MSc Data Science

Athens University of Economics and Business

Deep Learning

Respiratory Sound Recognition Using Deep Learning:

A Comparative Study of CNN, RESNET50, and DENSENET121 Architectures

Professor: Stafylakis Themos

Giagkos Stylianos | f3352410

June 2025

Table of Contents

1. Abstract

The automatic recognition of lung sounds plays a vital role in early diagnosis and management of respiratory conditions such as asthma, chronic obstructive pulmonary disease (COPD), and pneumonia. This study explores the application of deep learning models—Custom Convolutional Neural Networks (CNN), ResNet50, and DenseNet121—for classifying lung sounds into seven diagnostic categories. Leveraging a carefully designed preprocessing pipeline involving resampling, Discrete Fourier Transform (DFT)-based baseline filtering, snippet generation, and spectrogram conversion, the raw lung sound signals are transformed into a representation suitable for image-based classification. Furthermore, robust data augmentation strategies are employed to enhance model generalization. Experimental evaluations are conducted using patient-wise splits and macro-averaged F1-scores to account for class imbalance. The results demonstrate that deeper pre-trained models such as DenseNet121 significantly outperform baseline CNNs, achieving the highest F1-score on the validation set. This work highlights the potential of transfer learning and spectrogram-based approaches in clinical auscultation tasks and outlines future directions for real-world deployment.

Keywords: Lung Sound Classification, Deep Learning, Spectrograms, CNN, ResNet50, DenseNet121, Biomedical Signal Processing, DFT Filtering

2. Introduction

The analysis of lung sounds through auscultation is a cornerstone of respiratory diagnostics in clinical practice. Traditionally performed using stethoscopes by trained clinicians, this technique is inherently subjective and can suffer from inter-observer variability and limited sensitivity to subtle acoustic indicators of disease. With the advent of digital health technologies and deep learning, the automated recognition of pathological lung sounds offers a promising avenue for supporting medical diagnosis—especially in resource-constrained environments and remote monitoring settings.

Lung sounds such as wheezes, crackles, rhonchi, and stridor are acoustic manifestations of various underlying respiratory abnormalities. Accurate identification and classification of these sounds can lead to timely intervention and improved patient outcomes. However, lung sounds are typically non-stationary and low in frequency, often occurring amidst substantial background noise. These properties pose significant challenges to classical signal processing approaches. Deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a robust tool for modeling complex patterns in both time-series and image representations of biomedical audio data.

In this study, we investigate a deep learning pipeline for the classification of lung sounds using three distinct architectures: a custom-built CNN, the ResNet50 model, and DenseNet121. Each model is trained and evaluated on spectrograms derived from raw lung sound recordings after a rigorous preprocessing pipeline. To enhance generalization and model robustness, several data augmentation strategies are employed, including noise injection, time shifting, and pitch alteration.

The main goals of this research are:

1. To implement a reproducible and robust preprocessing pipeline specifically designed for lung sound signals.
2. To construct spectrogram-based datasets that enable deep image-based classification of respiratory audio.
3. To compare baseline and transfer learning models in terms of classification performance in a multi-class setting.
4. To discuss the practical applications and limitations of deploying deep learning-based auscultation systems in real-world clinical practice.

Automated lung sound classification has become increasingly feasible due to the integration of machine learning techniques in biomedical signal processing. Early approaches focused on handcrafted feature extraction using time-frequency transforms such as the Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCCs), and wavelet transforms. Although these techniques allowed for basic differentiation between healthy and pathological sounds, they struggled to generalize across varying recording conditions and patient characteristics.

With the rise of deep learning, CNNs have demonstrated substantial success in audio classification tasks, particularly when applied to spectrogram representations of sound signals. These 2D visualizations of frequency content over time provide an ideal input format for image-based neural networks. In the context of respiratory analysis, CNNs have been employed to detect specific pathologies, such as wheezes and crackles, by learning local and global patterns from spectrogram images.

Transfer learning has further enhanced the performance of deep learning models for medical audio classification. Architectures such as ResNet50 and DenseNet121—originally designed for large-scale image recognition tasks—have shown remarkable results when fine-tuned on spectrogram data. ResNet50 incorporates residual connections that alleviate vanishing gradient issues in deep networks, while DenseNet121 connects each layer to all subsequent layers, promoting feature reuse and efficient training. These characteristics make them highly suitable for medical applications where labeled data is limited and noise is prevalent.

Moreover, the combination of signal denoising, baseline wandering removal, and temporal segmentation has proven beneficial in preprocessing pipelines. These techniques help standardize input quality and mitigate confounding factors,

allowing the models to focus on clinically relevant acoustic features. Despite these advancements, few studies have systematically compared standard CNN architectures with pre-trained transfer learning models under a unified experimental setting. Our work aims to address this gap by evaluating and benchmarking all three models—CustomCNN, ResNet50, and DenseNet121—on the same dataset, using consistent preprocessing, augmentation, and evaluation protocols.

To provide medical context to the classification task, below is a brief explanation of each of the ten diagnostic labels used in this study:

- **Healthy:** Represents individuals with no detectable respiratory abnormalities, serving as the control group for classification.
- **Asthma:** A chronic inflammatory disease of the airways characterized by recurrent episodes of wheezing, breathlessness, and chest tightness.
- **Pneumonia:** An acute lung infection that inflames the air sacs, often filled with pus or fluid, leading to productive cough, fever, and difficulty breathing.
- **Chronic Obstructive Pulmonary Disease (COPD):** A progressive disease causing airflow limitation due to chronic bronchitis or emphysema, often linked to long-term smoking.
- **Bronchiectasis:** A condition marked by the irreversible dilation of bronchi, resulting in mucus buildup and frequent lung infections.
- **Upper Respiratory Tract Infection (URTI):** Infections affecting the nasal passages, pharynx, or larynx, commonly caused by viruses and characterized by coughing, congestion, and sore throat.
- **Lung Fibrosis:** A group of conditions that cause scarring (fibrosis) of the lung tissue, leading to chronic respiratory impairment and reduced lung elasticity.
- **Bronchiolitis:** A viral infection that affects the bronchioles, commonly seen in infants and young children, and characterized by wheezing and rapid breathing.
- **Pleural Effusion:** The accumulation of fluid between the layers of tissue lining the lungs and chest cavity, potentially compressing lung function.
- **Lower Respiratory Tract Infection (LRTI):** Infections involving the bronchi and lungs, including bronchitis and pneumonia, usually associated with productive cough and chest discomfort.

These categories form the basis of our multi-class classification task, with the goal of distinguishing between diverse respiratory pathologies based solely on lung sound recordings.

3. Related Work

Automated lung sound classification has gained traction in recent years due to the integration of machine learning techniques into biomedical signal processing. Early methods relied heavily on handcrafted features extracted from time-frequency representations such as the Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCCs), and wavelet transforms. These classical approaches often suffered from limited generalizability due to the inherent variability in recording conditions and patient-specific acoustic characteristics.
With the rise of deep learning, Convolutional Neural Networks (CNNs) have emerged as a dominant paradigm for audio classification tasks. CNNs are particularly effective in capturing spatial hierarchies in spectrograms, which represent acoustic signals as 2D images. Applications in respiratory sound classification have leveraged CNNs for detecting wheezes and crackles. These studies highlight the value of transforming lung sounds into visual domains to enable effective deep learning-based recognition.

Transfer learning with pre-trained image classification models such as ResNet and DenseNet has further improved performance in medical audio classification tasks. ResNet uses residual connections to address the vanishing gradient problem in deep networks. DenseNet connects each layer to every other layer in a feed-forward fashion, facilitating feature reuse and improved gradient flow. Both architectures have demonstrated superior accuracy and sample efficiency in medical image classification, and recent studies have successfully adapted them for audio-based diagnoses by training on spectrogram inputs.
In the respiratory domain, research has increasingly turned to hybrid pipelines combining advanced preprocessing techniques with deep learning models. Signal denoising, baseline wandering removal, and temporal segmentation have been shown to enhance input quality and model performance. However, studies comparing classical CNNs to state-of-the-art pre-trained models in a controlled setup remain limited. This work fills that gap by evaluating a custom CNN, ResNet50, and DenseNet121 under a unified pipeline for the task of seven-class lung sound classification.

4. Dataset and Preprocessing

The lung sound dataset used in this study consists of audio recordings in WAV format, each associated with one of seven diagnostic labels. These recordings represent various respiratory pathologies including wheezing, crackles, and bronchial breathing patterns. The preprocessing pipeline transforms these raw recordings into spectrogram images suitable for input to deep learning models.

4.1 Resampling and Standardization

All audio recordings are first downsampled to a uniform sampling rate of 4,000 Hz to reduce computational cost and ensure consistency across samples. Resampling helps mitigate discrepancies introduced by differing acquisition devices or sampling conditions.

4.2 DFT-Based Baseline Removal

To remove baseline wandering and low-frequency drift artifacts that may interfere with meaningful acoustic features, a Discrete Fourier Transform (DFT)-based filtering approach is employed. Specifically, the baseline component is approximated by summing the DFT coefficients corresponding to the lowest 8 frequencies and subtracting this estimate from the original signal. This method preserves the respiratory acoustic features while eliminating undesirable trends.

$$x_{\text{filtered}}(t) = x(t) - \sum_{k=0}^{K} \text{Re}[\hat{x}(f_k)]$$

where $\hat{x}(f_k)$ denotes the DFT coefficient at frequency fk and K is a low-frequency threshold (here K=8).

4.3 Snippet Generation and Overlapping Windows

Given the variability in signal length across recordings, each processed signal is segmented into 5-second windows with a 50% overlap. This windowing strategy increases the number of training samples and enables localized analysis of signal characteristics. Only windows exceeding a minimum energy threshold are retained, ensuring the removal of silent or irrelevant segments.

4.4 Data Augmentation

To enhance model generalization and mitigate overfitting, the following augmentation techniques are applied:

- **White noise injection**: Adds low-level Gaussian noise to simulate environmental variability.
- **Time shifting**: Randomly shifts the signal temporally by a few milliseconds to introduce temporal variance.
- **Pitch shifting and time stretching**: Simulates variations in vocal tract characteristics by altering the frequency content or speed of playback.

Each augmentation is applied probabilistically to expand the diversity of the training set.

4.5 Spectrogram Generation

Finally, each normalized audio snippet is converted into a log-mel spectrogram using a window length of 512 samples and 50% overlap. These spectrograms are resized to 224×224 resolution to match the input dimensions required by pre-trained CNN architectures. The mel-spectrograms are stored as PNG images, forming the final input dataset for classification models.

