Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Deep Learning for Chest X-Ray Classification: A Comprehensive Review of Methods, Advances, and Clinical Integration

Kawther Samer Ali [a*] **,** ***Ahmad Shaker Abdalrada*** [b]

[a]*College of Computer Science and Information Technology , Wasit university, Iraq, kawthers206@uowasit.edu.com*
[b]*College of Computer Science and Information Technology , Wasit university, Iraq, aabdalra@uowasit.edu.ia*

ARTICLE INFO

Article history:
Received: 22 /07/2025
Revised form: 21 /08/2025
Accepted : 01 /09/2025
Available online: 30 /12/2025

Keywords: Deep Learning, Chest X-Ray (CXR), Pneumonia, COVID-19, Multi-class Classification, EfficientNet, Grad-CAM , Hybrid CNN-Transformer

ABSTRACT

Deep learning transformed medical imaging and enabled the accurate and scalable diagnosis of respiratory infections such as COVID-19 and pneumonia. Unlike traditional radiology's weaknesses owing to overlapping characteristics and observer variability, CNNs are able to identify intricate visual patterns and achieve near-radiologist performance in multi-class CXR classification. Methodological innovations such as transfer learning, data augmentation, attention, and ensemble learning continue to enhance performance, with techniques to enhance interpretability like Grad-CAM advancing clinical trust. With significant advances already achieved, key challenges remain dataset imbalance, domain generalization, and computational cost. Directions for future research include the creation of standardized large-scale datasets, efficient model design for low-resource settings, and the fusion of imaging with clinical metadata. This review highlights recent achievements, current limitations, and potential directions in capitalizing on deep learning innovations into clinically reliable diagnostic tools.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.42537

1. Introduction

Medical imaging has been one of the most significant pillars of medicine for decades as an essential imaging modality for screening, diagnosis, treatment planning, and disease monitoring[1] . Of all the various imaging modalities, chest X-ray (CXR) is most frequently employed due to its low cost, accessibility, and the benefit of a relatively low dose of radiation compared with computed tomography (CT) scans [2]. It is routinely the first investigation of preference for patients with respiratory symptoms, providing valuable information on diseases of the lungs ranging from pneumonia and tuberculosis to (Chronic Obstructive Pulmonary Disease) COPD and, more recently, COVID-19 [3]. Although despite being cheap and readily available, CXR imaging has serious limitations, most significantly in the differential diagnosis between diseases with a similar appearance [2]. Radiologists tend to be unable to distinguish viral from bacterial pneumonia or detect subtle COVID-19 characteristics, especially when large volumes of diagnostic workloads are dealt with and inter-observer variability is a problem [4].

∗Corresponding author : Kawther Sameer Ali

Email addresses: *Kawthers206@uowasit.edu.com*

Communicated by 'sub editor'

With the advent of deep learning, and more specifically artificial intelligence (AI), a revolution in medical image analysis has been triggered [5]. Deep learning models—are basically convolutional neural networks (CNNs)—can learn automatically hierarchical feature representations from raw image data without handcrafted features typical of earlier computer-aided diagnosis systems [6]. This automation facilitates fast and highly scalable diagnosis systems that have the ability to supplement radiologists by providing second opinions, highlighting suspicious regions, and reducing the likelihood of missing something [7]. Most significantly, deep learning algorithms have demonstrated performance levels close to, and in some cases superior to, human experts for many CXR classification tasks , thereby solidifying their roles as revolutionary health care technologies [8].

The last few years have witnessed a surge of research in employing deep learning in CXR classification, propelled significantly by the acute need for accurate detection of COVID-19 in the international pandemic [9]. Beyond COVID-19, research has worked toward comprehensive multi-class classification models for diagnosing more than one respiratory disease simultaneously, e.g., bacterial pneumonia, viral pneumonia, and normal cases [10]. The integration of transfer learning from large-scale natural image databases, application of cutting-edge data augmentation pipelines, and application of sophisticated optimization algorithms have rendered the models significantly more generalizable and robust [11]. Moreover, introduction of attention mechanisms and ensemble learning protocols has added greater interpretability and stability, bridging the gap between black-box algorithms and clinician-interpretability systems [12].

Despite these stupendous developments, a number of challenges remain. Imbalance in the data is a persistent challenge since the availability of well-annotated medical images of certain diseases is limited [13]. Variability in image acquisition across hospitals, patient demographic variability, and the quality of the annotations are additional challenges in model generalizability [14]. The high computational cost of training and deploying deep learning models is also a scaling concern, especially in resource-limited healthcare environments [15]. Ethical considerations—algorithmic bias, data privacy, and regulatory requirements for approval, for example—also complicate the path to real-world clinical adoption [16].

Through these opportunities and challenges, an extensive overview of the state of affairs is crucial. The aim of this paper is to synthesize past knowledge by summarizing the state-of-the-art in deep learning for CXR classification, with a focus on methodological advances, performance milestones, interpretability frameworks, and clinical insights [17]. Through careful examination of trends, challenges, and solution paths, this review offers a roadmap for researchers and practitioners to guide the field toward reliable, transparent, and clinically integrated deep learning systems. Framed thus, the discussion not only identifies the revolutionary potential of AI in medical imaging but also the responsibility to ensure that these systems are safe, fair, and beneficial for global health.

This review (i) synthesizes recent advances in CXR-based deep learning for multi-class diagnosis (COVID-19, bacterial pneumonia, viral pneumonia, and normal) across architectures (CNNs, attention modules, and hybrid CNN–Transformer designs); (ii) proposes a practical taxonomy spanning datasets, preprocessing/augmentation, model families, training/regularization strategies, and interpretability; (iii) critically appraises persistent challenges—class imbalance, domain shift/generalization, and compute/latency constraints—with mitigation options; and (iv) distills actionable guidance for clinical integration and reporting transparency (e.g., patient-level splits, calibrated probabilities, and per-class error analysis). The main contributions and structure of this review are outlined as follows : Section 2 surveys related work and positions this review. Section 3 covers model methodologies (CNNs, transfer learning, augmentation, attention, ensembles, and hybrid CNN–Transformer). Section 4 summarizes major public CXR datasets and their labeling/limitations. Section 5 discusses evaluation metrics and reporting caveats. Section 6 focuses on interpretability (e.g., Grad-CAM and attention-based views). Section 7 analyzes challenges and limitations with mitigation strategies. Section 8 concludes with future directions and clinical translation pathways.

