Pneumonia Detection Using Deep Learning Models: A Comparative Study on the RSNA Dataset

Abdul Haseeb
Department of Computer Science
UET Lahore, New Campus
Lahore, Pakistan
2021se22@student.uet.edu.pk

Zain Ali
Department of Computer Science
UET Lahore, New Campus
Lahore, Pakistan
2021se23@student.uet.edu.pk

Umar Waris
Department of Computer Science
UET Lahore, New Campus
Lahore, Pakistan
2021se28@student.uet.edu.pk

***Abstract*— Pneumonia is a serious lung infection that affects millions worldwide and is often diagnosed using chest X-ray images. However, interpreting X-rays manually can be challenging and time-consuming. This study presents a deep learning approach to automatically detect pneumonia from chest X-rays using the RSNA Pneumonia Detection Challenge dataset. We started by building a simple convolutional neural network (CNN) as a baseline and gradually moved toward advanced transfer learning models such as ResNet, EfficientNet, Xception, and DenseNet. Each model was trained using a standardized pipeline with proper image preprocessing and class imbalance handling. After several rounds of fine-tuning and evaluation, DenseNet201 achieved the best results with an AUC of 0.85, recall of 0.77, and accuracy of 80%. These results show that deep learning, when carefully designed and fine-tuned, can provide reliable support for pneumonia detection and assist medical professionals in making quicker decisions.**

Keywords—Pneumonia Detection, Deep Learning, Chest X-ray, RSNA Dataset, DenseNet201, Medical AI, Transfer Learning

I. Introduction

A. Background

Pneumonia is a serious lung infection that can affect anyone, but it's especially dangerous for children, older adults, and those with weak immune systems. It causes the air sacs in the lungs to fill with fluid, making it hard to breathe and if not diagnosed early, it can become life-threatening. According to the World Health Organization, pneumonia accounts for millions of hospitalizations and deaths each year, particularly in low-resource settings [1].

Chest X-rays are one of the most commonly used tools for detecting pneumonia. However, interpreting X-ray images can be challenging, even for experienced radiologists. Factors such as image quality, overlapping structures in the lungs, and subtle visual patterns can make accurate diagnosis difficult. In busy clinical environments, human fatigue and time constraints can also lead to delayed or missed diagnoses. This opens the door for automated systems to assist with the diagnostic process.

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have shown strong potential in medical image analysis. Several studies have demonstrated that CNN-based models can achieve expert-level performance in classifying chest X-rays [3]. These models are capable of learning complex patterns in the images and making consistent predictions. However, the effectiveness of such models greatly depends on the quality and structure of the dataset used for training. After evaluating several datasets, we selected the RSNA Pneumonia Detection Challenge dataset due to its clean annotations, pneumonia-specific labeling, and realistic clinical imaging conditions [4]. This dataset served as a reliable base to train, fine-tune, and evaluate deep learning models for pneumonia detection.

B. Research Question

Which deep learning model provides the most reliable and accurate detection of pneumonia in chest X-rays when trained on the RSNA Pneumonia Detection Challenge dataset, considering both performance metrics and real-world clinical relevance?

C. Hypothesis

We believe that deep learning models, especially those using transfer learning, can learn to identify subtle signs of pneumonia in chest X-rays if trained on a well-structured and balanced dataset. Specifically, we expect DenseNet201 to outperform other architectures due to its unique dense connections that help preserve feature information across layers. By combining this model with proper preprocessing, class weighting, and fine-tuning techniques, we hypothesize that it can achieve strong recall and AUC scores, making it a reliable choice for real-world pneumonia detection tasks.

D. Objectives

- To compare multiple deep learning models for pneumonia detection using chest X-rays.
- To apply consistent preprocessing, including normalization, resizing, and RGB conversion.
- To address class imbalance through class weighting.
- To train models using both frozen and fine-tuned transfer learning approaches.
- To evaluate performance using AUC, recall, precision, F1-score, and accuracy.
- To identify the most reliable model for practical and clinical use.

II. Literature Review

A. Deep Learning in Medical Image Analysis

Deep learning, especially convolutional neural networks (CNNs), has transformed how medical images are analyzed. These models are capable of identifying patterns and features

in X-ray images that may not be obvious to the human eye. One of the landmark contributions in this field was CheXNet, which used a DenseNet121 model trained on over 100,000 chest X-rays and achieved radiologist-level performance in pneumonia detection [2]. This showed that deep CNNs can support or even match clinical expertise in image-based diagnosis.

