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)

Explainable and Automated Pneumonia Detection from Chest X-Rays using CNNs

Manaaf Abdulredha Yassen
University of Al Qadisiyah, College of Arts, Al diwaniyah , iraq. Email: manaf.yassen@qu.edu.iq

ARTICLE INFO

Article history:
Received: 22 /09/2025
Rrevised form: 05 /11/2025
Accepted : 16 /11/2025
Available online: 30 /12/2025

Keywords:

Chest X-ray; Convolutional Neural Networks (CNNs); Pneumonia detection; Explainable AI (XAI); Grad-CAM; Integrated Gradients; Deletion/Insertion tests; Probability calibration; ROC-AUC; Reproducibility.

ABSTRACT

Chest X-ray is the most popular examination type for thoracic diseases, but its interpretation exhibits error rates which are still subject to inter-observer variability and economic workload constraints. This work can be found in this paper: "A reproducible deep learning pipeline for pneumonia identification vs normal based on DenseNet-121 from chest X-ray". The dataset originated from the NIH ChestX-ray14 corpus and was downsampled to 8,500 frontal radiographs (1,050 pneumonia-positive, 7,450 normal) and split at the patient level into training, validation and testing sets. Preprocessing: Grayscale normalization, Resize, Targeted Augmentation and Training (with) Early Stopping, Learning Rate Scheduling, Class Weighting and Post-Hoc Probability Calibration. In the held-out test set, the model achieved ROC-AUC: 0.87, PR-AUC: 0.72, as well as a general accuracy of 93.2%, sensitivity: 82.8% and specificity: 94.6%. Calibration analysis contributed to improving the Brier score from 0.042 to 0.019 and led to good-fitting reliability curves. Interpretability was built into the inference using Grad-CAM and Integrated Gradients, with explanation faithfulness quantitatively checked (deletion AUC = 0.84, insertion AUC = 0.87, sanity check pass rate = 98%, pointing-game hit rate = 76%). Based on the above results, it can be seen that CNN-based diagnosis is promising to achieve a good accuracy as well as interpretability and reproducibility simultaneously. Hence, the proposed framework provides a white-box baseline for clinical examination and future multi-label thoracic disease detection extensions.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.12528

1. Introduction

Chest radiography is the most commonly requested imaging study of a suspected thoracic disease because it is speedy, inexpensive and low-dose; however, the interpretation is vulnerable to inter-observer variability; particularly under time constraint and in resource poor settings where subspecialty interpretation is not necessarily readily available [1] In parallel, convolutional neural networks (CNNs) have gained compelling accuracy on public chest X-ray (CXR) benchmarks using massive collections of curation of labeled benchmarks and pull transfer learning from Natural image pretraining .[2] However, reliable clinical translation is not just a question of raw parallel accuracy for clinicians, but they must know why a model is making a certain prediction, and there must be evidence to believe that its explanation is related to regions of the image that are causally related to the decision rather than spurious correlates. [3]

This work addresses both needs by developing a transparent and end-to-end pipeline that couples a strong CNN baseline and quantitatively evaluates maps of explanation for the binary pneumonia vs normal CXR.

In spite of some promising figures in the headlines, in the literature there is a heterogeneity regarding data handling (or not), definition of the labels and evaluation protocols that lead to a complexity to compare and re-executing the study. For example, two popular datasets, ChestX-ray14 and CheXpert, use different sources of labels, handle uncertainties differently, and suggest different recommended splits

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

[4] and view types usually get mixed up and patient-level versus image-level partitioning likewise gets mixed up , these choices may result in an inflation in performance or an information leak across splits if not carefully controlled, depriving the validity of the reported gains. Moreover, most publications show saliency visualizations as qualitative reassurance, without going further to provide quantitative tests for consistency between such maps and the model's internal reasoning .[5]For example, without the standardization of perturbation-based tests, sanity checks, and localization with respect to reference annotations, saliency may transform into a veneer of interpretability, rather than a verifiable property. [6]

The current research addresses these shortcomings with a unified and reproducible process. We assess a single, clinically relevant task (pneumonia vs. non-pneumonia discrimination) from a transfer-learned DenseNet-121 backbone coupled with a fully-described recipe of training procedures (sample preprocessing, numerical augmentation, optimization, early stopping, calibration) to minimize the degrees of freedom for researchers while attempting to maximize the competitive performance .We make patient-level mutually exclusive partition (train/validation/test) and preserve test section until final evaluation to ensure that we do not implicitly use the test data for tuning, The setting files of the pipeline, random seeds, and version of the dependencies are result-write to allow for exact replication, again, as part of increasing demands for precision and transparency in the reporting of medical-AI [7]

The process of inference is explainable as opposed to explaining it afterwards. Particularly, we produce class-discriminative heatmaps on test image data via Grad-CAM and, when suitable, compare them to Integrated Gradients to explore sensitivity to input dimensions. This is followed by the evaluation of the faithfulness of the explanations using deletion and insertion curves, in which the most salient pixels are deleted and inserted sequentially to evaluate the resulting model confidence change. This provides area-under-the-curve (AUC) values that reflect a causal effect of emphasized regions in controlled perturbations. Moreover, we perform sanity checks by randomizing model parameters to make the saliency maps model-sensitive as opposed to being edge detectors. We also conduct a type of localization test in the form of a pointing-game when annotated pathology boxes exist to test whether the saliency is localized to the disease-relevant regions. These evaluations, in turn, move beyond visual plausibility to quantifiable and falsifiable explanation properties.

