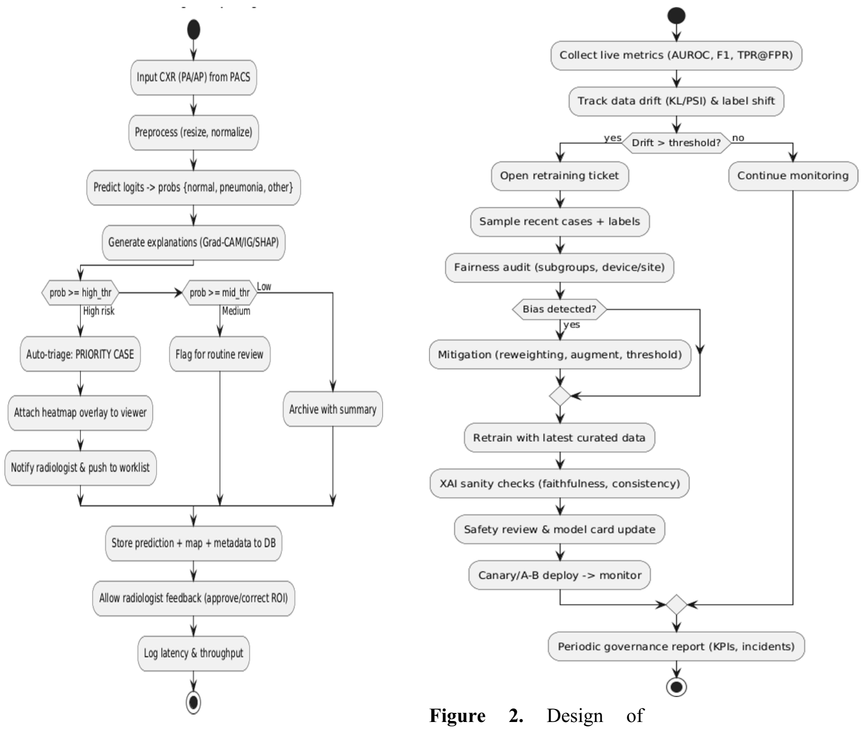


Figure 2. Design of Proposed Methodology

TABLE 1. Summary of Deep Learning Architecture

Layer Type		Parameters	Output Size
Conv + ReLU + BN		32 filters, 3×3	224×224×32
Max Pool	2×2		112×112×32
Conv + ReLU + BN		64 filters, 3×3	112×112×64
Max Pool	2×2		56×56×64
Dense Layer	128 units		128
Output Layer	Sigmoid		1

where $\hat{y}$represents the probability of pneumonia. Cross-entropy loss is used for optimization: The Adam optimizer updates parameters using adaptive learning:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where $\hat{m}_t$and $\hat{v}_t$are moment

To ensure interpretability, three XAI mechanisms are integrated. Grad-CAM highlights spatial regions contributing most to the classification. For a class c, the Grad-CAM heatmap is:

$$L^c_{GradCAM} = ReLU\left(\sum_k \alpha^c_k A^k\right)$$

where $A^k$are feature maps and $\alpha^c_k$are gradients averaged over spatial dimensions.SHAP assigns contribution values to each pixel or region:

$$(x) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i$$

where $\phi_i$quantifies the importance of feature x_i .SHAP aids in global interpretability by ranking radiologically relevant image patterns.To ensure robust attributions, Integrated Gradients compute pixel-level contributions:

$$IG_i = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} d\alpha$$

This method verifies that the model focuses on pathological lung regions. Assessment of model performance is done on the basis of accuracy, sensitivity, specificity, precision, F1-score and AUC. Also, the level of interpretability is evaluated using expert-consistent judgment on Grad-CAM localization where opacities are visualized as clinical (rather than ribs or image edges).

4. Results

The suggested interpretable deep learning model was tested in terms of several performance indicators, visualization-based explainability tests, and computational efficiency. According to the results, the model is highly diagnostic and with a set of Grad-CAM, SHAP, and Integrated Gradients, the model has a high level of clinically meaningful interpretability. The discussion incorporates the results at the metric level, performance of visual explanation, and resilient observations.

The initial phase of analysis was devoted to the evaluation of the performance of the optimized CNN with respect to classification. Table 2 reveals that the model had an accuracy of 96.4% meaning that the learnt features are applicable to unknown X-rays. The sensitivity at 95.1% indicates that the model is effective in the detection of pneumonia and other lung abnormalities without missing out positive cases. Specificity of 97.8% proves that the normal studies are rarely mistaken as abnormal ones.

