CXR-Next: An Explainable Multi-Class Deep Learning Framework for Thoracic Disease Classification from Chest X-Rays

Md Faisal Hasan[1], Mst Rokshanara Toma[1], Md Ataullha[1,2] *, Sharifur Rahman[1], and M. Shahidur Rahman[2]

[1] Department of Computer Science and Engineering, Green University of Bangladesh, Narayanganj 1461, Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh

{faisalhasan.work, tomasheikh29, ataullha00}@gmail.com, sharifur@cse.green.edu.bd, rahmanms@sust.edu

Abstract. Chest X-ray imaging remains a frontline tool for diagnosing thoracic diseases, yet manual reading is labor-intensive and susceptible to inter-reader variability. This work proposes **CXR-Next**, an explainable deep learning framework built upon a ConvNeXt-Base backbone to perform six-class classification—Normal, Viral Pneumonia, Bacterial Pneumonia, COVID-19, Tuberculosis, and Emphysema—from chest radiographs. We curate an 18,036-image subset ("ChestX6") and apply standardized preprocessing, cross-split de-duplication via MD5 hashing, class-balanced sampling, and data augmentation. Our model achieves **94.99% accuracy**, **95.11 macro F1**, and an **AUC-ROC of 0.98**, outperforming ResNet-50 and EfficientNet baselines by 5–7%. To enhance interpretability, Grad-CAM heatmaps highlight imaging regions that influence class decisions, facilitating clinical review and trust. While results are promising, further validation on larger and more diverse datasets, along with prospective clinical trials, is necessary before deployment. CXR-Next represents a step toward transparent, automated screening in resource-constrained settings.

Keywords: Deep learning, Chest X-ray, ConvNeXt, Explainable AI, Grad-CAM

1 Introduction

Chest radiography is a widely used diagnostic tool for detecting thoracic diseases such as pneumonia, tuberculosis, and COVID-19 [8]. It is inexpensive, fast, and broadly accessible, making it indispensable in both tertiary hospitals and resource-limited settings. However, manual interpretation is time-consuming

* Corresponding author: ataullha00@gmail.com

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_7

and subject to inter-observer variability, and radiologist availability is often constrained in high-volume or under-resourced environments.

Deep learning has significantly advanced chest X-ray (CXR) analysis by enabling models to learn disease-specific patterns directly from large datasets [2]. Convolutional neural networks (CNNs) have achieved performance on par with expert radiologists for several diagnostic tasks, and multi-class classification across multiple thoracic diseases is now practical at scale. However, two major challenges continue to limit clinical deployment: first, a lack of interpretability, as many models behave like *black boxes* that offer little insight into their decision-making; and second, weaknesses in dataset handling and training pipelines, including inconsistent preprocessing, inadequate integrity checks (e.g., cross-split duplicates), and imbalanced sampling—all of which reduce generalizability and hinder fair performance across disease classes.

We address these gaps with **CXR-Next**, an explainable multi-class framework based on a finetuned ConvNeXt-Base backbone [7] and enhanced with Gradient-weighted Class Activation Mapping (Grad-CAM) [10]. The framework employs a curated version of the publicly available ChestX6 dataset, consisting of 18,036 images. Preprocessing is standardized, cross-split duplicates are identified and removed using MD5 hashing, class imbalance is mitigated through weighted sampling, and robustness is improved with data augmentation.

CXR-Next attains **94.99% accuracy**, **95.11% macro F1-score**, and an **AUC-ROC of 0.98**, surpassing ResNet-50 and EfficientNet baselines by 5, and 7% respectively. In addition to strong classification performance, the framework incorporates built-in explainability via Grad-CAM, which produces class-discriminative heatmaps that localize radiographic cues relevant to each prediction. These visualizations provide clinically meaningful support for radiologist review and improve interpretability of model decisions.

In summary, our contributions are as follows:

- We introduce **CXR-Next**, an explainable deep learning framework that unifies state-of-the-art architecture design (ConvNeXt-Base) with rigorous dataset curation and built-in interpretability for thoracic disease classification.
- We curate a reliable six-class ChestX6 dataset with standardized preprocessing, explicit cross-split de-duplication (MD5), class-balanced sampling, and augmentation—establishing a reproducible benchmark-ready resource.
- We design a robust training strategy combining weighted, label-smoothed cross-entropy, staged fine-tuning, and OneCycleLR scheduling, yielding significant performance gains over strong CNN baselines.
- We integrate Grad-CAM explanations directly into the workflow, producing clinically meaningful heatmaps that improve transparency, radiologist trust, and reviewability.