Another notable effort was CheXpert, a large-scale dataset and benchmark developed to address uncertainty in medical image labels. The project helped highlight how deep learning models can be trained on noisy, real-world data and still achieve reliable results [3].

B. Transfer Learning and Pretrained Models

Training CNNs from scratch often requires massive datasets and long training times, which aren't always practical in healthcare research. That's why many studies now use transfer learning—where models pre-trained on general image datasets like ImageNet are fine-tuned for medical tasks [5]. This approach works well when medical datasets are smaller or when training time is limited. Models like ResNet, MobileNet, and EfficientNet have all been tested with this technique, often yielding promising results for detecting diseases such as pneumonia, tuberculosis, and COVID-19 in X-rays [6].

C. Challenges with Dataset Selection

Many previous studies used general-purpose datasets that either lacked proper labeling or included multiple overlapping disease labels. This often led to noisy learning and poor generalization. That's why more recent research has shifted toward better-curated datasets like the RSNA Pneumonia Detection Challenge dataset, which includes bounding box annotations and pneumonia-specific labels [4]. Its quality and focus have made it a go-to choice for researchers interested specifically in pneumonia detection.

D. Model Performance in Pneumonia Detection

Research comparing different architectures, like VGG, ResNet, EfficientNet, and DenseNet, has shown that deeper and more structured models often perform better in extracting relevant features from chest X-rays. A study by Islam et al. [6] demonstrated how an ensemble of CNNs improved overall accuracy, while Chouhan et al. [7] found that fine-tuned transfer learning models were more robust and generalized better across unseen data. However, success still relies heavily on careful preprocessing, proper data splitting, and handling of class imbalance all areas our study paid special attention to.

E. Contribution of This Study

Our work builds on this existing research by comparing several deep learning models using a carefully constructed pipeline. Instead of focusing only on accuracy, we prioritize recall and AUC—two metrics that reflect real clinical needs, such as minimizing missed pneumonia cases. We also ensure that all models were trained under the same preprocessing and evaluation setup to allow a fair comparison. This approach helps identify the most reliable model for pneumonia detection using the RSNA dataset under practical conditions.

III. Dataset and Preprocessing

A. Dataset Description

- This study uses the RSNA Pneumonia Detection Challenge dataset, which contains over 26,000 frontal-view chest X-ray images in DICOM format.
- Each image corresponds to a unique patient and is labeled as either Normal (0) or Pneumonia (1).
- For pneumonia cases, the dataset also provides bounding box annotations to highlight the affected lung regions.
- The dataset includes two main CSV files:
    - **stage_2_train_labels.csv**, which contains image-level labels and bounding boxes.
    - **stage_2_detailed_class_info.csv**, which provides detailed class descriptions.
- After merging both files using patient IDs, we created a simplified version of the dataset where each image was labeled as either pneumonia-positive or normal.
- We also confirmed that all records were unique after applying a duplicate check.
- One challenge we noticed early was the class imbalance: about 78% of images are normal, while only 22% show pneumonia.
- This imbalance required careful handling to prevent the model from favoring the majority class during training.

B. Data Preprocessing

Since the original images were in DICOM format, we performed several preprocessing steps to prepare them for deep learning models:

1) DICOM to PNG Conversion:

- All X-rays were converted to PNG format to make them easier to process using standard image libraries.

2) Image Resizing:

- Depending on the model architecture, images were resized to standard input shapes, such as 224×224 for ResNet and DenseNet, 240×240 for EfficientNetB1, and 299×299 for Xception.
- This made the inputs compatible with ImageNet pre-trained weights..

3) Normalization:

- Pixel values were normalized to a range between 0 and 1 to help stabilize model training.

4) Channel Conversion:

- Although chest X-rays are grayscale, most pre-trained CNNs expect three-channel (RGB) input.

- We handled this by stacking the grayscale image into three identical channels.

5) Stratified Train-Test Split:

- The dataset was split into 80% training and 20% validation sets using stratified sampling, which preserved the class ratio in both sets and helped avoid bias during model evaluation.

C. Dataset Handling Class Imbalance:

To reduce the effect of class imbalance, we applied a combination of techniques.

1) Class Weighting:

- During model training, we assigned higher weights to pneumonia samples so that the model treated them as equally important despite their lower count.