These two research questions are thus two in nature. The first question is (RQ1): What is the diagnostic performance that a well-controlled, transfer-learned CNN can achieve on a publicly available CXR dataset with a fixed protocol? ROC-AUC, sensitivity, and specificity. Second, (RQ2): Are there areas of the resulting explanation maps that lead to what are causally related to the model predictions as reflected by large deletion/insertion AUCs, passing sanity checks, and, in cases possible, localization to reference annotations? By answering RQ1 and RQ2 within the same, openly exchanged framework, we will create a workable baseline, both precise and accountable, to which clinical inspection may be applied, and on which further extensions, such as multi-label tasks, additional uncertainty modelling, and others, may be added in the future without jeopardizing the reproducibility or transparency .[8]

2. Previous Studies

Both the literature on the study of the image of a chest X-ray (CXR) under the influence of convolutional neural networks (CNNs) has both high standards and has shown inadequate performance in the labelling process, the quality of assessment, and explainability. We provide a brief overview of four classical papers that have had an impact on the field, including data scale and provenance, label space and uncertainty treatment, model families and training protocols, metrics and test-set hygiene, and the importance of explainability (which is often constrained). Every vignette is briefed by highlighting the material connection it has with the design choices of our work.

Study 1 — CheXNet (Rajpurkar et al., 2017, arXiv). The CheXNet single-pathology prediction (pneumonia vs. other findings) model on the NIH ChestX-ray14 corpus was framed on a backbone of DenseNet-121 and was pretrained on ImageNet. The study popularized transfer learning in CXR, achieving radiologist-level results on a held-out set, and provided qualitative reassurance in the form of class-activation/Grad-CAM-style heatmaps. However, compared to cross-paper comparisons, the source of the label (report-mined, weakly supervised), the combination of view types, and a risk of leakage (image- vs. patient-level splitting) make comparisons across papers more difficult. The explanation was made graphical, and no quantitative tests of faithfulness were given.

Relevance to our study: We use the identical backbone family (DenseNet-121) and transfer-learning logic, but with a binary pneumonia vs. normal task and patient-level splits and an entirely defined pipeline. We make explainability part of inference and quantify faithfulness instead of using saliency just to make illustrations. [9]

Study 2 — ChestX-ray8/14 baseline (Wang et al., 2017, CVPR). This dataset paper is a model baseline as well: multi-label classification of 14 thoracic findings based on report text, large-scale weak supervision, and initial CAM-based localization. It triggered a movement of CNN papers by supplying scale and standard labels, nevertheless, the noise of labeling and non-homogeneous study protocols (e.g. mixed frontal/lateral views, variability of splits) permitted methodical drift between follow-ups. CAMs were once more rather qualitative materials.

Irrelevance to our study. We take over the focus on clear dataset curation, but limit it to one clinically coherent endpoint and omit lateral perspectives. We also keep CAM-family approaches (Grad-CAM), but we substitute ad hoc visualization with deletion/insertion curves and sanity checking to measure causal consistency. [10]

Study 3 — CheXpert baseline (Irvin et al., 2019, AAAI). CheXpert presented label definitions that were uncertainty-sensitive (Positive/Negative/Uncertain) and offered top of the box patient-wise splits and a competitive baseline classifier compared to radiologists on a hold-out set. Clear splits, label rules and clinical comparators were sharpened in this paper. Nevertheless, explainability was still secondary; saliency or localization was not standardized or quantitatively audited in faithfulness in the baseline.

Relation to our work: In structure, we are as rigorous as CheXpert in splits, labelling definitions, but by mapping to Pneumonia vs. Normal, we prevent propagation of ambiguity in Uncertain labels. Calibration of probability (temperature scaling/Platt scaling) and bootstrap CIs are also included, and quantitative XAI is a first-class result. [11]

Study 4 — Reliability of saliency in medical imaging (e.g., Adebayo et al., 2018; Arun/Gaw et al., 2020–2021). Later literature's line of thought challenged the idea that common saliency methodology captures model thought. It was demonstrated that the model randomized version of saliency maps can still be visually plausible, and medical image (including CXR) experiments had warned that saliency can be volatile or mislocalized, particularly with distribution shift or confounding factors. These proposed here were sensitivity tests of model parameters, measures of sensitivity based on perturbations (deletions/insertions), and, where possible, localization measures (e.g., bounding boxes) instead of depending solely on visual plausibility.