2 Related Work

Many researchers have applied deep learning to chest X-ray (CXR) classification for detecting lung diseases. Architectures such as DenseNet [4] and ConvNeXt [5] have shown promising results. For example, Fujiya et al. (2025) [3] proposed a Double-TL model combining VGG and DenseNet to classify lung masses at radiologist-level performance. However, their system was limited to predicting a single condition and lacked interpretability, which restricts clinical trust.

More recent studies have addressed multi-class classification. Kim et al. (2022) [6] developed an EfficientNet v2-M model with transfer learning, achieving around 80% accuracy for lung disease detection. While efficient, the approach treated classification as a single-step black-box process without interpretability, and the accuracy was limited. Abad et al. (2024) [1] analyzed ResNet50, DenseNet121, and Inception-ResNet-v2 for COVID-19 detection using large datasets and introduced an ensemble method with uncertainty-based weighting, which is resource hungry, and the main focus remained on binary classification rather than multiple thoracic diseases.

Explainability has also emerged as a key concern. Yao et al. (2024) [14] introduced EVA-X, a self-supervised foundation model requiring minimal labeled data, but its evaluation was limited to a single dataset and lacked per-class reporting. Strick et al. (2025) [12] improved CheXNet with Vision Transformers, achieving an average AUC-ROC of 0.85 across 14 disease classes; however, the moderate F1-score (0.39) and signs of overfitting highlighted room for improvement. Uddin et al. (2025) [13] proposed a radiologist-guided few-shot learning model that improved interpretability by focusing on expert-annotated regions, but its reliance on small datasets limited generalizability.

Our approach differs by combining a strong modern backbone (ConvNeXt-Base) with explicit dataset curation and integrated Grad-CAM for six-way thoracic classification. The result—**CXR-Next**—exhibits state-of-the-art performance while providing clinician-aligned explanations.

3 Dataset

We used a curated version of the ChestX6 multi-class X-ray dataset containing 18,036 radiographs across six thoracic conditions: Normal, Viral Pneumonia, Bacterial Pneumonia, COVID-19, Tuberculosis, and Emphysema. The dataset was split into 14,551 training images, 1,748 validation images, and 1,737 test images, as summarized in Table 1 and visualized in Fig. 1. Images were provided in both `.png` and `.jpeg` formats.

To ensure data integrity, we applied MD5 hashing, a cryptographic hash function that generates a unique 128-bit signature for each file. Identical images across different splits yield the same hash, allowing efficient detection of duplicates. Using this method, we identified 17 cross-split duplicates, which were removed from the validation and test sets to eliminate data leakage and ensure unbiased evaluation.

Table 1. Class-wise distribution of the curated ChestX6 dataset

Class	Train	Validation	Test	Total
Normal	2,671	300	300	**3,271**
Pneumonia-Bacterial	2,400	300	300	3,000
Pneumonia-Viral	2,413	300	300	3,013
COVID-19	2,417	300	300	3,017
Tuberculosis	2,600	298	287	3,185
Emphysema	2,050	250	250	**2,550**
Total	14,551	1,748	1,737	18,036

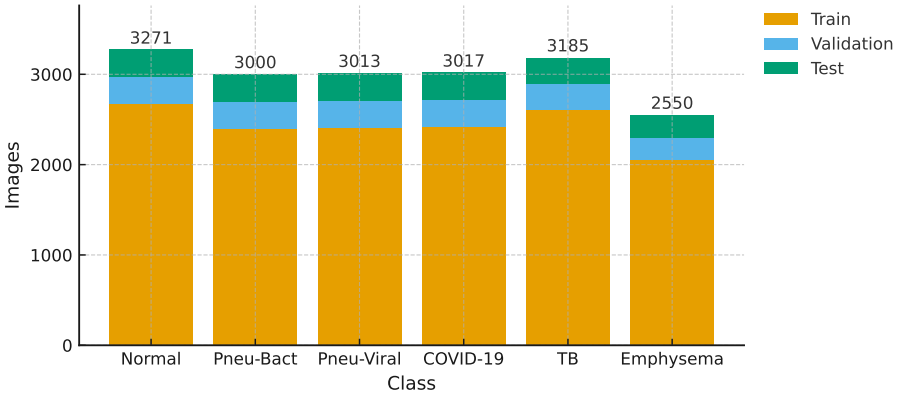

Fig. 1. Class-wise distribution across Train, Validation, and Test splits in the curated ChestX6 dataset (total 18,036 images). Stacked bars show per-split composition; totals are annotated above each bar. Tuberculosis is the largest class (3,185) and Emphysema the smallest (2,550).