2. Related Work

This article is a narrative review focused on CXR-based deep learning for multi-class thoracic diagnosis. We prioritize breadth and methodological synthesis over quantitative pooling because the literature exhibits substantial heterogeneity in tasks (binary vs multi-class), datasets and splits (often lacking patient-level separation), preprocessing pipelines, and reporting metrics. In the spirit of transparency, we summarize search channels (academic databases and leading preprint servers), representative inclusion criteria (deep-learning works on CXR with explicit task definition and evaluation on public datasets), and exclusion criteria (non-DL or purely CT/brain-MRI studies, duplicates). A PRISMA-style systematic review or meta-analysis is therefore out of scope, but we highlight where standardization could enable future quantitative synthesis.

Deep learning studies in chest X-ray classification have increased exponentially over the past few years, and a heterogeneous body of literature now exists that differs in terms of methodology, datasets, and applications. Early studies were almost exclusively focused on binary classification tasks, i.e., discriminating between normal and pneumonia cases, with relatively shallow CNN models learned from small-sized datasets[18]. While these investigations demonstrated the promise of AI application in radiographic interpretation, they were often hampered by low generalizability and susceptibility to overfitting.

With the publication of big publicly available datasets such as the ChestX-ray14, CheXpert, and COVID-19 Radiography Database, subsequent research gravitated toward more complex multi-class classification schemes. These efforts endeavored to differentiate between viral and bacterial pneumonia, in addition to COVID-19, to develop a more clinically relevant diagnostic tool [19]. As a point of highlight, transfer learning emerged as a dominant strategy, whereby ImageNet pre-trained models were fine-tuned for medical image applications. This approach significantly reduced computational requirements and improved convergence while maintaining competitive accuracy [20].

At the same time, methodological improvements introduced advanced data augmentation pipelines, ensemble methods, and attention mechanisms. For instance, attention modules were incorporated into CNN backbones to improve feature localization and interpretability [21], while ensemble methods have enhanced model robustness and stability [22]. Interpretability techniques, most prominently Gradient-weighted Class Activation Mapping (Grad-CAM), also enabled clinicians to visualize decision-making, facilitating trust in automated predictions [23].

Recent work has also investigated hybrid approaches that combine CNNs and transformers, taking advantage of both local feature extraction and long-range dependency modeling [24]. Such approaches exhibit encouraging performance, particularly in improving generalization on diverse datasets. Researchers have also begun to investigate domain adaptation methods, federated learning frameworks, and weakly supervised approaches to address the limited availability of fully annotated data [25]. Several benchmarks have reported >98% accuracy on within-dataset evaluations (e.g., the COVID-19 Radiography Database), [26] underscoring progress but also highlighting the need for patient-level splits and external validation before drawing broader generalization claims. .

Despite these advances, dataset imbalance, domain adaptation, and computational overhead remain pervasive issues in relevant research **[21]**. These limitations need to be overcome to ensure the transition of deep learning models from experimental setups to real-world clinical application. By the critical review of the existing works, the chapter presents a background context to the following methodological discussion, describing both the progress achieved and the issues still pending in the application of deep learning to chest X-ray analysis. **Table 1***:* compiles representative works selected for methodological diversity (binary vs multi-class; transfer learning; attention/ensembles; hybrid CNN–Transformer) and dataset coverage, to surface strengths and limitations relevant to CXR DL research.. Notably, reported gains often depend on dataset curation and split protocols. Studies using image-level (rather than patient-level) splits and overlapping sources may overestimate generalization. Conversely, works that adopt external validation or cross-institutional testing report more conservative yet realistic performance. Heterogeneous preprocessing (e.g., normalization, segmentation) further complicates head-to-head comparisons across papers.

Table 1: Summary of existing review papers and surveys on deep learning for medical imaging

Study & Year	Scope	Datasets Covered	Key Contributions	Limitations
Siddiqi et al.	Comprehensive review	Multiple small-scale	Reviewed early binary	Limited generalizability;

(2024) [18]	on AI for pneumonia diagnosis from CXR	pneumonia datasets	classification studies, shallow CNNs, and diagnostic challenges	small datasets; overfitting issues
Rajpurkar et al. (2017) [19]	Large-scale classification of chest diseases (CheXNet)	**ChestX-ray14** (112,120 images, 14 diseases)	First demonstration of CNNs at radiologist-level for pneumonia detection	Focused mainly on pneumonia; not multi-class (COVID not included)
Rahman et al. (2020) [20]	Multi-class classification with Transfer Learning	ChestX-ray14, COVID-19 Radiography Database	Demonstrated effectiveness of ImageNet pre-trained CNNs for chest disease classification	Dependence on transfer learning; dataset imbalance issues
Ait Nasser & Akhloufi (2023) [21]	Survey on deep learning methods for chest disease classification	CheXpert, ChestX-ray14, COVID-19 Radiography Database	Discussed augmentation pipelines, attention mechanisms, and interpretability	Lack of clinical validation; models not deployed in real-world
Nahiduzzaman et al. (2023) [22]	Ensemble + hybrid approaches (CNN + Transformer)	COVID-19 Radiography Database, pneumonia datasets	Improved robustness via ensembles; hybrid CNN-ViT architectures	High computational cost; complexity in training
Rajpoot et al. (2024) [23]	Multi-modal deep learning + Grad-CAM interpretability	CXR + CT datasets	Enhanced explainability using Grad-CAM, improved clinician trust	Requires multimodal data; less generalizable
Javed et al. (2024) [25]	Comprehensive survey on pneumonia detection	ChestX-ray14, CheXpert, COVID-19 Radiography Database	Discussed federated learning, weakly supervised methods, domain adaptation	Data imbalance remains unsolved; federated learning not widely tested
Rahman et al. (2022) [26]	Transfer learning in COVID-19 Radiography Database	COVID-19 Radiography Database (15,000+ images)	Achieved >98% accuracy in multi-class classification	Dependent on curated datasets; limited external validation

