Deep Learning for Chest X-Ray Classification: A CNN Approach for Pneumonia and Tuberculosis Detection

Kerem Gursoy
University of Houston
kgursoy@uh.edu

Jonathan Poffenberger
University of Houston
jkpoffen@uh.edu

AyaanAli Lakhani
University of Houston
aslakha8@uh.edu

Abstract

We build a convolutional neural network (CNN) to classify frontal chest X-ray images into three categories: normal, pneumonia, and tuberculosis. Automated analysis is attractive because X-rays are cheap and widely available, but human interpretation is time-consuming and inconsistent across readers. Our dataset contains thousands of labeled radiographs with an imbalanced class distribution, where tuberculosis is the majority class and pneumonia is the smallest. After resizing images, normalizing intensities, and applying simple data augmentation, we train a compact CNN composed of four convolutional blocks and a fully connected classification head. The final model is selected using validation accuracy and loss over training epochs. On a held-out test set of 2,569 images, the network achieves about 75% overall accuracy. Class-wise metrics show relatively high precision for pneumonia (0.80) and tuberculosis (0.87), with the highest recall for pneumonia (0.91) and more moderate recall for normal (0.76) and tuberculosis (0.65) cases. The confusion matrix reveals that some tuberculosis images are misclassified as normal, likely due to class imbalance and overlapping radiographic appearances. These results demonstrate that a small, interpretable CNN can capture useful diagnostic signals from chest X-rays while leaving room for improvement through better handling of data imbalance and model complexity.

1 Introduction

Respiratory diseases such as pneumonia and tuberculosis remain a major global health burden, particularly in low-resource settings where access to expert radiologists is limited. Chest X-rays are one of the most widely used tools for screening and diagnosis because they are fast, inexpensive, and broadly available, but interpreting them is challenging: image quality can vary, subtle findings are easy to miss, and human readings are subject to inter-reader variability and fatigue. These challenges motivate automated systems that can assist clinicians by flagging suspicious studies and providing a second opinion.

In this project, we build a convolutional neural network (CNN) to classify frontal chest X-ray images into three categories: *normal*, *pneumonia*, and *tuberculosis*. Our goal is to construct a compact, interpretable baseline model that demonstrates how deep learning can capture clinically meaningful patterns in radiographs. We evaluate performance using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrices, and we further investigate model interpretability using Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps that show which image regions influence the model's predictions. Through these analyses, we highlight both the promise of CNN-based classification and the limitations that must be addressed before real-world deployment.

Conference on Neural Information Processing Systems (NeurIPS).

2 Background

Convolutional neural networks have become the dominant paradigm for image analysis and have achieved strong performance in medical imaging tasks ranging from skin lesion classification to detection of thoracic abnormalities on chest X-rays. Prior work on chest radiography has shown that deep CNNs trained on large labeled datasets can approach radiologist-level performance for certain findings, and that visual explanation methods such as Grad-CAM can help clinicians understand and critique model behavior.

Most existing systems, however, either focus on binary classification problems (for example, pneumonia versus normal) or on multi-label prediction of many findings simultaneously, often using large, pre-trained architectures. In contrast, our project focuses on a simpler three-class setting—normal, pneumonia, tuberculosis—and a comparatively lightweight CNN architecture designed from scratch. This setting aligns with the educational goals of the course: it allows us to explicitly apply core ideas from the semester, including convolution and pooling layers, non-linear activations, dropout regularization, and softmax classification, while still tackling a realistic and clinically relevant task. By analyzing our model's successes and failures in detail, we also connect to broader topics in the course such as class imbalance, generalization, and model interpretability.

3 Data

Our dataset consists of approximately 25,000 anonymized frontal chest X-ray images, each labeled as one of three mutually exclusive classes: *normal*, *pneumonia*, or *tuberculosis*. The class distribution in the training set is imbalanced: about 35.5% of images are normal, 22.9% are pneumonia, and 41.6% are tuberculosis. This imbalance is clinically plausible—tuberculosis is relatively common in certain screening populations—but poses challenges for learning and evaluation because models can become biased toward majority classes.