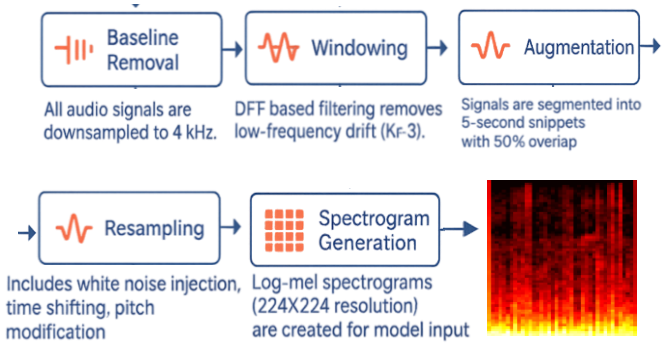

Figure 1 Lung Sound .wav Files Preprocessing

5. Models Methodology

5.1 Baseline Models

5.1.1 Custom CNN

The **CustomCNN** model is a deep convolutional neural network specifically designed for lung sound classification using 2D spectrogram inputs. The model accepts RGB spectrogram images with a fixed size of 224×224, and performs hierarchical feature extraction through a series of convolutional and pooling layers before feeding the features into fully connected layers for classification.

Architecture Overview

- Input Layer:
  - Shape: 3×224×224 (RGB spectrogram images)
- Convolutional Block 1:
  - Conv2D: 3 input channels, 32 output channels, 3×33 kernel, stride 1, padding 1.
  - Activation: ReLU
  - MaxPool2D: 2×22 with stride 2
  - Output shape: 32×112×112
- Convolutional Block 2:
  - Conv2D: 32 input channels, 64 output channels, 3×33 kernel
  - Activation: ReLU
  - MaxPool2D: 2×22
  - Output shape: 64×56×56
- Convolutional Block 3:
  - Conv2D: 64 input channels, 128 output channels, 3×33 kernel
  - Activation: ReLU
  - MaxPool2D: 2×22
  - Output shape: 128×28×28
- Flatten Layer:
  - Flattens the tensor from 128×28×28 to 100352
- Fully Connected Layer 1:
  - Linear: 100352 input features → 512 output features
  - Activation: ReLU
- Fully Connected Layer 2 (Output Layer):
  - Linear: 512 → num_classes (7 for lung sound categories)

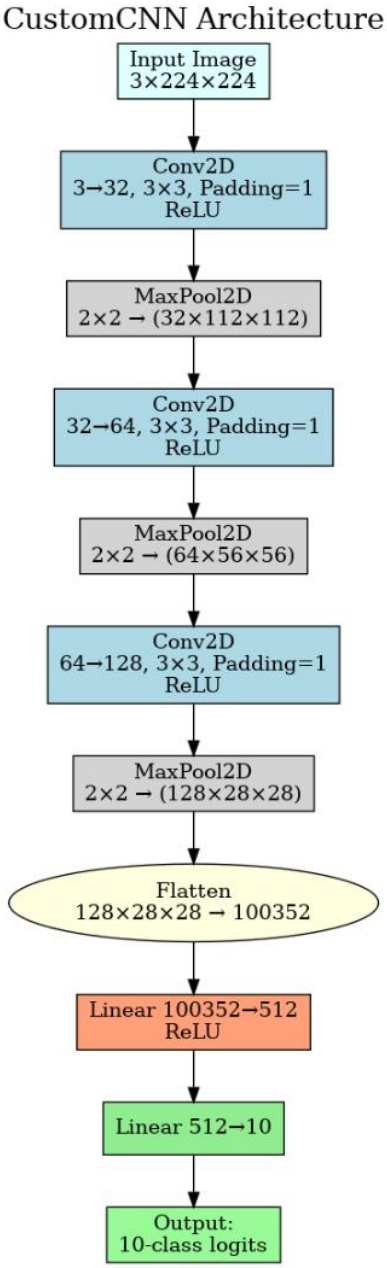

Figure 2 Custom CNN Architecture

5.1.2 Pretrained ResNet

The **ResNet50-based model** utilizes a **pretrained ResNet50** from torchvision as a feature extractor, which is a deep residual convolutional neural network consisting of 50 layers and residual connections that mitigate vanishing gradient issues during training (He et al., 2016). The model is adapted to handle single-channel (grayscale) spectrogram inputs by modifying the first convolutional layer. All layers except the final fully connected classification layer are used for feature extraction. These extracted features are then passed to a **custom residual classifier block**, improving the model's learning capacity and generalization.

Architecture Overview

- Input Layer:
  - Shape: 1×224×224 (Grayscale spectrogram images)
- Modified Initial Convolution:
  - Conv2D: 1 input channel → 64 output channels, 7×7 kernel, stride 2, padding 3
  - Activation: ReLU
  - Followed by MaxPool2D: 3×3 kernel, stride 2
  - Output shape: 64×56×56 (approx.)
- ResNet50 Backbone (Pretrained):
  - 4 Residual stages with identity and bottleneck blocks
  - Includes BatchNorm, ReLU, and skip connections in each residual block
  - Produces high-level feature maps
  - Output shape before pooling: 2048×7×7
- Global Average Pooling:
  - AdaptiveAvgPool2D: Output size 1×1
  - Output shape: 2048-dimensional vector
- Residual Block (Custom Classifier):
  - Linear: 2048 → 1024
  - BatchNorm + ReLU + Dropout
  - Linear: 1024 → 1024
  - Residual skip connection: 2048 → 1024
  - Output shape: 1024
- Fully Connected Layer (Output):
  - Linear: 1024 → num_classes (e.g., 10 for lung diseases)

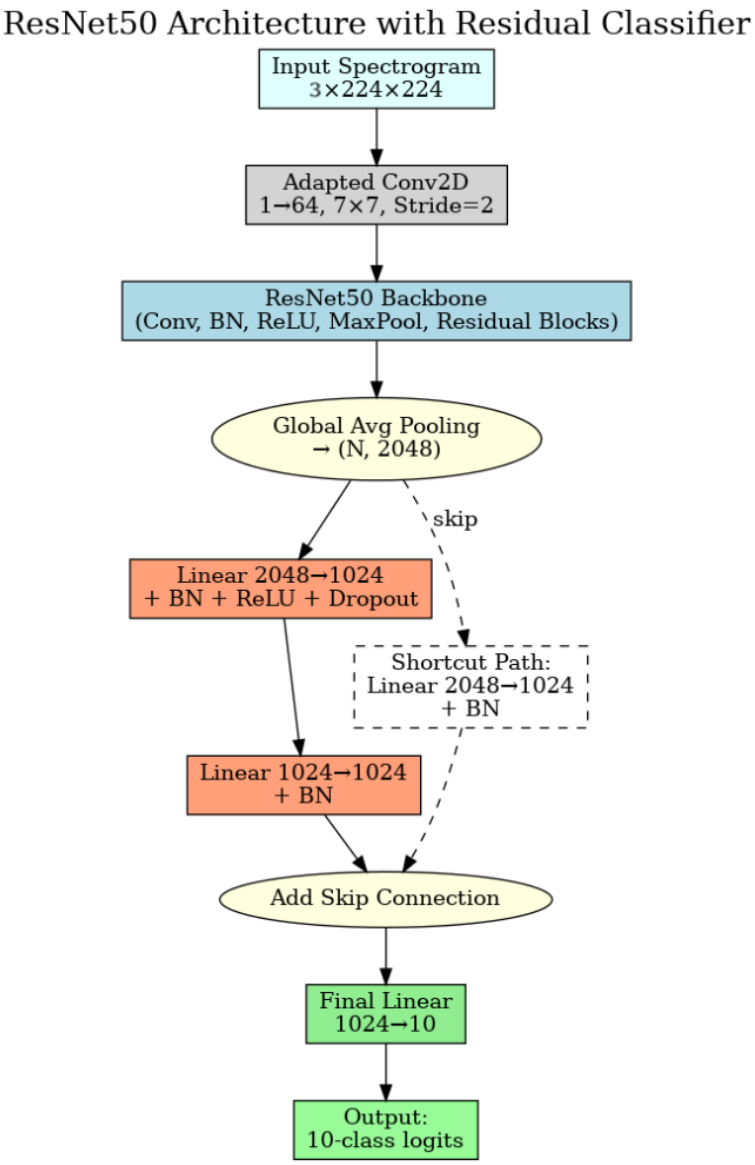

Figure 3 ResNet Architecture

5.1.3 Pretrained DenseNet

The **DenseNet121-based model** uses a pretrained **DenseNet121** backbone, known for its **dense connections** where each layer receives input from all preceding layers (Huang et al., 2017). This structure encourages feature reuse, reduces the number of parameters, and improves gradient flow. As with the ResNet variant, the input layer is modified for grayscale spectrograms. The output from the convolutional feature maps undergoes global average pooling and then flows through the same **custom residual block classifier**.

Architecture Overview

- Input Layer:
  - Shape: 1×224×224 (Grayscale spectrogram images)
- Modified Initial Convolution:
  - Conv2D: 1 input channel → 64 output channels, 7×7 kernel, stride 2, padding 3
  - Activation: ReLU
  - Followed by MaxPool2D: 3×3 kernel, stride 2
  - Output shape: 64×56×56 (approx.)
- DenseNet121 Backbone (Pretrained):
  - Dense blocks with dense connectivity:
    - Each layer receives feature maps from all preceding layers
  - Growth rate: 32
  - Transition layers downsample feature maps with 1×1 conv and pooling
  - Output shape before pooling: 1024×7×7
- Global Average Pooling:
  - AdaptiveAvgPool2D: Output size 1×1
  - Output shape: 1024-dimensional vector
- Residual Block (Custom Classifier):
  - Linear: 1024 → 512
  - BatchNorm + ReLU + Dropout
  - Linear: 512 → 512
  - Residual skip connection: 1024 → 512
  - Output shape: 512
- Fully Connected Layer (Output):
  - Linear: 512 → num_classes (e.g., 10 for lung diseases)