4 Methodology

Our methodology is structured as a multi-stage pipeline designed to deliver both high predictive accuracy and model interpretability. As shown in Fig. 2, the framework begins with dataset integrity verification and class imbalance handling, followed by systematic preprocessing and augmentation to ensure robust learning. CXR-Next adopts a custom ConvNeXt-Base model then fine-tuned for multi-class classification across six thoracic conditions. Finally, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to generate visual explanations of predictions, making the model's decisions transparent to clinicians.

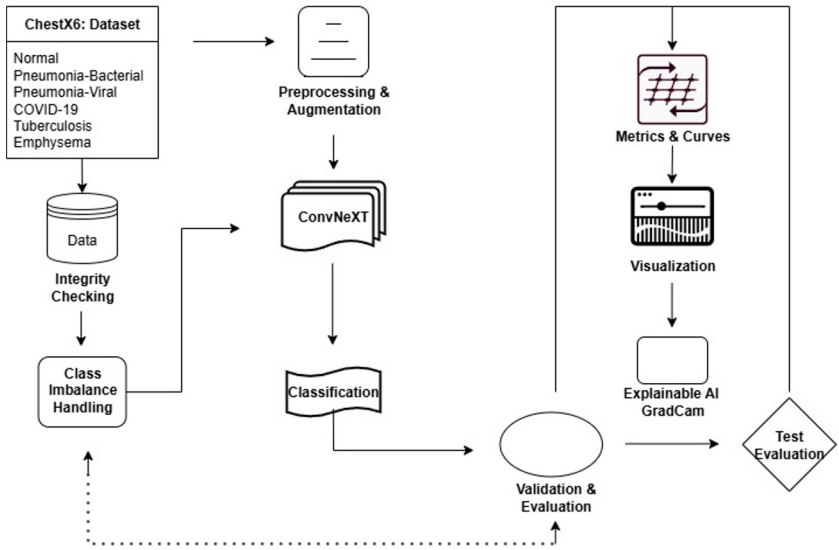

Fig. 2. Workflow of CXR-Next. Steps include dataset integrity checks (de-duplication), class imbalance handling, preprocessing and augmentation, ConvNeXt-based classification, and Grad-CAM visualization for interpretability.

4.1 Data Preprocessing

Prior to training, all radiographs were standardized. Each image was resized to 224×224 pixels, the input resolution of ConvNeXt-Base, converted to three-channel RGB, and normalized using ImageNet mean and standard deviation values.

To improve generalization, we adopted an augmentation pipeline applied only to the training set as mentioned in Table 2, while validation and test images were kept minimally processed for unbiased evaluation. The augmentation strategy included geometric and photometric transformations such as random horizontal flipping, color jitter operations. These augmentations introduce controlled variability in appearance and orientation, enabling the model to learn more robust and invariant representations.

Class imbalance—particularly between frequent categories such as Normal and rarer ones like Emphysema—was mitigated using a *WeightedRandomSampler*. This sampling strategy assigns higher probabilities to underrepresented classes during training, ensuring balanced exposure and preventing bias toward majority categories.

4.2 Model Architecture

CXR-Next adopts *ConvNeXt-Base* [7], a state-of-the-art convolutional neural network (CNN) with 88 million parameters, originally pre-trained on ImageNet-1K [9]. This large-scale pretraining provides a strong initialization for extracting

generic image features, which are then adapted to the domain-specific characteristics of chest radiographs. The overall block structure of ConvNeXt is illustrated in Fig. 3.

Formally, let $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ denote an input chest X-ray. The ConvNeXt-Base backbone yields a feature map

$$\mathbf{h} = f_\theta(\mathbf{x}) \in \mathbb{R}^{C' \times H' \times W'}, \tag{1}$$

where θ are pre-trained weights. A global average pooling (GAP) layer reduces $\mathbf{h}$ to a feature vector

$$\mathbf{z} = \mathrm{GAP}(\mathbf{h}) \in \mathbb{R}^d, \quad d = C'. \tag{2}$$

