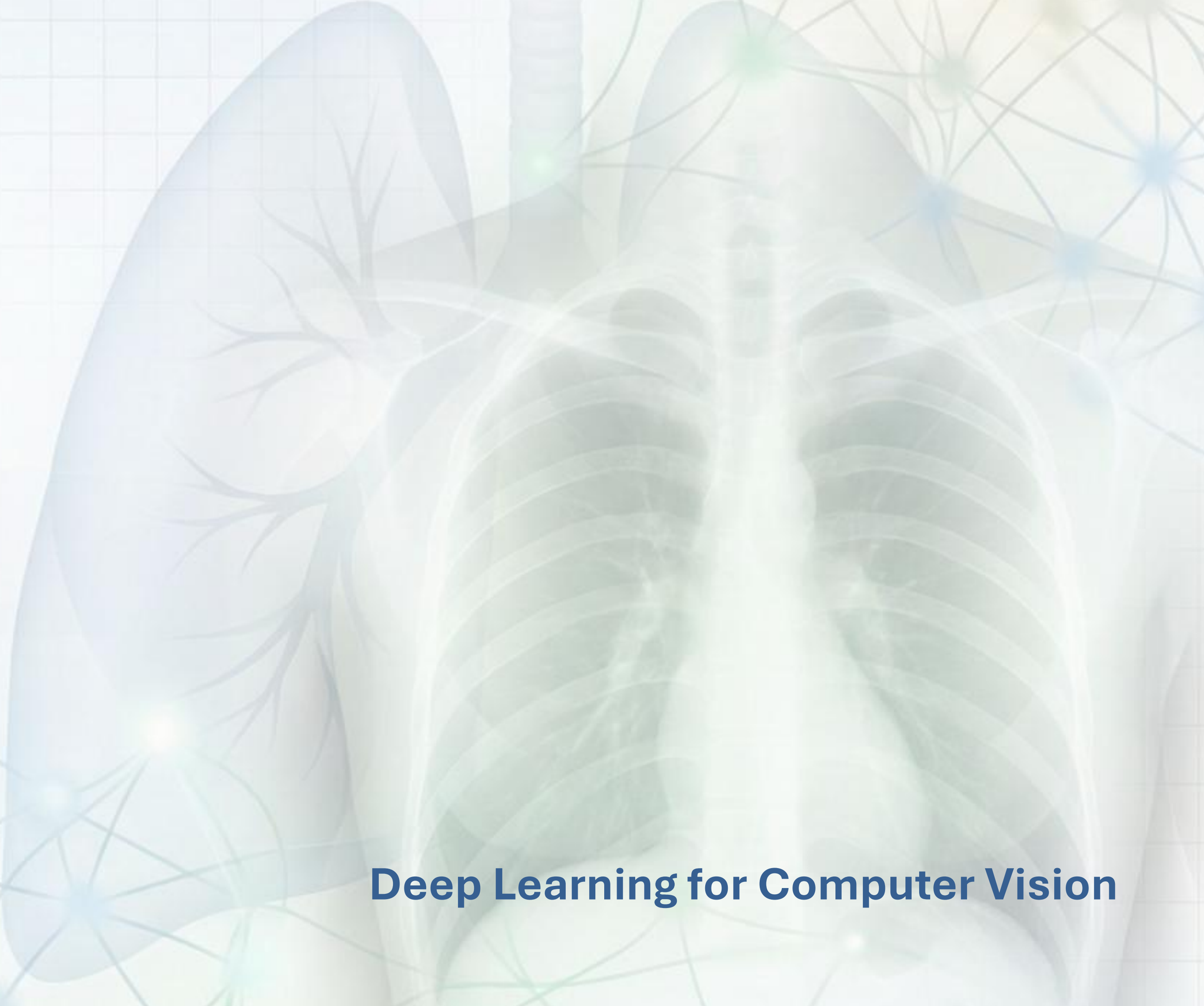
iscte INSTITUTO UNIVERSITÁRIO DE LISBOA

APVC



Deep Learning for Computer Vision

PROJECT 2

USE OF DATA AUGMENTATION AND TRANSFER LEARNING

CRISP-DM Phases 1-5

António Cruz (140129), **David Isaac** (120064), **Erik Daskalyuk** (120062), **Ivan Magalhães** (106586), **Ricardo Pereira** (120052)

Table of Contents

0. Environment Setup

0.1 Global Configuration & Reproducibility

This section establishes the environment settings and ensures the experiment is **deterministic** (repeatable).

- **Centralized Constants:** Defines file paths (DATASET_ROOT, TRAIN_DIR) and the ACTIVE_SCENARIO flag (e.g., Development vs. Production) in one place for easy management.
- **Model Checkpoint:** Sets the filename (best_model_checkpoint.keras) to automatically save the best performing model during training.
- **Reproducibility:** By fixing the SEED to **42** across the standard library (random), numpy, and tensorflow, we ensure that weight initialization and data shuffling remain consistent across every run.

0.2 Data Loading & Flow Control Utilities

This section defines two essential helper functions to manage workflow and ingest data:

- **scenario(*modes)**: A flow control switch. It validates if specific code blocks (like heavy visualizations or debug logs) should run based on the ACTIVE_SCENARIO.
  - *Logic:* "PRODUCTION" acts as a wildcard (always returns True), while other scenarios are checked explicitly against the allowed modes.
- **load_dataset_df(...)**: Transforms the raw directory structure into a structured **Pandas DataFrame**. It recursively scans the Train and Test folders to collect:
  - **image_path**: The file path to the image.
  - **label**: Derived from the subfolder name (the class).
  - **split**: Tags the entry as "train" or "test" for easy filtering later.

0.3 DataFrame Initialization & Separation

This step establishes a clean data lineage by enforcing a **separation of concerns** between analysis and modeling:

- **DF_ORIGIN (Source of Truth):** The immutable reference dataset. It serves as a permanent backup and is never modified directly.
- **DF_VIEW (Exploratory Sandbox):** A dedicated copy for **EDA** (Exploratory Data Analysis). This allows for safe experimentation and visualization without affecting the training data.
- **DF_MODEL (Training Pipeline):** The working copy destined for **Preprocessing** (resizing, augmentation, encoding) and feeding the neural network.

1. Business Understanding

1.1 Scope

This project focuses on developing an **AI-powered classification model** using **transfer learning** to **analyze chest X-ray images** and identify patients with pneumonia. Building on the foundational work of Project 1, this project explores advanced deep learning techniques to improve model performance, generalization, and clinical utility.

The project leverages pre-trained convolutional neural networks (CNNs), such as **VGG16, ResNet50, and EfficientNet**, to extract meaningful features from chest X-ray images.

By fine-tuning these models and applying data augmentation, the goal is to enhance the model's ability to **generalize to unseen data** while maintaining a **strong focus on high recall**, ensuring that pneumonia cases are not missed.

The system will: - Classify chest X-ray images into **Normal** and **Pneumonia** categories. - Enable healthcare professionals to retrospectively identify at-risk patients for targeted preventive care, reducing the risk of pneumonia recurrence and complications.

1.2 Objectives

1. **Implement Transfer Learning**

   - Develop a classification model using pre-trained CNNs (e.g., VGG16, ResNet50, EfficientNet) as a baseline, leveraging their ability to extract high-level features from medical images.

2. **Experiment with Data Augmentation**

   - Apply transformations such as rotation, zoom, and flipping to artificially expand the training dataset, improving model robustness and reducing overfitting.

3. **Explore Feature Extraction and Fine-Tuning**

   - Feature Extraction: Use pre-trained models as fixed feature extractors, training only the top classifier layers.
   - Fine-Tuning: Unfreeze and retrain later layers of the pre-trained models to adapt them specifically to the chest X-ray classification task.

4. **Evaluate Model Performance**

   - Compare the performance of different architectures and techniques using clinically relevant metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, with an emphasis on recall to ensure that pneumonia cases are not overlooked.

5. **Justify the Best-Performing Model**

   - Provide a detailed analysis of why the selected model (e.g., ResNet50 with fine-tuning) outperforms others, considering trade-offs such as training time, computational resources, and clinical utility.

6. **Deliver a Functional and Documented Solution**

   - Provide a well-documented Python codebase and a comprehensive report detailing the dataset, methodology, model architecture, training process, evaluation results, and the rationale behind the selected model and techniques.

1.3 Clinical and Operational Impact

The project aims to deliver a practical, recall-focused AI model that assists healthcare professionals in identifying patients with a history of pneumonia from historical X-ray records. By enabling targeted preventive interventions, the model supports secondary prevention efforts, reducing the risk of recurrence and improving patient outcomes.

2. Data Understanding

2.1 Introduction

2.1.1 Analytical Objectives

1. **Intensity and Contrast Analysis**
   - **Global Contrast (Std Dev):** Measures the spread of pixel intensities to identify low-contrast images that may hinder model convergence.
   - **Dynamic Range:** Calculates the difference between maximum and minimum pixel values to detect **exposure anomalies** (e.g., washed-out or overly dark scans).
2. **Texture and Sharpness Evaluation**
   - **Histogram Entropy:** Quantifies the information density within the image. Higher entropy often correlates with complex lung structures or pathologies, while low entropy may indicate poor capture quality.
   - **High-Frequency Energy (Laplacian):** Acts as a proxy for **image sharpness**. This metric helps identify and potentially filter out blurry images that lack the edge definitions required for accurate diagnosis.
3. **Statistical Distribution Metrics**
   - **Kurtosis and Skewness:** Analyzes the shape of the pixel intensity histogram.
     - *Skewness* indicates if the image is dominated by dark (air) or bright (bone/tissue) regions.
     - *Kurtosis* helps identify outliers in pixel distribution, flagging images with unusual artifact patterns.
4. **Geometric and Spatial Analysis**
   - **Centering Offsets (x, y):** Calculates the deviation between the center of the lung content and the center of the image frame.
   - **Content-to-Image Ratio:** Determines how much of the image frame is occupied by relevant anatomical data versus background.
   - **Impact:** These metrics are critical for defining **safe augmentation limits**. For example, if patients are already off-center, aggressive RandomTranslation could crop out essential lung tissue.

2.1.2 Clinical and Technical Impact

By systematically computing these metrics for the entire dataset (Train/Test), we transition from “guessing” augmentation parameters to a **data-driven configuration strategy**.

- **Operational Efficiency:** Automatically flags low-quality images that require manual review or exclusion (Data Cleaning).
- **Model Robustness:** Ensures that the neural network trains on high-quality, relevant signals rather than learning from artifacts or noise.
- **Reproducibility:** Provides a standard, numerical profile of the dataset's characteristics, facilitating comparison with future datasets or external benchmarks.

2.2 Global Contrast

A stable model performance requires a dataset with relatively consistent exposure and contrast characteristics. Large fluctuations in global contrast, whether caused by acquisition variability, compression artifacts, or post-processing, can lead to heterogeneous feature distributions during training.

Such variability increases the risk of the model focusing on contrast artifacts rather than clinically relevant patterns. Quantifying global contrast supports the identification of low-quality or overprocessed images that may require correction, filtering, or consistent preprocessing prior to model training.

Global contrast, expressed as the standard deviation of pixel intensities, is a fundamental indicator of image quality in chest radiography. Diagnostic information in X-ray images depends heavily on subtle grayscale variations that delineate lung fields, soft-tissue boundaries, vascular structures, and osseous anatomy. Low global contrast typically indicates underexposure or suboptimal acquisition, which can obscure anatomical details and reduce the discriminative capacity available to a learning algorithm.

Conversely, excessively high global contrast often suggests overprocessing or artificial enhancement, potentially introducing edges and textures not present in standard clinical imaging.

As part of dataset characterization, this metric establishes a baseline reference for expected exposure and contrast levels across the available images. This baseline assists in determining whether contrast normalization methods (e.g., histogram normalization, CLAHE, or no enhancement) are necessary or potentially counterproductive. It also provides a reference against which future inference-time images can be compared, helping to detect contrast outliers that may result from differing acquisition settings, device variability, or prior preprocessing.