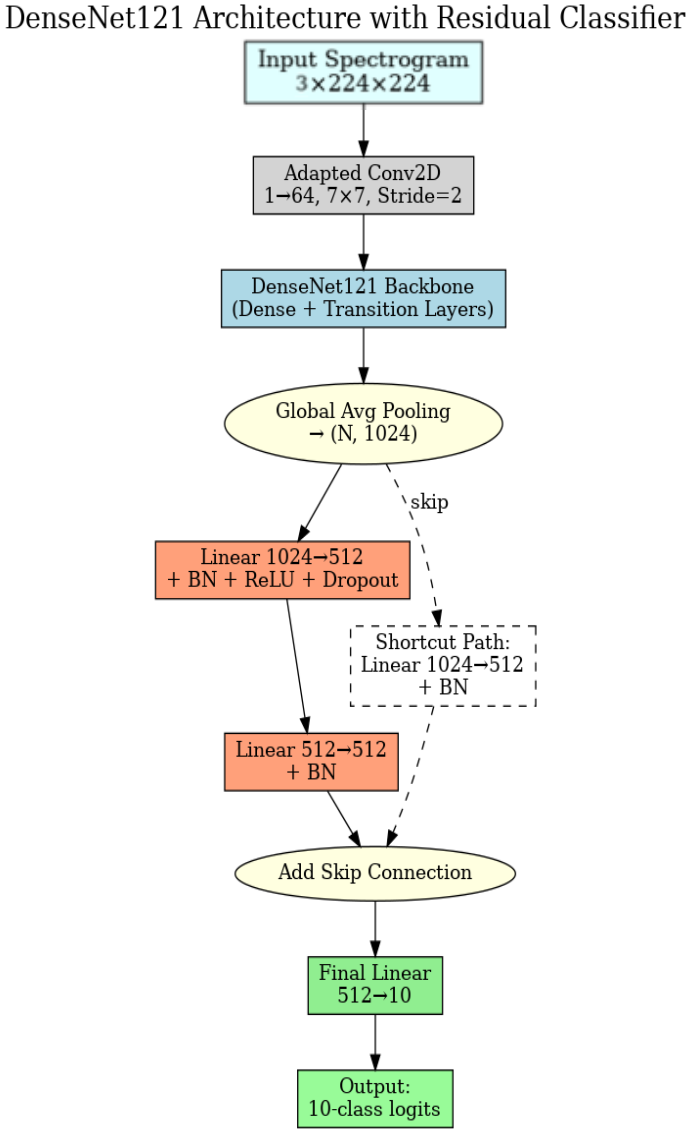

Figure 4 DenseNet Architecture

5.1.4 CORAL: CORrelation ALignment for Domain Adaptation

To address domain shift between source and target distributions, we integrate the **CORrelation ALignment (CORAL)** loss (Sun & Saenko, 2016) into our training objective. CORAL aims to minimize the difference in second-order statistics—specifically, the covariance—between deep feature representations extracted from the source and target domains. This promotes **domain-invariant feature learning** without requiring target labels.

Given two batches of extracted features, $Xs \in R^{ns \times d}$ from the source domain and $Xt \in R^{nt \times d}$ from the target domain, CORAL first **centers** the features by subtracting the mean of each dimension:

$$\tilde{\mathbf{X}}_s = \mathbf{X}_s - \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{x}_s^{(i)}, \quad \tilde{\mathbf{X}}_t = \mathbf{X}_t - \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{x}_t^{(i)}.$$

It then computes the **empirical covariance matrices**:

$$\mathbf{C}_s = \frac{1}{n_s - 1}\tilde{\mathbf{X}}_s^\top \tilde{\mathbf{X}}_s + \epsilon\mathbf{I}, \quad \mathbf{C}_t = \frac{1}{n_t - 1}\tilde{\mathbf{X}}_t^\top \tilde{\mathbf{X}}_t + \epsilon\mathbf{I},$$

where ϵ is a small constant added to ensure numerical stability, and I is the identity matrix.

The CORAL loss is defined as the squared **Frobenius norm** between the two covariance matrices:

$$\mathcal{L}_{\text{CORAL}} = \|\mathbf{C}_s - \mathbf{C}_t\|_F.$$

This formulation aligns the feature distributions in a **non-adversarial and efficient** manner. The loss is backpropagated along with the primary classification loss, and the total training objective becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{CORAL}},$$

where λ is a hyperparameter balancing classification performance with domain alignment.

The CORAL implementation used in our experiments closely follows this formulation and is applied to the feature representations immediately prior to the classification head.

5.1.5 CustomCNN CorAl

The **CustomCNN** is a lightweight, fully convolutional neural network designed specifically for classification of spectrogram images. It is not pretrained and is trained from scratch on the task-specific dataset. This model supports **CORAL loss** for domain adaptation through its extract_features() method, which exposes internal features before classification.

Architecture Overview

- **Input Layer:**
  - Shape: 3×224×224 (RGB spectrogram images)
- **Convolutional Block 1:**
  - Conv2D: 3 → 32, kernel: 3×3, stride 1, padding 1
  - Activation: ReLU
  - MaxPool2D: 2×2, stride 2
  - Output shape: 32×112×112
- **Convolutional Block 2:**
  - Conv2D: 32 → 64, kernel: 3×3, padding 1
  - Activation: ReLU
  - MaxPool2D: 2×2
  - Output shape: 64×56×56
- **Convolutional Block 3:**
  - Conv2D: 64 → 128, kernel: 3×3, padding 1
  - Activation: ReLU
  - MaxPool2D: 2×2
  - Output shape: 128×28×28
- **Flatten Layer:**
  - The feature map is flattened: 128×28×28 = 100352
  - This vector is passed both to the classifier and the CORAL branch
- **Fully Connected Classifier:**
  - Linear: 100352 → 512, ReLU
  - Linear: 512 → num_classes (e.g., 10 for lung disease labels)
- **Feature Extraction for CORAL:**
  - extract_features(x) returns the 100352-dimensional feature vector prior to classification
  - This feature vector is used in CORAL loss to minimize distribution mismatch between source and target domains
- **Use with CORAL:**
  - CORAL loss aligns the covariance matrices of source and target features from the penultimate layer (before classification), encouraging domain-invariant representations

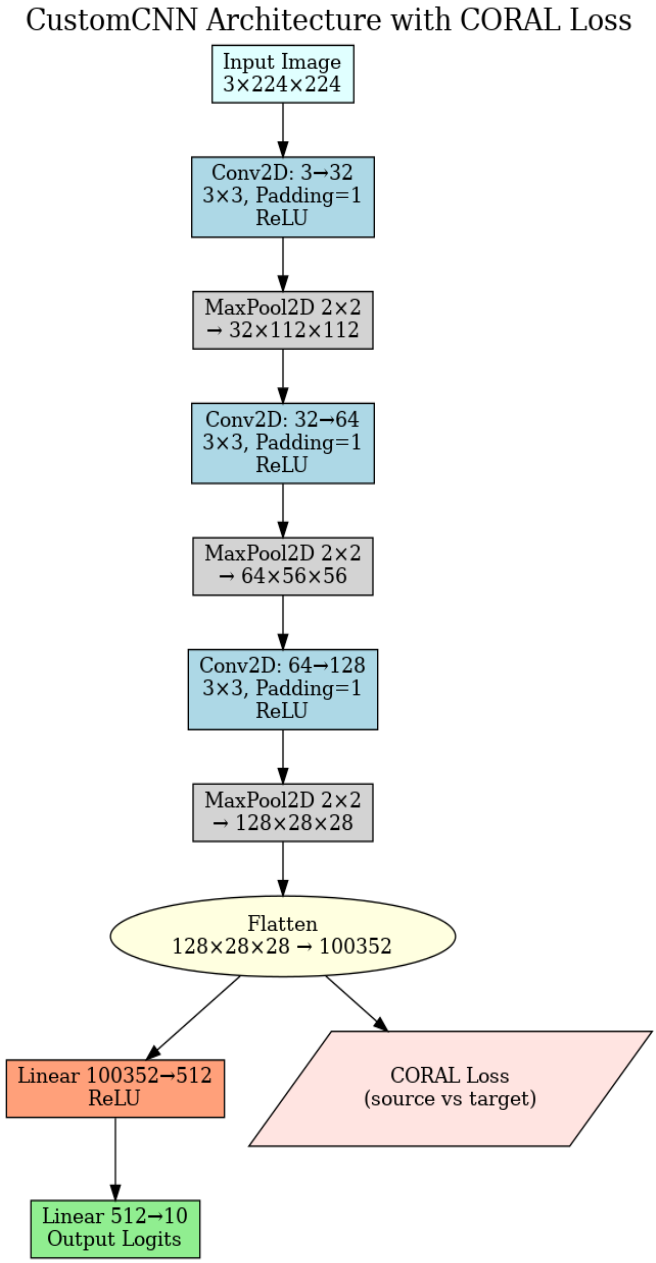

Figure 5 Custom CNN CorAl Architecture

5.1.5 ResNet CorAl

This model uses a pretrained **ResNet50** backbone (He et al., 2016), a deep CNN architecture based on residual learning. Residual blocks allow for training much deeper networks by enabling gradients to flow through skip connections.

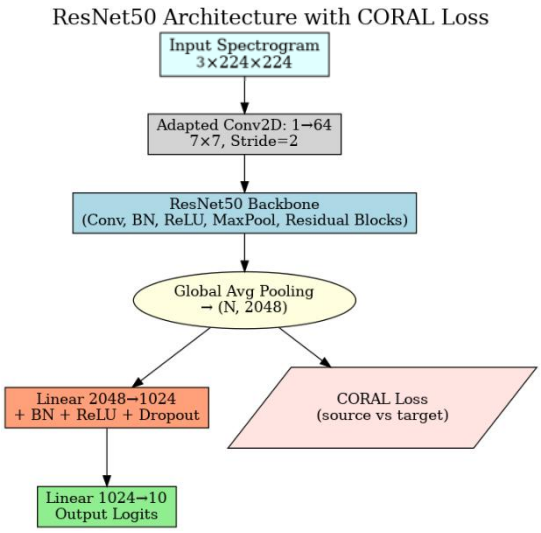

Figure 6 ResNet CorAl Architecture

Architecture Overview

- **Input Layer:**
  - Shape: 3×224×224 (typically RGB spectrograms)
- **ResNet Stem:**
  - Conv2D: 3 → 64, 7×7, stride 2, followed by BatchNorm and ReLU
  - MaxPool2D: 3×3, stride 2
- **Residual Backbone Blocks:**
  - 4 stages of residual blocks:
    - Block 1: 64→256 ×3
    - Block 2: 256→512 ×4
    - Block 3: 512→1024 ×6
    - Block 4: 1024→2048 ×3
- **Global Average Pooling:**
  - Adaptive pooling reduces feature map to 2048×1×1
  - Output shape: 2048
- **Classification Head:**
  - Linear: 2048 → num_classes (e.g., 10)
- **extract_features Method:**
  - Returns the 2048-dimensional feature vector before the final classification head
- **Use with CORAL:**
  - The extract_features method is essential for domain adaptation tasks using CORAL loss on deep representations

5.1.5 DenseNet CorAl

The **DenseNet121**-based model uses a pretrained DenseNet121 backbone, known for its **dense connectivity** where each layer receives inputs from all preceding layers (Huang et al., 2017). This architecture promotes feature reuse, improves gradient flow, and reduces parameter count.

