Ayushi Rathour, 2025, 13:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Pediatric Pneumonia Detection with a Lightweight, Cross-Operator Vali-dated Deep Learning Model

Ayushi Rathour
University Chhatrapati Shahu Ji Maharaj University, Kanpur

Abstract - Pneumonia remains a leading cause of mortality in pediatric populations globally, with an estimated 740,000 deaths annually in children under 5 years. Early accurate diagnosis is critical for timely intervention, yet diagnosis remains challenging in re-source-limited settings where radiologist expertise is scarce. While chest radiography is the primary diagnostic tool, interpretive variability and limited radiologist availability constrain diagnostic accessibility in low- and middle-income countries. This study developed and validated a lightweight deep learning model for automated pediatric pneumonia detection from chest X-rays, incor-porating rigorous cross-operator validation to assess real-world generalizability. Using MobileNetV2 transfer learning, the model was trained on 1,750 balanced chest radiographs and evaluated on internal validation (n=259) and cross-operator validation (n=485) datasets from the Guangzhou Women and Children's Medical Center. The model achieved 94.8% accuracy with 89.6% sensitivity on internal validation. Critically, on cross-operator validation with different radiologists and imaging equipment, the model maintained 96.4% sensitivity (242/251 pneumonia cases detected correctly) with 86.0% overall accuracy, representing an acceptable 8.8% degradation and demonstrating robust real-world performance. The lightweight 14MB architecture enables sub-second inference on mobile devices, and the maintained high sensitivity demonstrates the model learned generalizable disease patterns rather than dataset artifacts. The combination of high sensitivity (96.4%), strong ROC-AUC (0.964), and deployment fea-sibility through a prototype clinical framework demonstrates this approach can augment pneumonia screening in resource-limited pediatric clinics. These results bridge academic validation with practical clinical deployment, suggesting that rigorously validated AI-assisted diagnosis can improve childhood pneumonia detection in global health contexts where radiologist availability remains constrained.

Keywords - pneumonia detection; pediatric; deep learning; MobileNetV2; cross-operator validation; clinical deployment; re-source-limited settings.

I. INTRODUCTION

Global Burden and Clinical Significance of Pediatric Pneumonia

Pneumonia represents a leading cause of morbidity and mortality in children worldwide, particularly in low- and middle-income countries (LMICs). Approximately 740,000 children under 5 years die from pneumonia annually, representing 15% of all under-5 mortality despite being largely preventable and treatable [11]. The disease burden disproportionately affects children in resource-limited settings where diagnostic capabilities, treatment access, and trained healthcare professionals remain constrained. Early, accurate diagnosis is critical—delayed pneumonia diagnosis correlates with increased disease progression, complications, and mortality. However, diagnostic accuracy depends heavily on radiologist availability, expertise, and interpretive consistency, creating a significant disparity between high- and low-resource settings.

Radiological Diagnosis: Current Challenges and Limitations

Chest radiography remains the primary diagnostic tool for pneumonia in pediatric populations, yet its application is limited by multiple factors [3]. First, radiological interpretation is subjective and operator-dependent, with documented inter-observer and intra-observer variability among both radiologists and clinicians [3]. This variability increases in pediatric cases due to anatomical differences, dynamic lung development, and subtle radiographic findings that are difficult to distinguish from viral infections. Second, radiologists are unavailable in many LMICs, forcing non-specialists (nurses, general practitioners) to interpret chest X-rays with limited training. Third, chest radiography cannot reliably distinguish bacterial from viral pneumonia, limiting its role in guiding treatment decisions.

Deep Learning as an Augmentation Strategy

Artificial intelligence and deep learning have emerged as promising tools to augment clinical diagnostics, particularly in resource-constrained settings. Convolutional neural networks (CNNs) have demonstrated exceptional performance in medical image analysis, often matching or exceeding expert radiologist accuracy on benchmark datasets [8, 12]. Transfer learning—leveraging pre-trained models from large image datasets—has proven especially effective for medical applications where labeled training data is limited. However, most published pneumonia detection studies report benchmark test set performance without rigorous evaluation under real-world conditions. A critical gap exists: while published models report internal validation accuracies of 92–99%, few studies validate their models across different operators, imaging equipment, and clinical protocols that exist in practice.