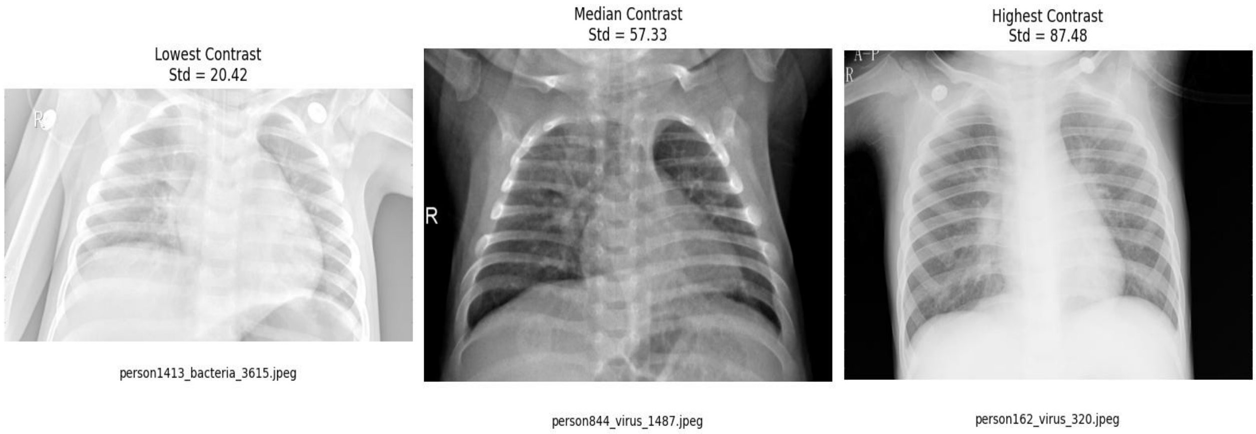


Figure 2.1 - Lowest, Median and Highest Contrast Images

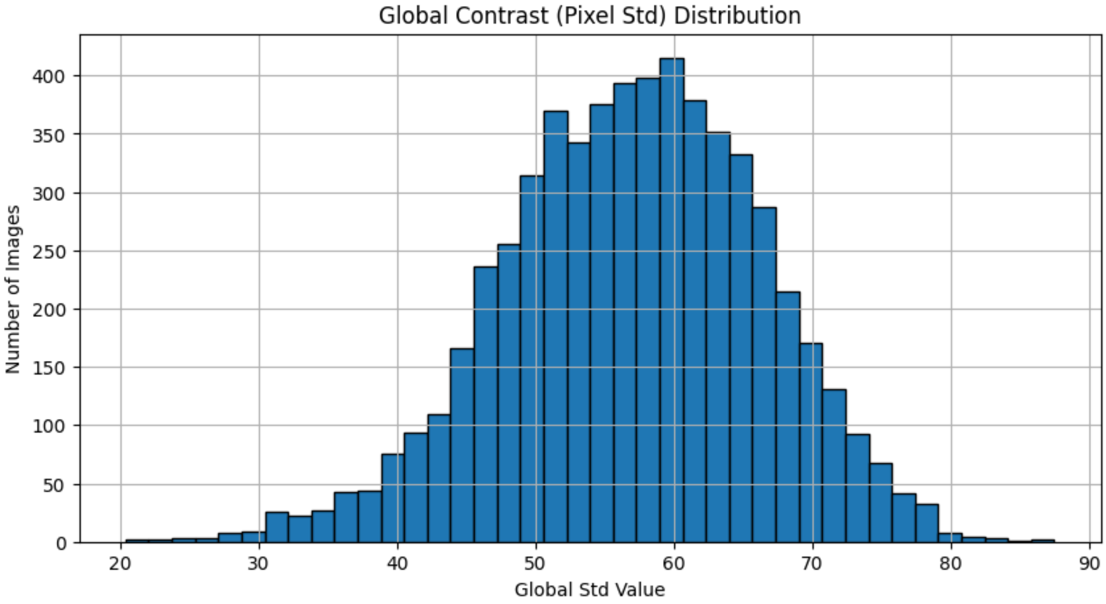


Figure 2.2 - Global Contrast Distribution

The distribution of global contrast values shows a well-formed, approximately Gaussian shape centered around the expected range for standard chest radiographs. This suggests that the majority of images share similar exposure and contrast characteristics, which is desirable for model training. The presence of a small number of low-contrast images at the left tail likely reflects instances of mild underexposure or limited dynamic range. Likewise, the small number of high-contrast cases at the right tail may correspond to images that have undergone stronger post-processing or contrast amplification.

Overall, the dataset appears broadly consistent in terms of contrast quality. The core of the distribution indicates that most images fall within a stable and clinically plausible contrast band, which supports robust model learning. The few outliers observed at the extremes do not dominate the dataset and may be reviewed individually during preprocessing decisions, but they do not suggest any structural imbalance or heterogeneous mixing of incompatible imaging sources.

2.3 Dynamic Range

Dynamic range, defined as the difference between the maximum and minimum pixel intensities in an image, serves as a measure of how fully the grayscale space is utilized during acquisition. Chest X-ray imaging relies on a broad representation of intensities to preserve the visibility of both low-attenuation regions such as lung fields and high-attenuation structures such as ribs or the mediastinum.

An image with a limited dynamic range may appear washed out or uniformly gray, which reduces the visibility of diagnostically relevant structures. Conversely, an abnormally wide dynamic range may indicate strong post-processing, unusual acquisition parameters, or inconsistent normalization steps applied before dataset construction.

A dataset with large variability in dynamic range can introduce significant inconsistency in the distribution of pixel intensities seen during training. This inconsistency makes it more difficult for a model to learn reliable patterns, as the same anatomical structure may appear substantially lighter or darker depending on acquisition conditions rather than actual physiology. Monitoring this metric makes it possible to identify images that deviate from standard radiographic appearance and that may require normalization or exclusion. A stable and well-defined dynamic range across samples contributes to more predictable model behavior and reduces the risk of learning contrast artifacts.

Dynamic range analysis complements global contrast analysis by providing a direct measurement of the span of intensities present in each image.

Whereas the global standard deviation characterizes the distribution's spread, dynamic range assesses the actual minimum and maximum gray levels used by the imaging system. Together, these metrics assist in determining whether preprocessing steps such as intensity clipping, histogram normalization, or contrast equalization are necessary to achieve a consistent representation of the dataset. They also support the identification of outliers that may result from non-standard imaging devices or prior enhancement.

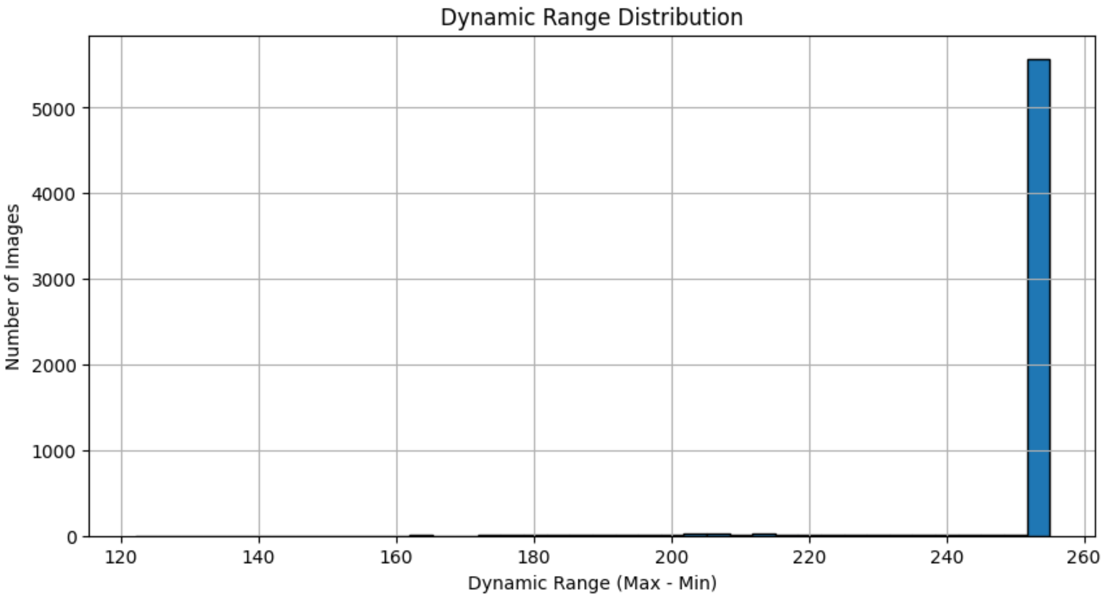


Figure 2.3 - Dynamic Range Distribution

Overall, the dataset appears highly uniform in terms of dynamic range. The dominant peak at 255 indicates consistent handling of intensity rescaling across images, which is advantageous for downstream modelling, as the input space remains stable. The small tail of lower dynamic-range images can be inspected individually during the data-quality review phase, but the distribution does not suggest any systemic inconsistency in acquisition or preprocessing.

2.4 Histogram Entropy

Histogram entropy measures the complexity and variability of the grayscale distribution in an image. Higher entropy values indicate that pixel intensities are more uniformly distributed across the available range, reflecting greater diversity in tonal values. Lower entropy values occur when intensities are concentrated in a narrower subset of values, producing flatter or more uniform images.

In the context of chest radiography, entropy provides an aggregate view of how much tonal information the acquisition contains. Images with low entropy may be underexposed, washed out, or lacking structural definition, while images with very high entropy may have undergone strong contrast equalization or other enhancement steps.

Stable entropy levels across a dataset contribute to consistent feature extraction during training. Variations in entropy reflect differences in perceptual complexity, structural prominence, and contrast distribution, all of which influence how convolutional layers respond to radiographic textures.

An image with unusually low entropy may lack sufficient local variation for robust learning, while an image with unusually high entropy may contain contrast artifacts that do not represent normal anatomy. By quantifying entropy across the dataset, it becomes possible to detect samples that deviate from the expected radiographic appearance and that may benefit from targeted normalization or exclusion.

Histogram entropy complements the global standard deviation and dynamic range metrics by capturing the distributional shape of image intensities rather than only the spread or the minimum and maximum values. Together, these metrics help characterize the global exposure profile of the dataset and guide decisions regarding preprocessing steps such as histogram normalization or contrast equalization. Entropy also establishes a reference distribution against which future inference-time inputs can be compared, supporting the identification of atypical acquisition conditions or unexpected preprocessing in production data.

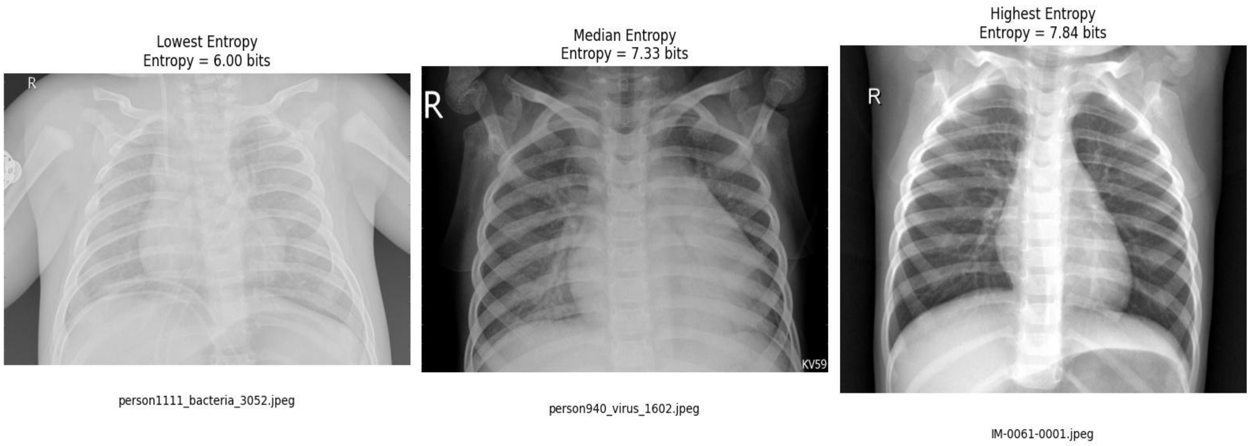


Figure 2.4 - Lowest, Median, and Highest Entropy Images

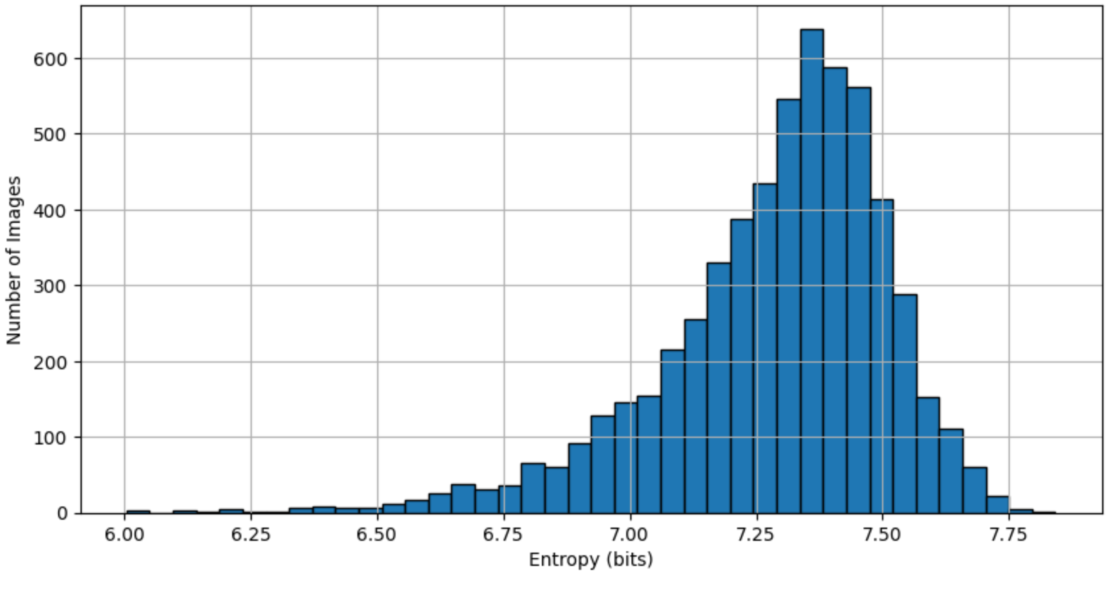


Figure 2.5 - Histogram Entropy Distribution

The entropy distribution shows a well-defined concentration around values between approximately 6.9 and 7.6 bits, indicating that most images exhibit a balanced spread of pixel intensities across the grayscale range.

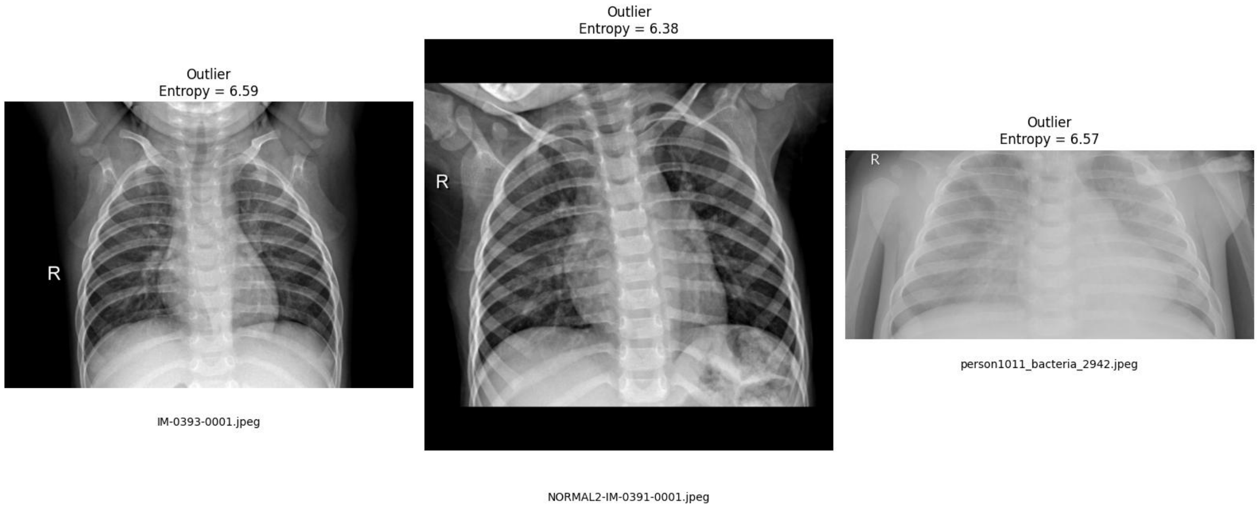
This range reflects an expected level of tonal complexity for chest radiographs in which both low-attenuation and high-attenuation structures are adequately represented. The presence of a smooth, unimodal shape suggests a consistent image formation process across the dataset, with no major groups of images derived from markedly different acquisition or post-processing pipelines.