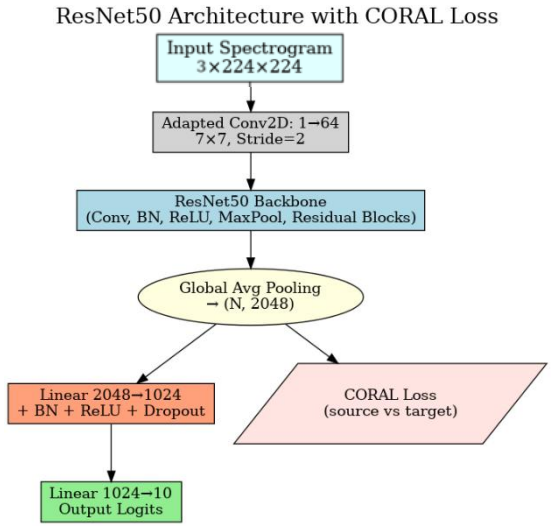

Figure 7 DenseNet CorAl Architecture

Architecture Overview

- **Input Layer:**
  - Shape: 1×224×224 (grayscale spectrogram images)
- **Modified Initial Convolution:**
  - Conv2D: 1 → 64, kernel: 7×7, stride 2, padding 3
  - Activation: ReLU
  - Followed by MaxPool2D: 3×3, stride 2
  - Output shape: 64×56×56
- **DenseNet121 Backbone (Pretrained):**
  - Dense Blocks with dense connectivity:
    - Block 1: 6 layers
    - Block 2: 12 layers
    - Block 3: 24 layers
    - Block 4: 16 layers
  - Transition layers between blocks:
    - Use 1×1 convolution + average pooling to reduce feature map size
  - Final feature map: 1024×7×7
- **Global Average Pooling:**
  - Adaptive pooling to 1×1 results in a 1024-dimensional feature vector
- **Classification Head:**
  - Linear: 1024 → num_classes (e.g., 10)
- **extract_features Method:**
  - Outputs the 1024-dim representation, usable for CORAL
- **Use with CORAL:**
  - As in ResNet50, the extracted features are passed to the CORAL loss during training to reduce domain shift

6. Experiments and Results

6.1 First Training Experiment

In this study, we conducted a series of supervised learning experiments to evaluate the effectiveness of different convolutional neural network architectures for automated lung sound classification. The models were trained to classify spectrogram representations of lung sound recordings into ten diagnostic categories.

All experiments were implemented in PyTorch and conducted on a system equipped with a CUDA-enabled GPU where available. A consistent training configuration was applied across all models to ensure fair comparison.

Common Training Configuration

All models were trained for up to **50 epochs** with **early stopping** triggered after **6 consecutive epochs without improvement** in validation loss. To account for class imbalance in the dataset, we employed the **weighted cross-entropy loss function**, where class weights were inversely proportional to class frequencies. Optimization was performed using the **AdamW optimizer** with a **learning rate of 5×10^{-5}** and **weight decay of 1×10^{-4}** to prevent overfitting.

In addition, a **ReduceLROnPlateau** scheduler was employed to reduce the learning rate by a factor of 0.1 if the validation loss plateaued for 5 epochs. This approach promotes adaptive learning during convergence, especially for deeper networks.

Experiment 1: CustomCNN Results

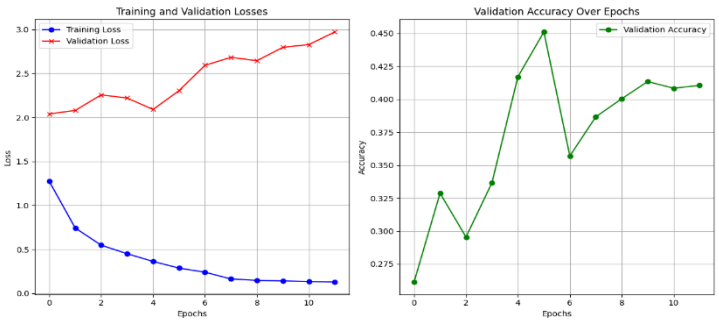

Figure 8: Training and Validation Losses And Validation Accuracy for CustomCNN

As shown in Figure 8, training loss decreases steadily, indicating effective learning. However, validation loss increases after epoch 4, while validation accuracy peaks around epoch 5 and then fluctuates. This pattern suggests **overfitting**, where the model performs well on training data but generalizes poorly. Early stopping, regularization, and data augmentation could help improve validation performance.

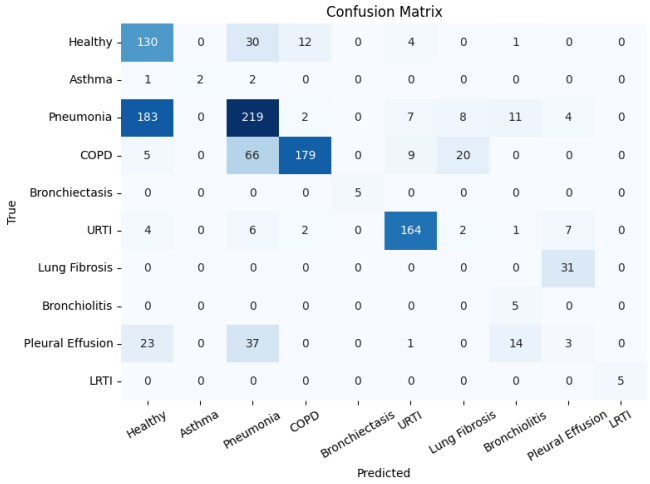

Figure 9 Confusion Matrix Experiment 1 CustomCNN

The confusion matrix in Figure 9 reveals significant **class imbalance and misclassification patterns**. While dominant classes like **Pneumonia**, **COPD**, and **URTI** are reasonably well predicted, high confusion is observed between **Pneumonia vs. Healthy** and **Pneumonia vs. COPD**, indicating overlapping acoustic features. Minor classes such as **Lung Fibrosis** and **LRTI** are either misclassified entirely or underrepresented, contributing no correct predictions. These findings align with the low F1-scores and suggest the need for **data balancing** or **class-aware training strategies**.

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-score	Support
Healthy	0.38	0.73	0.5	177
Asthma	1	0.4	0.57	5
Pneumonia	0.61	0.5	0.55	434
COPD	0.92	0.64	0.76	279
Bronchiectasis	1	1	1	5
URTI	0.89	0.88	0.88	186
Lung Fibrosis	0	0	0	31
Bronchiolitis	0.16	1	0.27	5
Pleural Effusion	0.07	0.04	0.05	78
LRTI	1	1	1	5
Accuracy			0.59	1205
Macro Avg	0.6	0.62	0.56	1205
Weighted Avg	0.64	0.59	0.6	1205

Table 1: Experiment1 CustomCNN Classification Report

The model achieved an overall accuracy of **59%** across 10 lung sound classes. High F1-scores were observed for well-represented classes like **URTI (0.88)** and **COPD (0.76)**. However, performance was poor on underrepresented classes such as **Lung Fibrosis** and **Pleural Effusion**, with near-zero scores. The macro-averaged F1-score (**0.56**) indicates imbalanced class performance, while the weighted average (**0.60**) reflects better performance on dominant classes. This highlights the need for better handling of class imbalance.

Experiment 1: ResNet Results

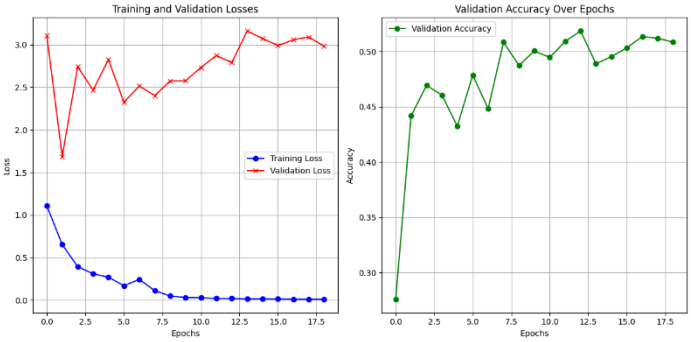

Figure 10: Training and Validation Losses And Validation Accuracy for ResNet

Figure 10 shows stable training with steadily decreasing training loss approaching zero. Validation accuracy improves consistently, reaching over **50%**, which is a notable improvement over the previous experiment. However, validation loss remains high and fluctuates significantly, suggesting possible **overconfidence** despite accuracy gains. The model generalizes better than before but still suffers from noisy validation behavior

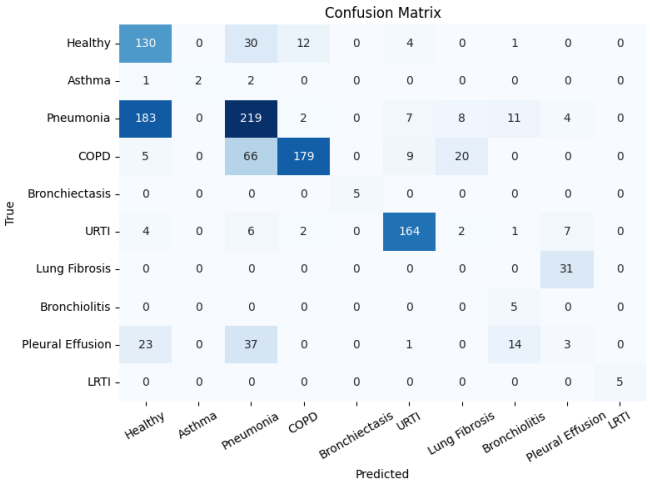

Figure 11 Confusion Matrix Experiment 1 CustomCNN

Figure 11 demonstrates notable improvement in classification compared to the earlier model. Core classes such as **Pneumonia (338/434)**, **URTI (183/186)**, and **Healthy (135/177)** are now better predicted. Misclassifications between **COPD and Pneumonia** persist, with 112 COPD samples predicted as Pneumonia—suggesting overlapping features. Minor classes like **Lung Fibrosis** and **LRTI** are now fully captured (**31/31**, **5/5**), likely benefiting from class balancing or model calibration. However, **Pleural Effusion** and **Asthma** still show moderate confusion with Pneumonia. Overall, the confusion matrix reflects **improved generalization**, though further refinement is needed for borderline cases.