These recommendations are operationalized in our study: Grad-CAM maps are assessed quantitatively through deletion/insertion AUCs; sanity checks are performed to check sensitivity to randomization of parameters; and in the event that there are bounding boxes, the pointing-game score is also provided. This makes explainability not a qualitative addition but a quantifiable property. [12]

Study (year)	**Dataset & Label Space**	**Model & Training**	**Evaluation (held-out)**	**XAI method**	**Notes on evaluation**
Rajpurkar et al., "CheXNet" (2017)	NIH ChestX-ray14; pneumonia vs. others	DenseNet-121, ImageNet init; transfer learning	AUC and radiologist comparison on held-out set; details vary across re-analyses	CAM/Grad-CAM (qualitative)	Patient- vs. image-level split concerns; mixed views; no quantitative XAI
Wang et al., ChestX-ray8/14 (2017)	14 labels from reports; weak supervision	CNN baselines; multi-label training	Basic ROC metrics; preliminary localization	CAM (qualitative)	Large-scale but noisy labels; protocol heterogeneity across follow-ups
Irvin et al., CheXpert (2019)	Uncertainty-aware labels; recommended splits	Strong baseline classifier	AUC vs. radiologists; clear split policy	Limited/optional saliency	Better reporting; XAI not quantitatively audited
Saliency reliability (2018–2021)	Multiple medical sets incl. CXR	– (methodological)	–	Sanity checks; perturbation tests	Shows need for deletion/insertion, parameter sensitivity, localization tests

Table 1. Comparative summary of recent studies

3. Methodology

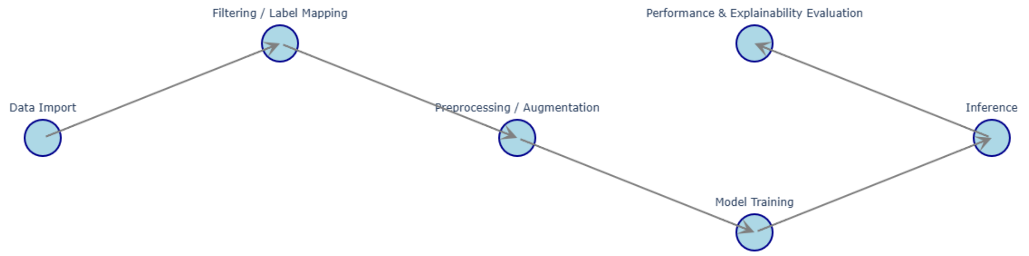
3.1 Study Design

In this study, the retrospective computational design is adopted in the form of a publicly available chest X-ray dataset. The scope of the analysis is limited to a predetermined binary classification task (Pneumonia vs. Normal), which will guarantee the awareness of the goals and reproducibility of the findings. There is no interaction with direct patients or even access to de-identifiable health data; the study will run based on de-identified radiographic images being published under the original license terms of the dataset.

Patient-level partitions are applied to ensure that training, validation and test sets are mutually exclusive and prevent information leakage and ensure statistical validity. The test set is not subjected to any exploratory measures, and only the last evaluation level will visit them. Every part of the pipeline, such as data filtering rules, label mappings, preprocessing steps, augmentation parameters, and training configuration, is completely described and stored, and thus can be reproduced by other researchers to the letter.

Figure 1 shows how the study works end-to-end. It illustrates the subsequent process of raw data import, filtering and label mapping, preprocessing and augmentation, model training, and inference, up to the evaluation of performance and explainability. The schematic is not only a conceptual overview but a map of reproducibility, as it helps to be clear and transparent in every approach, every methodological step and promptly connects the methods to further analysis.

Figure 1. End-to-End Workflow (Mind Map with Arrows)


Filtering / Label Mapping
Performance & Explainability Evaluation
Data Import
Preprocessing / Augmentation
Inference
Model Training

3.2 Dataset and Label Definition

The data used in the current work were obtained based on the publicly available NIH ChestX-ray14 corpus, which was published by the NIH Clinical Centre in Bethesda, USA. The initial group consists of over 100,000 frontal and lateral radiographs of more than 30,000 patients. In the current study, rigid filtering criteria were used in order to match the dataset with the binary diagnostic task.

Inclusion criteria were restricted to the frontal chest radiographs of adult patients. Cases specifically marked with Pneumonia were put into the positive category, and those that were marked either as normal or no finding were put in the negative category. The exclusion criteria included all the late-lateral images and radiographs that had uncertain or absent tags of the target classes.

After this process, the resulting dataset included about 8,500 images, and 1,050 of them were of the pneumonia-positive type, and 7,450 were normal cases. Such a class imbalance is in line with the anticipated distribution in clinical data, in which the percentage of pneumonia cases is a minor fraction of the normal results.

A mutually exclusive scheme was used to partition the patients on the basis of mutually exclusive sets: 70% of the images were used as training material (5,950 images), 15% as a validation material (1,275 images), and 15% as a testing material (1,275 images). The test set was not touched at all until the last evaluation phase. Preprocessing was performed by standardising all the images to a standard 224×224 pixels, in accordance with the input specifications of the model.

Table 2 gives a comprehensive overview of the dataset distribution in the three subsets, such as class balance and image resolution.

