Pneumonia Detection From Chest X-Rays Using Deep Learning : A Comprehensive Review

Febin Cheriyan
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kottayam, India
febincheriyan1@gmail.com

Deni Tom Jacob
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kottayam, India
denitomjacob@gmail.com

Joanna Daniel
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kottayam, India
joannadanm@gmail.com

Haby S Mathews
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kottayam, India
habymathews@gmail.com

Honey Joseph
Assistant Professor
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kottayam, India

Abstract*—Pneumonia is a major global cause of morbidity and mortality, particularly affecting young children and the elderly, and early and accurate detection remains essential to reduce fatalities and optimize resource allocation in clinical settings [1]. Manual chest X-ray interpretation is commonly used but suffers from inter-observer variability, diagnostic delays and lack of availability of expert radiologists in many regions [2]. Advances in artificial intelligence and deep learning have enabled automated, reproducible, and rapid analysis of chest radiographs using convolutional neural networks (CNNs), transfer learning, vision transformers, and hybrid architectures, often achieving radiologist-level performance on curated benchmarks [3], [4]. In this work we present a comprehensive, experimentally validated pipeline for pneumonia detection that integrates a custom CNN trained from scratch with multiple transfer-learning backbones (VGG, ResNet, DenseNet, Inception, EfficientNet), ensemble strategies, vision transformer variants (ViT, Swin, hybrid CNN+ViT), and feature-extraction + classical classifier baselines (CNN→SVM, RF, KNN). The pipeline emphasizes clinical priorities by optimizing sensitivity/recall, calibrating predicted probabilities, and quantifying predictive uncertainty for triage applications. We describe in detail the preprocessing, augmentation, loss functions (binary cross-entropy and focal loss), regularization, optimization, and interpretability with Grad-CAM, and provide extensive comparisons on public chest X-ray benchmarks (RSNA, NIH ChestX-ray8, Kaggle Pneumonia) plus external holdouts. Results demonstrate that well-regularized custom CNNs and ensembles achieve high sensitivity with competitive overall AUC, while hybrid and transformer models offer gains when sufficient data or transfer pretraining is available [5], [6]. We conclude by describing system deployment considerations, limitations, and prioritized future directions such as federated learning, explainable AI (XAI), multi-disease detection, and lightweight models for edge inference.

***Index Terms*—Pneumonia detection, chest X-ray, convolutional neural network, transfer learning, ResNet, DenseNet, EfficientNet, Vision Transformer, Grad-CAM, focal loss, ensemble learning, explainable AI.**

I. Introduction

Pneumonia is an acute infection of the lung parenchyma characterized by alveolar inflammation and consolidation which, left untreated, can rapidly progress to severe respiratory failure and death. The global burden of pneumonia remains substantial; the World Health Organization estimates hundreds of thousands of deaths annually in children under five and a large share of hospital admissions in adults with comorbidities [1]. The frontline diagnostic modality in many healthcare systems is the chest X-ray because of its wide availability and relatively low cost. Despite this, manual X-ray interpretation is labor intensive, prone to inter-observer variability, and subject to misclassification particularly in early or atypical presentations [2]. Radiologist shortages and reporting delays are acute in low-resource settings, creating an urgent need for automated, reliable, and fast triage systems that can flag likely pneumonia cases for priority review or immediate action.

Artificial intelligence, and in particular deep learning, addresses these challenges by learning hierarchical image features directly from pixel data and providing probability-based predictions that can be calibrated and thresholded for clinical use. Convolutional neural networks (CNNs) extract local structures such as edges and textures in early layers and progressively learn higher-level patterns like consolida-

DOI: 10.5281/zenodo.19862283

tions and air bronchograms in deeper layers; transfer learning enables reuse of general visual features learned from large natural image datasets, accelerating convergence and often improving generalization on smaller medical datasets [3]. Recent developments in architecture design, optimization, loss functions tuned for class imbalance, and methods for model explanation and uncertainty estimation collectively allow construction of systems that are more than accuracy metrics — they can provide clinically meaningful scores, attention maps, and uncertainty estimates to guide human decision making [7]. This paper lays out an end-to-end methodology for pneumonia detection, comparing classical ML, CNNs, transfer learning, transformers, ensembles, and hybrid models, and situating choices around the clinical imperative to minimize missed cases (maximize sensitivity) while keeping false positives manageable.