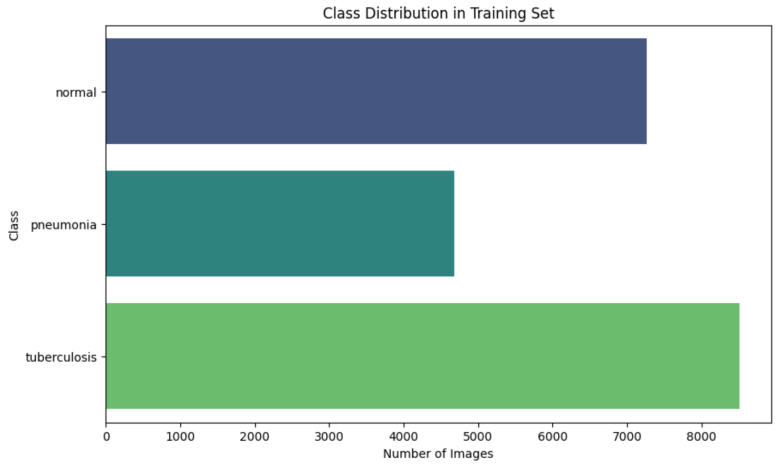

Figure 1: Class Distribution

The data are split into training, validation, and test subsets. The held-out test set used for final evaluation contains 2,569 images: 925 normal, 580 pneumonia, and 1,064 tuberculosis cases. All images are stored as grayscale radiographs. Before training, we resize each image to a fixed spatial resolution and normalize pixel intensities to a common scale, which standardizes input dimensions and reduces variability due to acquisition settings. Representative examples of each class illustrate the visual patterns the model must learn: normal images show clear lung fields and normal anatomy; pneumonia images exhibit patchy opacities and consolidation; and tuberculosis images often contain more focal lesions or chronic scarring patterns.

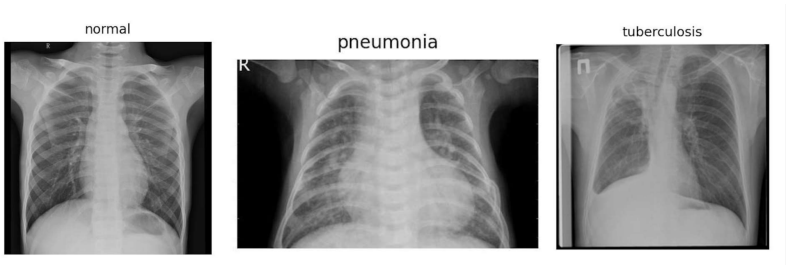

Figure 2: Example chest X-rays from each class

4 Methods

4.1 CNN architecture

To classify chest X-ray images, we design a compact CNN composed of four convolutional blocks followed by a small fully connected classification head. The architecture is summarized below:

- **Convolutional blocks.**
  - Block 1: 32 filters, 3×3 kernels, ReLU activation, followed by 2×2 max pooling and dropout with rate 0.1.
  - Block 2: 32 filters, 3×3 kernels, ReLU, 2×2 max pooling, dropout with rate 0.01.
  - Block 3: 64 filters, 3×3 kernels, ReLU, 2×2 max pooling.
  - Block 4: 64 filters, 3×3 kernels, ReLU, 2×2 max pooling.
- **Classification head.** The output of the final convolutional block is flattened and passed through a fully connected layer with 128 units and ReLU activation, followed by dropout with rate 0.3. The final layer is a dense layer with 3 units and softmax activation, producing class probabilities for normal, pneumonia, and tuberculosis.

This architecture reflects design principles discussed in class: early layers with fewer filters capture low-level edges and textures, while deeper layers with more filters capture higher-level structures. Max pooling progressively reduces spatial resolution and introduces translational invariance, and dropout acts as regularization to reduce overfitting.

4.2 Training procedure

Images are batched and fed to the CNN to minimize categorical cross-entropy loss using mini-batch stochastic gradient descent (implemented with a standard optimizer). We train the model for nine epochs, monitoring accuracy and loss on both the training and validation sets. The validation metrics guide model selection and provide an early indication of overfitting: a large gap between training and validation performance would suggest poor generalization, whereas closely tracking curves indicate that the model generalizes reasonably well.

Given the class imbalance in the dataset, we pay particular attention to per-class metrics such as precision, recall, and F1-score rather than relying solely on overall accuracy. During training and evaluation, predictions are computed by taking the argmax of the softmax output, and confusion matrices are used to quantify how often each true class is mapped to each predicted class.

4.3 Model interpretability with Grad-CAM

To better understand what the CNN learns, we apply Grad-CAM to generate heatmaps for individual predictions. For a given input image and target class, Grad-CAM computes the gradient of the class score with respect to the activations of the final convolutional layer, aggregates these gradients, and projects them back onto the image as a coarse localization map. Regions with high activation (shown as red or bright areas in the heatmap) are those that most strongly influence the model's decision.

We generate Grad-CAM visualizations for both correctly classified and misclassified examples. For correctly classified normal images, we expect strong activation over the lung fields, indicating that

the model is focusing on clinically relevant regions. For misclassified tuberculosis cases, we inspect whether the model is ignoring subtle lesions or focusing on irrelevant structures such as the heart or diaphragm. These qualitative analyses complement quantitative metrics and connect directly to the interpretability topics from the course.

5 Experiments

5.1 Training dynamics

Our first set of experiments evaluates how the model's performance evolves over training. We train the CNN for nine epochs and record accuracy and loss on both the training and validation sets. The resulting learning curves show a steady increase in accuracy from roughly 70% to about 75%, with training and validation accuracies remaining very close throughout. Similarly, training and validation loss decrease in parallel from around 0.60 to approximately 0.46. The absence of a large gap between training and validation curves suggests that the model is not severely overfitting and that our architecture and regularization are sufficient for this level of model capacity.