A smaller number of images appear in the lower-entropy tail, with values between 6.0 and 6.5 bits. These cases may correspond to underexposed or low-contrast scans in which intensity values are compressed into a narrow portion of the grayscale space. Conversely, the images near the upper end of the distribution reach entropy values close to 7.8 bits, which may indicate stronger histogram equalization or more pronounced tonal variation introduced during preprocessing.

Despite these isolated outliers, the overall profile of the entropy distribution reflects a dataset with broadly consistent radiographic complexity, supporting stable feature extraction and learning during model development.

2.4.1 Histogram Entropy Outliers

Images with unusually low or high histogram entropy, indicating abnormal pixel-value distributions. Low-entropy examples correspond to radiographs with reduced textural variability or mild posterization, while high-entropy examples reflect excessive noise or aggressive post-processing. These samples lie statistically outside the expected entropy range of the dataset.



Outlier
Entropy = 6.59
IM-0393-0001.jpeg

Outlier
Entropy = 6.38
NORMAL2-IM-0391-0001.jpeg

Outlier
Entropy = 6.57
person1011_bacteria_2942.jpeg

Figure 2.6 - Histogram Entropy Outliers

2.5 Kurtosis and Skewness

Kurtosis and skewness provide statistical descriptions of the shape of the grayscale intensity distribution. Kurtosis quantifies the heaviness of the histogram tails relative to a normal distribution, while skewness measures the asymmetry of the histogram. In chest radiography, the grayscale distribution is typically skewed toward lower intensities due to the predominance of air-filled lung regions. Abnormally low kurtosis may indicate a flatter distribution caused by contrast equalization or smoothing, while unusually high kurtosis suggests a histogram dominated by sharp peaks or extreme intensity clustering.

Atypical kurtosis or skewness values may signal that certain images have undergone nonstandard processing, contain noise patterns, or result from acquisition variations not present in the majority of the dataset. These deviations alter the histogram shape in ways that can mislead convolutional filters during training, especially in the earliest layers where pixel-level distributions influence the learned feature maps. By measuring kurtosis and skewness across the dataset, it becomes possible to detect images whose tonal composition differs substantially from typical chest radiographs and that might warrant further inspection.

Together with global contrast, dynamic range, and entropy, kurtosis and skewness offer a complementary view of image quality by characterizing the higher-order properties of the intensity distribution. These statistics allow the dataset's overall histogram shape to be compared against expected radiographic norms, supporting the identification of histograms that are unusually flat, unusually peaked, or strongly asymmetric. As part of the data understanding phase, these metrics contribute to determining whether preprocessing steps should normalize intensity distributions or whether the dataset already exhibits stable and consistent characteristics.

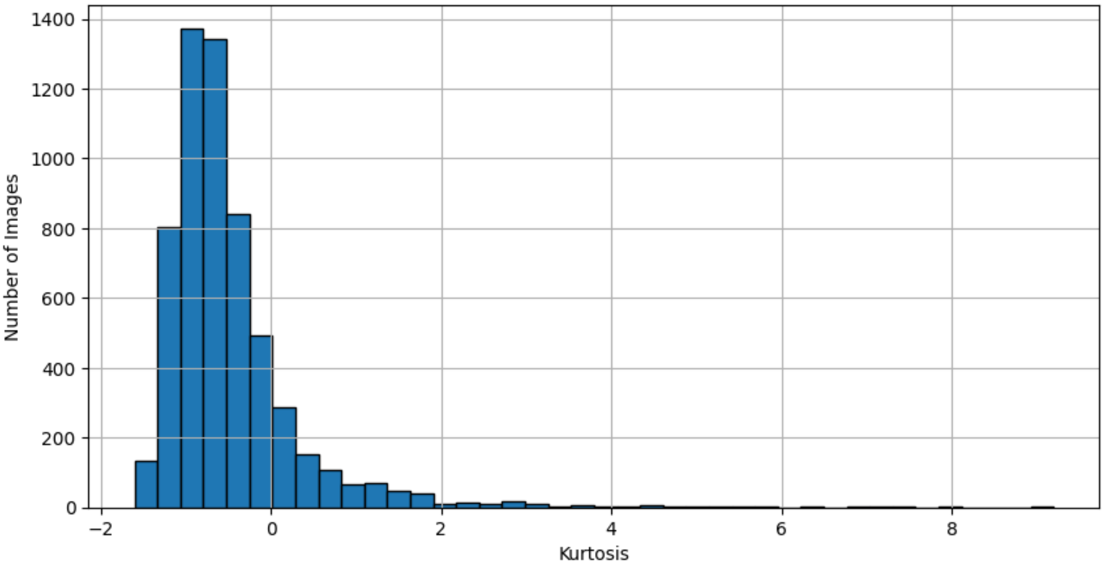


Figure 2.7 - Kurtosis Distribution

The kurtosis values form a distribution concentrated primarily between approximately -1.5 and +1.0, with a mean near -0.49. This pattern indicates that most images exhibit flatter histograms than a normal distribution would suggest, which is expected for chest radiographs. Air-filled lung fields occupy a substantial portion of the image area and contribute to a large concentration of darker pixel values, while anatomical structures with higher attenuation appear less frequently.

A limited number of images show higher kurtosis values extending above 2.0, and a few outliers reach values above 5.0. These cases may correspond to scans with pronounced intensity clustering, possibly caused by strong smoothing, localized contrast amplification, or specific acquisition characteristics. However, these remain isolated and do not indicate systemic inconsistency within the dataset.

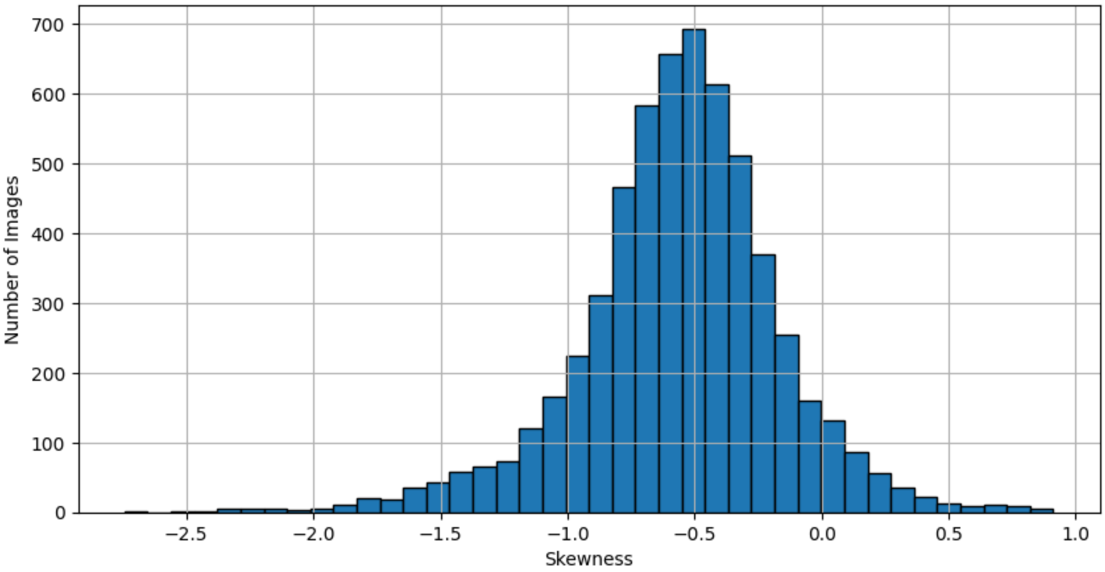


Figure 2.8 - Skewness Distribution

The skewness distribution is centered around negative values, with a mean near -0.56 and the majority of images falling between approximately -1.2 and -0.3. Negative skewness reflects the expected dominance of lower intensities in chest radiographs, as the lungs represent the largest anatomical region and naturally appear darker relative to bones and mediastinal structures. This asymmetry is characteristic of properly acquired radiographs and confirms that the dataset exhibits the typical tonal bias found in clinical imaging. A small number of images approach or exceed symmetric or positive skewness values, which may indicate unusual exposure conditions, differing preprocessing pipelines, or reduced representation of darker regions. These cases can be reviewed individually but do not appear frequently enough to affect the overall dataset composition.

2.5.1 Kurtosis Outliers

Radiographs exhibiting extreme kurtosis values, representing atypical tail behavior in intensity distributions. High-kurtosis images often contain sharp peaks or heavy-tailed noise patterns, while low-kurtosis images show overly flattened histograms. These deviations suggest non-standard imaging characteristics or post-acquisition alterations.

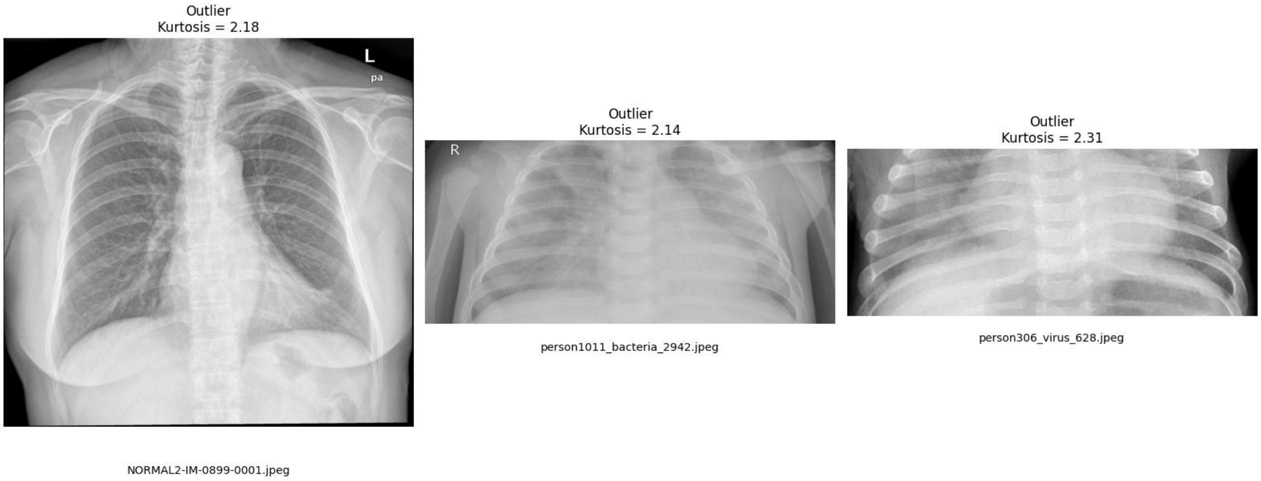


Figure 2.9 - Kurtosis Outliers

2.6 Local Contrast Uniformity

Local contrast uniformity describes how contrast varies across different regions of an image. It is computed by dividing the image into smaller tiles and calculating the standard deviation of pixel intensities within each tile. These tile-level measurements reveal how much local structure and intensity variation is present in different parts of the radiograph. Chest X-rays naturally contain heterogeneous regions: lung fields exhibit soft-tissue textures, the mediastinum is denser, and bony structures introduce sharp boundaries. A realistic radiograph therefore displays moderate variation in local contrast between regions.

Significant deviations in local contrast uniformity may indicate that an image has undergone strong local enhancement or smoothing. For example, local histogram equalization methods such as CLAHE (Contrast Limited Adaptive Histogram Equalization) increase local contrast uniformly across the entire image, which reduces the variability between tiles and often raises the average local contrast. Conversely, excessive smoothing lowers local contrast and may disproportionately affect diagnostically important areas. Monitoring this metric helps identify images that display atypical spatial contrast patterns, either due to acquisition characteristics or prior preprocessing, preventing such outliers from influencing the learned feature representations.

Local contrast analysis provides a spatially aware extension of the earlier global metrics. While global contrast and entropy measure aggregation over the entire intensity distribution, local contrast uniformity reveals whether the radiograph preserves appropriate regional differences. This information is especially useful when deciding whether the dataset requires additional normalization steps or whether any images need to be excluded. The metric also establishes a reference for assessing future inference images, helping detect preprocessed or enhanced scans that may not align with the training distribution.