3. Deep Learning Methodologies for Cxr Classification

3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have emerged as the backbone of medical image analysis due to their ability to learn automatically hierarchical features from chest X-ray (CXR) images **[27].** Unlike traditional hand-engineered methods, CNNs learn raw image spatial representations directly, which enhances stronger detection of complex radiographic patterns such as ground-glass opacities and lung consolidations **[28].** Several studies have established the superiority of CNN-based techniques to differentiate between COVID-19, bacterial pneumonia, viral pneumonia, and normal status with remarkable performances over conventional machine learning models **[29].**

3.2 Transfer Learning Strategies

Transfer learning leverages pre-trained CNN models like VGG, ResNet, and EfficientNet, pre-trained on large image databases like ImageNet and then fine-tuned for the targeted task like CXR classification **[30].** This strategy significantly reduces computational expenses and also combats the lack of annotated medical data . EfficientNet demonstrates strong generalization via compound scaling that jointly and systematically balances network depth, width, and input resolution, rather than compromising any of them , becoming a popular model in recent COVID-19 detection literature **[31].**

3.3 Data Augmentation Methods

Preprocessing choices have inherent trade-offs. For example, converting images to grayscale reduces dimensionality and training cost but may remove subtle radiodensity cues (e.g., vascular textures) that could aid discrimination. Normalization and histogram equalization improve global contrast but sometimes amplify noise. Cropping/lung segmentation can focus learning on clinically relevant areas but risk discarding contextual signs (e.g., pleural effusion). Reporting such trade-offs is critical to guide reproducibility and prevent over-interpretation of reported accuracy. Data augmentation significantly contributes to enhancing model robustness and class imbalance issues associated with medical image data sets . Rotation, flipping, scaling, and contrast alteration artificially expand the data set, allowing the model to learn more robustly **[32]**. More sophisticated augmentation methods such as AugMix and Mixup have been integrated into CXR classification pipelines, providing significant performance improvements in noisy or imbalanced training scenarios **[33]**.

3.4 Attention Mechanisms

Attention mechanisms have emerged as an important extension to CNN models via their capacity to make the model focus on clinically relevant regions on CXR images. Modules such as the Convolutional Block Attention Module (CBAM) and Squeeze-and-Excitation (SE) fortify feature recalibration with greater sensitivity towards regions related to disease [34]. Recently, Transformer-based attention models have been proposed that have been shown to excel at capturing long-distance dependencies and contextual cues in medical images . They not only improve the classification performance but also provide better interpretability of the model by better localizing pathological regions [35].

3.5 Ensemble Learning Techniques

Ensemble learning combines multiple classifiers to produce a more generalized and stronger model [36]. Ensemble techniques in CXR analysis have a tendency to pool predictions from different CNN models or folds of cross-validation and attain improvements in terms of accuracy, sensitivity, and specificity [22]. Model averaging, weighted voting, and stacking were employed for the detection of COVID-19 and pneumonia and performed significantly better compared to single-model baselines [37].

3.6 Hybrid Architectures (CNN + Transformer)

Hybrid architectures are a new development that combines CNNs and Transformer-based models to leverage both local representation of features and global contextual awareness **[24]**. In CXR classification, the low- to mid-level vision is dealt with by CNN layers while Transformer modules detect long-range relations between lung regions **[38]**. Hybrid architectures have been promising in achieving state-of-the-art performance in multi-class CXR classification tasks and a route toward more precise and interpretable diagnostic tools **[39]**.

To provide a clear overview of the end-to-end pipeline, a schematic flowchart is illustrated in **Figure1** , showing the progression from data acquisition to interpretability.

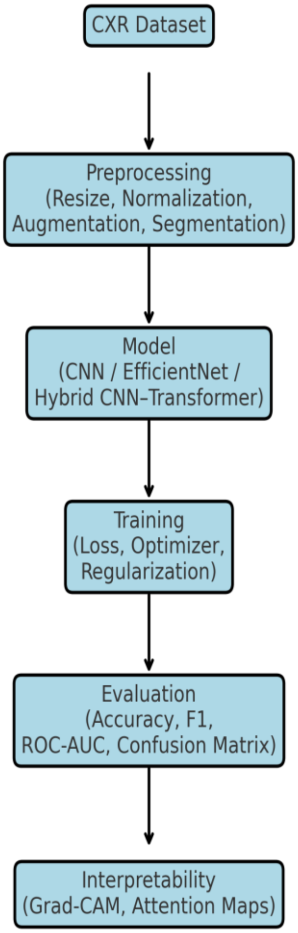

Figure 1. Pipeline flowchart for chest X-ray deep learning studies

4. Datasets for Chest X-Ray Classification

4.1 ChestX-ray14

The ChestX-ray14 dataset, released by the National Institutes of Health (NIH), is a large-scale publicly available CXR dataset, consisting of over 100,000 frontal-view X-ray images from over 30,000 patients . The dataset has 14 disease labels such as pneumonia, cardiomegaly, and effusion, where annotations were acquired through natural language processing of radiology reports . Due to its heterogeneity and size, ChestX-ray14 has served as a benchmark for model training and evaluation of deep learning models in the task of classifying thoracic disease. The dataset, however, suffers from label noise and class imbalance, which must be managed delicately while developing the model **[40]**.