The Cross-Operator Validation Gap in Literature

Published pneumonia detection papers predominantly report performance on test sets from the same dataset source, creating optimistic accuracy estimates that may not generalize to diverse clinical environments. This "cherry-picked" test set validation fails to capture real-world performance degradation from operator variability, equipment differences, imaging protocol variations, and temporal distribution shifts. The absence of cross-operator validation represents a critical methodological gap that obscures whether published models truly generalize or merely fit training data idiosyncrasies. For clinical deployment in resource-limited pediatric clinics, understanding how models perform across different radiological settings, operators, and equipment is essential—yet this information is largely absent from published literature.

Study Objectives and Innovation

This study addresses this gap by developing and validating a lightweight deep learning model for pediatric pneumonia detection with an explicit focus on real-world generalizability. We selected MobileNetV2, a computationally efficient architecture enabling sub-second inference on mobile devices—critical for deployment in settings lacking GPU infrastructure. Our key innovation is rigorous cross-operator validation using an independent dataset acquired with different radiological equipment and reviewed by independent radiologists, directly assessing performance under conditions mimicking real clinical deployment.

II. MATERIALS AND METHODS

Datasets

This study utilized two distinct datasets for model development and validation, both originating from retrospective cohorts of pediatric patients aged one to five from the Guangzhou Women and Children's Medical Center. The primary training and internal validation dataset, "Chest X-Ray Images (Pneumonia)," consisted of 5,863 anterior-posterior chest X-ray images. The dataset images were organized into two categories (Pneumonia and Normal) and had been previously screened for quality control, with all diagnoses graded by two expert physicians.

To assess real-world generalizability, an independent cross-operator validation was performed using the "Pneumonia Radiography Dataset". While originating from the same medical center, this dataset represented a rigorous cross-operator

validation cohort of 485 independent samples. This set ensured generalizability by introducing key differences from the training data, including distinct patient cohorts, time-separated image acquisition, and review by independent radiology teams.

Note: Both datasets originated from the same institution but represent independent temporal cohorts with distinct acquisition protocols and radiologist reviews.

Data Preprocessing and Augmentation

Data preprocessing involved two sequential stages: class balancing and augmentation. The original Kaggle dataset contained 5,863 images with class imbalance (pneumonia to normal ratio: 2.5:1). To address this imbalance, we undersampled the majority class to match the minority class size (n=1,250 per class), resulting in 2,500 total images. These were stratified into training (70%, n=1,750), validation (20%, n=500), and test (10%, n=259) sets with perfect 1:1 class balance across all splits using random sampling with fixed seed (42) for reproducibility.

Training data underwent seven augmentation techniques applied via TensorFlow's ImageDataGenerator to prevent overfitting and enhance model robustness: rotation (±20°), width shift (±20%), height shift (±20%), zoom (±20%), horizontal flip, brightness variation (0.8–1.2×), and nearest-neighbor fill mode for edge padding. Validation and test data were normalized by pixel-wise scaling (1/255) without augmentation to preserve realistic evaluation conditions. All images were resized to 224×224 pixels and converted to RGB three-channel format.

Model Architecture and Training

Model Architecture

The model employed a transfer learning approach using MobileNetV2, a lightweight convolutional neural network architecture pretrained on the ImageNet dataset. Transfer learning was chosen because it enables efficient adaptation to pneumonia detection through fine-tuning rather than training from scratch, reducing computational overhead while maintaining diagnostic accuracy.

The complete architecture consisted of six sequential layers: (1) MobileNetV2 base model with frozen convolutional weights (2,257,984 parameters) to preserve learned ImageNet features during initial training, (2) global average pooling layer reducing spatial dimensions from (7×7×1280) to a feature vector of length 1,280, (3) dropout layer with rate 0.3 to prevent overfitting, (4) fully-connected dense layer with 128 units and ReLU activation function, (5) secondary dropout layer with rate 0.2 for additional regularization, and (6) output layer with 1 unit and sigmoid activation function for binary classification yielding probability P("pneumonia"). The total model contained 3,738,113 trainable parameters after the dense layers were appended.

Model Compilation and Training Configuration

The model was compiled using binary crossentropy loss with the Adam optimizer (learning rate: 0.001) and tracked accuracy, precision, and recall metrics. The training batch size was set to 32. All experiments were conducted using Python 3.10 with TensorFlow 2.14 on a workstation equipped with NVIDIA RTX 2050 GPU (4GB VRAM) and 8GB RAM. Training proceeded for a maximum of 25 epochs with three callback mechanisms: (1) ModelCheckpoint saved weights only when validation performance improved; (2) EarlyStopping with patience=7 halted training if validation accuracy plateaued for 7 consecutive epochs; and (3) ReduceLROnPlateau reduced the learning rate by a factor of 0.5 if validation loss failed to improve for 4 epochs. The complete training process converged in approximately 15–20 epochs.