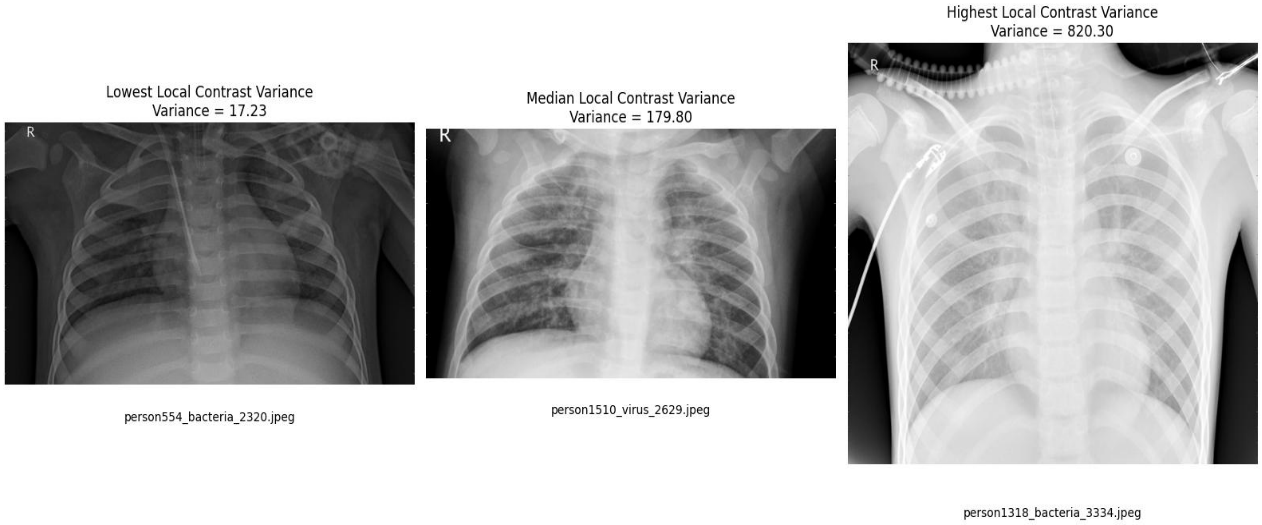

Lowest Local Contrast Variance
Variance = 17.23
R
person554_bacteria_2320.jpeg
Median Local Contrast Variance
Variance = 179.80
R
person1510_virus_2629.jpeg
Highest Local Contrast Variance
Variance = 820.30
R
person1318_bacteria_3334.jpeg

Figure 2.10 - Lowest, Median, and Highest Local Contrast Variance Images

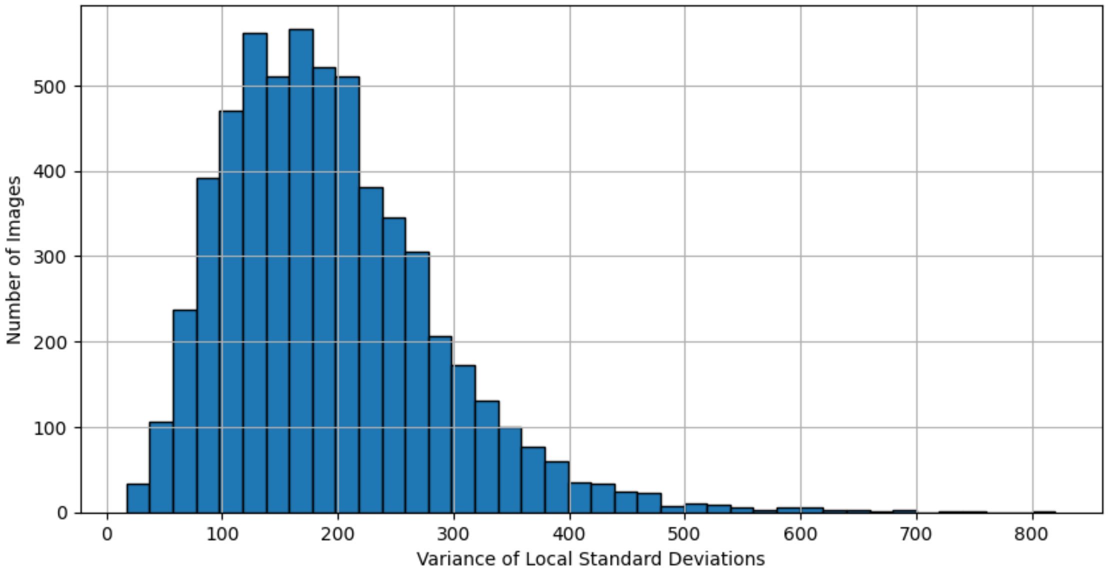


Figure 2.11 - Local Contrast Variance Distribution

The distribution of local contrast variance displays a broad range of values, with most images concentrated between approximately 80 and 300. This behaviour is consistent with chest radiographs, which naturally contain regional differences in contrast: lung fields present relatively homogeneous textures, bony structures introduce strong edges, and the mediastinum contains dense tissue with intermediate variability. The long tail extending toward higher variance suggests that a subset of images exhibits unusually heterogeneous local contrast. These outliers may correspond to radiographs with sharper boundaries, stronger acquisition noise, or more pronounced tonal differences, but they do not form a cluster that would indicate systematic overprocessing or local equalization.

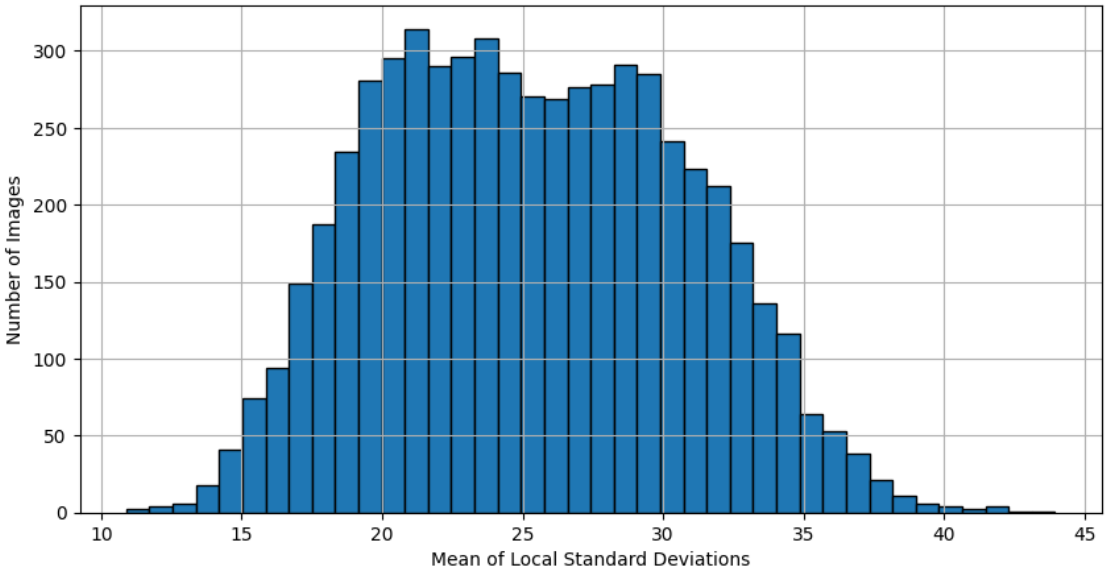


Figure 2.12 - Local Contrast Mean Distribution

The distribution of mean local contrast shows a dense peak between roughly 20 and 32, reflecting a stable representation of structural detail across the dataset. This profile is characteristic of radiographs that preserve adequate local variation without excessive enhancement. Only a small number of images appear at the lower end of the scale near 11 or at the higher end near 40, which likely represent underexposed scans or images with unusually strong spatial contrast characteristics. These extremes are isolated and do not suggest the presence of a distinct subgroup of preprocessed or contrast-equalized images.

2.6.1 Local Contrast Variance Outliers

Images with unusually high or low variance in tile-level standard deviations. High-variance samples typically indicate regions of intensified noise or sharpness inconsistencies, whereas low-variance samples correspond to overly smoothed or low-detail radiographs. These outliers reflect local structural irregularities not representative of the dataset's general behavior.

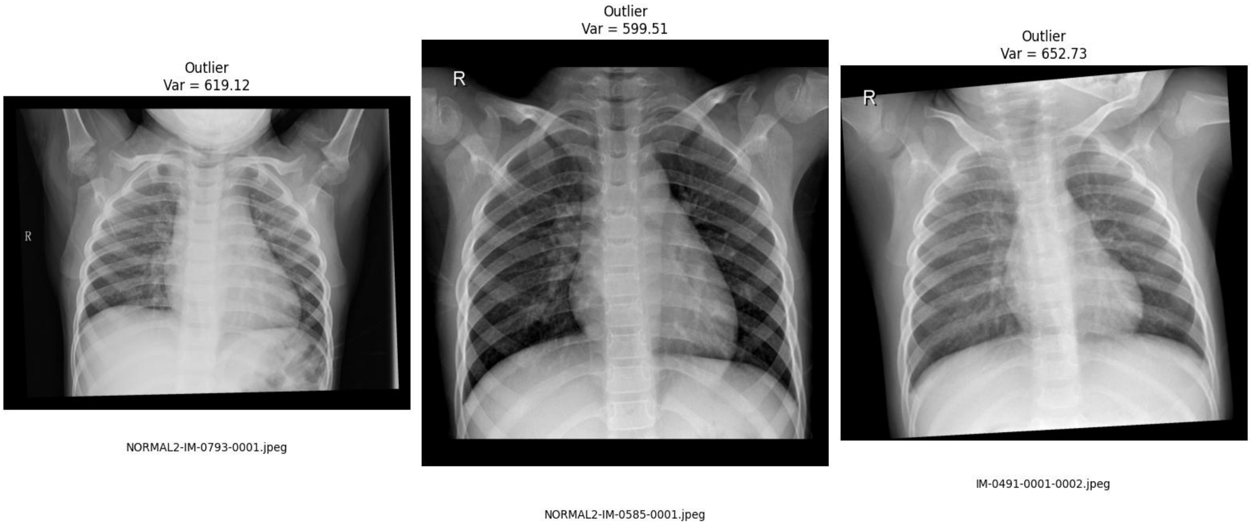


Figure 2.13 - Local Contrast Variance Outliers

2.7 High-Frequency Content

High-frequency Content (Edge Energy) measures the amount of rapid intensity variation present in an image. It is typically computed by applying a high-pass operator, such as the Laplacian filter, and quantifying the average magnitude of the resulting response. In chest radiography, high-frequency information corresponds to edges and fine structural details such as rib boundaries, vascular markings, clavicle contours, and soft-tissue texture. A well-formed radiograph maintains a balanced level of high-frequency content: excessive smoothing significantly reduces edge energy, while aggressive sharpening or noise amplification increases it.

Large deviations in high-frequency content indicate potential issues in image acquisition or preprocessing. Overly smoothed images may result from denoising pipelines, motion blur, or low-dose acquisition, all of which suppress clinically relevant textural cues. Conversely, images with unusually high edge energy may exhibit strong sharpening, residual noise, or compression artifacts.

These variations affect the early convolutional layers of a neural network, which rely heavily on the presence and stability of local gradients to extract meaningful features. By characterizing high-frequency behaviour across the dataset, it becomes possible to identify images whose sharpness profile differs substantially from standard radiographic texture.

High-frequency analysis acts as a complementary measure to the earlier contrast-related metrics by focusing on structural detail rather than tonal distribution. Examining edge energy helps determine whether the dataset contains imaging outliers that may need normalization, exclusion, or preprocessing adjustments. It also establishes a reference for future inference images, enabling the identification of scans that underwent different noise-reduction or sharpening pipelines. Taken together with the preceding metrics, high-frequency content provides a comprehensive overview of the dataset’s radiographic fidelity.

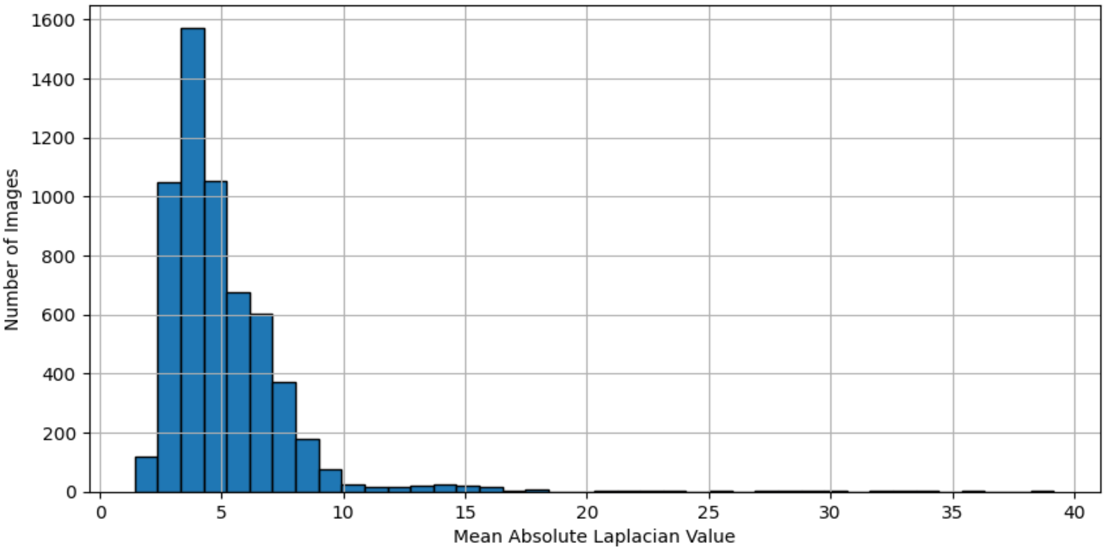


Figure 2.14 - Lowest, Median, and Highest High-Frequency Content Images

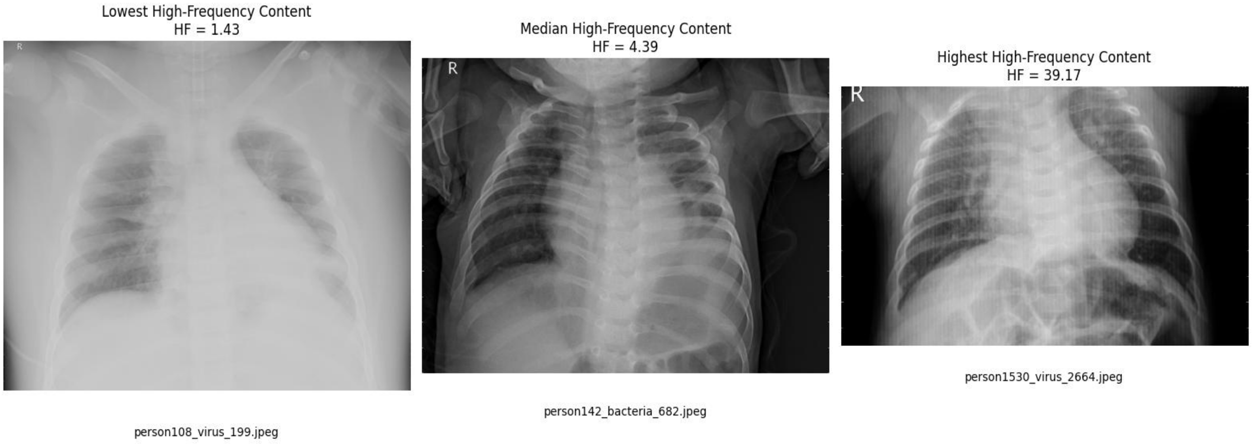

Lowest High-Frequency Content
HF = 1.43
person108_virus_199.jpeg
Median High-Frequency Content
HF = 4.39
person142_bacteria_682.jpeg
Highest High-Frequency Content
HF = 39.17
person1530_virus_2664.jpeg

Figure 2.15 - High-Frequency Content Distribution

The distribution of high-frequency content is concentrated between roughly 3.0 and 7.0, indicating that most images exhibit a normal level of structural detail and edge sharpness consistent with standard chest radiographs. This pattern suggests a uniform acquisition process with stable preservation of fine anatomical boundaries. A smaller number of images fall below this range, which may reflect smoothing, reduced dose, or minor motion effects, while a long upper tail contains isolated cases with unusually strong edges or noise. These outliers remain rare and can be reviewed individually if needed.