4.2 CheXpert

CheXpert was developed by Stanford University, comprising more than 220,000 chest radiographs of 65,000 patients. It has 14 common thoracic conditions with labels derived from radiology reports by an advanced rule-based labeler . CheXpert is particularly helpful due to its uncertainty labels ("positive," "negative," and "uncertain"), allowing researchers to cope with ambiguity in medical image labeling. It has been used intensively in deep learning studies, with state-of-the-art classification in multi-label CXR **[41]** . One unique aspect of CheXpert is its handling of uncertainty labels, which are annotated as “positive,” “negative,” or “uncertain.” Researchers often adopt strategies such as U-Ignore (excluding uncertain labels), U-Zero (treating uncertain as negative), or U-One (treating uncertain as positive). The choice of strategy can substantially affect both training dynamics and reported performance, underscoring the importance of transparent reporting.

4.3 COVID-19 Radiography Database

The COVID-19 Radiography Database, curated across the pandemic from all over the globe, consists of over 21,000 CXR images categorized under COVID-19, normal, and pneumonia (viral and bacterial) cases **[42]**. The dataset has played a pivotal role in the development and validation of deep learning-based models for COVID-19 detection, often serving as a benchmark standard in related literature. The excellent representation of diverse diagnostic classes in this dataset renders it a cornerstone for multi-class CXR classification tasks [29].

4.4 Other Publicly Released Datasets

Apart from the well-known datasets mentioned above, several smaller but notable datasets have been released to facilitate CXR classification research. These include Montgomery and Shenzhen for detecting tuberculosis, which contain handpicked images with ground-truth annotations **[43]**. Other datasets such as RSNA Pneumonia Detection and BIMCV-COVID19+ also facilitate large-scale benchmarking on multiple diagnostic tasks **[44, 45]**. While smaller in sample size compared to ChestXray14 and CheXpert, both of these datasets are essential to model generalizability estimation across populations and imaging conditions. Table 2: Comparison of major publicly available chest X-ray datasets, including dataset size, number of classes, labeling method, and limitations. To provide a visual understanding of the different diagnostic categories, representative chest X-ray samples from the major publicly available datasets are illustrated in **Figure 2**. These examples highlight typical radiographic patterns observed in COVID-19, viral pneumonia, bacterial pneumonia, and normal cases.

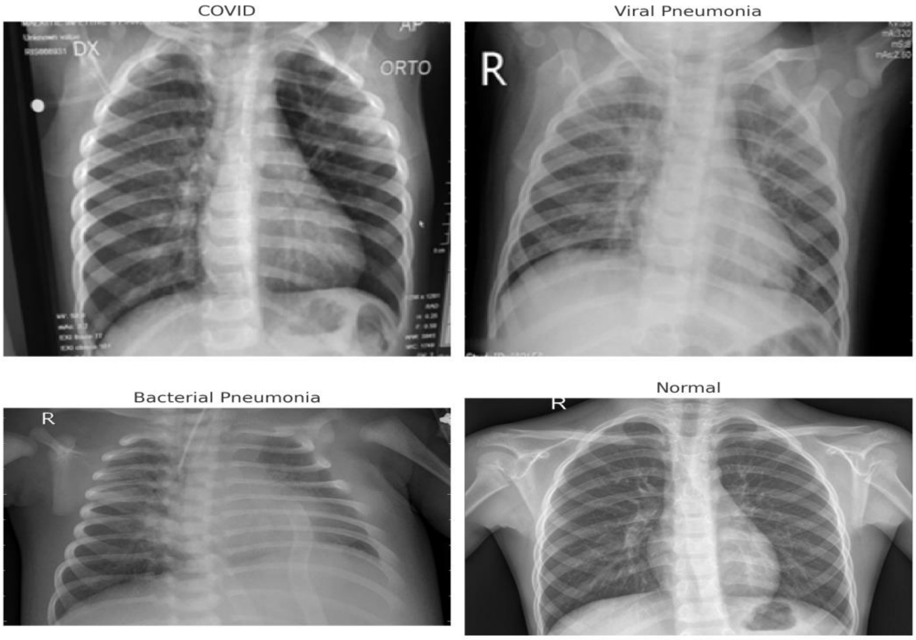

Figure 2. Representative Chest X-Ray Samples from Public Datasets.

Table 2: Comparison of major publicly available chest X-ray datasets

Dataset	Size (# Images / Patients)	Classes (Labels)	Labeling Method	Limitations
ChestX-ray14	112,120 images /	14 thoracic diseases (e.g.,	NLP from radiology	Label noise, class

(NIH)	30,000 patients	pneumonia, effusion, cardiomegaly)	reports	imbalance
CheXpert (Stanford)	224,316 images / 65,000 patients	14 thoracic conditions + Uncertainty labels	Rule-based labeler from radiology reports	Ambiguity in "uncertain" labels
COVID-19 Radiography Database	21,000+ images	3 classes: COVID-19, Normal, Pneumonia (bacterial, viral)	Expert curation & global sources	Smaller scale than NIH/Stanford sets
Montgomery & Shenzhen	~1,000 images (combined)	Tuberculosis (TB) detection	Expert-annotated CXR with masks	Small sample size
RSNA Pneumonia Detection	30,000+ images	Pneumonia (bounding-box annotations)	Radiologist-verified bounding boxes	Limited to pneumonia only
BIMCV-COVID19+	100,000+ images (multiple modalities)	COVID-19, pneumonia, normal + metadata	Hospital-sourced, radiologist reports	Class imbalance, variable quality

5. Performance Metrics and Evaluation

5.1 Accuracy

Accuracy is one of the most commonly reported metrics, representing the proportion of correctly predicted samples over the total. While useful, accuracy alone can be misleading under class imbalance, since a classifier may achieve high accuracy by favoring majority classes [46].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

5.2 Precision

Precision measures the proportion of correctly predicted positive observations to the total predicted positives. It is especially useful in scenarios where the cost of false positives is high [47].

$$Precision = \frac{TP}{(TP + FP)}$$

5.3 Recall (Sensitivity)

Recall, or sensitivity, is the proportion of true positive cases that the model correctly predicts [30]. In healthcare diagnosis, high recall is essential since false negative cases of pneumonia or COVID-19 would have serious health consequences [48].