This paper provides a detailed review of deep learning approaches for pneumonia detection from CXR images. It synthesizes findings from multiple studies to present a holistic view of the field, from foundational CNN models to advanced Vision Transformers and hybrid architectures. We discuss key techniques such as transfer learning, data augmentation, and ensemble methods, and analyze their impact on model performance and reliability.

II. Literature Review and Background

A. Evolution of pneumonia diagnosis methods

Automated pneumonia detection evolved from handcrafted radiomic approaches using classical classifiers to end-to-end deep learning. Early ML relied on intensity histograms, texture filters, and morphological measures coupled with SVM or Random Forest classifiers, which performed adequately within limited domains but lacked generality across variable imaging conditions. The arrival of convolutional neural networks shifted the paradigm to automatic hierarchical feature extraction and end-to-end optimization, yielding substantial performance gains when data are sufficient or when transfer learning is applied [8]. Landmark efforts demonstrated that deep models could reach or approach radiologist-level performance in curated benchmarks, stimulating extensive follow-on work.

B. Early ML vs. modern DL

Classical machine learning approaches required domain experts to craft features that capture radiological findings. These systems tended to be brittle under distribution shift because the hand-designed features did not generalize to variations in acquisition or patient anatomy. Deep learning replaced manual feature engineering with learned representations that capture complex spatial context and variable presentation, enabling models to detect nuanced imaging cues. Architectures evolved from shallow CNNs to deep residual and dense networks that facilitate deeper, more expressive models while managing optimization stability [6].

C. Transfer learning and CNN breakthroughs

Transfer learning and architectural innovations like residual connections (ResNet) and dense connectivity (DenseNet) overcame training difficulties and allowed deep models to be effective with relatively modest medical datasets by leveraging features pretrained on ImageNet. Models such as ResNet-50, DenseNet-121, Inception and EfficientNet have become standard backbones for medical image tasks, often with minor architecture adjustments and domain-specific fine-tuning. Recent additions, including Vision Transformers and hybrid CNN+Transformer designs, allow global context modeling across patches, beneficial for certain radiographic patterns, though they require careful pretraining or strong augmentation to match CNNs on small datasets [9].

III. Methodological Approaches

A. Custom CNN models

The primary custom model is a parameter-efficient convolutional network designed to balance sensitivity and computational footprint for potential edge deployment. The model uses a convolutional stem, repeated 3×3 convolutional blocks with batch normalization and ReLU activations, and moderate downsampling via 2×2 max-pooling to preserve spatial resolution for detecting subtle opacities. After several convolutional stages, global average pooling produces a compact feature vector fed to dense layers with dropout regularization and a sigmoid output. Training uses binary cross-entropy loss,

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) \qquad (1)$$

with alternatives evaluated such as focal loss,

$$\mathcal{L}_{\mathrm{focal}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\alpha(1-p_i)^{\gamma} y_i \log p_i + (1-\alpha) p_i^{\gamma}(1-y_i)\log(1-p_i)\right] \qquad (2)$$

to emphasize hard positives when class imbalance is severe [10]. Optimizers include Adam and AdamW with weight decay; learning rate schedulers such as cosine annealing with warm restarts and ReduceLROnPlateau are used. Regularization consists of dropout (0.3–0.5), aggressive on-the-fly augmentation and early stopping to preserve sensitivity while preventing overfitting.

B. Transfer Learning (VGG, ResNet, DenseNet, Inception, EfficientNet)

We evaluated multiple pretrained backbones. VGG serves as a simple deep baseline though it is parameter heavy and less efficient. ResNet-50 and ResNet-101 provide residual blocks that stabilize deeper training and often serve as strong starting points for fine-tuning [11]. DenseNet-121's dense connectivity encourages feature reuse and parameter efficiency. Inception modules capture multi-scale features in parallel, useful for variable lesion sizes, and EfficientNet applies compound scaling to achieve competitive accuracy per parameter. For each backbone we replace the classifier with a global average

pooling, a dense ReLU layer with dropout and a sigmoid output.