Overall, the dataset presents a coherent sharpness profile with only limited deviations, indicating that the radiographs are largely consistent in their high-frequency characteristics and appropriate for further analysis without major preprocessing adjustments.

The image quality metrics collectively indicate that the dataset exhibits a high degree of tonal and structural consistency. Global contrast and dynamic range measurements show that most radiographs maintain appropriate exposure levels, covering nearly the full available grayscale range with limited variation. Histogram entropy reinforces this finding by demonstrating that grayscale complexity remains stable across the dataset, with only a small number of images displaying reduced or unusually elevated entropy.

The distribution of skewness and kurtosis further supports the conclusion that the dataset adheres to typical radiographic intensity profiles. The negative skewness values reflect the expected predominance of darker lung regions, while kurtosis values indicate slightly flattened but otherwise normal histogram shapes. Together, these global metrics suggest that the images have not undergone diverse or aggressive post-processing procedures and that acquisition characteristics are relatively homogeneous.

Spatial metrics derived from local contrast and high-frequency content show comparable stability. Tile-based measures reveal appropriate regional variability in contrast without signs of widespread local enhancement, and the high-frequency content distribution indicates consistent sharpness and preservation of anatomical detail. Only a small number of images deviate noticeably from these patterns, and these cases appear isolated rather than systematic. Overall, the metrics suggest that the dataset is well suited for downstream modelling without the need for substantial corrective preprocessing.

2.7.1 High-Frequency Content Outliers

Radiographs with extreme high-frequency responses based on the mean absolute Laplacian. These samples exhibit disproportionate levels of edge enhancement or high-frequency noise, often indicating non-standard sharpening, low-dose acquisition artifacts, or compression effects. Their frequency profiles fall well outside the distribution observed in the dataset.

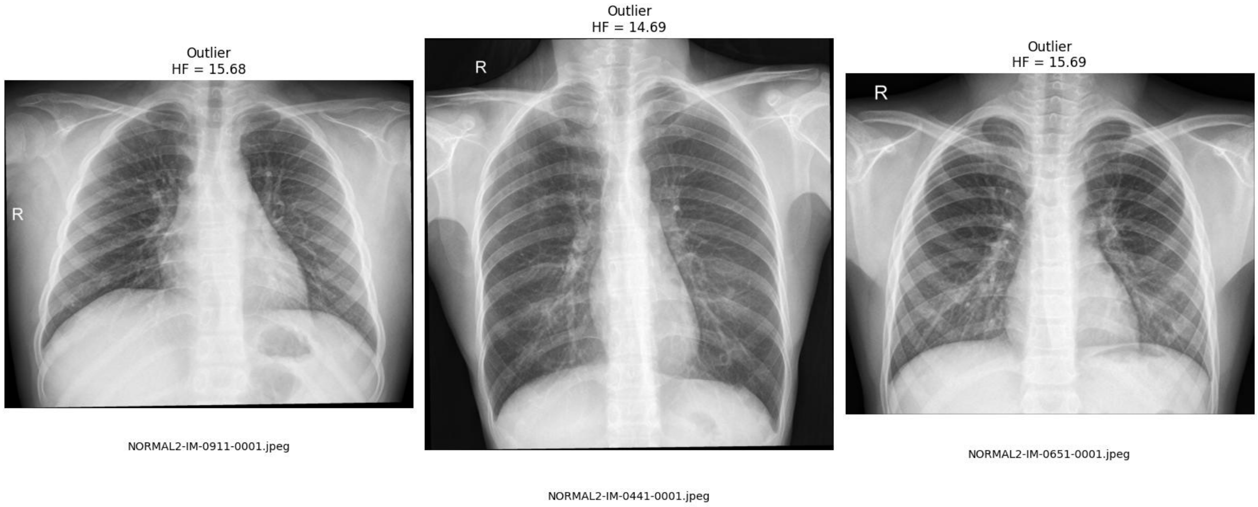

Outlier
HF = 15.68
R
NORMAL2-IM-0911-0001.jpeg
Outlier
HF = 14.69
R
NORMAL2-IM-0441-0001.jpeg
Outlier
HF = 15.69
R
NORMAL2-IM-0651-0001.jpeg

Figure 2.16 - High-Frequency Content Outliers

2.8 Geometric Metrics

The preceding contrast and frequency domain metrics characterize tonal and textural properties of the dataset but do not address spatial positioning or image composition. Geometric metrics quantify how X-ray content is positioned and framed within each image, capturing variations in patient centering, field of view, and anatomical coverage. These measurements are essential for deriving data-driven augmentation parameters, particularly for spatial transformations such as translation, zoom, and rotation.

Chest radiographs naturally exhibit variability in patient positioning during acquisition. Even with standardized protocols, factors such as patient cooperation, technician technique, and equipment constraints introduce small deviations in centering and framing. Quantifying these natural geometric variations establishes realistic bounds for augmentation: transformations that exceed the dataset's observed variability risk generating implausible training samples, while transformations that fall short of natural variation fail to improve model robustness.

The computed geometric metrics include horizontal and vertical centering offsets, which measure how far the radiographic content deviates from the image center, and content width and height ratios, which indicate what fraction of the image frame is occupied by actual diagnostic information. Additionally, the aspect ratio of each image is recorded to support quality control and to inform decisions regarding resizing or padding strategies during preprocessing. Together, these metrics enable the derivation of translation and zoom augmentation parameters that reflect the dataset's intrinsic spatial characteristics rather than arbitrary defaults.

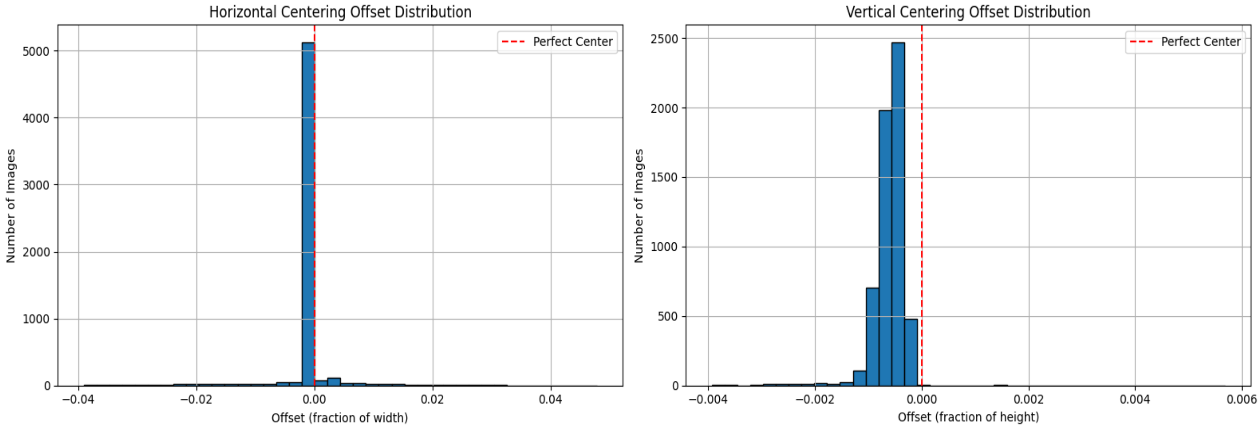


Figure 2.17 - Horizontal and Vertical Centering Offset Distribution

The horizontal and vertical centering offset distributions show that most images are well-centered, with offsets concentrated tightly around zero. The standard deviations of 0.5% horizontally and near 0% vertically indicate highly consistent patient positioning across the dataset, suggesting minimal natural variation in image composition.

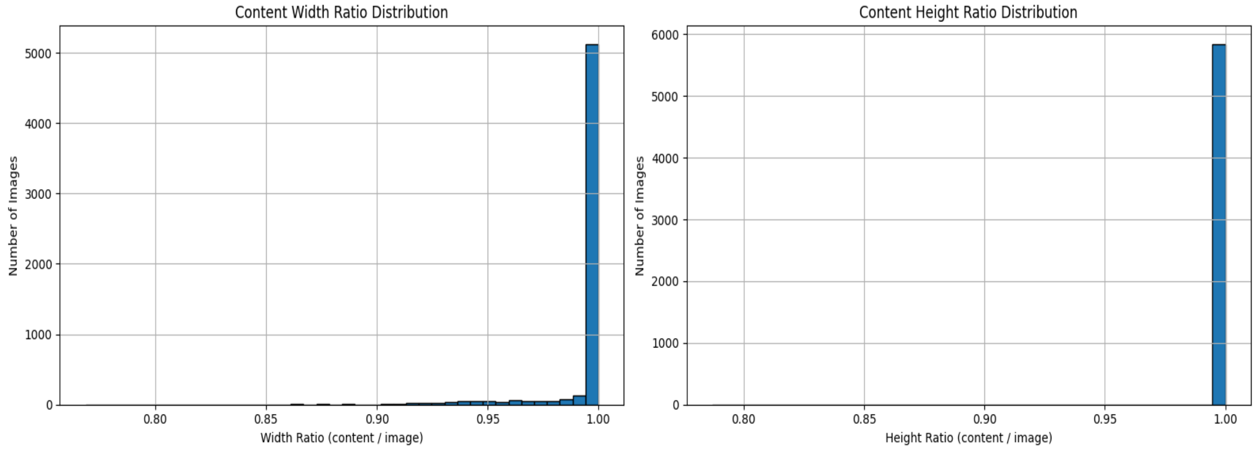


Figure 2.18 - Content Width and Height Ratio Distributions

Content width and height ratios reveal that radiographic content occupies approximately 95-98% of the frame in most images, indicating minimal black borders or empty space. This consistency simplifies preprocessing decisions and suggests that the dataset was acquired with standardized framing protocols.

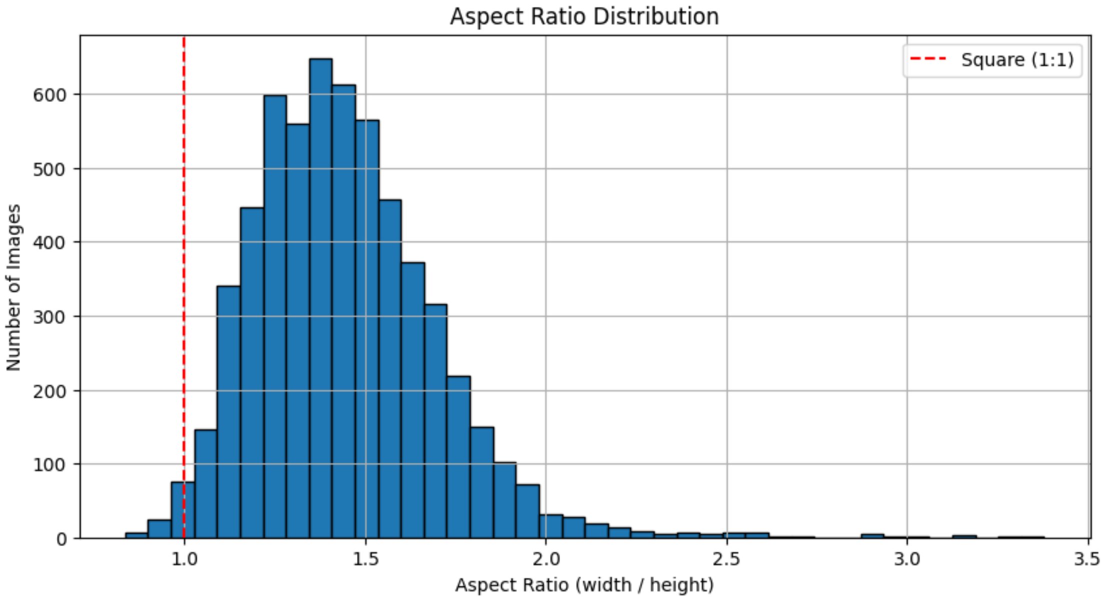


Figure 2.19 - Aspect Ratio Distribution

The aspect ratio distribution is concentrated around 1.0-1.4, with most images being roughly square or slightly wider than tall. The tight distribution without extreme outliers confirms consistent image acquisition geometry across the dataset, reducing the need for aspect-ratio-specific preprocessing strategies.