To adapt the model for thoracic disease classification, we append a fully connected layer parameterized by $W \in \mathbb{R}^{K \times d}$ and $b \in \mathbb{R}^K$, where $K = 6$ is the number of disease classes:

$$\mathbf{o} = W\mathbf{z} + b. \tag{3}$$

The predicted class probabilities are obtained via the softmax function:

$$p(y = k \mid \mathbf{x}) = \frac{\exp(o_k)}{\sum_{j=1}^{K} \exp(o_j)}, \quad k = 1, \ldots, K. \tag{4}$$

To address class imbalance and improve generalization, we optimize a weighted, label-smoothed cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} w_k \, \tilde{y}_{i,k} \, \log p_{i,k}, \tag{5}$$

where w_k is the class weight, and the smoothed target label is

$$\tilde{y}_{i,k} = (1 - \varepsilon) \mathbb{1}[y_i = k] + \frac{\varepsilon}{K}. \tag{6}$$

Training was carried out using staged fine-tuning: the classification head was trained initially with the backbone frozen, followed by full network training. We

Table 2. Data preprocessing and augmentation strategies

Phase	Category	Operation
Training	Geometric	Resize to 224 × 224
	Geometric	RandomHorizontalFlip (50%)
	Photometric	ColorJitter (max ±0.2)
	Photometric	RandAugment (N=2, M=7)
	Normalization	ToTensor [0, 1]
	Normalization	$\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$
Val/Test	Geometric	Resize to 224 × 224
	Normalization	ToTensor [0, 1]
	Normalization	$\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$

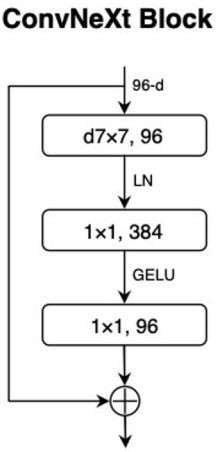

Fig. 3. Illustration of a ConvNeXt block. Each block consists of a depthwise convolution (7×7) followed by layer normalization (LN), a pointwise convolution expanding the feature dimension (1×1, 384), a GELU activation, and another pointwise convolution reducing it back to the original dimension (1×1, 96). The input is added to the block output through a residual connection, enabling stable training and efficient feature propagation.

employed AdamW optimization with weight decay 1×10^{-2}, excluding normalization and bias parameters. The OneCycleLR scheduler [11] was used with a maximum learning rate of 1×10^{-4}, along with mixed-precision training and gradient clipping at 1.0. A weighted random sampler was applied to balance classes across mini-batches (See Table 3). Data preprocessing and augmentation strategies applied during training, validation, and testing are summarized in Table 2.

Table 3. Training hyperparameters.

Parameter	**Value**
Optimizer	AdamW
Learning Rate	1×10^{-5} (base), 1×10^{-4} (peak)
Weight Decay	1×10^{-2}
Scheduler	OneCycleLR
Epochs	15
Batch Size	32
Precision	Mixed (FP16)
Gradient Clip	1.0

4.3 Explainability with Grad-CAM

To make the decision process of our model transparent, we employed Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM generates class-discriminative heatmaps (see Fig. 4) that highlight the regions of an image most influential in driving a prediction. In chest radiographs, this corresponds to localizing anatomical structures or abnormalities that the network associates with specific thoracic conditions.

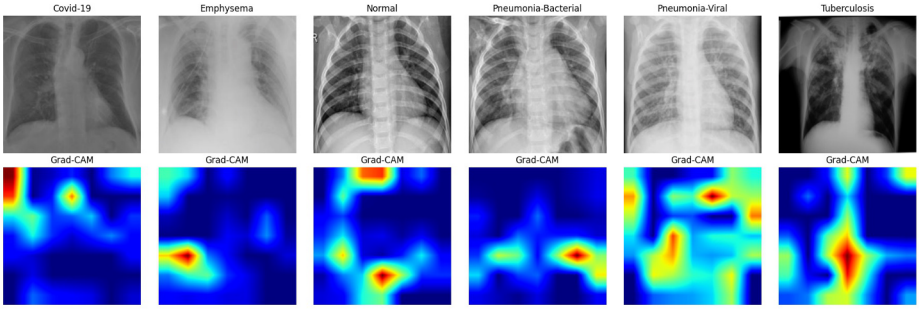

Fig. 4. Grad-CAM visualizations for the six thoracic conditions: original X-rays (top) and corresponding heatmaps (bottom). Red indicates strong model attention.