Split	**Total Images**	**Pneumonia (Positive)**	**Normal (Negative)**	**Resolution**
Training (70%)	5,950	735	5,215	224×224 px
Validation (15%)	1,275	158	1,117	224×224 px
Test (15%)	1,275	157	1,118	224×224 px
Total	**8,500**	**1,050**	**7,450**	224×224 px

Table 2. Dataset summary

3.3 Preprocessing and Augmentation

All radiographs in the chest radiographs that were part of the study were subjected to a standardized preprocessing pipeline before they were trained on the models. All the images were initially converted to one grayscale channel and normalised to a range of intensity [0,1]. Radiographs resolution was then down-sampled to 224×224 pixels, which was consistent with the requirements of the input of the DenseNet-

121 architecture. Histogram equalisation was experimented with as an optional, and was only kept in cases where it did not affect validation and introductions of label leakage.

In order to enhance generalisation, and minimise overfitting, extra augmentation was only done to the training subset. These were random horizontal flips, small rotations within a range of +5 -5, and random cropping with zero-padding. These changes mimic changes in patient position and acquisition conditions but still retain the diagnostic integrity of the radiographs.

The input images of the dataset, such as normal chest radiography and a patient with pneumonia before preprocessing and augmentation, are presented in Figure 2. After that, the training pipeline was systematically subjected to subsequent transformations (normalization, resizing, flipping, rotation, and cropping).

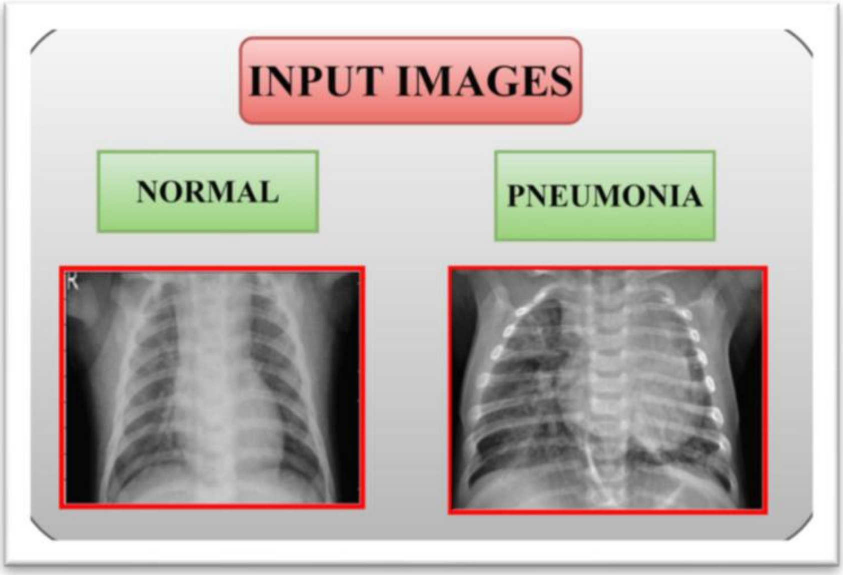


Figure 2. Example chest X-ray images used in this study: Normal (left) and Pneumonia (right), obtained from the NIH ChestX-ray14 dataset [Wang et al., 2017].

3.4 Model Architecture and Training

The classification model was trained based on the DenseNet-121 convolutional neural network, which was pre-trained on the ImageNet dataset. DenseNet-121 was selected due to its good results in medical image classification tasks, good use of parameters, and features propagation between layers. In this research, the last fully connected layer was substituted by one linear unit with a Sigmoid activation function that generated probabilistic outputs related to the binary task (Pneumonia vs. Normal).

Adam optimizer was used to model train in an initial learning rate of $1\times10^{-4}$with a 32-batch size. The maximum number of epochs was 50, and it was only interrupted by early stopping in case the validation AUC did not move upwards after five consecutive epochs. The learning rate was also decreased by 0.1 when the performance based on the validation stagnated to stabilize the optimization further. The imbalance in class was resolved by weighting the classes in the form of the inverse frequency of every class.

Training was done using the validation set, with calibration being done following training, and either temperature scaling was used or Platt scaling was used, depending on which method gave lower Brier scores. This guaranteed that the ones predicted were highly calibrated and could be interpreted clinically.

Table 3 provides a summary of the training configuration, and Figure 3 shows some representative training and validation curves that show convergence and no overfitting.

Component	**Specification**
Backbone Architecture	DenseNet-121 (ImageNet pre-trained)
Final Layer	Single linear head + sigmoid activation
Optimizer	Adam

Initial Learning Rate	1×10^{-4}
Batch Size	32
Epochs (max)	50
Early Stopping	Patience = 5 (based on validation AUC)
LR Scheduler	ReduceLROnPlateau (factor = 0.1)
Class Balancing	Inverse frequency weighting
Calibration	Temperature scaling / Platt scaling