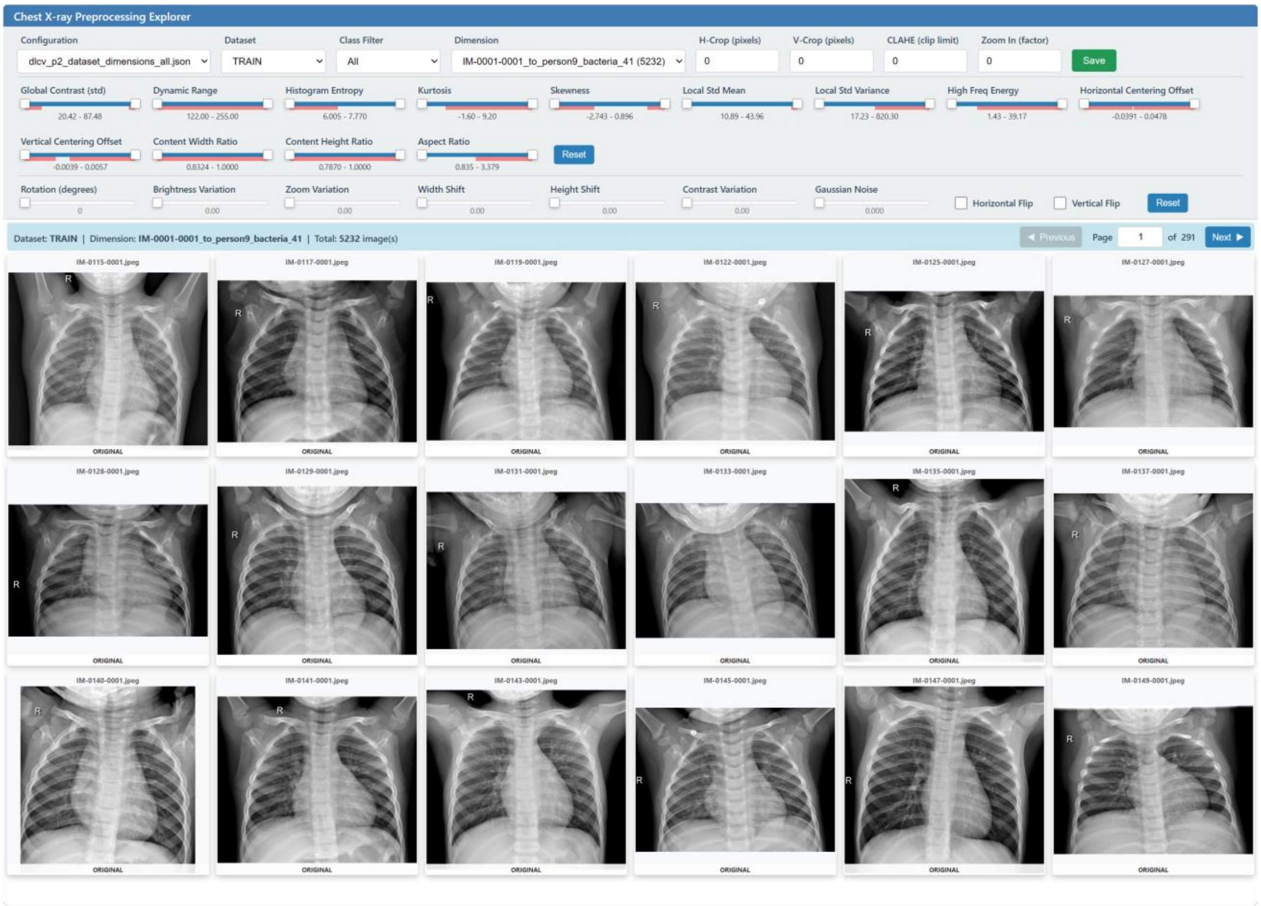
2.9 Data Understanding Metrics

2.9.1 Metrics Export

After computing image-level characteristics in the Data Understanding phase, the notebook exports a consolidated JSON file that contains these metrics for every image in the dataset. The file includes the image path, class label, dataset split, all computed quantitative metrics, and the corresponding outlier indicators. It serves as an external representation of the dataset's quality profile, derived directly from the analytical steps documented in this section.

2.9.2 CXRAY - Chest X-ray Preprocessing Explorer

This exported file is consumed by ***CXRAY***, an interactive preprocessing exploration application we developed specifically for Project 2. We made this tool available online at: https://logus2k.com/cxray. It provides a visual environment for examining the effects of different preprocessing operations, such as cropping strategies, zoom adjustments, and contrast modifications. When supplied with the exported metrics, the tool allows the user to relate visual inspection to the quantitative properties measured during Data Understanding. This connection supports more informed and consistent Data Preparation decisions, helping ensure that the preprocessing applied during model training is both statistically justified and visually validated.



Chest X-ray Preprocessing Explorer

Configuration: dlcv_p2_dataset_dimensions_all.json
Dataset: TRAIN
Class Filter: All
Dimension: IM-0001-0001_to_person9_bacteria_41 (5232)
H-Crop (pixels): 0
V-Crop (pixels): 0
CLAHE (clip limit): 0
Zoom In (factor): 0

Global Contrast (std): 20.42 - 87.48
Dynamic Range: 122.00 - 255.00
Histogram Entropy: 6.005 - 7.770
Kurtosis: -1.60 - 9.20
Skewness: -2.743 - 0.896
Local Std Mean: 10.89 - 43.96
Local Std Variance: 17.23 - 820.30
High Freq Energy: 1.43 - 39.17
Horizontal Centering Offset: -0.0391 - 0.0478
Vertical Centering Offset: -0.0039 - 0.0057
Content Width Ratio: 0.8324 - 1.0000
Content Height Ratio: 0.7870 - 1.0000
Aspect Ratio: 0.835 - 3.379

Rotation (degrees): 0
Brightness Variation: 0.00
Zoom Variation: 0.00
Width Shift: 0.00
Height Shift: 0.00
Contrast Variation: 0.00
Gaussian Noise: 0.000
□ Horizontal Flip
□ Vertical Flip

Dataset: **TRAIN** | Dimension: **IM-0001-0001_to_person9_bacteria_41** | Total: **5232** image(s)

Page 1 of 291

Figure 2.20 – Chest X-ray Preprocessing Explorer

3. Data Preparation

3.1 Data Splitting Strategy

The dataset's predefined train and test partitions are preserved to avoid leakage. A validation set is created by reserving a fixed proportion of the training data, using a deterministic random seed to ensure reproducibility. Split sizes and class distributions are documented to support transparent evaluation and to confirm that each subset remains representative of the overall dataset.

3.2 Data-Driven Augmentation Parameter Analysis

Data augmentation is a critical component of deep learning pipelines for medical imaging, particularly when working with datasets of limited size or class imbalance. However, the selection of augmentation parameters is often based on ad-hoc experimentation, domain conventions, or arbitrary defaults rather than on systematic analysis of the dataset's inherent characteristics. This approach risks introducing either insufficient variation, failing to improve model generalization, or excessive distortion, creating unrealistic samples that mislead the learning process and degrade performance on genuine clinical data.

The metrics computed in the preceding sections provide a quantitative foundation for deriving augmentation parameters directly from the dataset's natural variation. By measuring the statistical properties of brightness, contrast, spatial positioning, content framing, and noise across the existing images, it becomes possible to establish augmentation ranges that reflect realistic variability rather than speculative guesses. This data-driven approach ensures that augmented samples remain within the distribution of plausible radiographic appearances while still introducing sufficient diversity to prevent overfitting.

Class-specific analysis is particularly important for imbalanced datasets such as this one, where the PNEUMONIA class outnumbers the NORMAL class by a ratio of approximately 2.7:1.

Applying uniform augmentation to both classes may not address the representational asymmetry effectively. Instead, examining each class independently allows for tailored augmentation strategies: more aggressive transformation of the minority class to increase its effective sample size, and conservative augmentation of the majority class to preserve its existing diversity without introducing artifacts. This targeted approach supports more balanced learning and reduces the risk of the model developing a systematic bias toward the overrepresented class.

The following analysis computes coefficient of variation (CV) for brightness, contrast, and noise metrics, and standard deviation for geometric positioning and content framing.

These statistics are then translated into recommended parameter ranges for Keras preprocessing layers (RandomRotation, RandomTranslation, RandomZoom, RandomBrightness, RandomContrast, and GaussianNoise), providing concrete, defensible values that can be directly applied during model training through a Sequential augmentation pipeline. The analysis is performed separately for NORMAL and PNEUMONIA images, and concludes with strategic recommendations for handling class imbalance through either differential augmentation or class weighting.

3.3 Data Augmentation

Augmentation is applied exclusively to the training set and kept consistent across all models to enable fair comparison. Transformations are limited to clinically realistic variations that simulate routine acquisition differences, such as small rotations, translations, mild zoom, and conservative brightness adjustments. Anatomically implausible transformations, including vertical flips or aggressive geometric distortions, are avoided. Augmentation ranges are selected to preserve diagnostic content while improving generalization under typical real-world imaging variability.

3.4 Model-Specific Preprocessing Pipeline

Each model architecture employs a defined preprocessing pipeline, including resizing, channel arrangement, and normalization. Augmentation is applied before model-specific normalization to avoid altering standardized distributions. Pipelines for convolutional networks replicate grayscale images into three channels.

4. Modeling

4.1 Checkpoint Saving

We set up a callback to save the best model during training based on validation PR-AUC rather than raw accuracy. This is more appropriate for an imbalanced, clinically oriented problem where the balance between precision and recall is more important than overall accuracy.

The ModelCheckpoint callback:

- Saves the model to the file path defined by CHECKPOINT_PATH.
- Monitors the val_pr_auc metric on the validation set.
- Keeps only the model corresponding to the highest observed val_pr_auc during training.

By saving the best validation-PR-AUC checkpoint instead of simply using the weights from the final epoch, we reduce the risk of overfitting. If performance on the validation set starts to degrade after a certain point, we still retain the version of the model that showed the best generalization according to the chosen metric.

4.2 Early Stopping

We also use an EarlyStopping callback to automatically stop training when further epochs no longer improve the model's performance on the validation set.

In this project:

- EarlyStopping monitors the validation PR-AUC (val_pr_auc), not the training metrics.
- If val_pr_auc does not improve for a specified number of epochs (the patience value), training is stopped.
- With restore_best_weights=True, the model weights are rolled back to those from the epoch with the highest val_pr_auc.

This prevents wasting epochs once the model has effectively converged and reduces overfitting by avoiding continued training after validation performance has stopped improving, while keeping the best-performing version of the model for subsequent evaluation.

4.3 Learning Rate

The learning rate is a key hyperparameter that controls how much the model weights are updated during training. If it is too high, the optimizer may overshoot good minima and converge to a suboptimal solution; if it is too low, training can become very slow or get stuck.

In this project, we use the AdamW optimizer with an initial learning rate defined by LEARNING_RATE, together with a learning rate scheduler based on the ReduceLROnPlateau callback. The scheduler:

- Monitors the validation loss (val_loss).
- If val_loss does not improve for a certain number of epochs (patience), it reduces the learning rate by a given factor.
- Ensures that the learning rate never goes below a specified minimum (min_lr).

This strategy allows the model to make relatively large updates in the early stages of training, then gradually switch to smaller, finer updates once progress slows. In practice, this helps the model converge more reliably and can improve final validation performance without manual tuning of the learning rate schedule.

4.4 Build and Compile

The classification model is built using Keras' Functional API to process 256×256 grayscale images for binary prediction. Since ResNet50 requires 3-channel input, the grayscale image is replicated across three channels before being passed into the pretrained ImageNet backbone. The ResNet50 feature extractor is frozen to preserve its learned representations, and a lightweight dense classification head with L2 regularization, Batch Normalization and Dropout is added. The architectural configuration is summarized in Table 2.

Table 2 - Summary of the ResNet50 architecture, training configuration, and optimization components

Component	Description
Input layer	Accepts grayscale images of size **256×256×1**.
Grayscale to RGB	The single grayscale channel is **replicated into three channels** using Concatenate(axis=-1), producing a **256×256×3** tensor required by ResNet50.
Backbone	**ResNet50** pretrained on ImageNet, configured with include_top=False and pooling='avg'. The backbone is **frozen** (trainable=False) so its pretrained weights are not updated during training.
Global average pooling	Provided by ResNet50 through the pooling='avg' configuration, which converts the final convolutional feature maps into a single feature vector per image.
Dense head	A lightweight classification head consisting of: a Dense layer with 128 units and ReLU activation; Batch Normalization; a Dropout layer with 0.2 rate; a second Dense layer with 64 units and ReLU activation; and a final sigmoid output neuron for binary prediction.
Regularization	**L2 regularization** (1e-4) applied to Dense layers; Batch Normalization stabilizes training; Dropout helps prevent overfitting.
Model checkpointing	A ModelCheckpoint callback that saves the **best model** to CHECKPOINT_PATH, monitoring **val_pr_auc** in max mode and storing full model weights (save_weights_only=False).
Early stopping	An EarlyStopping callback that monitors **val_pr_auc** and restores the best weights after training stops. Uses **patience=20**, suitable for slower convergence when training with weight decay or schedule-based optimizers.

4.5 Two-Stage Training - Feature Extraction and Fine-Tuning

To leverage the power of pre-trained models like EfficientNet, we employ a **Two-Stage Transfer Learning** strategy. This approach maximizes performance while protecting the knowledge learned from the massive ImageNet dataset.

- **Stage 1: Feature Extraction (Base Frozen):** The pre-trained convolutional base (EfficientNet) is **frozen** to lock its weights. Only the newly added classification head is trained. This "warms up" the dense layers to interpret the base's features using a standard learning rate.
- **Stage 2: Fine-Tuning (Base Unfrozen):** After the head is stable, the entire model is **unfrozen**, but the learning rate is dramatically reduced ($\mathbf{10^{-5}}$). This allows the base

model to slightly adapt its features specifically to the texture and anatomy of X-ray images, ensuring stable convergence and optimal results.

4.5.1 Feature Extraction (Base Frozen)

In this initial stage, the base model (EfficientNetV2B0 or ResNet50V2) is set to **un-trainable**. We compile the model and train it for a small number of epochs (15 in this case) using a standard learning rate ($\mathbf{10^{-3}}$).

The primary goal is to **train the new classification head** (GlobalAveragePooling + Dense layers) to correctly map the frozen features produced by the base to our specific binary labels (Normal/Pneumonia). This stabilizes the head before we attempt fine-tuning.

4.5.2 Fine-Tuning (Base Unfrozen)

After Stage 1 stabilizes the classification head, we proceed to **unfreeze the entire base model** (base_model.trainable = True).