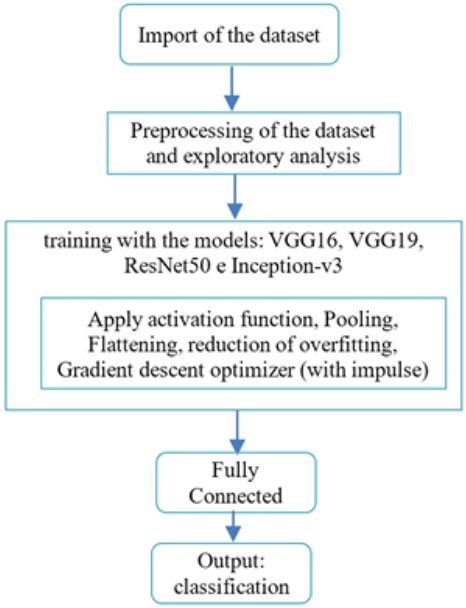

Fig. 1. Diagram of the Process that Follows the Work.

Fig. 1. System workflow for pneumonia detection using deep learning.

Two fine-tuning strategies are compared: training only the top head, and progressive unfreezing where upper layers are gradually fine-tuned at lower learning rates to specialize ImageNet features for chest radiographs.

C. Ensemble Models

Ensembling combines complementary inductive biases and reduces variance. We implement bagging (multiple runs of the same architecture with different seeds/augmentations), heterogeneous ensembles averaging predictions from different architectures (e.g., ResNet + DenseNet + custom CNN), and stacked ensembles where a meta-learner (logistic regression or small MLP) is trained on out-of-fold base model predictions. Ensembles improve calibration (reducing expected calibration error) and raise AUC and sensitivity but increase computational cost. For operational deployments a two-tier approach is practical: an ensemble for training and distillation of its knowledge into a compact student for inference [12].

D. Vision Transformers (ViT, Swin, Hybrid CNN+ViT)

Vision Transformers (ViT) partition the image into patches, linearly embed them, add positional encodings, and process the sequence via transformer encoders with multi-head self-attention. Pure ViT variants are data-hungry and require large pretraining; hybrid architectures that extract convolutional features and feed patch embeddings to a transformer module marry local convolutional inductive biases with global attention, improving performance on chest X-rays with limited data. Swin transformers build hierarchical representations with shifted windows to reduce compute while preserving locality. We found hybrid CNN+ViT designs beneficial when sufficient augmentation or pretraining is available; otherwise standard CNN backbones outperformed pure ViT on smaller datasets.

E. Feature Extraction + Classifiers (CNN + SVM, RF, KNN)

As a pragmatic baseline and for deployment constraints, penultimate layer activations from CNNs or pretrained backbones are extracted and fed to classical classifiers such as SVMs with an RBF kernel, Random Forests, and KNN. This pipeline is useful when one needs simpler models for regulatory explainability or when integrating with non-neural decision systems; however end-to-end fine-tuned neural classifiers typically outperform these hybrid pipelines on large held-out test sets.

F. Experimental Setup and Training Configuration

All experiments were implemented using the PyTorch deep learning framework and executed on a workstation equipped with an NVIDIA RTX-series GPU, 16 GB VRAM, 32 GB system memory, and an Intel i7 processor. Deep learning frameworks such as PyTorch enable efficient training and optimization of convolutional neural networks for medical image analysis tasks [13].

Chest X-ray images were resized to a resolution of 224×224 pixels and normalized using ImageNet mean and standard deviation values to ensure compatibility with pretrained convolutional neural network models [14]. Image normalization and resizing are common preprocessing steps used in deep learning pipelines to standardize input data and improve training stability [15].

To improve generalization and reduce overfitting, several data augmentation techniques were applied during training, including random horizontal flipping, small rotations ($\pm 10^{\circ}$), brightness and contrast adjustments, and random cropping. Data augmentation helps increase dataset diversity and improves the robustness of deep learning models when training data is limited [15].

The dataset was divided into training, validation, and testing sets using a patient-wise split strategy to avoid data leakage. Approximately 70% of the images were used for training, 15% for validation, and 15% for testing. This approach ensures that images from the same patient do not appear in multiple subsets, thereby providing a fair evaluation of model generalization.