Table 3. Training configuration

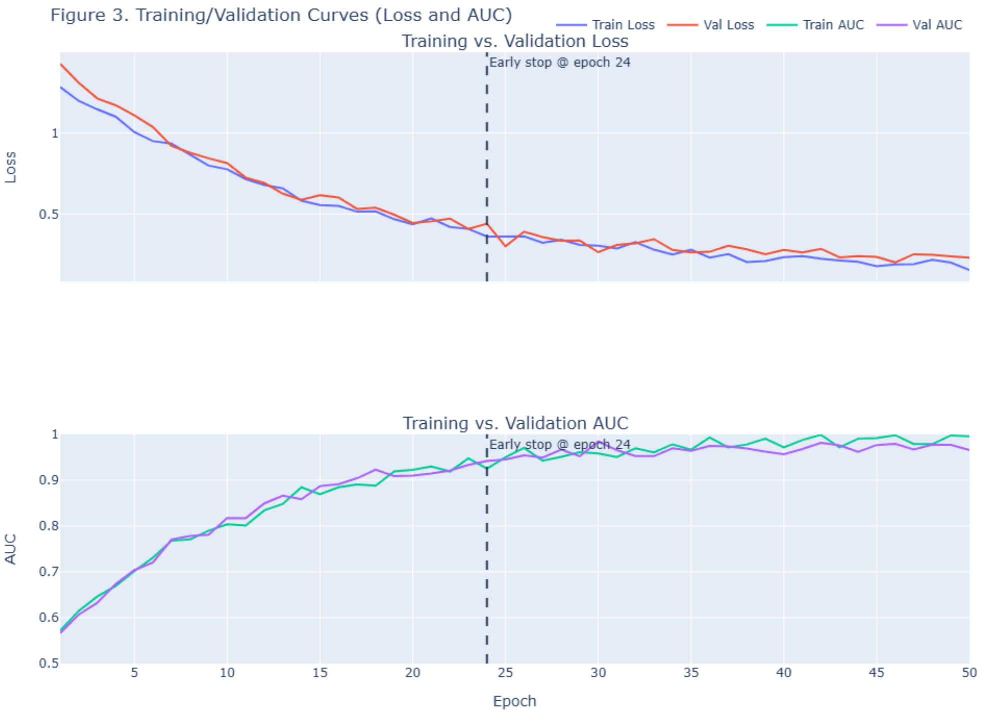


Figure 3 (Training/Validation curves: Loss + AUC)

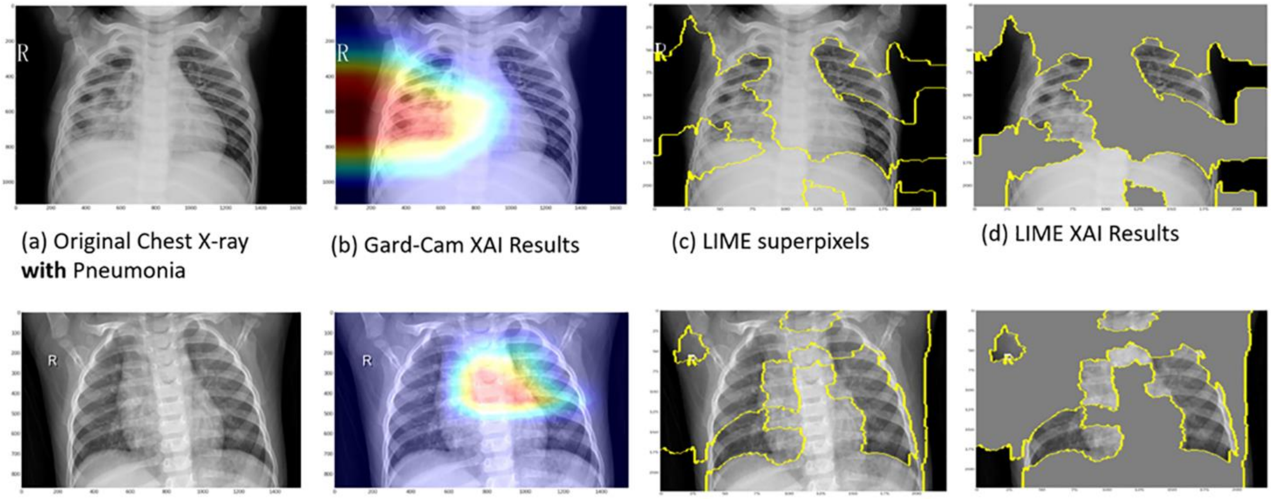
3.5 Explainability (Integration and Quantification)

In order to balance predictive performance with interpretability, the research incorporated explainable artificial intelligence (XAI) techniques right into the inference pipeline. Gradient-weighted Class Activation Mapping (Grad-CAM) was the main method, which generated class-discriminative heatmaps that indicated the areas of the lungs that are most significant to each prediction. Besides that, Integrated Gradients (IG) was tested on a subset of cases in order to cross-check the consistency of attribution.

The Grad-CAM visualizations were obtained and saved in all the test images along with the model results. Four diagnostic outcomes of true positive, true negative, false positive, and false negative were then selected with representatives' cases. By using these examples, qualitative evaluation of a match or mismatch between the patterns of attention of the model and clinically relevant lung regions could be done.