IV. Methodology

A. Model Selection Strategy

To evaluate how well deep learning models can detect pneumonia from chest X-rays, we followed an iterative development approach. We started by building a basic CNN from scratch to establish a performance baseline. Once that was in place, we moved on to transfer learning using pre-trained models such as ResNet50, ResNet101, EfficientNetB0, EfficientNetB1, Xception, and DenseNet201.
Each model was trained in two stages:

- First, we froze the base layers and trained only the top classification layers to take advantage of pre-learned features.
- Next, we unfroze selected layers and fine-tuned the entire model using a lower learning rate, allowing the model to adapt to pneumonia-specific features in the RSNA dataset.

This two-stage training helped improve performance while preventing overfitting in the early phases.

B. Models Used

- Basic CNN:
  A simple architecture with a few convolutional and pooling layers, followed by fully connected layers. This served as our baseline.

- ResNet50 & ResNet101:
  Known for their residual connections which help prevent vanishing gradients in deep networks. We experimented with both frozen and fine-tuned versions.

- EfficientNetB0 & B1:
  Lightweight models that balance depth, width, and resolution. EfficientNetB0 achieved high recall but lower accuracy overall.

- Xception:
  A deeper model that uses depthwise separable convolutions. Performed consistently across both frozen and fine-tuned settings.

- DenseNet201 (Final Model):
  This model uses dense connections between layers, allowing better gradient flow and feature reuse. It showed the most balanced performance after fine-tuning and became our final choice.

C. Training Configuration

Model performance was evaluated using the following metrics:

- Accuracy: Overall correctness of predictions
- Precision: How many predicted pneumonia cases were correct
- Recall: How many actual pneumonia cases were detected
- F1-Score: Balance between precision and recall
- AUC (Area Under the Curve): Measures overall model confidence and discrimination ability

Special focus was placed on recall and AUC, since missing a pneumonia case is more critical in clinical practice than a false positive.

D. Tools and Environment

All experiments were implemented using Python with TensorFlow and Keras. The models were trained using Google Colab, which provided free GPU access and enough memory to handle the dataset efficiently.

V. Results and Analysis

A. Model Performance

We evaluated all models using five key metrics: accuracy, precision, recall, F1-score, and AUC. These metrics help give a complete picture of how well each model performed, especially in detecting pneumonia cases. The results are summarized in the table below.

- Basic CNN achieved 77.5% accuracy, 0.71 recall, and 0.82 AUC — decent baseline performance.
- Fine-Tuned CNN reached 80.76% accuracy but recall dropped to 0.32, showing overfitting.
- ResNet50 (Frozen) showed limited learning with an AUC of 0.72.
- ResNet101 (Frozen) had 74.61% accuracy, 0.46 recall, and 0.74 AUC.
- ResNet101 (Fine-Tuned) gave unstable results with 22.52% accuracy and 1.00 recall but only 0.55 AUC.

- EfficientNetB0 recorded 72.6% accuracy and high recall of 0.83, though other metrics were inconsistent.
- EfficientNetB1 failed to generalize with 77.48% accuracy but 0.00 on all other key metrics.
- Xception (both frozen and fine-tuned) maintained consistent performance with 78.34% accuracy and 0.84 AUC.
- DenseNet201 (Final Model) delivered the best results: 80.0% accuracy, 0.77 recall, 0.69 F1-score, and 0.85 AUC.

B. *Final Model: DenseNet201*

Among all the tested models, DenseNet201 stood out as the most balanced and reliable. It achieved the highest AUC (0.85), which indicates strong discrimination between pneumonia and normal cases. The recall of 0.77 is particularly important, as it shows the model was able to correctly detect most pneumonia cases, which is critical in a medical setting where missing a case can be dangerous.

The model also had a solid F1-score (0.69), reflecting a good trade-off between precision and recall. While some other models like EfficientNetB0 showed slightly higher recall, they lacked in consistency across other metrics.

C. *Learning Curves and Training Behavior*

During training, we observed that models with frozen layers often plateaued early. Once we switched to fine-tuning (unfreezing layers and lowering the learning rate), performance improved significantly. This was especially true for DenseNet201, where both the training and validation loss steadily decreased, and accuracy improved over time.

The learning curves also helped us catch overfitting early in some models, like ResNet101, which showed unusually high recall but very low overall accuracy — a sign that the model had memorized the training data but failed to generalize.