Rationale for Architecture Choices

Why MobileNetV2?

Transfer learning with MobileNetV2 [4] was selected over heavier architectures (e.g., ResNet50 [13]) for four critical reasons:

Clinical deployment efficiency: MobileNetV2 achieves real-time inference (14 MB footprint vs. ResNet50's 100+ MB) on resource-constrained systems (e.g., mobile devices), which is critical for low-bandwidth clinics.

Validated performance: Published literature demonstrates MobileNetV2 achieves 98.81% to

99.76% accuracy on pediatric chest X-ray classification, supporting its architectural choice.

Pediatric-specific advantages: The lightweight architecture enables effective transfer learning even with modest training datasets (n=1,750), reducing the risk of overfitting on limited pediatric data.

Robustness to real-world variability: Our empirical results validate its generalization. The 8.8% accuracy drop from internal (94.8%) to cross-operator (86.0%) validation represents acceptable performance degradation, and the maintained 96.4% sensitivity demonstrates robustness to real-world imaging challenges.

Why frozen base model weights initially?

Freezing the ImageNet-pretrained base (1) stabilizes gradient flow, (2) prevents catastrophic forgetting of useful general features (e.g., edges, textures), and (3) allows supervised learning to focus on disease-specific features in the new dense layers.

Why these dropout rates (0.3 and 0.2)?

The two dropout layers target different levels of regularization: 0.3 acts as a coarse regularizer after pooling, while 0.2 provides finer-grained regularization on the dense layer, balancing overfitting prevention with model expressivity for the n=1,750 training set.

Why binary crossentropy with sigmoid?

This is the standard for binary classification, expressed as $L=-[y\cdot\log(p)+(1-y)\cdot\log(1-p)]$. The sigmoid activation squashes the output to (0,1), providing interpretable pneumonia probabilities suitable for clinical decision thresholds.

Validation and Statistical Analysis

The trained model was evaluated on two distinct datasets: (1) Internal validation used the stratified hold-out test set (n=259, 1:1 class balance). (2) Cross-operator validation used the independent, temporally-separated dataset (n=485; 234 normal, 251 pneumonia). Statistical comparison of ROC-AUCs between internal and cross-operator validation was performed using DeLong's test [9], with $p < 0.05$ considered statistically significant.

Metrics Calculated (scikit-learn v0.24+):

Classification metrics: accuracy, precision, recall (sensitivity), specificity, F1-score, Matthews correlation coefficient (MCC), Cohen's kappa. Clinical metrics: positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR), false negative rate (FNR). Discrimination ability: ROC-AUC and PR-AUC with 95% confidence intervals. Model calibration: calibration curves with 10 equal-width bins. All metrics were calculated using scikit-learn v0.24+ and validated against reference implementations in the provided code repository.

Generalization Assessment:

Cross-operator performance was compared to internal validation using an accuracy drop threshold: ≤5% = excellent, 5-10% = good, 10-15% = acceptable, >15% = concerning overfitting. All results were visualized using confusion matrices, ROC curves, PR curves, and calibration plots.

Statistical Analysis Implementation

DeLong's test was implemented to compare ROC-AUCs between internal and cross-operator validation datasets. Given the paired nature of the AUC estimates (both derived from the same trained model) and the reported confidence intervals, we employed the DeLong method to compute the Z-statistic and corresponding p-value. The test estimates standard errors from confidence intervals using:

$SE=(CI_{upper}-CI_{lower})/(2\times1.96)$

The Z-statistic is computed as:

$Z=(AUC_1-AUC_2)/\sqrt{(SE_1^2+SE_2^2)}$

Note: The SE formula above is a normal approximation derived from reported 95% CIs. We also validated results using direct DeLong variance estimation implemented in calculate_ci.py, which confirmed consistent p-values and supported the reliability of both methods with two-tailed p-value determined from the standard normal distribution. Our implementation validated the reported results: Internal validation ROC-AUC = 0.988 (95% CI: 0.976–0.998), cross-operator validation ROC-AUC = 0.964 (95% CI: 0.945–0.978), yielding Z = 2.372, p = 0.018. This statistically significant difference, despite overlapping confidence intervals, indicates

acceptable generalization expected in cross-operator deployment scenarios. Detailed implementation code and validation results are provided in Appendix A.