To prevent the destruction of the pre-trained weights, the model is **recompiled** with a **significantly lower learning rate** ($\mathbf{10^{-5}}$). This low rate ensures the training process only makes small, careful adjustments to the base model’s weights, adapting them to the X-ray domain while maintaining a focus on high validation PR-AUC. The learning rate scheduler and early stopping continue to manage the process.

5. Evaluation

In this section, the final model is evaluated on a held-out test set using clinically relevant classification metrics (accuracy, precision, recall, F1-score and AUC), together with confusion matrices, Precision–Recall analysis and threshold-based error summaries. We also inspect the loss and accuracy curves over epochs to verify that training converged without severe overfitting before selecting the best checkpoint for testing.

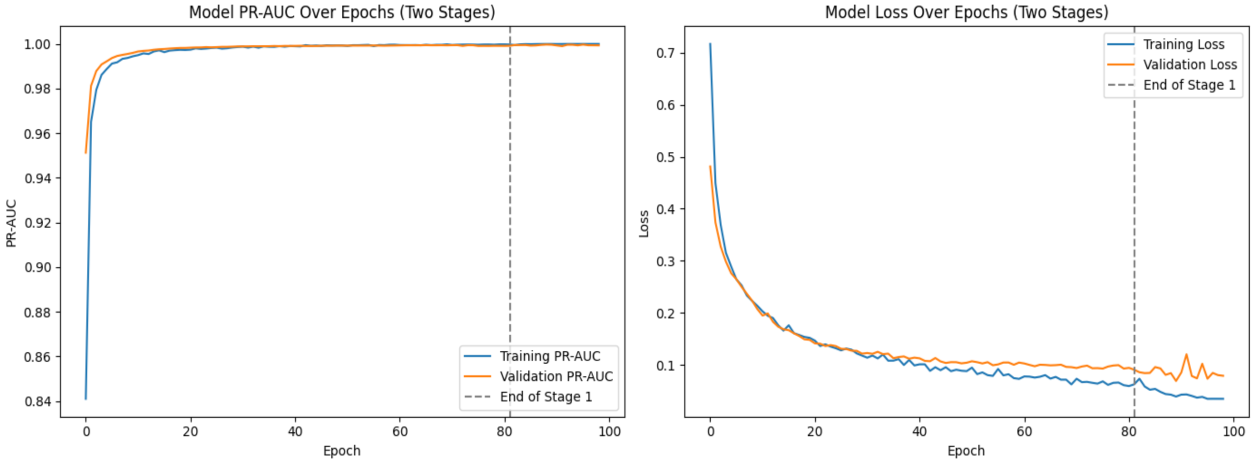


Figure 5.1 - Training and validation accuracy and loss curves over epochs for the CNN model.

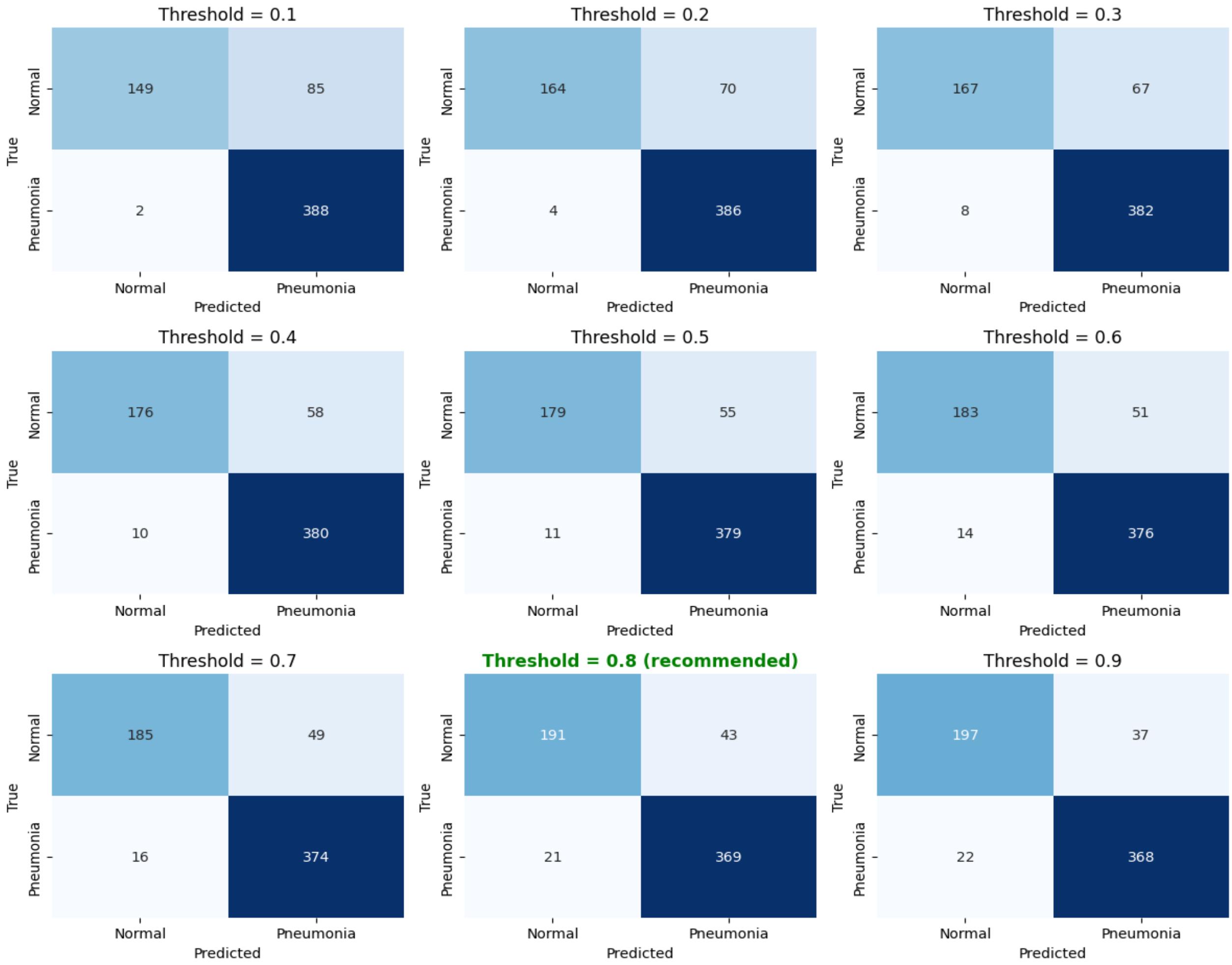


Figure 5.1 - Confusion matrices on the test set for thresholds

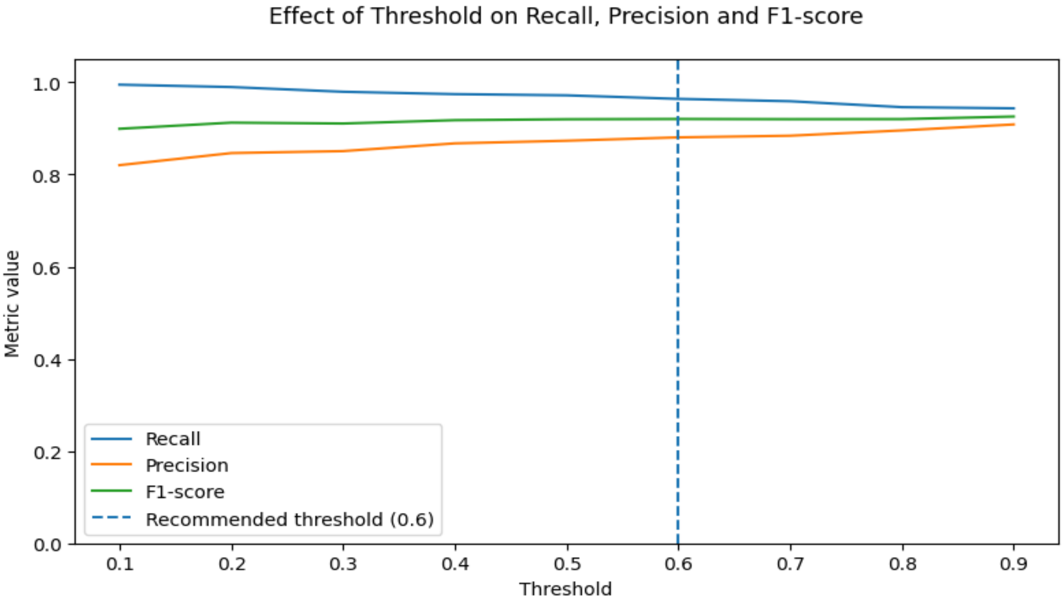


Figure 5.2 - Recall, precision and F1-score on the test set as a function of threshold.

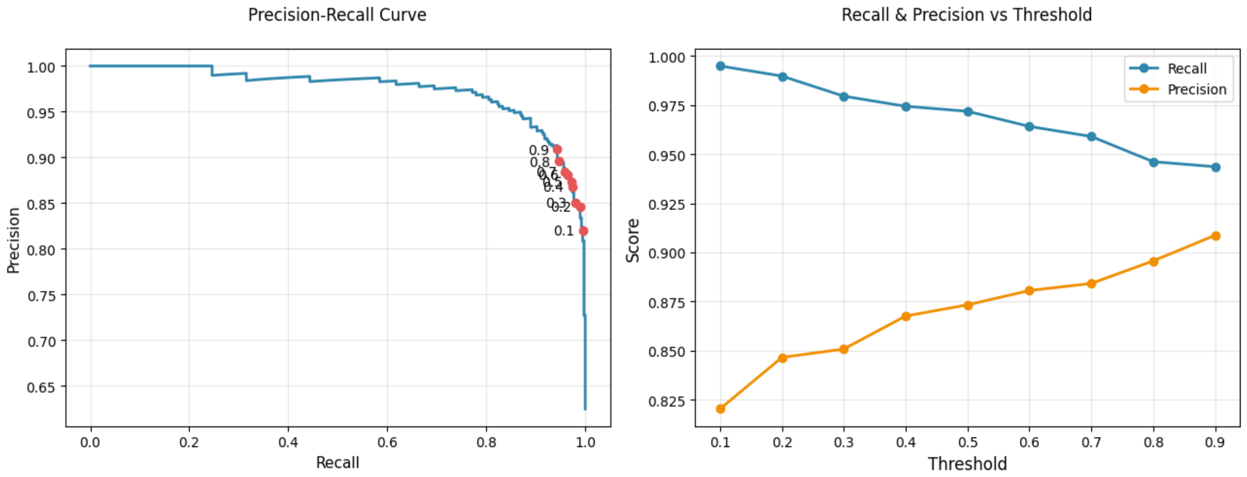


Figure 5.3 - Precision–recall curve and recall/precision vs threshold on the test set.

5.1 Pneumonia Cases Missed

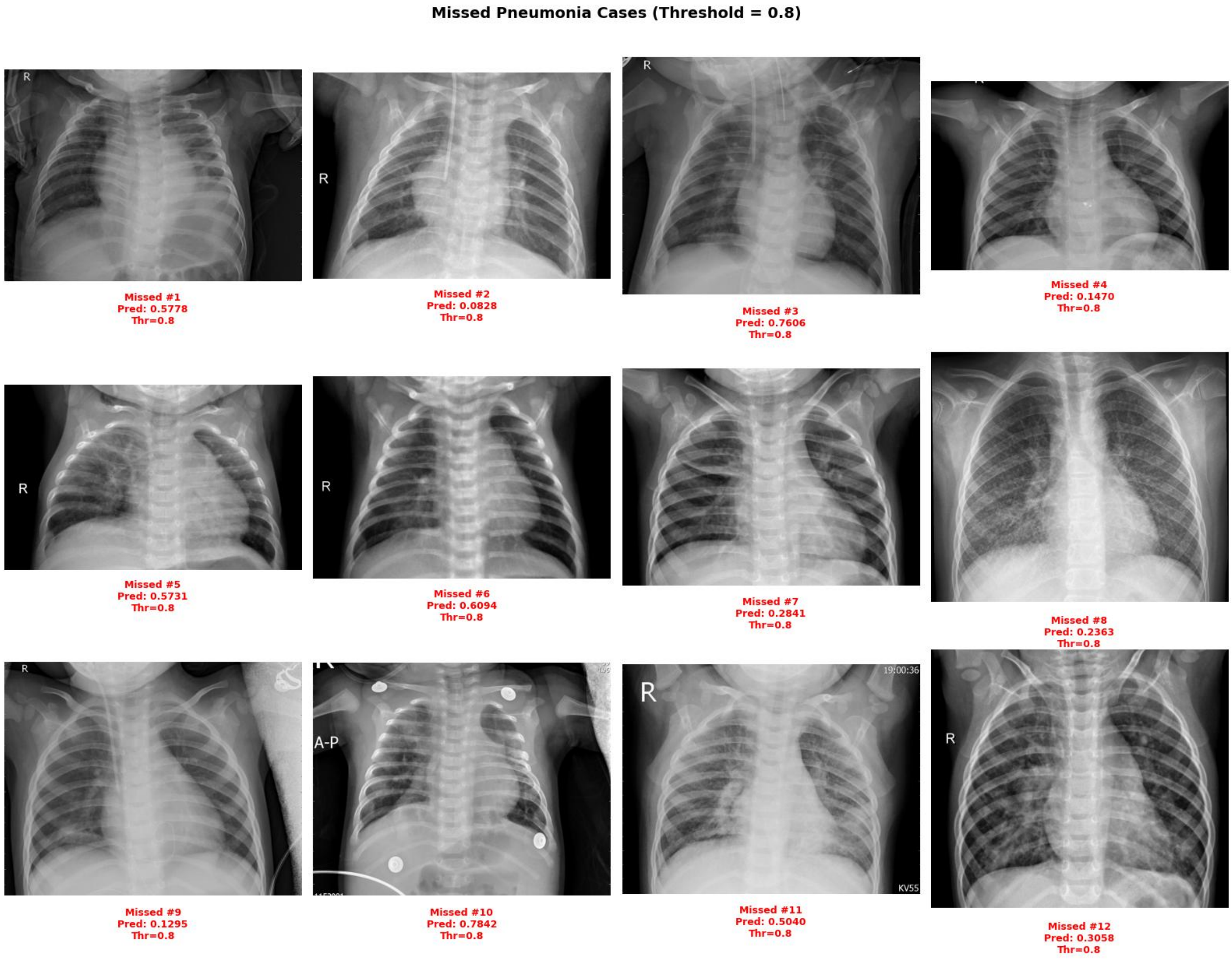


Figure 5.4 - False Negatives: Missed Pneumonia Cases (Top 12)