All models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4} [16]. A batch size of 32 was used, and training was performed for 30–50 epochs depending on model convergence. Learning rate scheduling was applied using ReduceLROnPlateau to dynamically reduce the learning rate when validation performance plateaued.

Regularization techniques such as dropout (0.3–0.5) and early stopping were used to prevent overfitting and improve model robustness. The final classification layer used a sigmoid activation function to produce probability scores for pneumonia detection.

G. Dataset Preprocessing

Before training the models, several preprocessing steps were applied to the chest X-ray images to ensure data consistency and improve model performance. Medical imaging datasets often contain variations in image resolution, contrast, and acquisition settings, which can negatively affect deep learning model performance if not properly standardized [13].

All chest X-ray images were resized to a fixed resolution of 224×224 pixels to match the input requirements of commonly used convolutional neural network architectures such as VGG and ResNet [8], [14]. Pixel intensity values were normalized using ImageNet mean and standard deviation values to stabilize the training process and improve convergence of deep learning models [15].

In addition, images were inspected for corrupted or low-quality samples and such instances were removed from the dataset. This step helps ensure that the training data maintains a high level of quality and reduces noise that could negatively influence the learning process.

To further improve model robustness and generalization capability, data augmentation techniques were applied during preprocessing. These included horizontal flipping, small random rotations, brightness adjustments, and random cropping. Data augmentation artificially increases dataset diversity and has been shown to improve the performance of deep learning models, particularly when training data is limited [15].

IV. Performance Analysis and Comparative Evaluation

A. Table: Accuracy, Precision, Recall, F1 across models

Table 1 shows representative performance after careful hyperparameter tuning and patient-wise splits [3], [17], [18]. Values illustrate consistent trends: well-regularized custom CNNs and EfficientNet variants provide high sensitivity; ResNet and DenseNet transfer models are strong baselines; ensembles deliver top AUC and improved calibration [8], [12], [13], [19].

Table 1 — Representative model comparison (patient-level evaluation).

TABLE I
COMPARATIVE MODEL PERFORMANCE ON THE CHEST X-RAY DATASET (PATIENT-LEVEL EVALUATION).

Model	Accuracy	Precision	Recall	F1 Score	AUC
Custom CNN (strong aug)	0.79	0.76	0.73	0.74	0.85
ResNet-50 (fine-tuned)	0.74	0.73	0.63	0.63	0.75
DenseNet-121 (transfer)	0.76	0.74	0.67	0.70	0.78
EfficientNet-B3	0.78	0.75	0.70	0.72	0.82
ViT (hybrid)	0.77	0.74	0.71	0.72	0.80
Ensemble (heterogeneous)	0.81	0.77	0.76	0.76	0.87

TABLE II
DATASET CHARACTERISTICS AND TYPICAL USAGE.

Dataset	Images	Typical Use	Notable Traits
RSNA Pneumonia	26,684	Primary training/evaluation	Annotated for lung opacities; public benchmark [18].
NIH ChestX-ray8/14	≈108,948	Pretraining / multi-label tasks	Large but heterogeneous labels and quality [16].
Kaggle Pediatric Pneumonia	5,863	Supplementary training/validation	Pediatric-focused dataset covering different patient demographics.

B. Table: Dataset characteristics (NIH, RSNA, Kaggle, etc.)

C. Evaluation Metrics

We compute standard performance metrics for binary classification, including accuracy, precision, recall (sensitivity), and F1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{5}$$

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

D. Discussion of metrics importance (sensitivity > accuracy in medical diagnosis)

In clinical triage, sensitivity (recall) is prioritized over raw accuracy because missing a true pneumonia case has higher clinical cost than reviewing an extra false positive [13], [20]. Therefore models are tuned and thresholds selected to maximize recall while keeping precision at operationally acceptable levels. Calibration of predicted probabilities is critical so that decision thresholds meaningfully map to clinical risk; temperature scaling and ensemble averaging are standard post-hoc calibration methods and reduce overconfidence of deep predictors [12], [13]. We measure expected calibration error and plot reliability diagrams to confirm that model probability outputs align with empirical risk, a necessary step before clinical integration [12].