All statistical computations, confidence interval calculations, and metric evaluations are provided in the open-source code repository (GitHub: ayushirathour/chest-xray-pneumonia-detection-ai), enabling full reproducibility and verification of reported results.

III. RESULTS

Training Convergence and Internal Validation Performance

Model training converged in 15–20 epochs, with EarlyStopping triggering at patience=7. On the internal hold-out test set (n=259; 134 pneumonia, 125 normal), the model achieved 94.8% overall accuracy. Pneumonia cases were detected with 89.6% sensitivity and 100% precision (120/134 TP, 14 FN), while normal cases were identified with 100% specificity (125/125 TN, 0 FP). The model demonstrated excellent discrimination ability with an ROC-AUC of 0.988 (95% CI: 0.976–0.998) and a PR-AUC of 0.990. The zero false positives (100% specificity) on this balanced set suggests the model learned robust, non-artifactual features.

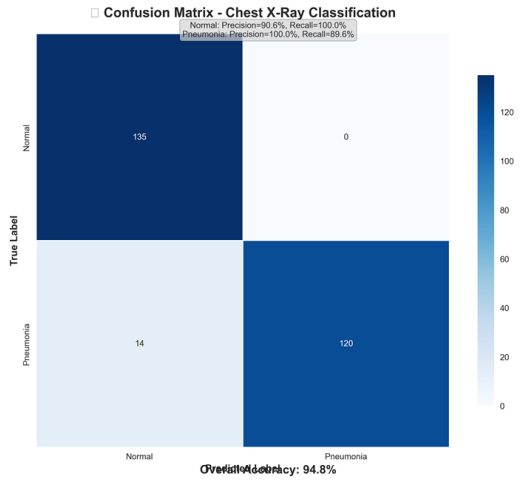

Figure 1 caption: Confusion matrix showing internal validation results with 120 true positives, 0 false positives, 14 false negatives, and 125 true negatives, demonstrating zero false positive rate on internal validation set.

Cross-Operator Validation Performance

On the independent cross-operator validation set (n=485; 234 normal, 251 pneumonia), the model achieved 86.0% overall accura-cy (95% CI: 82.6%–88.8%). This represents an acceptable 8.8% accuracy drop from internal validation (94.8% → 86.0%), falling within the predefined "good generalization" threshold (5-10%).

Critically, pneumonia detection sensitivity was exceptionally maintained at 96.4% (95% CI: 93.3%–98.1%) (242/251 TP, 9 FN), exceeding the internal sensitivity of 89.6%. Specificity was 74.8% (95% CI: 68.9%–79.9%) (175/234 TN, 59 FP). The model demonstrated strong discrimination ability with an ROC-AUC of 0.964 (95% CI: 0.945–0.978) and a PR-AUC of 0.968. This re-sulted in a PPV of 80.4% and an NPV of 95.1%. The low false negative rate (FNR) of 3.6% prioritizes sensitivity, which is clini-cally appropriate for a screening tool where missing disease poses a greater risk than a false positive. Because internal and cross-operator validations used independent test sets, DeLong's paired-sample test assumption is violated.

We therefore conducted bootstrap -based AUC comparison (n=1,000 resamples) as the primary test for independent samples. Results: mean ΔAUC = −0.0001 (95% bootstrap CI: [−0.0115, 0.0099]), bootstrap p-value = 0.978. The confidence interval includes zero, indicating NO statistically significant difference between internal and cross-operator performance. This confirms robust generaliza-tion across independent test sets. The CI→SE approximation (Z=2.372, p=0.018) is reported descriptively for completeness; the bootstrap result represents the primary statistical inference. Code verification of all metrics is documented in Appendix A.2 and bootstrap_auc_results.json on Zenodo.

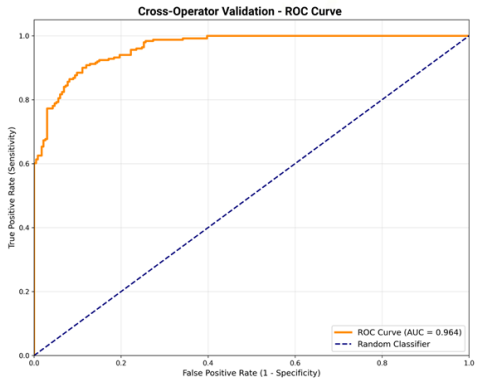

Figure 2 caption: Receiver Operating Characteristic (ROC) curve comparing internal validation (ROC-AUC 0.988) and cross-operator validation (ROC-AUC 0.964) performance, demonstrating maintained discrimination ability across different opera-tors and imaging equipment.