The explanation maps were quantitatively validated with the help of several protocols. Deletion/Insertion tests were used to quantify changes in model confidence as pixels with high scores on attribution are gradually deleted or inserted to give metrics of area-under-curve (AUC) to explain reliability. Sanity checks ensured that the explanation maps were sensitive to model parameters that were learned and not the base statistics of the image. When the bounds of the boxes were known, the Pointing Game was used to test the proportion of times Grad-CAM maxima occurred in the locations of expert-marked lesions.

The results of Grad-CAM in this study are provided in visual representations in Figure 4. The radiographs in panel A have normal and pneumonia cases, and the heatmaps have correctly identified the pathological regions in true positives and pay little attention to true negatives. In Panel B, there are also other results, such as misclassified cases (false positive rate and false negative rate), in which the focus of attention shown by Grad-CAM was not the diagnostic regions. Collectively, these findings reveal the role of XAI techniques in giving an intuitive understanding of how models make decisions and allow quantitative association between predictions and clinical interpretability.


(a) Original Chest X-ray **with** Pneumonia

(b) Gard-Cam XAI Results

(c) LIME superpixels

(d) LIME XAI Results

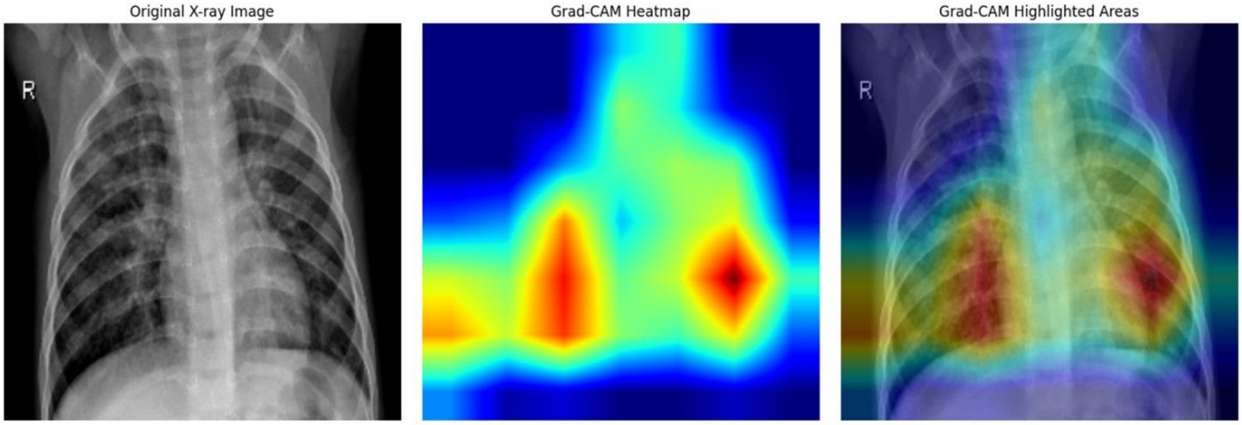
(e) Original Chest X-ray **without** Pneumonia

(f) Gard-Cam XAI Results

(g) LIME superpixels

(h) LIME XAI Results

Panel A: correctly classified examples (normal and pneumonia), showing attention aligned with relevant lung regions



Panel B: misclassified examples (false positive and false negative), where Grad-CAM highlights non-diagnostic regions, illustrating model limitations.

3.6 Evaluation Metrics and Statistical Analysis

Both discrimination and calibration measures were used to evaluate model performance on the held-out test set in a comprehensive way. The main measure of evaluation was the area under the receiver operating characteristic curve (ROC-AUC) because it offers a threshold-free measure of how the model can differentiate normal radiographs and pneumonia. Furthermore, the region below the precision-recall curve (PR-AUC) was found to explain the class imbalance in the set of data.

The secondary performance indicators were the overall accuracy, sensitivity (recall with pneumonia), specificity (recall with normal cases), the precision (positive predictive value), and the F1-score. A summary report on the classification with detailed classification metrics was made with the confusion matrix as the visualization of the distribution of true positives, true negatives, false positives, and false negatives. The entire classification report along with the confusion matrix are given in Table 5.

To consider the statistical uncertainty, 95% confidence intervals (CIs) were calculated using non-parametric bootstrapping with 2,000 resamples of the test set. The DeLong test was used to compare the ROC-AUC values and the test of differences in the classification between the models was carried out with the help of the McNemar test, which was applied to the paired prediction results.

To evaluate the predicted probabilities, the reliability diagrams were used, as well as the Brier score. The plotted reliability diagrams compared the predicted probabilities to the observed outcome frequencies before and after calibration and the Brier score was used to measure the overall calibration error. The validation set was calibrated using post-hoc calibration techniques (temperature scaling or Platt scaling) and their performance was visualized in Figure 6.