$$Recall = \frac{TP}{(TP + FN)}$$

5.4 F1-Score

The F1-score is the harmonic mean between precision and recall, averaging both of their values into one number . The measure is particularly used in class imbalance because it provides a better estimate compared to accuracy measurements alone [48].

$$F1\,Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

5.5 ROC AUC

The Receiver Operating Characteristic Curve Area Under the Curve (ROC AUC) quantifies the model to separate positive and negative cases at various classification thresholds . High AUC suggests high class separability and is thus one of the best measures of diagnostic performance for deep learning-based CXR classification [48]. In addition to ROC-AUC, the Precision–Recall Area Under the Curve (PR-AUC) is particularly useful in highly imbalanced datasets such as CXR classification, where the number of normal cases often dominates pathological ones. PR-AUC provides better insight into the trade-off between precision and recall for the minority classes. Furthermore, probability calibration techniques (e.g., Platt scaling or isotonic regression) ensure that the predicted confidence scores align more closely with the true likelihood of disease, which is critical for clinical decision-making.

5.6 Confusion Matrix

The confusion matrix provides an exact breakdown of the classification outcomes, i.e., true positives, false positives, true negatives, and false negatives [49]. It is an effective tool for per-class performance analysis and identifying systematic model prediction errors, e.g., widespread misclassification between bacterial and viral pneumonia .

It is also essential to avoid data leakage when designing experimental protocols. For medical imaging datasets, this often requires ensuring patient-level splitting, such that images from the same patient do not appear in both training and testing sets. Failure to enforce this may lead to overly optimistic results that do not generalize to real-world clinical settings.

6. Interpretability and Explainability in Deep Learning

6.1 Grad-CAM and Visualization Techniques

Interpretability is an important requirement for the use of deep learning models in medical images. Gradient-weighted Class Activation Mapping (Grad-CAM) has been one of the most widely used techniques to visualize the regions of chest X-ray images that contribute the most to the model's decision . By generating heatmaps over the input images, Grad-CAM allows clinicians and researchers to confirm whether the model is paying attention to pathologically meaningful areas of the lung, such as infiltrates or ground-glass opacities . Other visualization techniques, such as saliency maps and Layer-wise Relevance Propagation (LRP), have been explored, although Grad-CAM is by far the most commonly utilized due to its efficacy and simplicity [50]. Beyond Grad-CAM, other explainability methods such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and Integrated Gradients have been explored in CXR classification. These techniques provide complementary perspectives on model decision-making and further enhance clinical trust by allowing radiologists to validate algorithmic focus against medical knowledge. **Figure 3** Grad-CAM visualizations for representative CXR images across four classes: Viral Pneumonia, Normal, Bacterial Pneumonia, and COVID-19. The second column shows enhanced Grad-CAM maps, while the third column overlays heatmaps onto the original images. These visualizations highlight that the model attends to clinically relevant lung regions, enhancing interpretability and trust in diagnostic predictions.

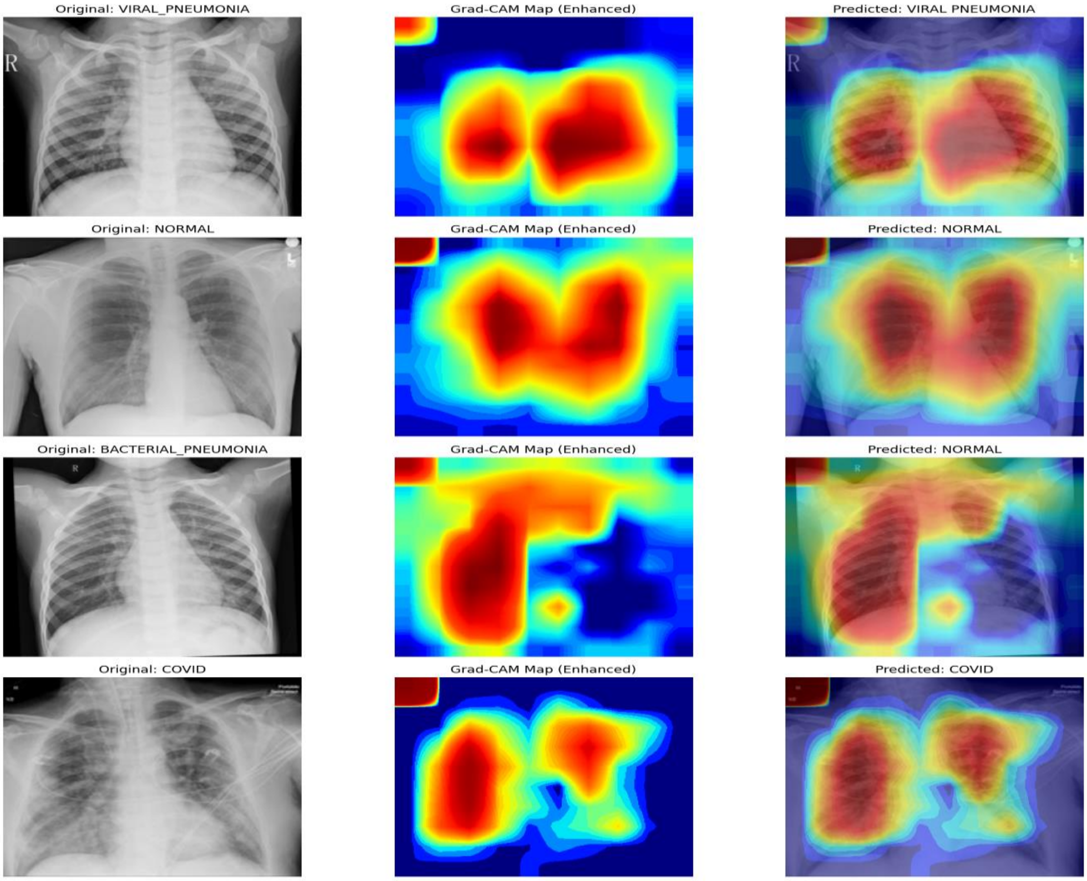

Figure 3. Grad-CAM visualizations for representative CXR images

6.2 Attention-based Interpretability

There have been recent advances in attention mechanisms, like Convolutional Block Attention Module (CBAM) and Squeeze-and-Excitation (SE) blocks, that suggest new approaches to enhancing interpretability . These modules allow the network to selectively prioritize the most informative spatial or channel-wise features, giving insight into which anatomical structures guided the decision-making process. Furthermore, transformer-based architectures inherently generate attention maps, which can highlight global dependencies between lung regions, making them a valuable tool for classification and interpretability [51].