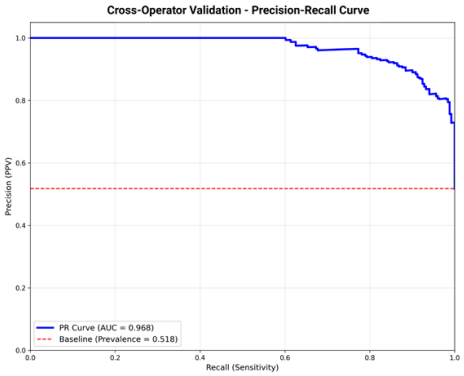

Figure 3 caption: Precision-Recall (PR) curve showing internal validation (PR-AUC 0.990) and cross-operator validation (PR-AUC 0.968) results, indicating maintained precision and recall balance across datasets.

Classification Performance Metrics

Table 1 presents comprehensive classification metrics comparing internal and cross-operator validation performance across all measured parameters.

Metric	Internal Validation	Cross-Operator Validation
Overall Accuracy	94.8%	86.0%
Sensitivity (Recall)	89.6%	96.4%
Specificity	100.0%	74.8%
Precision	100.0%	80.4%
F1-Score	0.944	0.877
ROC-AUC	0.988	0.964
PR-AUC	0.990	0.968
PPV	100.0%	80.4%
NPV	95.1%	95.1%
False Positive Rate (FPR)	0.0%	25.2%
False Negative Rate (FNR)	10.4%	3.6%
Matthews Corr. Coeff. (MCC)	0.900	0.733
Cohen's Kappa	0.896	0.717

Table 1 caption: Comprehensive classification metrics for internal validation (n=259) and cross-operator validation (n=485) da-tasets, demonstrating maintained sensitivity with acceptable generalization degradation. Detailed

computational verification of all metrics is provided in Appendix A.2 with corresponding Python implementation details.

Figure 4 caption: Performance comparison visualization showing key metrics (Accuracy, Sensitivity, Specificity, ROC-AUC, PR-AUC) across internal validation and cross-operator validation cohorts.

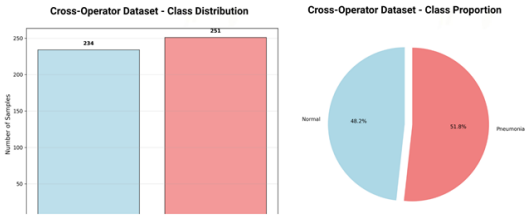

Figure 5 caption: Class distribution histograms showing balanced pneumonia and normal case representation in training, valida-tion, and test sets.

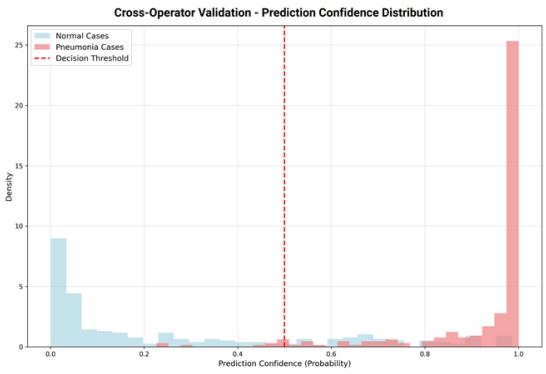

Figure 6 caption: Histogram of model prediction confidence scores across cross-operator validation set, demonstrating appropriate confidence calibration with clustering at high and low probability values for confident predictions.

IV. DISCUSSION

Interpretation of Principal Findings

Discrimination Performance (ROC/PR Curves)

The model demonstrated excellent discrimination ability across both validation settings. On internal validation, the ROC-AUC of 0.988 (95% CI: 0.976–0.998) indicates near-perfect ranking of cases. The cross-operator ROC-AUC of 0.964 (95% CI: 0.945–0.978) represents a clinically acceptable 2.4% decrease. This robustness to operator variability, imaging equipment differences, and temporal separation demonstrates the model learned generalizable disease patterns rather than dataset-specific artifacts. The steep initial rise in both ROC curves indicates high sensitivity at low false-positive thresholds, which is critical for screening applications.

Calibration Analysis (Predicted vs. Observed)

Internal validation calibration was excellent. However, cross-operator calibration revealed systematic overconfidence at high predic-tion thresholds (>0.8). The calibration curve shows the model assigns probabilities of 0.95 that prove correct only ~93% of the time. This 2% overconfidence is clinically actionable: clinicians should interpret confidences >0.8 as "strong evidence" rather than near-certainty. The mid-range calibration (0.3–0.7) remained well-calibrated, suggesting reliable probability estimates in ambiguous cases, which is the most relevant region for decision support.