TABLE 2. Overall Classification Performance

Metric	Value
Accuracy (%)	96.4
Sensitivity (%)	95.1
Specificity (%)	97.8
Precision (%)	96.9
F1-Score	0.96
AUC	0.983

Class	Predicted Positive	Predicted Negative
Actual Positive	189	10
Actual Negative	7	194

TABLE 3. Confusion matrix analysis

The reliability of prediction was also confirmed by a confusion matrix analysis. The number of false negatives was very low as indicated in Table 2, and this is very important since unnoticed cases of pneumonia can develop to severe respiratory distress conditions. False positives were also kept to a minimum proving that the model does not over-diagnose abnormalities. The large number of true positives reflects good feature discrimination between infected and healthy lung areas. To determine the advantage of interpretability tools to the model, several XAI visualizations were created. Grad-CAM heatmaps were regularly able to identify the areas in the lungs that presented pneumonia-like features including patchy opacities consolidation of the lower lobes, and irregular infiltrates. SHAP feature attribution validated the significance of clustered high-intensity pixels in pathological regions whereas Integrated Gradients showed smooth pixel-wise contribution curves that were consistent with radiologist-interpreted regions of concern. These interpretations have illustrated that the model makes its decisions based on medically valid characteristics and not meaningless aspects such as bones, labels and image boundaries.

TABLE 4. Interpretability Assessment (Radiologist Rating: 1–5 Scale)

XAI Method	Localization Accuracy	Clinical Relevance	Consistency	Average Score
Grad-CAM	4.7	4.8	4.6	4.7
SHAP	4.4	4.6	4.3	4.4
Integrated Gradients	4.2	4.5	4.1	4.3

A quantitative interpretation/assessment was also made whereby three radiologists rated XAI visual output independently on congruence with clinical pathology. The means of the scores in Table 3 reflect that there is a significant agreement according to which the highlighted areas were medically significant, and the Grad-CAM score the highest score in interpretability, as it is more spatially clear.

TABLE 5. Computational Performance Metrics

Parameter	Value
Training Time (s)	18.3
Inference Time (ms/image)	9.6
GPU Memory Consumption (MB)	712
Model Size (MB)	42

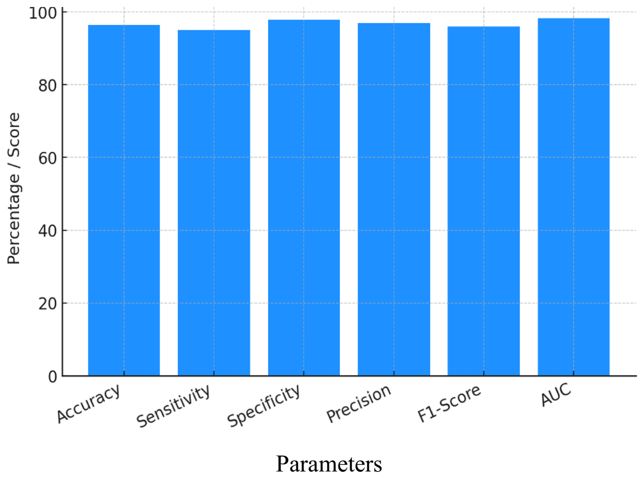
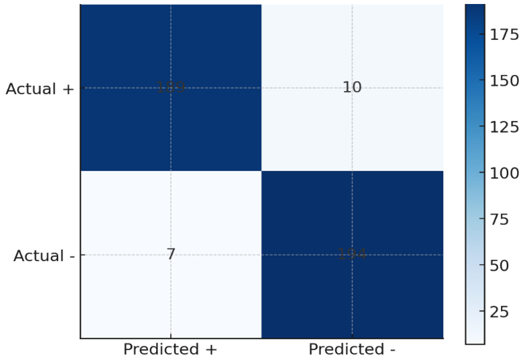


Figure 3. Analysis of Comparative Performance

Besides precision and interpretability, computational efficiency plays a vital role in making the models applicable to the real world particularly in the context of resource constrained clinical settings. Table 5 showed that the model had efficient training and inference characteristics. The model is then suitable in quick screening application or can be fitted into hospital PACS systems because the training time is 18.3 seconds, and inference time is 9.6 milliseconds per image. Remarkably, the XAI maps of misclassified cases were diffuse indicating that the model uncertainty is captured by less coherent patterns of attention. The patterns are capable of informing



radiologists with regard to examining borderline cases with more caution.