For example, in pneumonia cases the Grad-CAM heatmaps consistently activated around lung regions containing consolidations, while in tuberculosis cases activations were concentrated in the upper lobes, where cavities and infiltrates are commonly observed. Such behavior indicates that the network is attending to clinically meaningful regions rather than irrelevant background patterns. This improves both the transparency of the system and the confidence of medical professionals in its predictions.

Beyond interpretability, Grad-CAM can function as a clinical support tool. It enables radiologists to quickly verify whether a prediction is consistent with radiological signs, and to challenge or reconsider predictions in borderline cases. Importantly, we applied Grad-CAM to the final convolutional layer of our proposed CXR-Next, which preserves high-level semantic information while maintaining sufficient spatial resolution for localization. This design choice ensures that the resulting heatmaps are both accurate in highlighting diagnostic cues and interpretable for medical validation.

5 Results

We evaluated the proposed framework on a held-out test set of 1,737 chest radiographs. Performance was assessed using accuracy, macro F1-score, area under the ROC curve (AUC-ROC), per-class precision/recall/F1, confusion matrix analysis, and qualitative interpretability with Grad-CAM. These metrics together

provide both a global view of classification ability and class-specific robustness, particularly for rare conditions.

Table 4. Performance on the curated ChestX6 test set.

Model	Acc. (%)	Macro F1 (%)	AUC-ROC
CXR-Next	**94.99**	**95.11**	**0.98**
ResNet-50	88.50	86.30	0.92
EfficientNet	87.67	86.70	0.91

Per-Class Metrics (CXR-Next)

Class	Prec. (%)	Rec. (%)	F1 (%)
Normal	96.4	97.7	97.0
Viral Pneumonia	88.9	88.3	86.0
Bacterial Pneumonia	89.9	85.7	87.7
COVID-19	97.6	95.3	96.5
Tuberculosis	100.0	100.0	100.0
Emphysema	96.8	97.2	97.0

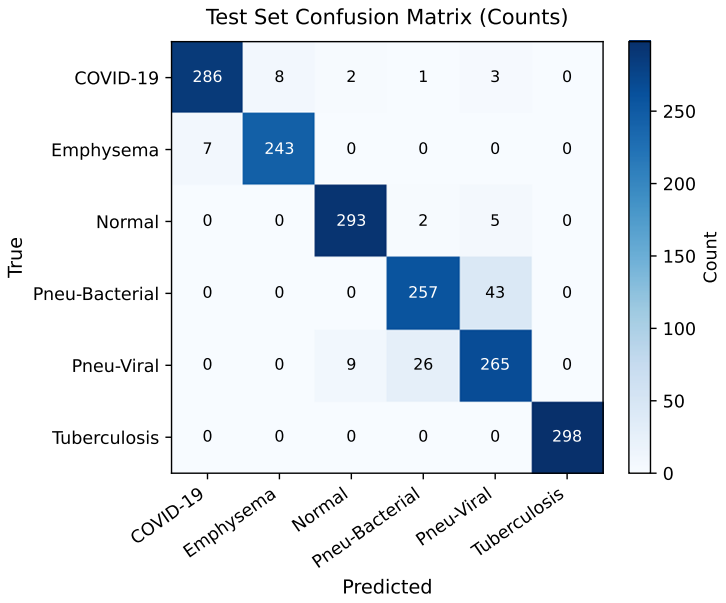

Fig. 5. Confusion matrix on the test set for **CXR-Next**. Most errors occur between Bacterial and Viral Pneumonia; other classes are well separated.

Table 4 compares **CXR-Next** with ResNet-50 and EfficientNet. **CXR-Next** achieves **94.99% accuracy**, **95.11% macro F1**, and a **0.98 AUC-ROC**, consistently outperforming the baselines by 5–7%. The per-class breakdown highlights several observations. Tuberculosis achieves perfect precision, recall, and F1 (100%), reflecting both dataset curation and balanced sampling. Emphysema, despite being the smallest class, also performs robustly (F1 = 97.0%), demonstrating resilience to imbalance. In contrast, the main challenge lies in distinguishing Bacterial and Viral Pneumonia, which exhibit overlapping radiographic patterns and yield modest drops in F1 (87.7% and 86.0%, respectively).

Figure 5 shows the confusion matrix, confirming that most errors are confined to pneumonia subtypes, while other classes remain well separated.

Figure 6 reports per-class ROC curves, all exceeding 0.98 AUC individually, with a macro-average AUC-ROC of 0.98. These results indicate excellent separability across all six thoracic conditions.