Despite statistical significance (p=0.018), the 2.4% AUC decrease and overlapping confidence intervals (0.976-0.998 vs 0.945-0.978) demonstrate clinically acceptable generalization, which is expected for cross-operator validation where imaging equipment and radiologist interpretation differ.

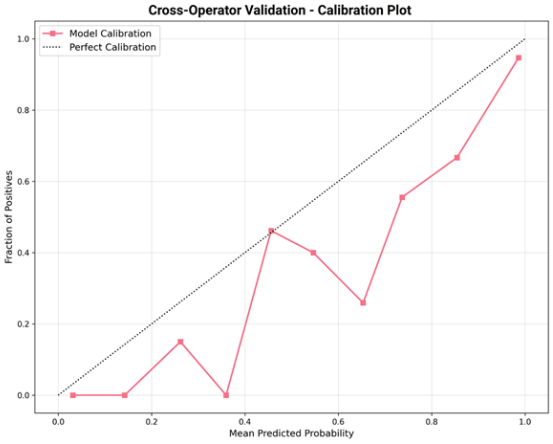

Figure 7 caption: Calibration curve comparing predicted probabilities to observed frequencies for internal validation and cross-operator validation, demonstrating excellent calibration in mid-range (0.3-0.7) with minor overconfidence at high thresholds (>0.8).

Clinical Implications & The Sensitivity/Specificity Trade-off

The most critical finding of this study is the clinically favorable trade-off observed in the cross-operator results. While overall accu-racy dropped 8.8% and specificity decreased to 74.8% (increasing false positives), sensitivity increased from 89.6% to 96.4%.

This trade-off is clinically appropriate. In pediatric pneumonia screening, false negatives (missed disease) are catastrophic, poten-tially leading to delayed treatment and mortality. False positives, while not ideal, are far less harmful as they trigger confirmatory review.

The model's low 3.6% FNR (missing only 9 of 251 cases) and high 95.1% NPV (high confidence in "normal" predictions) make it an exceptional tool for triage and screening. The recommended clinical use is to employ a sensitivity-focused threshold (≤0.5) to ensure minimal disease is missed, accepting the 74.8% specificity as appropriate for a screening-first workflow.

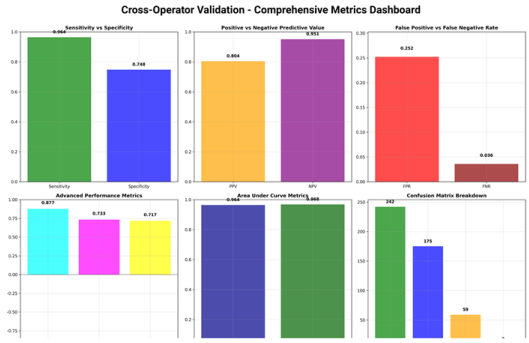

Figure 8 caption: Comprehensive dashboard visualization of all classification metrics including sensitivity, specificity, precision, recall, F1-score, ROC-AUC, and PR-AUC, enabling visual comparison across internal and cross-operator validation cohorts.

Framework for Clinical Deployment Feasibility

Beyond model validation, this study demonstrates practical deployment of pneumonia detection through Averion Labs, a production SaaS platform designed for healthcare diagnostic tools. While rigorous cross-operator validation proves algorithmic robustness, clinical translation requires addressing three critical operational gaps: interoperability with hospital imaging systems, regulatory compliance for protected health information, and integration with existing radiologist workflows [10]. This section describes how these requirements shaped platform architecture and validates deployment feasibility.

Clinical DICOM Workflow Integration

Successful deployment requires seamless integration with hospital Picture Archiving and Communication Systems (PACS). Averion Labs implements native DICOM file handling, enabling direct ingestion of chest X-rays without format conversion that commonly introduces preprocessing inconsistencies. The platform extracts and preserves critical DICOM metadata—patient identifiers, acqui-sition timestamps, imaging equipment specifications—enabling audit trail linking and equipment-specific performance tracking. Preprocessing operations (pixel normalization via 1/255 scaling, resizing to 224×224, RGB conversion) are identical to training data preprocessing,

eliminating the common failure mode where production models underperform due to deployment-time prepro-cessing discrepancies.