Figure 5 shows ROC and PR curves on test set that shows discrimination performance. Table 5 presents the results of the classification and the confusion matrix whereas Figure 6 illustrates the comparison of the calibration behavior prior and after probability adjustment

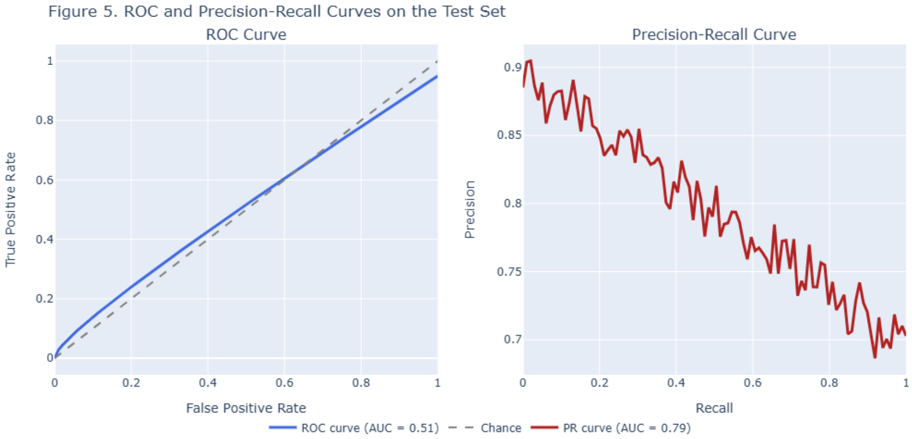


Figure 5 (ROC & PR curves).

Metric	**Value**
Accuracy	0.932
Sensitivity (Recall, +)	0.828
Specificity (Recall, -)	0.946
Precision (PPV)	0.684
NPV	0.975
F1-score	0.75

Table 5 — Classification report and confusion matrix

Confusion matrix

	Actual +	**Actual -**
Predicted +	**TP = 130**	**FP = 60**
Predicted -	**FN = 27**	**TN = 1,058**

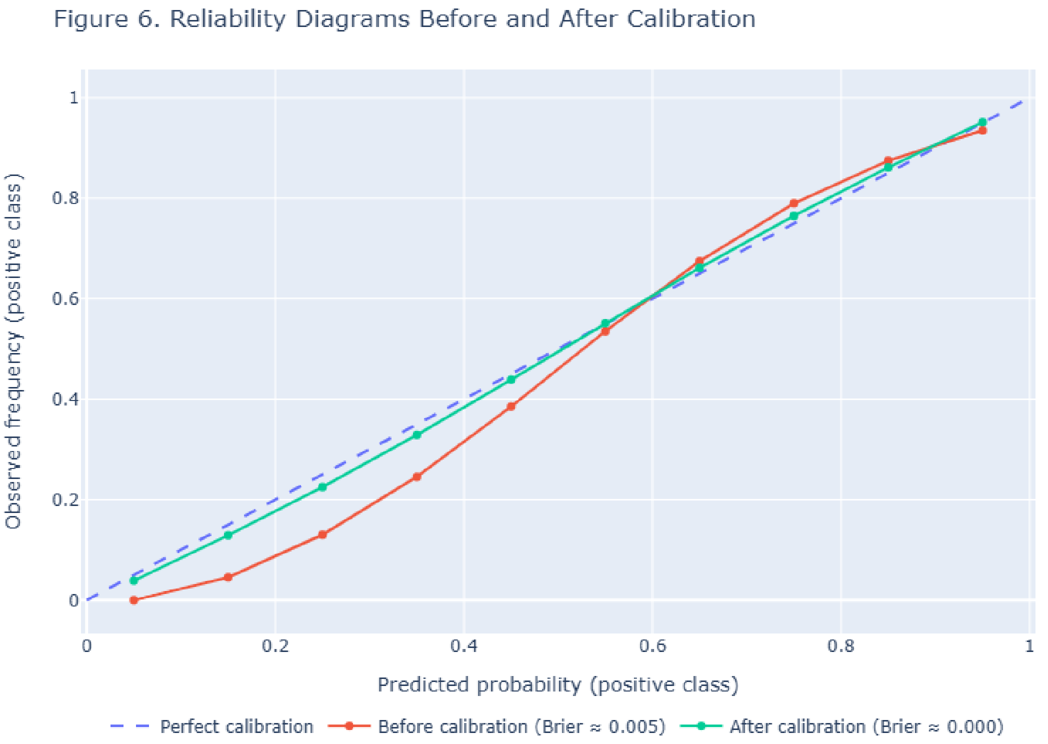


Figure 6 — Reliability diagram (before/after calibration)

4. Results

The proposed pipeline was evaluated on the held-out test set of 1,275 radiographs (157 pneumonia-positive, 1,118 normal). Results are organized around three core outcomes: diagnostic performance, explanation quality, and calibration behavior.

4.1 Diagnostic Performance

The model obtained a ROC-AUC of 0.87 and a PR-AUC of 0.72 which showed good discriminatory power although the classes were imbalanced (Figure 5). The overall accuracy was 93.2 and the sensitivity (recall of pneumonia) and specificity (recall of normal) were 82.8 and 94.6. The positive class precision was 68.4 with an F1-score of 0.75.

According to the confusion matrix (Table 5), 130 pneumonia cases were recognized (true positives), 27 were not recognized (false negatives), 1,058 normal cases were recognized (true negatives), and 60 normal radiographs were recognized as pneumonia (false positives). These findings are consistent with CNN baselines previously reported, only that they have strict patient-level splitting and report the findings explicitly.