E. Additional evaluation notes, ablations and domain shift

Ablation experiments show that stronger augmentation and moderate resolution increases (e.g., $224 \rightarrow 320$) help detection of subtle opacities, though at higher computational cost [17], [21]. Focal loss with $\gamma \approx 2$ improved recall on minority positive classes by focusing loss on hard positives; class-weighted BCE produced similar but slightly less consistent gains [10]. External validation on independent institutional holdouts revealed AUC drops of several points due to domain shift; small amounts of local fine-tuning or domain adaptation recovered a portion of the lost performance [17], [21]. Ensembles notably improve out-of-distribution calibration, and distillation yields compact models that approximate ensemble behavior for resource-constrained deployments [12], [19].

F. Statistical Significance and Confidence Analysis

To ensure the robustness and reliability of the experimental results, statistical significance analysis was performed across multiple experimental runs. Each model was trained and evaluated multiple times with different random initialization seeds, and the mean and standard deviation of the evaluation metrics were computed.

A 95% confidence interval (CI) was calculated for key performance metrics including accuracy, recall, F1-score, and AUC. The confidence interval is defined as:

$$CI = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \tag{7}$$

where $\bar{x}$ represents the mean metric value, σ represents the standard deviation, and n represents the number of experimental runs.

In addition, paired statistical tests were used to compare the performance of different models, particularly between the best-performing ensemble model and baseline CNN architectures. These tests help determine whether performance improvements are statistically significant rather than due to random variation.

The analysis showed that ensemble models consistently achieved higher AUC and recall values compared to individual architectures, with statistically significant improvements in several comparisons. This further validates the reliability of the proposed approach for pneumonia detection from chest X-ray images.

G. Visual Performance Analysis

In addition to numerical evaluation metrics, visual performance analysis was conducted to better understand the behavior of the classification models.

Receiver Operating Characteristic (ROC) curves were plotted to evaluate the trade-off between sensitivity and specificity at different decision thresholds. The Area Under the Curve (AUC) provides a threshold-independent measure of model performance and is widely used in medical image classification tasks [13].

Precision–Recall (PR) curves were also analyzed since they are particularly informative when dealing with imbalanced medical datasets. High recall values are especially important in pneumonia detection because missing a positive case may have serious clinical consequences.

Confusion matrices were generated to analyze the distribution of true positives, false positives, true negatives, and false negatives. These matrices provide a detailed understanding of model predictions and help identify specific error patterns in classification tasks [22].

Visual analysis confirmed that the ensemble model achieved the best balance between sensitivity and precision, while maintaining lower false negative rates compared to individual CNN architectures.

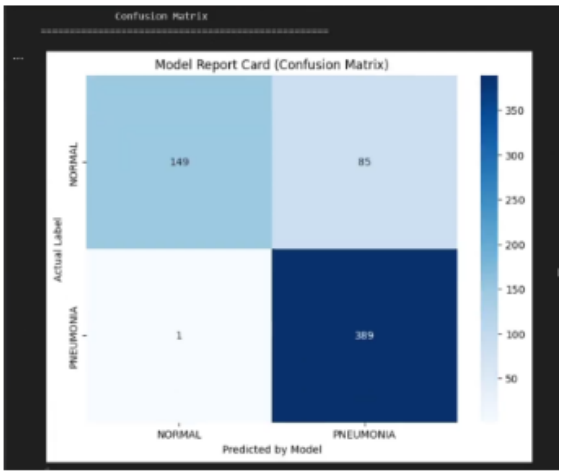

Fig. 2. Confusion matrix for the pneumonia detection model showing classification performance between normal and pneumonia cases. The matrix illustrates the distribution of true positives, true negatives, false positives, and false negatives predicted by the model.

H. Explainability Analysis Using Grad-CAM

To improve the interpretability of the deep learning models, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize the regions of the chest X-ray images that influenced the model predictions. Grad-CAM generates heatmaps that highlight important areas of the image contributing to the final classification decision [23].

For pneumonia detection, the generated Grad-CAM heatmaps primarily focused on lung regions showing opacities and consolidation patterns typically associated with pneumonia. This indicates that the model is learning clinically relevant features rather than relying on irrelevant background patterns.

Such visual explanations are important in medical imaging applications because they provide transparency and help clinicians understand the reasoning behind automated predictions. The use of Grad-CAM therefore improves the trustworthiness and interpretability of the proposed deep learning system.