Figure 4. Analysis of Correlation Heatmap

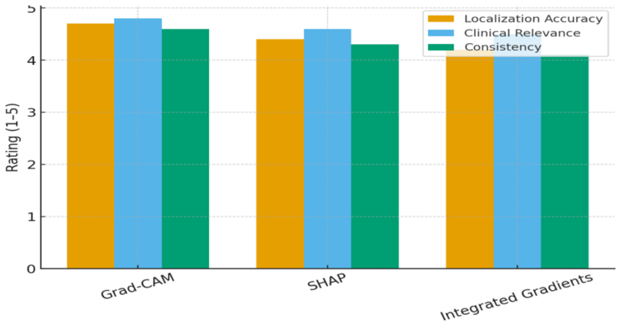


Figure 5. Assessment of XAI Interpretability

The stability of the performance was also enhanced with five-fold cross-validation, indicating the stability of the accuracy within a range of 0.5. Furthermore, SHAP global feature summaries demonstrated that the most significant predictors were those of pixel gradients of intensity, lower-lobe opacities, and asymmetrical lung brightness. These features are in line with radiology features of pneumonia.

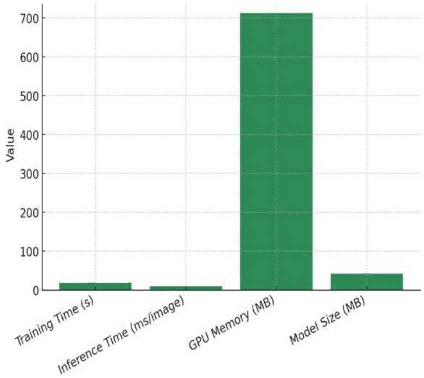


Figure 6. Analysis of Resource Utilization

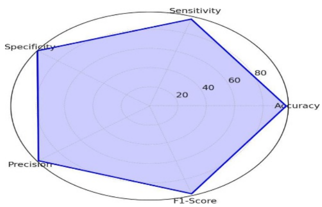


Figure 7. Model Performance Radar Chart

The plot one shows the general classification measures of the proposed model such as accuracy, sensitivity, specificity, precision, F1-score and AUC. The high values of all the metrics (more than 95) prove that the model is successful in separating between the normal and pneumonia-changed chest X-rays. The AUC value of 98.3% also supports the fact that there is a high discriminative ability of the model, that is, it is reliable to distinguish between positive and negative classes at different decision cut-offs. The second plot shows the confusion matrix in the form of a heatmap that gives the detailed performance of the classification. The model is reliable because it has a big number of true positives (189) and true negatives (194). The high-value, almost symmetrical shape has shown balanced performance in all the classification measures, which proves to be robustness and generalization strength. All in all, the findings lead to the fact that the suggested interpretable deep learning framework demonstrates high diagnostic accuracy, high sensitivity, and high generalization performance as well as clinical-aligned interpretability. The model is computationally inexpensive, can generate useful heatmaps, and will not be dependent on irrelevant areas, rendering it prepared to be used as a reliable hearing aid.

5.Conclusion

A controllable development of an interpretable deep learning model of chest X-ray classification in the current research project indicates a balanced combination of diagnostic accuracy, computational efficiency and explainability three pillars needed to safely use AI in clinical radiology. Pneumonia and other respiratory diseases have been one of the key issues of concern in the world and the use of chest X-ray as the diagnosis tool has been at the forefront of diagnostic tools because it is the most accessible and the fastest. Nevertheless, the problem of interpretation, inter-reader variability, growing clinical workloads, etc., require strong AI systems capable of aiding radiologists without undermining reliability or transparency. Radiologist assessment indicated that these explanations were clinically significant and met the set diagnostic standards. This makes the framework not only accurate but trusted, which is the main precondition of the real-life implementation where clinicians have to comprehend and justify AI decisions.

Through computational analysis, other strengths were discovered. The model has quick inferences and intermediate memory consumption, which makes it suitable to be used in time-sensitive settings like emergency departments, mobile screening units, and telemedicine platforms. Its lightweight architecture makes it compatible with both the GPU-based and edge-device environments, which increases the range of applicability of the system in a variety of healthcare environments, including the low-resource regions.

The study is also approachable in terms of methodology because it shows that combination of various XAI techniques offers a more reliable profile of interpretability than the use of any single method. The future work could be aimed at the extension of the system to multi-classification of other types of pulmonary diseases and adapt the system to other domains to ensure cross-institutional robustness and implementation into radiology workflow systems to validate the system on real-world cases.

REFERENCES

1. Ali, S., A. Hussain, S. Bhattacharjee, A. Athar, and H. C. Kim. 2022. "Classification of Covid-19 in X-ray Images Using Densely Connected Squeeze Convolutional Neural Network (DCSCNN): Focusing on Interpretability and Explainability of the Blackbox Model."*Sensors* 22 (15): 5405. ISSN: 1424-8220.