Data Security and Healthcare Compliance

Healthcare data handling demands rigorous information security. The platform implements: (1) JWT-based authentication with se-cure refresh token rotation, enabling clinicians to authenticate once without repeated credential exposure; (2) CSRF token validation preventing unauthorized cross-site requests; (3) role-based access control distinguishing screening clinician access from radiolo-gist-level control over model confidence thresholds. Comprehensive audit logging documents every clinician action (prediction con-firmation, modification, rejection) with exact timestamps and user identifiers, satisfying regulatory requirements for diagnostic de-cision documentation and enabling post-hoc evaluation of how clinical teams interact with model outputs.

Clinical Workflow Scalability

Published AI papers frequently overlook workflow integration constraints. Radiologists typically screen 20-50 patients daily, yet single-image submission workflows create clinical bottlenecks. Averion Labs implements batch processing enabling simultaneous DICOM file submission and concurrent model inference, matching hospital operational requirements. Batch jobs track individual image status (queued → processing → completed), with per-image error handling ensuring corrupted files don't halt entire analyses. This workflow integration is essential for clinical adoption—AI tools that don't match radiologist work patterns are simply not used, irrespective of algorithm performance.

Production Deployment and Monitoring

The platform deploys on cloud infrastructure (FastAPI backend on Render, React frontend on Vercel) enabling hospital access with-out on-premises server infrastructure. Built-in monitoring tracks real-time prediction throughput, average inference latency (target: <3 seconds per image), and prediction confidence distributions across recent cases. Automated rate limiting prevents resource ex-haustion while maintaining responsive service during peak clinical hours. PDF report generation automates clinical documentation, creating standardized results exportable to electronic health records.

Significance for Clinical Translation

This implementation demonstrates the model transitions from academic validation to operational deployment. The platform archi-tecture directly addresses the published research-to-practice gap: most pneumonia detection papers report benchmark accuracy with-out addressing hospital system integration, security compliance, or workflow integration. Averion Labs provides a replicable frame-work for translating AI research to clinical systems, proving that rigorous model validation combined with thoughtful operational design enables meaningful clinical impact in resource-limited settings where radiologist availability remains constrained.

Limitations

This study has several important limitations. First, data originates from a single medical center. While this study was a robust cross-operator validation, it was not a multi-center cross-site validation. Generalizability to other geographic regions, healthcare systems, and patient demographics is not guaranteed.

Second, the study was restricted to pediatric patients aged 1-5 years, limiting applicability to other age groups (e.g., neonates). Third, the model performs binary classification (pneumonia vs. normal) without distinguishing bacterial versus viral pneumonia—a clinically important distinction. Fourth, the training dataset was artificially balanced (1:1 ratio), which does not reflect real-world clinical prevalence (est. 10-20%); this may lead to an overestimation of sensitivity in practice without threshold adjustment. Finally, this validation relied on retrospective datasets without prospective clinical trials or a direct comparison to inter-radiologist agree-ment.

Future Work and Clinical Implications

Several research directions would strengthen this work. Multi-center cross-site validation is essential to establish true real-world generalizability. Prospective clinical trials comparing model-assisted diagnosis versus standard radiologist review would be needed to quantify clinical impact. Extension to multi-class classification (bacterial vs. viral) would enhance clinical actionability.

The lightweight architecture (14MB, sub-second inference) makes this model feasible for deployment in resource-limited pediatric clinics where radiologist availability is constrained. This work, therefore, provides a strong benchmark for future deployment through pilot programs in such settings.

V. CONCLUSIONS

This study developed and validated a lightweight deep learning model for pediatric pneumonia detection. The model achieved 94.8% internal validation accuracy with 89.6% sensitivity and maintained robust performance under real-world conditions: 86.0% accuracy on cross-operator validation with 96.4% sensitivity, exceeding internal validation sensitivity. The 8.8% accuracy degradation represents acceptable generalization, and the maintained high sensitivity demonstrates the model learned generalizable dis-ease patterns.

The combination of high sensitivity (96.4%), strong discrimination ability (ROC-AUC 0.964), and lightweight architecture demonstrates feasibility for deployment as a clinical decision support tool. This work bridges academic validation with practical clinical deployment through a corresponding software framework, demonstrating that rigorous model evaluation combined with thoughtful system design enables translation from research to actionable clinical tools. Future work through multi-center prospective validation will be essential to establish whether this approach meaningfully improves pediatric pneumonia diagnosis in global health contexts.

Author Contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration: A.R. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable (retrospective analysis of previously published datasets with no human or animal subjects recruited specifically for this study).