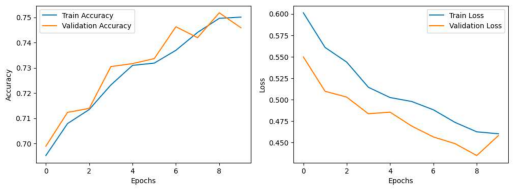

Figure 3: Accuracy & Loss Curves

5.2 Test performance and error analysis

We then evaluate the final model on the held-out test set of 2,569 images. Overall accuracy reaches 75%. A more detailed classification report reveals the following per-class metrics:

- Normal: precision 0.62, recall 0.76, F1-score 0.68 (925 images).
- Pneumonia: precision 0.80, recall 0.91, F1-score 0.85 (580 images).
- Tuberculosis: precision 0.87, recall 0.65, F1-score 0.74 (1,064 images).

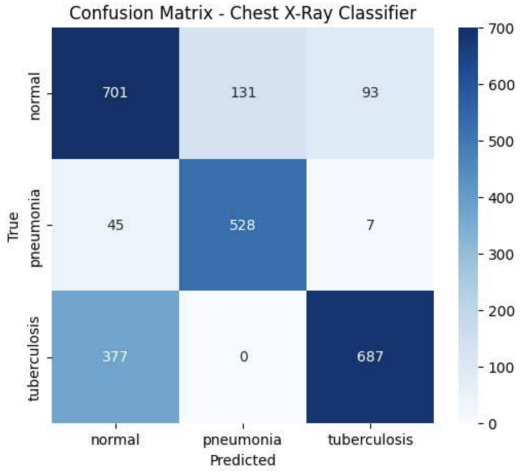

Figure 4: Confusion Matrix

The macro-averaged F1-score is 0.76, indicating reasonably consistent performance across classes despite imbalance. Pneumonia is the best-classified class, with very high recall and relatively few

missed cases. Normal images are handled moderately well but are sometimes labeled as pneumonia or tuberculosis. Tuberculosis is the most challenging class: while precision is high (few non-TB cases are incorrectly labeled as TB), recall is only about 65%, meaning that many true TB cases are missed.

A confusion matrix provides a more fine-grained view of errors. The most striking pattern is that 377 tuberculosis images are misclassified as normal, which significantly contributes to TB's low recall. In contrast, pneumonia has relatively low off-diagonal counts, consistent with its strong F1-score. These findings suggest that the model is conservative in assigning the tuberculosis label and may struggle with subtle or atypical TB presentations.

5.3 Qualitative analysis with Grad-CAM

To complement quantitative metrics, we perform qualitative experiments using Grad-CAM. For correctly classified normal cases, the Grad-CAM heatmaps show strong activation over the lung regions and relatively low activation over the heart and abdomen. This behavior is consistent with how clinicians interpret chest X-rays and suggests that the CNN has learned to focus on clinically meaningful structures.

We also examine a tuberculosis example that the model incorrectly predicts as normal. The corresponding Grad-CAM heatmap reveals that the model's attention is concentrated in regions that are less informative for TB diagnosis, while the areas containing TB-related abnormalities receive weaker activation. This misalignment between the model's focus and clinically relevant regions helps explain the misclassification and illustrates how Grad-CAM can be used to diagnose model failure modes and guide future improvements.

6 Conclusion

In this project, we developed and analyzed a compact CNN for three-class chest X-ray classification (normal, pneumonia, tuberculosis). The model achieves approximately 75% accuracy on a held-out test set and exhibits balanced performance across classes, with a macro F1-score of 0.76. Pneumonia is detected most reliably, with high recall and F1-score, while tuberculosis remains the most challenging class, with many TB cases misclassified as normal. Training and validation curves indicate that the model generalizes reasonably well without severe overfitting.

Beyond raw performance, Grad-CAM visualizations show that the CNN often attends to clinically meaningful lung regions when making correct predictions, supporting its potential usefulness as an assistive tool. At the same time, misclassified tuberculosis examples reveal that the model can sometimes focus on irrelevant areas, leading to missed cases. These observations highlight the dual importance of quantitative metrics and interpretability in evaluating medical AI systems.

Looking forward, several extensions could improve performance and robustness. Addressing class imbalance through techniques such as reweighting or targeted data augmentation may help boost TB recall. Exploring deeper architectures or transfer learning from large medical imaging models could further enhance accuracy. Finally, combining Grad-CAM with other explanation methods and involving clinicians in the review process would strengthen trust and help identify remaining failure modes. Overall, our results illustrate both the promise and the challenges of applying deep learning to chest X-ray classification in support of AI-assisted diagnosis.