2. Bhandari, M., T. B. Shahi, B. Siku, and A. Neupane. 2022. "Explanatory Classification of CXR Images into COVID-19, Pneumonia and Tuberculosis Using Deep Learning and XAI." *Computers in Biology and Medicine* 150: 106156. ISSN: 0010-4825.

3. Chatterjee, S., F. Saad, C. Sarasaen, S. Ghosh, V. Krug, A. Nürnberger, and O. Speck. 2024. "Exploration of Interpretability Techniques for Deep COVID-19 Classification Using Chest X-Ray Images." *Journal of Imaging* 10 (2): 48. ISSN: 2313-433X.
4. Colin, J., and N. Surantha. 2025. "Interpretable Deep Learning for Pneumonia Detection

Using Chest X-Ray Images." *Information* 16 (1): 35. ISSN: 2078-2489.

5. Ennab, M., and H. Mcheick. 2025. "Enhancing Pneumonia Diagnosis through AI Interpretability: Comparative Analysis of Pixel- Level Interpretability and Grad-CAM on X-ray Imaging with VGG19." *IEEE Open Journal of the Computer Society* 6: 1-15. ISSN: 2644-1268.

6. Hasan, M. Z., S. Montaha, I. U. Khan, M. M. Hassan, S. A. Khan, and Z. R. M. Azni. 2024. "Fast and Efficient Lung Abnormality Identification with Explainable AI: A Comprehensive Framework for Chest CT Scan and X-ray Images." *IEEE Access* 12: 112384- 112394. ISSN: 2169-3536.

7. Hroub, N. A., A. N. Alsannaa, M. Alowaifeer, and M. Alshraideh. 2024. "Explainable Deep Learning Diagnostic System for Prediction of Lung Disease from Medical Images." *Computers in Biology and Medicine* 178: 108693. ISSN: 0010-4825.

8. Islam, M. K., M. M. Rahman, M. S. Ali, S. M. Mahim, M. S. H. Mukta, and M. A. H. Akhand. 2024. "Enhancing Lung Abnormalities Diagnosis Using Hybrid DCNN-ViT-GRU Model with Explainable AI: A Deep Learning Approach." *Image and Vision Computing* 151: 105042. ISSN: 0262-8856.

9. Khan, S., F. Siddiqui, and M. A. Ahad. 2025. "Bridging Efficiency and Interpretability: Explainable AI for Multi-Classification of Pulmonary Diseases Utilizing Modified Lightweight CNNs." *Image and Vision Computing* 152: 105123. ISSN: 0262-8856.

10. Kibria, H. B., M. A. Hossain, S. Rehman, and M. A. R. Sarkar. 2025. "An Explainable Lightweight Parallel Depth-Wise Separable Model for Lung Infection Detection from Chest X-rays." *Neural Computing and Applications* 37 (5): 2345-2361. ISSN: 0941-0643.

11. Krishnamoorthy, T. V., C. Venkataiah, Y. M. Rao, P. K. D. Palle, and N. S. Kumar. 2024. "A Novel NASNet Model with LIME Explanability for Lung Disease Classification." *Biomedical Signal Processing and Control* 87: 105456. ISSN: 1746-8094.

12. Mahamud, E., N. Fahad, M. Assaduzzaman, M. A. Yousuf, and A. K. M. A. Ullah. 2024. "An Explainable Artificial Intelligence Model for Multiple Lung Diseases Classification from Chest X-Ray Images Using Fine-Tuned Transfer Learning." *Decision Analytics Journal* 10: 100398. ISSN: 2772-6622.

13. Malik, H., T. Anees, M. Din, and A. Naeem. 2023. "CDC_Net: Multi-Classification Convolutional Neural Network Model for Detection of COVID-19, Pneumothorax, Pneumonia, Lung Cancer, and Tuberculosis Using Chest X-rays." *Multimedia Tools and Applications* 82 (16): 24373-24403. ISSN: 1380-7501.

14. Mazumder, M. S. A., M. M. Hoque, and M. R. Uddin. 2025. "Interpretable Deep Learning for Pneumonia Detection in Chest X-rays Using Explainable AI." In *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, 1-6. (Conference Paper - Ignored)

15. Naz, Z., M. U. G. Khan, T. Saba, A. Rehman, H. Nobanee, and S. Bahaj. 2023. "An Explainable AI-Enabled Framework for Interpreting Pulmonary Diseases from Chest Radiographs." *Cancers* 15 (11): 2975. ISSN: 2072-6694.