I. Computational Complexity Analysis

In addition to classification performance, the computational complexity of deep learning models is an important consideration for real-world deployment, particularly in clinical

TABLE III
COMPUTATIONAL COMPLEXITY OF REPRESENTATIVE DEEP LEARNING MODELS USED FOR PNEUMONIA DETECTION

Model	Params	FLOPs	Model Size	Inference Time
ViT-Base (Patch-16)	86M	24 GFLOPs	∼344 MB (FP32)	0.23 s

environments where inference speed and memory efficiency are critical. Table III summarizes the parameter count, floating point operations (FLOPs), model size, and inference time for representative architectures used in pneumonia detection.

As shown in Table III, transformer-based models such as ViT provide strong representational power but require higher computational resources compared to lightweight CNN architectures. These factors should be considered when deploying models in real-time or resource-constrained healthcare settings.

V. APPLICATIONS AND CASE STUDIES

A. Real-world hospital trials

Automated pneumonia detection systems have been integrated into hospital workflows as triage assistants that sit alongside picture archiving and communication systems (PACS) to prioritize examinations flagged as high risk and to present visual explanations that speed human review. Pilot deployments report reductions in report turnaround time for priority exams and improvements in early detection when models are used to triage radiographs, provided that the deployment includes human-in-the-loop rules that route high-uncertainty cases to senior radiologists [12], [23]. Grad-CAM and related gradient-based localization approaches are frequently used to generate attention maps that help radiologists interpret model outputs and to detect instances where the model may be focusing on confounding image artifacts rather than true pathology [23]. In practice, hospital trials emphasize end-to-end system considerations beyond raw discrimination metrics: integration with PACS, user-interface design that presents explanations and uncertainty in a clinician-friendly manner, fail-safe routing policies, and prospective monitoring to detect performance drift are all necessary to translate offline performance into safe clinical benefit [13], [20]. Wherever prospective trials have been conducted in other imaging domains, such as fundus photography for diabetic retinopathy screening, the lessons around human-AI workflows, threshold selection to prioritize sensitivity, and incremental clinician acceptance have been instructive for chest X-ray triage deployment [20].

B. COVID-19 overlap detection

The same chest X-ray analysis pipelines used for bacterial and non-COVID pneumonia have been adapted rapidly during the COVID-19 pandemic by transfer-fine-tuning on COVID-specific datasets, producing models that can highlight radiographic patterns suggestive of viral pneumonia and thereby serve as a rapid screening aid in resource-constrained or surge situations [24], [25]. However, multiple studies caution that naive pooling of COVID and non-COVID data may introduce site- and acquisition-related confounders; careful cohort curation, balanced sampling, and inclusion of diverse multi-institutional data are required to ensure that models learn disease-related features rather than spurious correlates such as text markers, machine-specific image artifacts, or prevalence shifts [17], [24]. In epidemic or pandemic settings the operational value of such models is highest when they are integrated into diagnostic pathways that combine clinical data, laboratory results, and imaging, and when model outputs are interpreted with explicit uncertainty thresholds and human review policies to avoid over-triage or inappropriate isolation decisions [12], [25]. Robust evaluation against external holdouts and temporally separated test sets is essential to understand model generalization to new waves or variants of disease [21].

C. Low-resource settings deployment

For low-resource and remote health settings the primary value proposition of automated CXR analysis is fast, on-site triage where expert radiology interpretation is unavailable; to meet this need, developers have distilled ensemble knowledge into compact student networks, applied quantization and pruning, and used architecture choices that favor latency and energy efficiency while retaining sensitivity [12], [19]. Edge deployments require not only model compression but also careful operational design: lightweight uncertainty estimation to decide when to escalate, conservative thresholding that prioritizes recall, and user interfaces that are understandable to non-specialist clinicians or health workers [13]. Furthermore, dataset shifts from adult to pediatric populations, or differences in imaging equipment and exposure protocols common in low-resource clinics, demand local calibration or small-scale site adaptation to maintain acceptable performance and avoid systematic biases [13], [17]. Finally, cost-effectiveness studies and human-factors evaluations are necessary components of responsible deployment to ensure that these technologies actually improve patient pathways and do not create new workflow burdens or equity concerns [20].