6.3 Clinician Trust and Transparency

For deep learning models to be integrated into clinical workflows, interpretability must result in trust and transparency . Attention mechanisms and visualization tools bridge the gap between black-box AI systems and clinical decision-making by providing intuitive model prediction explanations. It is important to build clinician trust since physicians must ensure whether model attention aligns with medical knowledge before the clinical adoption of AI-supported diagnostic tools. Lastly, explainability not only promotes accountability and ethical AI research but also strengthens the possibilities for effective deployment in real healthcare systems [52].

7. Challenges and Limitations

Given the sensitive nature of medical imaging, privacy and ethical considerations are critical. Data de-identification, patient consent, and institutional data-sharing agreements are vital to ensure compliance with regulations such as HIPAA and GDPR. Furthermore, dataset biases—such as unequal representation across age, gender, or ethnicity—must be acknowledged to prevent inequitable performance in real-world clinical practice.

7.1 Dataset Imbalance

One of the most persistent challenges in chest X-ray (CXR) classification is dataset imbalance, where certain classes (e.g., normal cases) significantly outnumber others (e.g., COVID-19 or rare pneumonia types). This imbalance often biases models toward the majority class, reducing sensitivity in detecting minority conditions [53]. Various solutions, such as oversampling, synthetic data generation using GANs, and class-weighted loss functions, have been proposed, but imbalance remains a critical limitation in real-world applications [54].

7.2 Domain Adaptation and Generalization

Deep learning models frequently suffer from performance degradation when tested on external datasets collected from different hospitals, imaging devices, or patient populations [55]. This lack of generalization highlights the domain adaptation problem, where models overfit to the training distribution but fail to transfer effectively to unseen clinical environments [56]. Approaches such as transfer learning, domain adversarial training, and federated learning are being explored to address this issue, but achieving robust generalization remains an open challenge [57].

7.3 Computational Costs and Scalability

Training state-of-the-art CNNs, transformers, or hybrid architectures on large-scale medical datasets requires significant computational resources, including high-end GPUs and large memory capacities [58]. Such requirements pose barriers for deployment in resource-limited healthcare systems, especially in developing regions. Scalability issues also emerge when attempting to process vast repositories of CXR images in real time, necessitating research into lightweight architectures and efficient inference techniques [59].

7.4 Ethical and Regulatory Considerations

Beyond technical limitations, ethical and regulatory challenges play a major role in the adoption of AI-based diagnostic tools. Issues of patient privacy, informed consent, and compliance with regulations such as HIPAA and GDPR must be addressed [60]. Furthermore, the black-box nature of many deep learning models raises accountability concerns in case of diagnostic errors. Regulatory approval processes demand rigorous validation and explainability to ensure safe integration into clinical practice, making this one of the most complex challenges in medical AI [61]. **Table 3**: Summary of key challenges and limitations in deep learning for CXR classification, with potential mitigation strategies. **Table 4:** Comparison of emerging methodologies (Federated Learning, Semi-supervised Learning, Vision Transformers, Self-supervised Learning) and their potential for CXR classification.

Table 3: Summary of key challenges and limitations in deep learning for CXR classification

Challenge	Description	Potential Mitigation Strategies
Dataset Imbalance	Certain diseases underrepresented	Data augmentation, re-sampling, synthetic data
Domain Adaptation	Models fail to generalize across institutions	Transfer learning, domain adaptation techniques
Computational Costs	High training and inference requirements	Model compression, efficient architectures
Ethical Issues	Bias, fairness, privacy concerns	Federated learning, explainability, regulatory frameworks

Table 4: Comparison of emerging methodologies

Methodology	Description	Potential for CXR Classification
Federated Learning	Training across decentralized data without sharing raw data	Preserves privacy, multi-institutional learning
Semi-supervised Learning	Uses small labeled + large unlabeled datasets	Reduces annotation costs, improves generalization
Vision Transformers	Attention-based architectures	Capture long-range dependencies, promising accuracy
Self-supervised Learning	Learns representations from unlabeled data	Effective in limited label scenarios

8. Conclusion and Future Perspectives

Deep learning has revolutionized the field of chest X-ray (CXR) analysis by offering powerful tools for automated disease detection, classification, and clinical decision support. Over the past decade, models such as convolutional neural networks, transfer learning frameworks, attention-enhanced architectures, and hybrid CNN-transformer approaches have demonstrated state-of-the-art performance in identifying conditions like COVID-19, bacterial pneumonia, viral pneumonia, and other thoracic diseases. The integration of interpretability techniques such as Grad-CAM and attention-based mechanisms has further strengthened the potential of these systems by enabling clinicians to better understand and trust model predictions.

Despite these advancements, significant challenges remain. Dataset imbalance, limited generalizability across domains, and high computational costs hinder the robustness and scalability of deep learning models in real-world clinical settings. Moreover, ethical and regulatory considerations continue to shape the trajectory of AI in healthcare, as transparency, accountability, and patient privacy are essential for clinical adoption.

Looking forward, the future of deep learning in CXR analysis will likely be defined by several key directions. First, larger and more diverse datasets, possibly enabled through global collaborations and federated learning frameworks, will improve model generalization across heterogeneous populations. Second, lightweight and efficient model architectures will make AI solutions more accessible in resource-limited environments, expanding their global impact. Third, the integration of multimodal data—such as combining radiographic findings with clinical, laboratory, or genomic information—may enable more comprehensive and precise diagnostics. Finally, explainable AI and human–AI collaboration will be crucial in building trust and ensuring the ethical use of deep learning in healthcare.