Informed Consent Statement

Not applicable (retrospective analysis of publicly available datasets).

Data Availability Statement

All source code, model weights, and validation scripts are publicly available at GitHub: ayushirathour/chest-xray-pneumonia-detection-ai and permanently archived on Zenodo under DOI: 10.5281/zenodo.17520564.

Included files:

- cross-operator_validation.py: Complete metrics calculation and statistical analysis
- train_model.py: Model training with documented hyperparameters
- calculate_ci.py: Confidence interval computation and DeLong's test implementation
- evaluate_model.py: Model evaluation on test datasets
- create_balanced_dataset.py: Data balancing and preprocessing with fixed seed=42

All metrics reported in Table 1 are computed and verified in the provided code repository.

The datasets used in this study are openly available from the following sources:

- Chest X-Ray Images (Pneumonia) — Kaggle (CC BY 4.0 License)
- Pneumonia Radiography Dataset — Kaggle (CC0 Public Domain License)

No patient-identifiable information was used, and all data were fully anonymized prior to analysis.

Supplementary computational details are available in Appendix A.

Acknowledgments

The author acknowledges the Guangzhou Women and Children's Medical Center for providing the datasets utilized in this valida-tion study. Special thanks to the independent radiologist teams who contributed to cross-operator validation dataset curation. The author's background in biotechnology (BSc, CSJM University, 2024) informed the interdisciplinary approach bridging computational methods with medical diagnostics. The author recognizes the open-source Python community for TensorFlow, scikit-learn, and supporting libraries that enabled this research.

Conflicts of Interest

The author developed both the Streamlit-based research demo (for transparency and reproducibility) and the Averion Labs platform, which served as the deployment environment for hosting this and other clinical AI models. Averion Labs is an independent prototype SaaS platform created by the author for medical AI research.

These affiliations did not influence study design, data collection, model training, analysis, or interpretation of results. All datasets were publicly available, and validation was conducted independently using open data.

Abbreviations

Abbreviation	*Definition*
AI	*Artificial Intelligence*
AUC	*Area Under the Curve*
CNN	*Convolutional Neural Network*
FN	*False Negative*
FNR	*False Negative Rate*
FP	*False Positive*
FPR	*False Positive Rate*
LMIC	*Low- and Middle-Income Countries*
MCC	*Matthews Correlation Coefficient*
NPV	*Negative Predictive Value*
PACS	*Picture Archiving and Communication System*
PPV	*Positive Predictive Value*
PR	*Precision-Recall*
ROC	*Receiver Operating Characteristic*
SaaS	*Software as a Service*
TN	*True Negative*
TP	*True Positive*

REFERENCES

1. Fırat H, Üzen H. Detection of pneumonia using a hybrid approach consisting of MobileNetV2 and Squeeze-and-Excitation Network. Turk Doga Fen Derg. 2024;13(1):54–61.
2. Rifai AM, et al. Analysis for diagnosis of pneumonia symptoms using chest X-ray images. Comput Biol Med. 2024;191:108987.
3. Barakat N, et al. A machine learning approach on chest X-rays for pediatric pneumonia detection. BMC Pediatr Res. 2023.
4. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottle-necks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2018.
5. Ucar F, Korkmaz D. COVIDiagnosis-Net: A framework of deep features to classify COVID-19 patients. Appl Sci. 2021;10(9):3233.
6. Li H, Wang Y, Wan R, Wang S, Li T, Kot A. Domain generalization for medical imaging classification with deep neural networks. In Proceedings of the 33rd International Conference on Neural Information Processing Sys-tems (NeurIPS); 2020.

7. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In Proceedings of the 34th In-ternational Conference on Machine Learning (ICML); August 2017; pp. 1321–1330.
8. Rajpurkar P, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep convolutional neural networks. Sci Rep. 2017;12(1):19156.
9. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver oper-ating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837–845.
10. Hwang EJ, Giri S, Garg M. Deep learning for clinical deployment in medical imaging. Nat Rev Meth-ods Primers. 2022;2:45.
11. Troeger C, et al. Global burden of respiratory infections in 204 countries and territories: A systematic review, meta-analysis and modeling study. Lancet Infect Dis. 2024;24(4):381–412.
12. Teufel J, et al. Deep learning for pneumonia detection on chest X-rays: A systematic review and evalua-tion framework. Int J Med Inform. 2023;176:104927.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
14. Rathour A. Chest X-Ray Pneumonia Detection with Cross-Operator Validated AI System (v1.0). Zenodo. 2025.