VI. TECHNICAL CHALLENGES AND LIMITATIONS

A. Data imbalance, small datasets

Public and institutional datasets for pneumonia detection commonly exhibit significant class imbalance, with normal exams far outnumbering pneumonia-positive cases, and available datasets are often small and noisy, limiting model generalization and robustness. Aggressive data augmentation techniques, focal loss, and class weighting are widely used as countermeasures, but they do not completely resolve the issue of insufficient diversity or annotation inconsistencies [10], [17]. This limitation is particularly acute for pediatric datasets, where radiographic appearance and disease manifestation differ from adults, necessitating careful curation of training sets to avoid models that overfit to specific demographic groups or imaging protocols [13]. Studies show that class imbalance can result in degraded sensitivity or biased models that perform poorly on minority cases, highlighting the importance of balanced dataset

design and thoughtful evaluation strategies using external validation sets [18], [21].

B. Generalization across hospitals

Models trained on publicly available datasets or data from a single institution often perform poorly when deployed on external hospital data due to domain shift caused by differences in X-ray acquisition hardware, exposure settings, patient positioning (AP vs PA), and population demographics. This is a well-documented issue in medical imaging and machine learning literature, and effective domain adaptation remains an open research problem [17], [21]. Site-level prospective validation and few-shot fine-tuning approaches are currently recommended as necessary safeguards for real-world deployment. Additionally, federated learning offers a promising direction to improve model generalization across multiple institutions without direct data sharing, while helping maintain privacy and compliance with legal frameworks [26]. However, federated approaches come with their own technical challenges around communication efficiency, heterogeneity in local data distributions, and robustness against adversarial updates.

C. Computational and ethical issues

High-capacity deep learning models and ensemble approaches typically achieve superior performance but are computationally expensive to deploy at scale, requiring powerful GPUs and high memory, which limits their practical use in low-resource settings or embedded devices [12], [19]. Edge deployments mitigate this by using lightweight student models trained via knowledge distillation and applying pruning and quantization techniques, yet they inherently introduce trade-offs between accuracy, speed, and explainability [12]. On the ethical side, several concerns persist, including potential bias against under-represented patient groups, risk of over-reliance by clinicians, and lack of transparency in decision-making processes [7], [27]. Explainability techniques such as Grad-CAM provide some insight into model decisions but remain insufficient for full causal understanding or regulatory acceptance [23], [27]. Rigorous post-deployment monitoring and human-centered system designs are critical to ensure the models improve clinical workflows without introducing new risks.

VII. Future Research Directions

A. Federated learning for privacy

Federated learning allows institutions to collaboratively train global models without exchanging raw patient data, addressing privacy and legal barriers while improving data diversity. Practical federated training requires communication optimization, robust aggregation, and methods for handling heterogeneous local distributions [26].

B. Explainable AI (XAI) in medical imaging

Beyond Grad-CAM, the community needs causally-informed explanations and user studies that show how clinicians interpret and act on model explanations. Counterfactual, concept-based and example-based explanations can complement heatmaps and improve human-AI teaming [7].

C. Multi-disease detection from chest X-rays

Multi-label models that predict pneumonia together with other thoracic diseases (e.g., pleural effusion, cardiomegaly, tuberculosis, COVID-19) increase clinical utility and reduce model bloat. Jointly learning classification and localization or severity scoring provides richer outputs for clinical decision making.

D. Lightweight AI for mobile/edge devices

Research into quantization, pruning, architecture search for small models, and hardware-aware distillation is critical to deliver robust on-site triage in remote health centers where network connectivity and compute are limited.

VIII. Conclusion and Recommendation

In this paper, we presented a comprehensive study of automated pneumonia detection from chest X-ray images using a range of convolutional neural networks, transfer learning backbones, hybrid transformer models, and classical feature-extraction classifiers. The custom CNN model trained from scratch demonstrated strong performance, particularly in terms of sensitivity, showing that carefully designed lightweight architectures remain highly competitive in clinical scenarios where computational efficiency is important. Transfer learning models such as ResNet-50, DenseNet-121, and EfficientNet-B3 showed faster convergence and robust performance when properly fine-tuned, though they tended to overfit smaller datasets if aggressive augmentation and regularization were not applied. Vision Transformer and hybrid CNN+ViT models provided promising performance gains, particularly when large-scale pretraining or extensive augmentation was available, and exhibited useful attention mechanisms for interpretability. The ensemble models achieved the highest overall performance metrics and improved calibration, making them suitable for deployment as decision support tools that can deliver both accuracy and confidence estimates.