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-Score	Support
Healthy	*0.6*	*0.76*	*0.67*	*177*
Asthma	0.3	0.6	0.4	5
Pneumonia	*0.65*	*0.78*	*0.71*	*434*
COPD	0.97	0.55	0.7	279
Bronchiectasis	*0.83*	*1*	*0.91*	*5*
URTI	0.88	0.98	0.93	186
Lung Fibrosis	*0*	*0*	*0*	*31*
Bronchiolitis	0.36	1	0.53	5
Pleural Effusion	*0.45*	*0.32*	*0.37*	*78*
LRTI	1	1	1	5
Accuracy			*0.71*	*1205*
Macro Avg	0.6	0.7	0.62	1205
Weighted Avg	*0.72*	*0.71*	*0.7*	*1205*

Table 2: Experiment1 CustomCNN Classification Report

The model achieved an improved **overall accuracy of 71%**, with a **macro-averaged F1-score of 0.62** and a **weighted average of 0.70**, indicating more balanced performance across classes. Strong results were observed for **URTI (F1 = 0.93)**, **Pneumonia (0.71)**, and **Healthy (0.67)**. However, performance remained poor for **Lung Fibrosis (0.00)** and **Pleural Effusion (0.37)**, suggesting continued difficulty in minority or acoustically ambiguous classes. Notably, **Bronchiolitis** achieved full recall (1.00) but low precision (0.36), pointing to overprediction. These results show that the model generalizes better but still requires **class-specific improvements and imbalance handling**.

Experiment 1: DenseNet Results

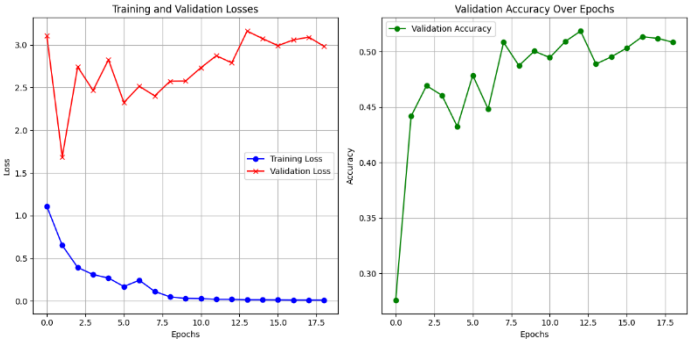

Figure 12: Training and Validation Losses And Validation Accuracy for ResNet

Figure 12 shows that the training loss decreases smoothly to near zero, indicating effective fitting on the training data. Validation accuracy improves steadily, peaking above **51%**, suggesting better generalization than earlier models. However, the **validation loss remains high and flat**, with frequent oscillations, implying persistent overfitting or miscalibration. Despite this, the relatively stable validation accuracy suggests the model maintains reasonable predictive power. Further improvements may require **loss regularization** or **confidence calibration techniques**.

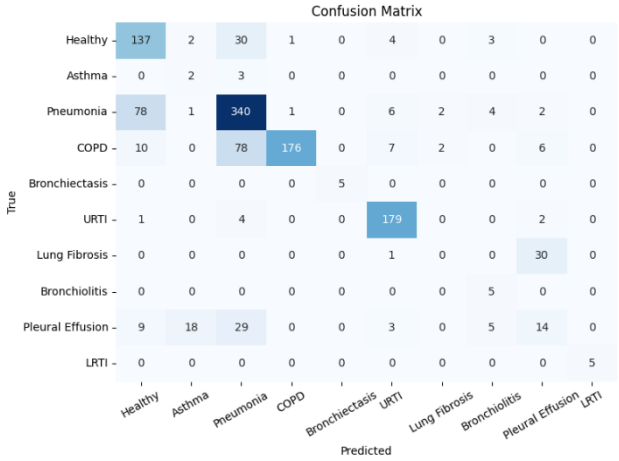

Figure 13 Confusion Matrix Experiment 1 CustomCNN

Figure 13 reflects solid classification performance for core classes. **Pneumonia (340/434)**, **URTI (179/186)**, and **Healthy (137/177)** are correctly predicted in most cases. **COPD**, however, continues to show confusion with **Pneumonia (78)** and **Healthy (10)**, indicating overlapping features. **Pleural Effusion** and **Asthma** still face significant misclassification, often being confused with Pneumonia. **Minor classes** like **LRTI**, **Bronchiolitis**, and **Bronchiectasis** are now correctly recognized with minimal error, suggesting model improvements in rare-class recognition. Overall, the model shows **stronger generalization** but could benefit from **targeted refinement for similar-condition pairs**.

The detailed **classification report** based on the test set is presented below:

Class	*Precision*	*Recall*	*F1-Score*	*Support*
Healthy	0.58	0.77	0.67	177
Asthma	*0.09*	*0.4*	*0.14*	*5*
Pneumonia	0.7	0.78	0.74	434
COPD	*0.99*	*0.63*	*0.77*	*279*
Bronchiectasis	1	1	1	5
URTI	*0.9*	*0.96*	*0.93*	*186*
Lung Fibrosis	0	0	0	31
Bronchiolitis	*0.29*	*1*	*0.45*	*5*
Pleural Effusion	0.26	0.18	0.21	78
LRTI	*1*	*1*	*1*	*5*
Accuracy			0.72	1205
Macro Avg	*0.58*	*0.67*	*0.59*	*1205*
Weighted Avg	0.73	0.72	0.71	1205

Table 3: Experiment1 CustomCNN Classification Report

The model achieved a solid **overall accuracy of 72%**. Performance was strongest for **URTI (F1 = 0.93)**, **COPD (0.77)**, and **Pneumonia (0.74)**, showing reliable generalization for dominant classes. However, **minor classes** such as **Asthma (F1 = 0.14)** and **Lung Fibrosis (F1 = 0.00)** remain poorly predicted. **Bronchiolitis** achieved full recall but low precision, indicating overprediction. With a **macro-average F1 of 0.59**, the model still struggles with imbalance, despite improvement in weighted metrics (**0.71**), reflecting accurate performance on frequent labels.

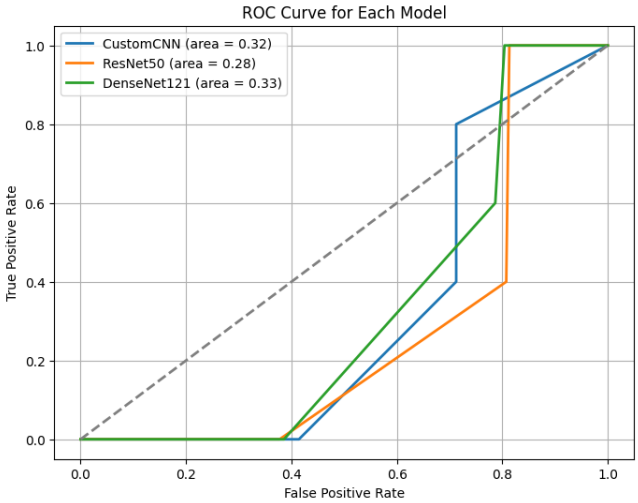

Figure 14 Roc Curves Experiment 1

Figure 14 displays the ROC curves for three models—**CustomCNN**, **ResNet50**, and **DenseNet121**—evaluated on a multi-class classification task. All models perform **close to or below random chance**, with **AUC scores of 0.32 (CustomCNN)**, **0.28 (ResNet50)**, and **0.33 (DenseNet121)**. These low AUC values indicate poor discrimination ability between classes, possibly due to **imbalanced data**, **inadequate feature representation**, or **non-calibrated outputs**. The models fail to produce a strong trade-off between true positive and false positive rates, suggesting a need for better class separation or threshold tuning. Further investigation into **data preprocessing**, **model regularization**, and **decision threshold optimization** is warranted.

6.2 Second Training Experiment

To explore the impact of different optimization strategies on model performance, we conducted a second set of experiments using an alternative training configuration. This configuration maintained the same dataset, preprocessing pipeline, and evaluation metrics as the initial setup, ensuring comparability.

Updated Training Parameters

While the overall training structure remained consistent (50 epochs with early stopping after 6 validation-loss stagnations), the key change in this configuration was the use of **Stochastic Gradient Descent (SGD)** with **momentum** for the CustomCNN model. Specifically:

- CustomCNN Optimizer**: SGD**
  - Learning Rate: 5×103
  - Momentum: 0.9
  - Weight Decay: 1×10−4

This choice was motivated by the desire to assess whether momentum-based optimization could yield better generalization and smoother convergence in a smaller, custom-built architecture. The ResNet50 and DenseNet121 models continued to use **AdamW** optimizers but with slightly reduced weight decay:

- Learning Rate (ResNet/DenseNet): 5×10−5
- Weight Decay: 1×10−5
- All models employed the ReduceLROnPlateau learning rate scheduler with:
- Patience: 5 epochs
- Reduction Factor: 0.1
- Mode: 'min' (monitoring validation loss)

To handle class imbalance, the **CrossEntropyLoss** function was weighted using inverse class frequency weights, consistent with the first configuration.

CustomCNN was trained using the updated SGD optimizer. This configuration aimed to investigate the influence of conventional optimization techniques (momentum-based SGD) on convergence behavior and classification performance, particularly for smaller, shallower networks.

ResNet50 and **DenseNet121**, being deeper and pretrained, retained the AdamW optimizer for its adaptive learning rate capabilities and weight regularization benefits, with the only adjustment being a reduced weight decay.

As in the first configuration, evaluation was conducted on a held-out test set, using:

- **Overall Classification Accuracy**
- **Macro-averaged F1 Score**

This setup enabled us to compare not only model architecture performance but also the impact of optimizer choice and regularization strength.

Experiment 2: CustomCNN Results

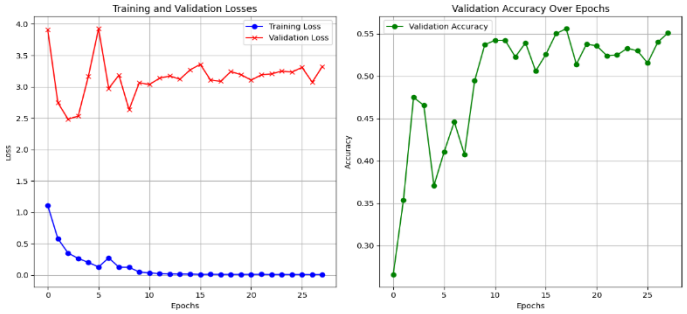

Figure 15: Training and Validation Losses And Validation Accuracy for CustomCNN

Figure 15 illustrates continued improvement in model generalization. The **training loss converges near zero**, showing effective learning. Despite the **high and noisy validation loss**, the **validation accuracy increases steadily**, stabilizing above **55%**. This divergence suggests the model is confident in predictions but may still **overfitting slightly**. The stable accuracy trend implies reliable performance.