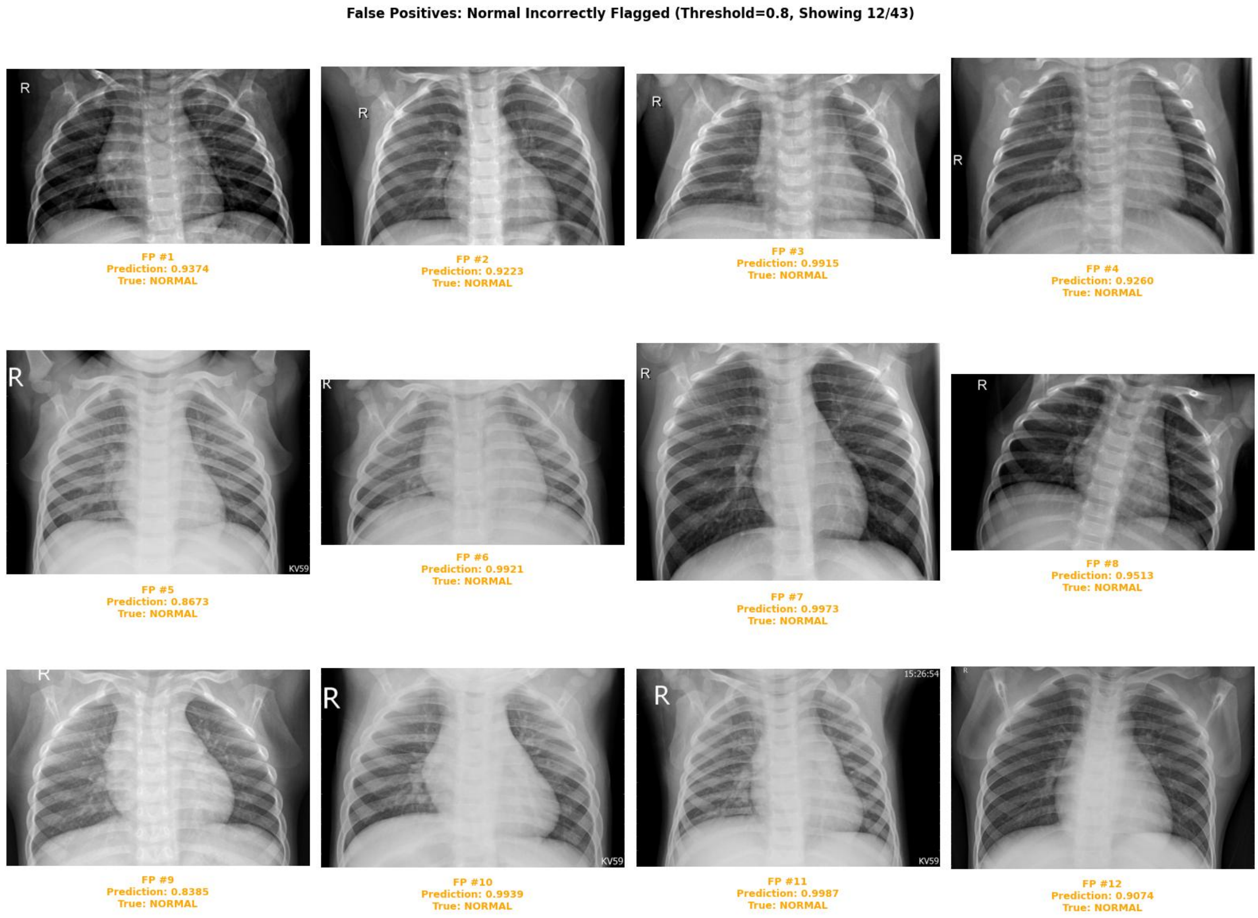


Figure 5.5 - False Positives: Normal Incorrectly Flagged (Top 12)

5.2 Specificity and Case Volume Reduction

5.3 Comparison of Model Performance to a Clinical AI Study

The following plot, compares our model results with the results reported in the article in the same case scenario. This plot remarks both the workload reduction in medical attention, unecessary contacts and reduction in false negatives.

The left plot compares specificity (the percentage of “unremarkable” or normal cases correctly excluded) between Our Model and Article AI at three very high recall (sensitivity) thresholds: 99.9%, 99.0%, and 98.0%. High recall thresholds mean the models are tuned to miss as few true positives as possible. Across all high-recall operating points, our model consistently achieves higher specificity than the model reported in the article. This mean it better filters out normal cases while still maintaining extremely strong sensivity.

In contrast, the right plot compares the Case Volume Reduction (CVR) achieved by Our Model versus Article AI at three high sensitivity (recall) thresholds: 99.9%, 99.0%, and 98.0%. CVR represents the estimated reduction in radiologists workload, that is, how many cases can be safely removed from manual review.

As the recall threshold decreases slightly (from 99.9% → 98.0%), both models are allowed to filter out more normal cases, resulting in higher CVR. Even at extremely high recall levels, both systems offer meaningful workload reduction. Across all high-sensitivity operating points, Our Model consistently achieves higher case-volume reduction than the model reported in the article. This means it can more effectively eliminate low-value, normal cases from radiologists' queues - improving efficiency without compromising diagnostic safety.

PubMed Article, in National Library of Medicine: Using AI to Identify Unremarkable Chest Radiographs for Automatic Reporting, Radiology 2024 Aug;312(2):e240272. doi: 10.1148/radiol.240272.

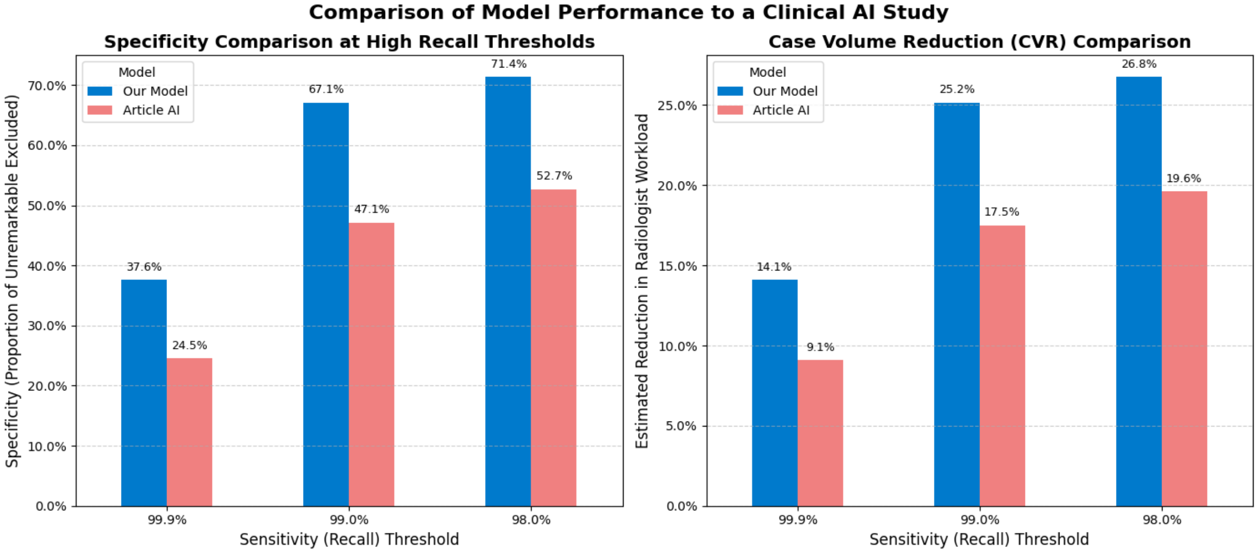


Figure 5.6 – Comparison of model performance to a clinical AI study

6. Conclusions

6.1 Introduction

This project compared two approaches for automated pneumonia detection from chest X-ray images: a custom CNN trained from scratch (Project 1) and a transfer-learning model based on ResNet50V2 with artifact-aware data augmentation (Project 2). The transfer-learning model achieved superior overall performance across accuracy, precision, and F1-score. At their recommended operating points, Project 2 reached approximately 90% accuracy compared with 81% for Project 1, and substantially improved precision (about 90% vs. 76%).

Important factors contributing to the performance of Project 2 included the pretrained feature representations of ResNet50V2, augmentation specifically designed to counteract dataset artifacts, conservative learning rates, and a two-stage fine-tuning strategy that preserved useful pretrained knowledge.

Project 1, although less precise, achieved perfect sensitivity at thresholds up to 0.6, making it well suited to settings where missing a pneumonia case is unacceptable.

From a clinical perspective, both models reached strong levels of sensitivity. Project 2, however, offered a far better balance between sensitivity and specificity. At threshold 0.8, it achieved approximately 95% sensitivity and 82% specificity, reducing false alarms by 36–43 cases compared with Project 1 at similar recall levels. This corresponds to a potential reduction of 6–7% in radiologist workload.

From a research standpoint, the results highlight the importance of targeted augmentation and carefully controlled transfer learning in improving generalization and mitigating the influence of dataset artifacts.

6.2 Model Architectures

Project 1: Custom CNN

- Three convolutional layers (64 → 128 → 256 filters)
- Approximately 1.2 million parameters
- Trained from scratch with no data augmentation
- Batch size 32
- Learning rate 1e-4 with adaptive reduction

Project 2: ResNet50V2 Transfer Learning

- ResNet50V2 backbone with a dense classification head (128 → 64 units)
- Approximately 25 million parameters, with 100k trainable during the initial stage
- Two-stage training (frozen backbone followed by full fine-tuning)
- Artifact-aware augmentation targeting centering, brightness, and framing differences
- Batch size 16
- Learning rates of 1e-5 and 1e-6

6.3 Performance Comparison

The recommended operating point for Project 1 is threshold 0.6. Project 2 achieves its most balanced performance at threshold 0.8.

Table 6.1 - Project 1 vs Project 2 performance comparison

Metric	Project 1 (θ = 0.6)	Project 2 (θ = 0.8)
Sensitivity	100.0%	94.62%
Specificity	48.7%	81.6%
Precision	76.47%	89.56%
F1-score	86.67%	92.02%
Accuracy	80.77%	89.74%

Project 1 displays an extreme sensitivity bias, missing no pneumonia cases at the selected threshold. Project 2 provides a significantly more balanced profile with higher precision and far greater specificity.

6.4 Precision-Recall Trade-off Analysis

The custom CNN in Project 1 yields perfect recall for thresholds up to 0.6, but at the expense of high false-positive rates and reduced specificity. This reflects the challenges of training a model from scratch on a modest dataset with no augmentation, coupled with class imbalance that encourages over-prediction of pneumonia.

Project 2 achieves high recall while maintaining substantially lower false-positive rates. Its balanced trade-off is driven by stronger pretrained features, targeted augmentation that reduced dependence on acquisition artifacts, and conservative fine-tuning.

6.5 False Positive Reduction Analysis

When comparing operating points with similar recall levels, Project 2 produces far fewer false positives.

Table 6.2 - Project 1 vs Project 2 recall levels

Operating Point	Project 1	Project 2	FP Reduction
High recall (~99%)	θ = 0.6 → 120 FP	θ = 0.2 → 70 FP	42%
Balanced recall (97–98%)	θ = 0.8 → 97 FP	θ = 0.6 → 51 FP	47%

Across the 624-patient test set, these reductions correspond to 36–43 fewer false alarms, reducing radiologist workload by approximately 6–7% while maintaining high sensitivity.

6.6 Impact of Data Augmentation

The dataset exhibited systematic differences between classes in centering, brightness, and framing. Project 2 applied targeted augmentations to reduce these biases through translation, brightness adjustments, and slight zooming. This strategy improved precision, decreased false-positive rates, produced smoother probability calibration, and strengthened generalization, leading the model to focus on anatomical pathology rather than acquisition artifacts.

6.7 Learning Rate Sensitivity

Transfer learning required careful control of the learning rate. Larger learning rates such as 1e-3 and 1e-4 caused catastrophic forgetting or unstable convergence. Learning rates of 1e-5 for the initial stage and 1e-6 for full fine-tuning produced stable improvements and preserved pretrained representations. This highlights the need for substantially lower learning rates when adapting pretrained networks to medical imaging.

6.8 Recommended Operating Points for Clinical Deployment

Emergency Screening (maximizing sensitivity)

- Model: Project 1
- Threshold: 0.6
- Performance: 100% recall, 76.47% precision, 120 false positives
- Appropriate when missing pneumonia cases must be avoided

General Radiology Workflow (balanced)

- Model: Project 2
- Threshold: 0.6
- Performance: 96.41% recall, 88.06% precision, 51 false positives
- Reduces workload while retaining high sensitivity

Pre-Diagnostic Triage (maximizing specificity)

- Model: Project 2
- Threshold: 0.8
- Performance: 94.62% recall, 89.56% precision, 43 false positives
- Suitable for prioritizing cases most likely to represent pneumonia

6.9 Future Directions

Future work may focus on enhanced interpretability through techniques such as Grad-CAM or attention visualizations. Uncertainty estimation could identify ambiguous cases suited for expert review. Model ensembles and exploration of advanced architectures such as Vision Transformers may further improve robustness and generalization.

6.10 Final Remarks

Both models demonstrated strong and clinically relevant performance. The transfer-learning model in Project 2 achieved a balanced combination of sensitivity, precision, and specificity suitable for routine workflows, whereas the custom CNN in Project 1 offered maximum sensitivity for primary triage scenarios.

These findings emphasize the importance of augmentation, careful fine-tuning, and appropriate threshold selection in the development of medical AI systems.