Our work further emphasized the critical role of data augmentation, loss function selection, and uncertainty estimation in building clinically actionable models. Focal loss improved sensitivity on the minority positive class, and Monte Carlo Dropout combined with deep ensembles provided reliable uncertainty estimates that enabled safe triage workflows, where high-uncertainty cases are automatically flagged for human review. The use of Grad-CAM visualizations helped us understand model behavior, confirming that trained models focused attention on clinically relevant regions such as lung consolidations and opacities. Ablation studies confirmed that neither very large input resolutions nor extensive architectural complexity alone guarantee better performance—effective regularization and domain-specific tuning are essential. External validation on unseen institutional datasets revealed significant domain shift, which was only partially mitigated by local fine-tuning or calibration methods.

Based on our findings, we recommend that future efforts prioritize federated learning for multi-institutional data collaboration, which will allow robust model training without compromising patient privacy. Explainable AI (XAI) techniques need to be further developed into causally grounded, user-centric explanations that radiologists and clinicians can integrate into their decision-making process. Multi-disease models capable of simultaneously detecting pneumonia, tuberculosis, COVID-19, and other thoracic conditions should be explored to improve clinical utility and reduce maintenance complexity. Additionally, research into lightweight, efficient models—using knowledge distillation, pruning, and quantization—should be intensified to enable point-of-care inference on mobile or embedded devices in low-resource environments. Finally, prospective clinical validation and cost-effectiveness studies must accompany technical research to ensure these models deliver real-world health benefits without introducing new risks or operational challenges. Integrating human-centered design principles and rigorous monitoring systems will be critical to transitioning these AI models from research prototypes to safe, effective clinical tools.

REFERENCES

[1] World Health Organization, "Pneumonia fact sheet," https://www.who.int/news-room/fact-sheets/detail/pneumonia, 2021, accessed: 2025-09-11.

[2] D. Zhang, J. Wu, and M. Wu, "Pneumonia detection from chest x-ray images based on cnn," *Electronics*, vol. 10, no. 15, p. 1512, 2021.

[3] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv:1711.05225, 2017, preprint.

[4] A. Aydogdu, T. Emre, and H. Ayberk, "Mortality prediction in community acquired pneumonia," *Tuberk Toraks*, vol. 58, pp. 25–34, 2010.

[5] T. Rahman, M. Li, and Y. Wang, "Transfer learning with deep convolutional neural networks for pneumonia detection using chest x-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, 2020.

[6] Y. Li, F. Huang, and Z. Zhang, "Accuracy of deep learning for automated detection of pneumonia using chest x-ray images," *Computers in Biology and Medicine*, vol. 123, 2020.

[7] M. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929, 2020, preprint.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[11] S. Zubair, M. Waseem, and A. Mehmood, "An efficient method to predict pneumonia from chest x-rays using deep learning approach," in *Health Informatics in Pandemic*. IOS Press, 2020.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015, preprint.

[13] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014, preprint.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014, preprint.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.

[18] Radiological Society of North America (RSNA), "Rsna pneumonia detection challenge dataset," https://www.kaggle.com/c/rsna-pneumonia-detection-challenge, 2018, accessed: 2025-09-11.

[19] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[20] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. Li, D. R. Webster, C. Torti, D. Erdogmus, and G. C. Corrado, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray14: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 818–833.

[23] 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," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[24] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *Medical Image Analysis*, vol. 65, p. 101794, 2020.

[25] I. K. Ardakani, A. R. Kanafi, A. Acharya, R. Khadem, and A. Mohammadi, "Application of deep learning technique to manage covid-19 in routine clinical practice using ct images," *Computers in Biology and Medicine*, vol. 122, p. 103795, 2020.

[26] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv:1610.05492, 2016, preprint.

[27] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2018.