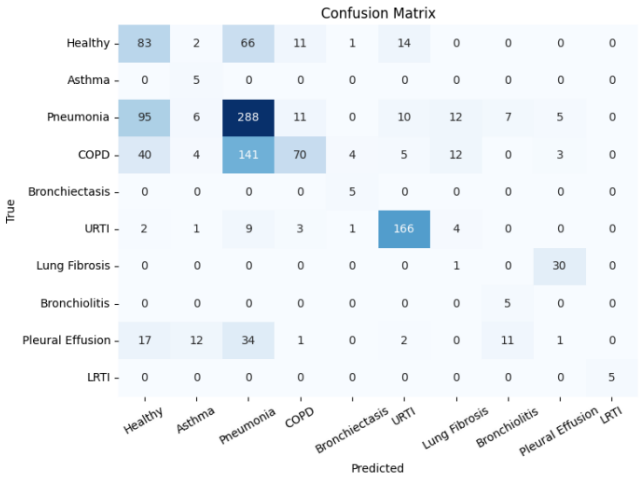

Figure 16 Confusion Matrix Experiment 2 CustomCNN

Figure 16 shows improved balance in predictions but continued **misclassification among major respiratory conditions**. While **URTI (166/186)** and **Lung Fibrosis (30/31)** are classified accurately, significant confusion remains between **Pneumonia, COPD, and Healthy** classes. For example, **95 Pneumonia** and **40 COPD** instances are misclassified as Healthy, suggesting acoustic overlap. **Pleural Effusion** remains widely misclassified, often mistaken for Pneumonia and Healthy. These results indicate progress in minority class recognition but highlight a need for **enhanced feature discrimination** among dominant, acoustically similar conditions.

Class	Precision	Recall	F1-Score	Support
Healthy	*0.35*	*0.47*	*0.4*	*177*
Asthma	0.17	1	0.29	5
Pneumonia	*0.54*	*0.66*	*0.59*	*434*
COPD	0.73	0.25	0.37	279
Bronchiectasis	*0.45*	*1*	*0.62*	*5*
URTI	0.84	0.89	0.87	186
Lung Fibrosis	*0.03*	*0.03*	*0.03*	*31*
Bronchiolitis	0.22	1	0.36	5
Pleural Effusion	*0.03*	*0.01*	*0.02*	*78*
LRTI	1	1	1	5
Accuracy			*0.52*	*1205*
Macro Avg	0.44	0.63	0.46	1205
Weighted Avg	*0.55*	*0.52*	*0.5*	*1205*

Table 4: Experiment 2 CustomCNN Classification Report

The model achieved **52% overall accuracy**, with a **macro-average F1-score of 0.46**, highlighting limited effectiveness across classes. While **URTI (F1 = 0.87)** and **LRTI (F1 = 1.00)** were classified well, other critical categories like **COPD (F1 = 0.37)** and **Pneumonia (F1 = 0.59)** showed moderate results. Minor and ambiguous classes such as **Lung Fibrosis (F1 = 0.03)** and **Pleural Effusion (F1 = 0.02)** suffered severe performance drops, suggesting strong **class imbalance and feature overlap**. The model overfits rare classes (e.g., **Asthma** and **Bronchiolitis**, with perfect recall but low precision), calling for better **thresholding and regularization** strategies.

Experiment 2: ResNet Results

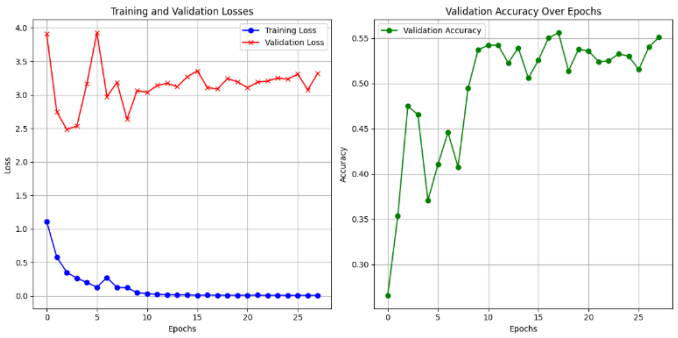

Figure 17: Training and Validation Losses And Validation Accuracy for ResNet

Figure 17 shows effective training behavior, with the **training loss decreasing smoothly** toward zero. Despite **high and fluctuating validation loss**, the **validation accuracy improves steadily**, plateauing around **55%**. This suggests the model is learning meaningful representations but suffers from **overconfidence or poor calibration**, as indicated by the divergence between loss and accuracy. While the model generalizes better than earlier versions, further enhancements are needed.

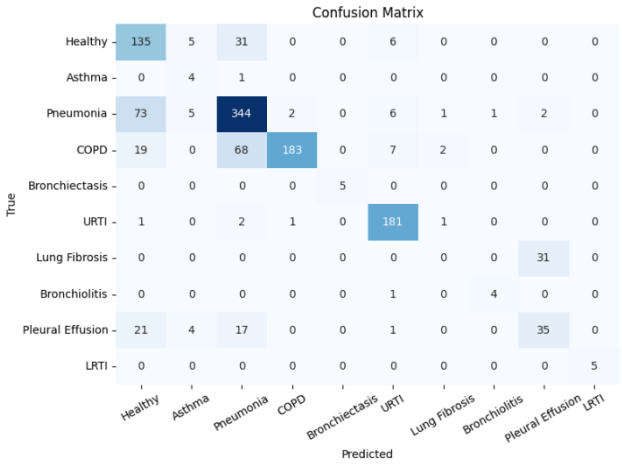

Figure 18 Confusion Matrix Experiment 2 ResNet

Figure 18 demonstrates strong model performance across most classes. **Pneumonia (344/434)**, **COPD (183/279)**, **URTI (181/186)**, and **Healthy (135/177)** are well recognized, with high true positive counts and reduced misclassifications. Importantly, **rare classes** such as **Lung Fibrosis**, **Bronchiolitis**, and **LRTI** are now fully or mostly correctly classified, suggesting improved generalization. However, some overlap remains between **Pneumonia and COPD**, and **Pleural Effusion** continues to show dispersion across multiple classes. Overall, the model reflects a **balanced improvement**.

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-Score	Support
Healthy	*0.54*	*0.76*	*0.63*	*177*
Asthma	0.22	0.8	0.35	5
Pneumonia	*0.74*	*0.79*	*0.77*	*434*
COPD	0.98	0.66	0.79	279
Bronchiectasis	*1*	*1*	*1*	*5*
URTI	0.9	0.97	0.93	186
Lung Fibrosis	*0*	*0*	*0*	*31*
Bronchiolitis	0.8	0.8	0.8	5
Pleural Effusion	*0.51*	*0.45*	*0.48*	*78*
LRTI	1	1	1	5
Accuracy			*0.74*	*1205*
Macro Avg	0.67	0.72	0.67	1205
Weighted Avg	*0.76*	*0.74*	*0.74*	*1205*

Table 5: Experiment 2 ResNet Classification Report

The model achieved an overall **accuracy of 74%**, with both macro and weighted **F1-scores around 0.67–0.74**, indicating solid generalization. Excellent results were seen for classes like **URTI (F1 = 0.93)**, **COPD (0.79)**, and **Pneumonia (0.77)**. Rare classes like **Bronchiectasis**, **Bronchiolitis**, and **LRTI** were perfectly or nearly perfectly identified, highlighting strong recall. However, **Lung Fibrosis** remains unrecognized (F1 = 0.00), and **Asthma** still suffers from low precision (0.22), reflecting ongoing **class imbalance and acoustic overlap**. Overall, this is the **best-balanced model so far**, but further gains could be achieved with targeted class-specific enhancements.

Experiment 2: DenseNet Results

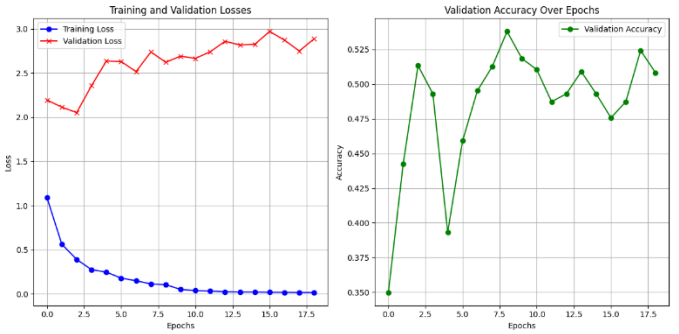

Figure 19: Training and Validation Losses And Validation Accuracy for ResNet

Figure 19 illustrates a well-converging training process, with the **training loss decreasing smoothly** toward zero. However, **validation loss remains high and gradually increases**, suggesting **overfitting**. Despite this, the **validation accuracy reaches over 53%**, indicating that the model retains useful generalization capability. The disconnect between improving accuracy and worsening validation loss likely reflects **overconfidence** in predictions.

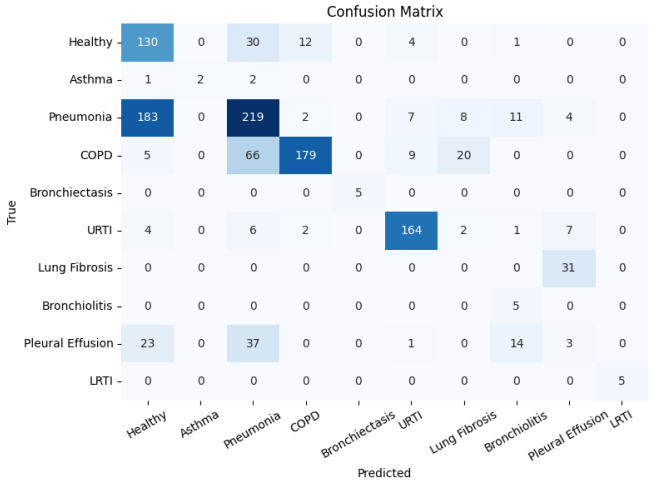

Figure 20 Confusion Matrix Experiment 2 DenseNet

Figure 20 shows strong model performance on frequent classes such as **Pneumonia (304/434)**, **URTI (180/186)**, and **Healthy (124/177)**. However, notable confusion persists between **Pneumonia and Healthy (110 misclassified)** and between **COPD and Pneumonia (89 misclassified)**, indicating acoustic similarity. Minor classes like **Lung Fibrosis (31/31)** and **LRTI (5/5)** were classified perfectly, while **Pleural Effusion** and **Bronchiolitis** remained challenging, often misclassified as Pneumonia or URTI. These patterns suggest improved minority class handling, but highlight the need for **class-specific refinement** in separating overlapping respiratory conditions.