In conclusion, while challenges persist, the trajectory of research and innovation suggests that deep learning will continue to transform chest X-ray interpretation and clinical diagnostics. By addressing current limitations and embracing interdisciplinary collaboration, the field can move closer to achieving reliable, transparent, and equitable AI-driven healthcare solutions that benefit patients worldwide.

References

[1] B. Abhisheka, S. K. Biswas, B. Purkayastha, D. Das, and A. Escargueil, "Recent trend in medical imaging modalities and their applications in disease diagnosis: a review," *Multimedia Tools and Applications,* vol. 83, no. 14, pp. 43035-43070, 2024.

[2] X. Ou *et al.*, "Recent development in x-ray imaging technology: Future and challenges," *Research*, 2021.

[3] A. T. Senno and R. K. Brannon, "Respiratory Diseases: Asthma, Pneumonia, Influenza, Tuberculosis, and COVID-19," in *Maternal-Fetal Evidence Based Guidelines*: CRC Press, 2022, pp. 269-296.

[4] D. L. Pezzutti, V. Wadhwa, and M. S. Makary, "COVID-19 imaging: Diagnostic approaches, challenges, and evolving advances," *World Journal of Radiology,* vol. 13, no. 6, p. 171, 2021.

[5] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval,* vol. 11, no. 1, pp. 19-38, 2022.

[6] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging,* vol. 35, no. 5, pp. 1285-1298, 2016.

[7] K. Chircop, "Predicted clinical impact of artificial intelligence in radiology: a rapid evidence assessment," 2025.

[8] C. Banapuram, A. C. Naik, M. K. Vanteru, V. S. Kumar, and K. K. Vaigandla, "A comprehensive survey of machine learning in healthcare: Predicting heart and liver disease, tuberculosis detection in chest X-ray images," *SSRG International Journal of Electronics and Communication Engineering,* vol. 11, no. 5, pp. 155-169, 2024.

[9] B. Terenzio, "Deep Learning Approaches for COVID-19 Detection: A Brief Review of Datasets, Techniques, and Challenges," 2024.

[10] A. P. Shinde, "Multiclass classification of Covid-19, Tb, Pneumonia, and health cases using Deep Learning," Dublin, National College of Ireland, 2022.

[11] P. Yadav *et al.*, "Investigation and empirical analysis of transfer learning for industrial IoT networks," *IEEE Access,* 2024.

[12] T. Shaik, X. Tao, H. Xie, L. Li, N. Higgins, and J. D. Velásquez, "Towards Transparent Deep Learning in Medicine: Feature Contribution and Attention Mechanism-Based Explainability," *Human-Centric Intelligent Systems,* pp. 1-21, 2025.

[13] R. Birjais, "Challenges and Future Directions for Segmentation of Medical Images Using Deep Learning Models," *Deep Learning Applications in Medical Image Segmentation: Overview, Approaches, and Challenges,* pp. 243-264, 2025.

[14] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports,* vol. 10, no. 1, p. 13724, 2020.

[15] C. Chen *et al.*, "Deep learning on computational-resource-limited platforms: A survey," *Mobile Information Systems,* vol. 2020, no. 1, p. 8454327, 2020.

[16] P. Goktas and A. Grzybowski, "Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI," *Journal of Clinical Medicine,* vol. 14, no. 5, p. 1605, 2025.

[17] N. R. Council, G. Affairs, O. o. I. Affairs, C. o. Science, and H. A. o. t. F. P. A. o. t. U. States, "The pervasive role of science, technology, and health in foreign policy: Imperatives for the department of state," 1999.

[18] Siddiqi and et al., "Artificial intelligence for pneumonia diagnosis from chest X-rays: a review," *Journal of Imaging,* 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11355845.

[19] P. Rajpurkar *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225,* 2017.

[20] T. Rahman and et al., "Transfer Learning with Deep CNNs for Multi-Class Classification of Chest Diseases," 2020. [Online]. Available: https://arxiv.org/abs/2004.06578. [Online]. Available: https://arxiv.org/abs/2004.06578

[21] Y. Ait Nasser and M. A. Akhloufi, "Chest diseases classification from X-ray images: a comprehensive review," *Applied Sciences,* 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9818166.

[22] J. A. Prakash, V. Ravi, V. Sowmya, and K. Soman, "Stacked ensemble learning based on deep convolutional neural networks for pediatric pneumonia diagnosis using chest X-ray images," *Neural Computing and Applications,* vol. 35, no. 11, pp. 8259-8279, 2023.

[23] K. Rajpoot and et al., "A multi-modal deep learning system with Grad-CAM visualization for lung disease detection," *Scientific Reports,* 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-75915-y.

[24] A. Khan *et al.*, "A survey of the vision transformers and their CNN-transformer based variants," *Artificial Intelligence Review,* vol. 56, no. Suppl 3, pp. 2917-2970, 2023.

[25] Javed and et al., "Deep Learning for Pneumonia Detection in Chest X-ray Images: A Comprehensive Survey," *Journal of Imaging (Survey, ResearchGate Preprint),* 2024. [Online]. Available: https://www.researchgate.net/publication/382520052.

[26] T. Rahman and et al., "Improved multi-class classification of chest X-rays with transfer learning on COVID-19 Radiography Database," *Journal of Imaging,* 2022. [Online]. Available: https://www.mdpi.com/2313-433X/10/10/250.

[27] S. S. Kshatri and D. Singh, "Convolutional neural network in medical image analysis: a review," *Archives of Computational Methods in Engineering,* vol. 30, no. 4, pp. 2793-2810, 2023.

[28] R. S. A. Al-bayatı, "Detection and Measurement of Multilevel COVID-19 Infection Using Gamma Correction and Features Extracted by CNN Enhanced with Xgboost from CT Scan Images," Kirsehir Ahi Evran University (Turkey), 2024.