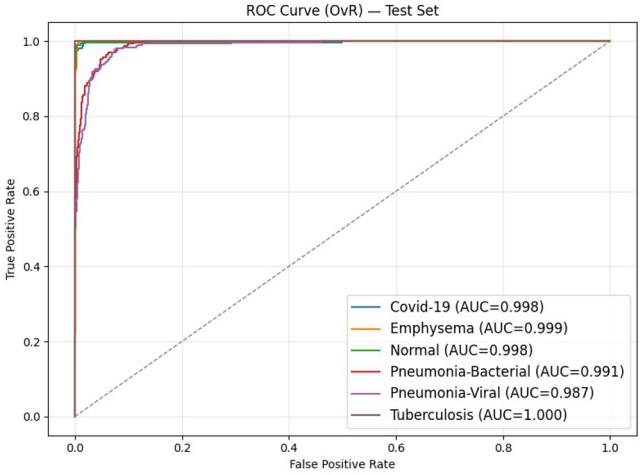

Fig. 6. ROC curves for six-class classification with **CXR-Next**. Macro-average AUC-ROC = 0.98.

Finally, Fig. 4 visualizes Grad-CAM heatmaps for each thoracic condition. The model consistently attends to anatomically meaningful regions: consolidations in pneumonia, upper-lobe involvement in tuberculosis, hyperinflation in emphysema, and ground-glass opacities in COVID-19. Normal scans remain largely inactive. These explanations enhance interpretability, enabling clinician verification and increasing trust in automated predictions.

6 Conclusion and Future Work

We presented **CXR-Next**, a customized ConvNeXt-based framework for explainable six-class thoracic disease classification from chest X-rays. Trained on a curated 18,036-image ChestX6 subset, CXR-Next achieved 94.99% accuracy, 95.11% macro F1, and 0.98 AUC-ROC, outperforming strong baselines by 5–7%. Integrated Grad-CAM heatmaps align with clinical cues, improving transparency and supporting review.

While promising, broader validation is needed to cover diverse demographics and imaging protocols. Future work includes evaluation on larger datasets (e.g., NIH ChestX-ray14), expansion to additional thoracic conditions, and prospective assessment within clinical workflows—steps toward reliable, real-world deployment.

References

1. Abad, M., Casas-Roma, J., Prados, F.: Generalizable disease detection using model ensemble on chest x-ray images. Scientific Reports **14**, 5890 (2024). https://doi.org/10.1038/s41598-024-56171-6
2. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L.: Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. Journal of Big Data **8**(1), 53 (2021)
3. Fujiya, G., Tang, M., Grasnick, A.: Assessment of mass classification from chest x-ray using double-transfer learning: a comparative study with diverse image selection for single-transfer learning. Clinical Radiology **89**, 107023 (2025). https://doi.org/10.1016/j.crad.2025.107023
4. Hage Chehade, A., Abdallah, N., Marion, J.M., Hatt, M., Oueidat, M., Chauvet, P.: Advancing chest x-ray diagnostics: A novel cyclegan-based preprocessing approach for enhanced lung disease classification in chestx-ray14. Computer Methods and Programs in Biomedicine **259**, 108518 (2025). https://doi.org/10.1016/j.cmpb.2024.108518
5. Jin, Y., Lu, H., Zhu, W., Huo, W.: Deep learning based classification of multi-label chest x-ray images via dual-weighted metric loss. Computers in Biology and Medicine **157**, 106683 (2023). https://doi.org/10.1016/j.compbiomed.2023.106683
6. Kim, S., Rim, B., Choi, S., Lee, A., Min, S., Hong, M.: Deep learning in multi-class lung diseases' classification on chest x-ray images. Diagnostics **12**(4), 915 (2022). https://doi.org/10.3390/diagnostics12040915
7. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (2022)
8. McAdams, H.P., Samei, E., Dobbins III, J., Tourassi, G.D., Ravin, C.E.: Recent advances in chest radiography. Radiology **241**(3), 663–683 (2006)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)

10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
11. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
12. Strick, D., Garcia, C., Huang, A.: Reproducing and improving chexnet: Deep learning for chest x-ray disease classification (2025), arXiv preprint arXiv:2505.06646
13. Uddin, I.I., Wang, L., Santosh, K.: Expert-guided explainable few-shot learning for medical image diagnosis (2025), arXiv preprint arXiv:2509.08007
14. Yao, J., Wang, X., Song, Y., Zhao, H., Ma, J., Chen, Y., Liu, W., Wang, B.: Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning (2024), arXiv preprint arXiv:2405.05237