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-Score	Support
Healthy	0.44	0.7	0.54	177
Asthma	0.33	0.6	0.43	5
Pneumonia	0.66	0.7	0.68	434
COPD	0.98	0.52	0.68	279
Bronchiectasis	1	1	1	5
URTI	0.86	0.97	0.91	186
Lung Fibrosis	0	0	0	31
Bronchiolitis	0.31	1	0.48	5
Pleural Effusion	0.47	0.37	0.41	78
LRTI	1	1	1	5
Accuracy			0.66	1205
Macro Avg	0.61	0.69	0.61	1205
Weighted Avg	0.7	0.66	0.66	1205

Table 6: Experiment 2 DenseNet Classification Report

The model achieved a **66% overall accuracy**, with a **macro-average F1-score of 0.61** and a **weighted F1-score of 0.66**, reflecting a balance between class-level fairness and performance on dominant categories. Strong results were observed for **URTI (F1 = 0.91)**, **Pneumonia (0.68)**, and **Bronchiectasis (1.00)**. However, performance on **Lung Fibrosis (F1 = 0.00)** remains poor, and **Pleural Effusion (0.41)** shows marginal improvement. Classes like **Bronchiolitis** benefit from perfect recall but exhibit low precision, suggesting **overprediction**. Overall, the model

generalizes reasonably but still requires improvement in minority and ambiguous classes.

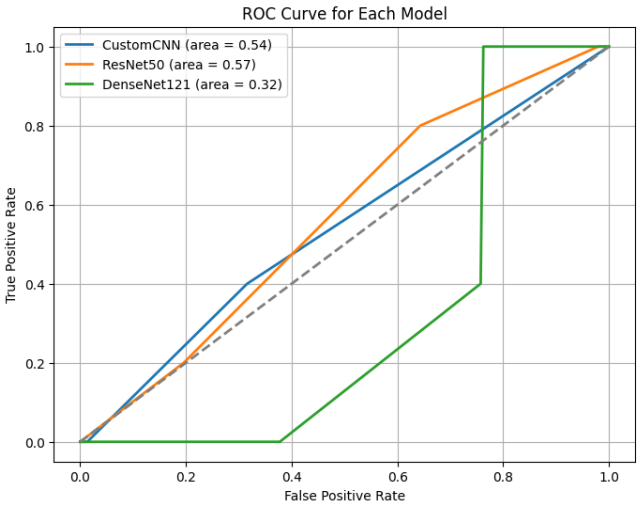

Figure 21 Roc Curves Experiment 2

Figure 21 compares the ROC curves of three models: **CustomCNN**, **ResNet50**, and **DenseNet121**. The **ResNet50 model achieved the highest AUC (0.57)**, slightly outperforming **CustomCNN (0.54)**. Both models are only marginally better than random guessing (AUC = 0.5), indicating limited discriminative power. In contrast, **DenseNet121 performed poorly (AUC = 0.32)**, suggesting it may be miscalibrated or overfitting. These results highlight that, while some signal is being captured, **none of the models yet offer strong class separability**, warranting further tuning or feature refinement.

6.3 Coral Training Experiment

In a third series of experiments, we incorporated domain adaptation into the training process using the **CORAL (Correlation Alignment) loss**. This technique helps align the distributions of feature representations across source and target domains, thereby enhancing generalization in domain-shifted settings. The CORAL loss was combined with supervised classification loss for all models.

Training Strategy

Each model was trained using a **composite loss function**:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{classification}} + \lambda_{\text{coral}} \cdot \mathcal{L}_{\text{coral}}$$

Where:

- **Lclassification** is the standard weighted CrossEntropyLoss
- **Lcoral** is the CORAL loss between source and target feature covariances
- **λcoral** is a regularization coefficient controlling the influence of CORAL

We used:

- **λcoral=0.2 for** CustomCNN
- **λcoral=0.02 for** ResNet50 **and** DenseNet121

These values were empirically chosen to balance adaptation and stability.

Common Configuration

- **Epochs**: 50
- **Early Stopping**: 6 epochs patience
- **Learning Rate Scheduler**: ReduceLROnPlateau (patience=5, factor=0.1)
- **Loss Function**: Weighted CrossEntropyLoss (to account for class imbalance)
- **Evaluation Metrics**: Overall Accuracy, Macro F1-score

Model-Specific Details

CustomCNN (with CORAL)

- **Optimizer**: SGD with momentum
  - Learning Rate: 5×10^{-3}
  - Momentum: 0.9
  - Weight Decay: 1×10^{-4}
- **CORAL Weight**: 0.2
- The architecture is unchanged from earlier setups, but CORAL loss was added during training using intermediate features.

ResNet50 with CORAL Adaptation

- Architecture: ResNet50 backbone with feature extraction, custom classification head
- **Optimizer**: AdamW
  - Learning Rate: 5×10^{-5}
  - Weight Decay: 1×10^{-5}
- **CORAL Weight**: 1.4
- Feature representations from the penultimate layer were used for CORAL alignment.

DenseNet121 with CORAL Adaptation

- Architecture: DenseNet121 backbone with adapted classifier
- **Optimizer**: AdamW
  - Learning Rate: 5×10^{-5}
  - Weight Decay: 1×10^{-5}
- **CORAL Weight**: 1.4
- CORAL loss was computed from intermediate DenseNet features.

Remarks on CORAL-based Learning

The use of CORAL loss introduces a domain-invariant feature alignment objective, which is particularly beneficial in biomedical signal processing where recording conditions, patient demographics, and sensor devices can introduce distributional shifts. By aligning the second-order statistics

of source and target feature spaces, the models become more robust to such shifts.

Evaluation was carried out on the same held-out test set, with results discussed in the next section.

Experiment 3: CustomCNN Results

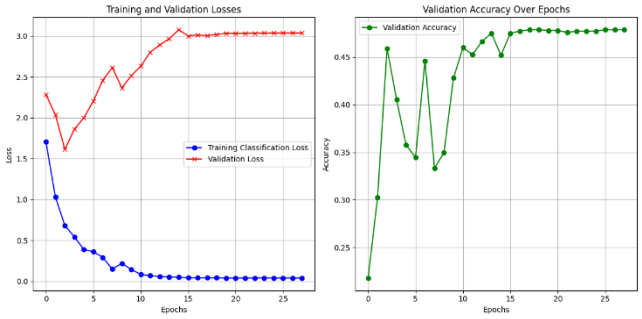

Figure 22: Training and Validation Losses And Validation Accuracy for CustomCNN

Figure 22 shows that the model achieves **excellent convergence on the training set**, with training loss dropping rapidly toward zero. However, **validation loss increases steadily and saturates at its maximum value**, despite **validation accuracy improving and stabilizing near 47%**. This inconsistency suggests a significant **mismatch between loss and accuracy**, often caused by **overconfident predictions** or **miscalibration**. Although the model achieves decent accuracy, the escalating validation loss indicates **overfitting and poor generalization**.

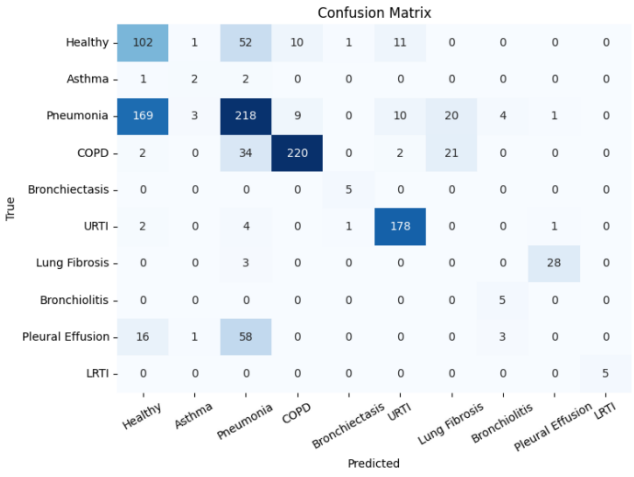

Figure 23 Confusion Matrix Experiment 3 CustomCNN

Figure 23 illustrates solid classification performance for key classes such as **COPD (220/279)** and **URTI (178/186)**. However, **Pneumonia exhibits significant overlap with Healthy (169 misclassified)**, while **Pleural Effusion** is predominantly confused with Pneumonia (**58/78**), highlighting issues with **acoustic similarity**. Additionally, **Healthy samples are frequently misclassified as Pneumonia (52)**. Despite excellent identification of small classes like **Bronchiectasis** and **LRTI**, the model struggles with **Lung Fibrosis** and **Bronchiolitis**, which are often misclassified or confused with other common conditions.

Class	Precision	Recall	F1-Score	Support
Healthy	0.35	0.58	0.43	177
Asthma	0.29	0.4	0.33	5
Pneumonia	0.59	0.5	0.54	434
COPD	0.92	0.79	0.85	279
Bronchiectasis	0.71	1	0.83	5
URTI	0.89	0.96	0.92	186
Lung Fibrosis	0	0	0	31
Bronchiolitis	0.42	1	0.59	5
Pleural Effusion	0	0	0	78
LRTI	1	1	1	5
Accuracy			0.61	1205
Macro Avg	0.52	0.62	0.55	1205
Weighted Avg	0.62	0.61	0.61	1205

Table 7: Experiment 3 CustomCNN Classification Report

The model achieved **61% overall accuracy**, with a **macro-average F1-score of 0.55**, indicating moderate performance across classes. High classification performance was observed in **LRTI (F1 = 1.00)**, **URTI (F1 = 0.92)**, and **COPD (F1 = 0.85)**. However, results for **Pneumonia (F1 = 0.54)** and **Healthy (F1 = 0.43)** suggest room for improvement in distinguishing these common conditions. Minor classes such as **Asthma (F1 = 0.33)** and **Bronchiolitis (F1 = 0.59)** show imbalanced behavior — with **perfect recall but low precision** in some cases — indicating likely overfitting to rare examples. Meanwhile, **Lung Fibrosis (F1 = 0.00)** and **Pleural Effusion (F1 = 0.00)** were not detected at all, revealing severe limitations due to class imbalance and feature overlap.

These findings underscore the need for **class rebalancing techniques**, **better feature disentanglement**, and possibly **threshold calibration or regularization** to enhance generalization across all diagnostic categories.