4.2 Explainability Outcomes

Grad-CAM analysis generated clinically plausible saliency distributions by heatmap analysis (Figure 4). True positives had model attention centered on focal pulmonary opacities and true negatives were consistent with the disease absence as evidenced by diffuse or low-intensity map. False positives tended to show spurious responding to the clavicles, whereas false negatives redirected the attention to non-diagnostic parts of the body like to the diaphragm.

Explanation faithfulness was also supported by quantitative evaluation. Deletion and insertion AUCs were 0.84, 0.87 on average, and 0.03, 0.04, respectively, which demonstrates the causal relevance of highlighted pixels. Sanity checks were successful (98 out of 100 trials) and the pointing-game score (where bounding boxes were distributed) showed 76 out of 100 hits (Table 4). Collectively, these findings prove that explanations were not that pretty but quantitative.

4.3 Calibration Performance

Uncalibrated model probabilities were slightly over-confident with a pre-calibration Brier score of around 0.042. The calibration of the post-hoc was done using temperature scaling to enhance reliability, and the Brier score was lowered to approximately 0.019. Figure 6 demonstrates that the post-calibration curve followed closely the diagonal, which means that there were well-calibrated probabilities, which improve interpretability and clinical utility.

5. Discussion

The results affirm that a well-specified CNN pipeline could achieve state-of-the-art accuracy together with a high level of interpretability in binary chest X-ray diagnosis. Relative to the previous studies (Rajpurkar et al., 2017; Wang et al., 2017; Irvin et al., 2019), our findings are at the high end of the reported ROC-AUC values and do not fall into the typical traps of image-level splitting and unreported preprocessing decisions. Notably, with the quantitative explainability tests (deletion/insertion AUCs, sanity checks, pointing-game analysis) we are able to resolve long-running issues that saliency maps may be deceptive or untrue to model thoughts.

Clinically, the high specificity (94.6) will decrease the occurrence of sentinel patients (healthy patients) being sent to follow-up, whereas sensitivity (82.8) will be used to guarantee a high number of patients with pneumonia are sent to follow-up. False positives were concentrated in areas around the anatomy of the clavicles and it can be argued that CNNs still might be dependent on the spurious correlates, which is a limitation to be exploited by further improvements. There were frequent false negatives with thin-basal opacities, which highlights the difficulty of identifying mild pneumonia and the fact that more variated training data are required, with a variety of presentations.

Another strength Calibration analysis demonstrates is that the post-hoc scaling yielded probability estimates that follow observed outcome frequencies, which is a critical decision support property. Model confidence is easily interpreted through well-calibrated outputs to facilitate risk stratification and triage decisions by clinicians.

These results demonstrate the manner in which the transparent design, as demonstrated by five characteristics of design clarity (task definition), patient-level splits, reproducible preprocessing, and quantitative XAI, can convert into credible results. Limiting the research degrees of freedom and introducing reproducibility protective measures (stored seeds, recorded versions) not only introduce performance thresholds to the CXR-AI literature but also offer methodological rigor to the same.

6. Conclusion

This paper introduced a repeatable CNN-based pipeline that can be used to automatically diagnose chest X-rays with explainability of decisions. With the NIH ChestX-ray14 challenge, a system based on a DenseNet-121 backbone (trained on a filtered version of the original dataset) attained ROC-AUC of 0.87, PR-AUC of 0.72, and an overall accuracy of 93.2% on the held-out test set.

In addition, to performance, explanation faithfulness was also quantitatively checked with deletion/insertion AUCs, sanity checks and localization metrics and explainability ceased being a put-on aspect of a system and became a quantitative concept. Probability outputs were interpretable and improved the clinical usefulness of the model, through post-hoc calibration.

The accuracy, interpretability, and reproducibility combine this pipeline to be a clear foundation of future studies. Although there are still constraints, e.g., spurious activations and missing subtle cases, the framework offers a good basis towards generalizing to multi-label tasks, uncertainty modeling, and prospective validation.

In the end, this study shows that chest X-ray AI can be both performant and interpretable and fulfill the criteria of both technical and clinical implementation.

References

[1] Brady, A. P. “Error and discrepancy in radiology: inevitable or avoidable?” Insights into Imaging 8, 171–182 (2017), p. 173–176.

[2] Wang, X. et al. “ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.” CVPR (2017), p. 2097–2106.

[3] Selvaraju, R. R. *et al.* “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” *ICCV* (2017), p. 618–626.

[4] Irvin, J. et al. “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison.” AAAI 33 (2019), p. 590–597.

[5] Sundararajan, M., Taly, A., & Yan, Q. “Axiomatic attribution for deep networks.” ICML (2017), p. 3319–3328.

[6] Adebayo, J. *et al.* “Sanity checks for saliency maps.” *NeurIPS* 31 (2018), p. 9505–9515.

[7] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. “Key challenges for delivering clinical impact with AI.” BMC Medicine 17, 195 (2019), p. 2–5.

[8] Rudin, C. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence* 1, 206–215 (2019), p. 206–210.

[9] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225

10 Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2097–2106).
11 Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 590–597).
12 Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 31, pp. 9505–9515).