[29] K. Kansal, T. B. Chandra, and A. Singh, "Advancing differential diagnosis: a comprehensive review of deep learning approaches for differentiating tuberculosis, pneumonia, and COVID-19," *Multimedia Tools and Applications,* vol. 84, no. 13, pp. 11871-11906, 2025.

[30] D. Mallick, A. Singh, E. Y.-K. Ng, and V. Arora, "Classifying chest x-rays for COVID-19 through transfer learning: a systematic review," *Multimedia Tools and Applications,* vol. 84, no. 2, pp. 689-748, 2025.

[31] M. A. Talukder, M. A. Layek, M. Kazi, M. A. Uddin, and S. Aryal, "Empowering covid-19 detection: Optimizing performance through fine-tuned efficientnet deep learning architecture," *Computers in Biology and Medicine,* vol. 168, p. 107789, 2024.

[32] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial intelligence review,* vol. 56, no. 11, pp. 12561-12605, 2023.

[33] X. L. Lan, "Traditional Augmentation Versus Deep Generative Diffusion Augmentation for Addressing Class Imbalance in Chest X-ray Classification," 2023.

[34] A. Laouarem, "Toward advanced deep learning techniques for medical Image analysis," 2024.

[35] A. Khan *et al.*, "A recent survey of vision transformers for medical image segmentation," *arXiv preprint arXiv:2312.00634,* 2023.

[36] M. Sewell, "Ensemble learning," *RN,* vol. 11, no. 02, pp. 1-34, 2008.

[37] R. Sadoon and A. Chaid, "Classification of pulmonary diseases using a deep learning stacking ensemble model," *Informatica,* vol. 48, no. 14, 2024.

[38] R. D. Bhosale and D. M. Yadav, "Multiple Pulmonary Disease Detection from CXR Images with Hybrid ViT and DTCWT Architecture," *International Journal of Computing,* vol. 18, no. 1, pp. 1-19, 2025.

[39] G. M. M. Alshmrani, Q. Ni, R. Jiang, H. Pervaiz, and N. M. Elshennawy, "A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images," *Alexandria Engineering Journal,* vol. 64, pp. 923-935, 2023.

[40] N. I. H. C. Center. *ChestX-ray14 Dataset*. [Online]. Available: https://nihcc.app.box.com/v/ChestXray-NIHCC

[41] G. Stanford Machine Learning. *CheXpert Dataset*. [Online]. Available: https://stanfordmlgroup.github.io/competitions/chexpert/

[42] T. Rahman. *COVID-19 Radiography Database*. [Online]. Available: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

[43] S. Jaeger, S. Candemir, S. Antani, Y. Wang, X. Lu, and G. Thoma, "Montgomery County and Shenzhen Chest X-ray Datasets for Tuberculosis Detection," *IEEE Transactions on Medical Imaging,* vol. 34, no. 1, pp. 108-119, 2014, doi: 10.1109/TMI.2014.2341155.

[44] G. Shih and et al., "RSNA Pneumonia Detection Challenge Dataset," *Radiological Society of North America (RSNA) Machine Learning Challenge,* 2019.

[45] M. de la Iglesia Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, and et al., "BIMCV-COVID19+: A large annotated dataset of COVID-19 chest X-ray images," *arXiv preprint,* vol. arXiv:2006.01174, 2020. [Online]. Available: https://arxiv.org/abs/2006.01174.

[46] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of big data,* vol. 6, no. 1, pp. 1-54, 2019.

[47] P. Galdi and R. Tagliaferri, "Data mining: accuracy and error measures for classification and prediction," *Encyclopedia of bioinformatics and computational biology,* vol. 1, pp. 431-436, 2018.

[48] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061,* 2020.

[49] J. T. Townsend, "Theoretical analysis of an alphabetic confusion matrix," *Perception & Psychophysics,* vol. 9, no. 1, pp. 40-50, 1971.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision,* vol. 128, no. 2, pp. 336-359, 2020, doi: 10.1007/s11263-019-01228-7.

[51] M. Rigotti, C. Miksovic, I. Giurgiu, T. Gschwind, and P. Scotton, "Attention-based interpretability with concept transformers," in *International conference on learning representations*, 2021.

[52] A. C. Bontempo, "Patient attitudes toward clinicians' communication of diagnostic uncertainty and its impact on patient trust," *SSM-Qualitative Research in Health,* vol. 3, p. 100214, 2023.

[53] F. Alshanketi *et al.*, "Pneumonia detection from chest X-Ray images using deep learning and transfer learning for imbalanced datasets," *Journal of Imaging Informatics in Medicine,* pp. 1-20, 2024.

[54] M. Altalhan, A. Algarni, and M. T.-H. Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access,* 2025.

[55] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health,* vol. 1, no. 6, pp. e271-e297, 2019.

[56] G. Sarafraz, A. Behnamnia, M. Hosseinzadeh, A. Balapour, A. Meghrazi, and H. R. Rabiee, "Domain adaptation and generalization of functional medical data: A systematic survey of brain data," *ACM Computing Surveys,* vol. 56, no. 10, pp. 1-39, 2024.

[57] A. T. Nguyen, P. Torr, and S. N. Lim, "Fedsr: A simple and effective domain generalization method for federated learning," *Advances in Neural Information Processing Systems,* vol. 35, pp. 38831-38843, 2022.

[58] K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani, "Neural architecture search for transformers: A survey," *IEEE Access,* vol. 10, pp. 108374-108412, 2022.

[59] B. Akdemir *et al.*, "From Technical Prerequisites to Improved Care: Distributed Edge AI for Tomographic Imaging," *IEEE Access,* 2025.

[60] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, "Ethical and regulatory challenges of AI technologies in healthcare: A narrative review," *Heliyon,* vol. 10, no. 4, 2024.

[61] A. Marey *et al.*, "Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology," *Egyptian Journal of Radiology and Nuclear Medicine,* vol. 55, no. 1, p. 183, 2024.