D. *Clinical Implications*

- In medical diagnosis, recall is more important than , missing a pneumonia case can be dangerous.
- A model that flags possible cases (even with some false positives) is safer than one that misses real cases.
- DenseNet201's high recall (0.77) makes it well-suited for clinical support, helping reduce missed diagnoses.
- Its AUC of 0.85 reflects strong confidence and decision-making ability, important in real-world usage.
- This balance of metrics positions DenseNet201 as a reliable screening tool to assist radiologists in hospitals.

VI. Discussion

A. *Why DenseNet201 Worked Best*

- DenseNet201 uses dense connections between layers, which helped preserve detailed pneumonia-related features.
- It performed well after a two-stage training process: frozen layers first, then careful fine-tuning.
- Class weighting and balanced preprocessing contributed to its stable and consistent results.
- It achieved the best balance of recall, F1-score, and AUC among all tested models.

B. *Challenges with Other Models*

- ResNet101 (fine-tuned) showed unstable results, with 1.00 recall but very poor accuracy (22.52%).
- EfficientNetB1 failed to learn meaningful patterns and performed poorly across all metrics.
- EfficientNetB0 had high recall (0.83) but lacked consistency in other scores.
- Xception delivered decent overall performance but never outperformed DenseNet201.

C. *Lessons from the Training Process*

- Fine-tuning worked best when paired with a low learning rate and gradual unfreezing of layers.
- Using a consistent preprocessing pipeline was crucial for fair and stable comparisons across models.
- Accuracy alone was not enough, recall and AUC were more helpful in understanding clinical performance.
- Class imbalance had a noticeable effect, but class weighting and augmentation helped counter it.

D. *Real-World Relevance*

- DenseNet201's strong recall (0.77) and AUC (0.85) make it a practical option for assisting radiologists.
- It can help flag high-risk pneumonia cases quickly, reducing diagnostic delays.
- Useful in resource-limited settings, where access to expert radiologists is limited.
- The model does not replace human decision-making but acts as a reliable support system.

VII. Conclusion and Future Work

A. *Summary of Findings*

- We compared several deep learning models for pneumonia detection using chest X-rays from the RSNA dataset.

- DenseNet201, after fine-tuning, achieved the best results with 80% accuracy, 0.77 recall, and 0.85 AUC.
- A consistent training pipeline and proper class imbalance handling were key to stable performance.
- Models like ResNet and EfficientNet showed potential but lacked the balance and reliability of DenseNet201.

B. *Practical Implications*

- Deep learning models like DenseNet201 can help reduce the time needed for diagnosis and catch pneumonia cases earlier.
- Such tools can be especially valuable in clinics with limited staff, providing automated second opinions.
- High recall ensures fewer missed diagnoses, which is vital in real-world healthcare applications.

C. *Future Work*

- Implement Grad-CAM or similar techniques to visualize and explain model predictions for better interpretability.
- Experiment with partial unfreezing of DenseNet layers to further reduce overfitting risks.
- Test ensemble methods by combining predictions from multiple high-performing models.
- Optimize performance using automated hyperparameter tuning tools like Keras Tuner.
- Explore real-time deployment in clinical environments, possibly with integration into radiology systems.

D. *Figures and Tables*

Table 1: Performance Metrics of Deep Learning Models for Pneumonia Detection

Model Variant	Accuracy (%)	Recall
Basic CNN	77.5	0.71
Fine-Tuned CNN	80.76	0.32
Fine-Tuned CNN (Retry)	75.33	0.74
ResNet50 (Frozen)	77.44	0.00
ResNet101 (Frozen)	74.61	0.46
ResNet101 (Fine-Tuned)	22.52	1.00
EfficientNetB0	72.6	0.83
EfficientNetB1	77.48	0.00
Xception	78.34	0.71
DenseNet201 (Final)	**80.0**	**0.77**

VIII. REFERENCES

[1] World Health Organization, "Pneumonia," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint* arXiv:1711.05225, 2017.

[3] J. Irvin et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[4] RSNA, "RSNA Pneumonia Detection Challenge Dataset," [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[5] M. Raghu et al., "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[6] M. S. Islam et al., "Pneumonia Detection from Chest X-ray Images Using an Ensemble of CNN Models," *Healthcare Technology Letters*, vol. 7, no. 5, pp. 149–156, 2020.

[7] J. Chouhan et al., "A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images," *Applied Sciences*, vol. 10, no. 2, pp. 559–572, 2020.