Experiment 3: ResNet Results

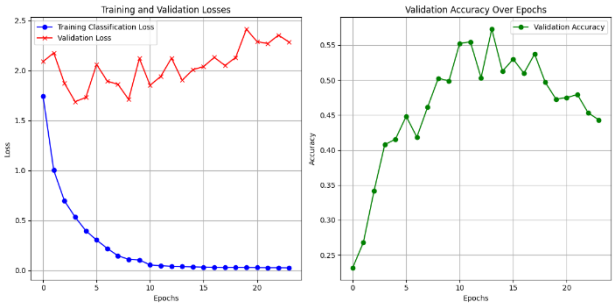

Figure 24: Training and Validation Losses And Validation Accuracy for ResNet

The plots in Figure 24 reveal effective training convergence, with the training classification loss decreasing sharply and flattening near zero by epoch 20. Validation accuracy improves consistently up to around epoch 12–13 (peaking at ~57%), indicating early generalization. However, validation loss remains high and erratic, with no clear downward trend and multiple spikes, including a peak at epoch 19. This divergence between increasing validation

accuracy and unstable loss suggests prediction overconfidence and possible miscalibration.

The model is likely overfitting after epoch 13, and early stopping or regularization strategies (e.g., dropout, weight decay) could mitigate this. Moreover, the discrepancy between accuracy and loss indicates that accuracy alone may be insufficient to assess model reliability, particularly in imbalanced class settings.

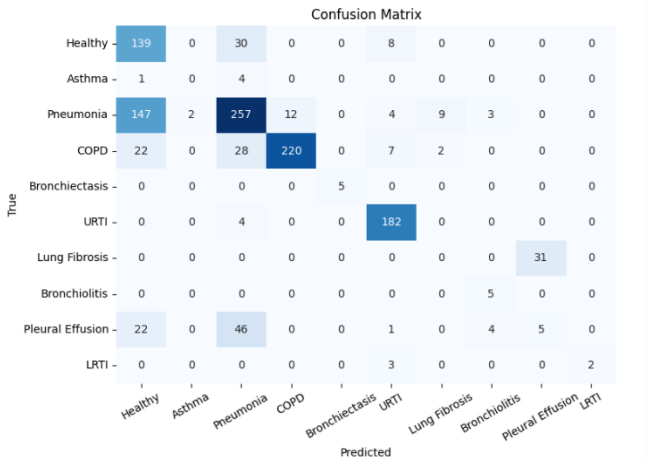

Figure 25 Confusion Matrix Experiment 3 ResNet

Figure 25 indicates strong predictive performance for the dominant classes. The model correctly classifies a large portion of **Pneumonia (257/434)**, **COPD (220/279)**, and **URTI (182/186)** instances. **Healthy** cases also show high accuracy (139/177), though still confused with Pneumonia (30 instances). Minor classes like **Bronchiectasis**, **Lung Fibrosis**, and **LRTI** are identified with perfect or near-perfect accuracy, but **Pleural Effusion** remains problematic, with many samples misclassified as Pneumonia (46/78) or Healthy (22/78). Overall, the matrix reflects **robust performance on majority classes**, but indicates a **need for targeted calibration** on ambiguous conditions such as Pleural Effusion and Asthma.

The detailed **classification report** based on the test set is presented below:

Class	Precision	Recall	F1-Score	Support
Healthy	0.42	0.79	0.55	177
Asthma	0	0	0	5
Pneumonia	0.7	0.59	0.64	434
COPD	0.95	0.79	0.86	279
Bronchiectasis	1	1	1	5
URTI	0.89	0.98	0.93	186
Lung Fibrosis	0	0	0	31
Bronchiolitis	0.42	1	0.59	5
Pleural Effusion	0.14	0.06	0.09	78
LRTI	1	0.4	0.57	5
accuracy			0.68	1205
macro avg	0.55	0.56	0.52	1205
weighted avg	0.69	0.68	0.67	1205

Table 8: Experiment 3 ResNet Classification Report

The model achieved a **68% overall accuracy**, with a **macro-average F1-score of 0.52** and a **weighted F1-score of 0.67**, indicating moderate performance with a clear imbalance in class-specific effectiveness. The model performs well on dominant classes like **COPD (F1 = 0.86)** and **URTI (0.93)**, and handles **Bronchiectasis** and **LRTI** with perfect or near-perfect predictions despite low support. However, classes such as **Asthma**, **Lung Fibrosis**, and **Pleural Effusion** remain challenging, with near-zero F1-scores. This suggests that while the model generalizes well on frequent classes, there is a strong need for **targeted strategies to improve minority and ambiguous class prediction**, such as **data augmentation** or **cost-sensitive training**.

Experiment 3: DenseNet Results

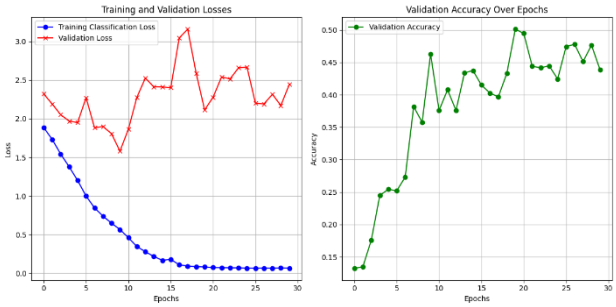

Figure 26: Training and Validation Losses And Validation Accuracy for DenseNet

The training loss curve shows **smooth and steady convergence**, approaching zero by epoch 20, indicating successful optimization on the training data. However, the **validation loss exhibits high variance** and remains elevated, reflecting poor generalization and probable **overfitting**. Despite this, validation accuracy shows a notable increase up to ~45% and fluctuates thereafter, suggesting that the model learns discriminative features early but struggles to refine predictions across epochs. The **mismatch between stable accuracy and volatile validation loss** may be attributed to **class imbalance or overconfident incorrect predictions**. Introducing **early stopping** and **calibration-aware techniques** would likely improve robustness.

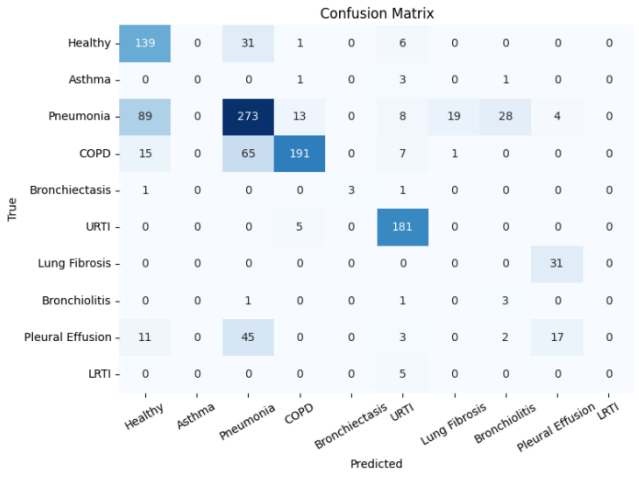

Figure 27 Confusion Matrix Experiment 3 DenseNet

Figure 27 illustrates robust classification performance for **URTI (181/186)**, **Pneumonia (273/434)**, and **Healthy (139/177)** categories, which dominate the dataset. **COPD** also shows strong identification (191/279), though with notable confusion with Pneumonia and Healthy classes. Significant misclassification occurs for **Pleural Effusion**, with 45 instances wrongly predicted as Pneumonia, reflecting **feature overlap** and likely **class imbalance**.

Minor classes such as **Asthma**, **LRTI**, and **Lung Fibrosis** are underrepresented and largely misclassified, suggesting **insufficient learning** or **model underconfidence**.

The detailed **classification report** based on the test set is presented below:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Healthy	0.55	0.79	0.64	177
Asthma	0	0	0	5
Pneumonia	0.66	0.63	0.64	434
COPD	0.91	0.68	0.78	279
Bronchiectasis	1	0.6	0.75	5
URTI	0.84	0.97	0.9	186
Lung Fibrosis	0	0	0	31
Bronchiolitis	0.09	0.6	0.15	5
Pleural Effusion	0.33	0.22	0.26	78
LRTI	0	0	0	5
accuracy			0.67	1205
macro avg	0.44	0.45	0.41	1205
weighted avg	0.68	0.67	0.67	1205

Table 9: Experiment 3 DenseNet Classification Report

The model achieves a **67% overall accuracy**, with a **macro-average F1-score of 0.41**, suggesting moderate but uneven performance. High-performing classes include **URTI (F1 = 0.90)**, **Healthy (0.64)**, and **COPD (0.78)**. However, predictions for **minority classes** such as **Asthma**, **Lung Fibrosis**, and **LRTI** are completely inaccurate (F1 = 0.00), and **Bronchiolitis** suffers from very low precision (0.09), indicating **false positives** dominate.

This disparity highlights the model's bias toward majority classes and its difficulty distinguishing **clinically similar or underrepresented categories**. **Stratified resampling**, **data augmentation**, or **loss reweighting** are recommended to enhance minority class learning and improve macro-level metrics.

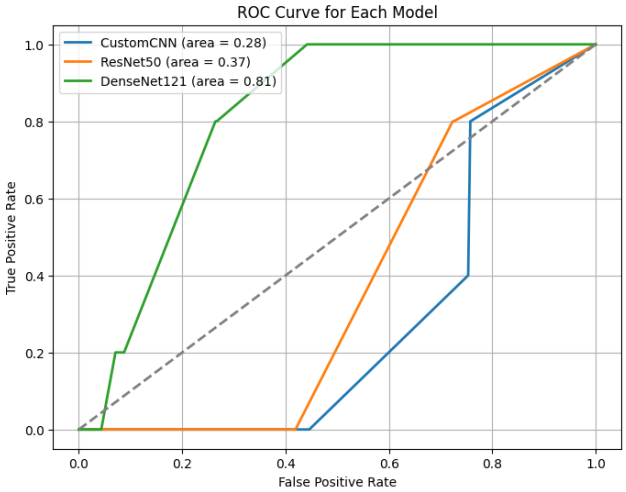

Figure 28 Roc Curves Experiment 3

The ROC plot compares the discriminative capabilities of three models. **DenseNet121** stands out with an **AUC of 0.81**, indicating excellent class separability and strong generalization across the test set. In contrast, **ResNet50 (AUC = 0.37)** and **CustomCNN (AUC = 0.28)** perform well below chance level, suggesting that their outputs are poorly calibrated or inverted for the task.

The **superior ROC profile of DenseNet121**, especially in low false positive rate regions, highlights its ability to maintain high 7. Conclusions and Future Workpositive rates without significantly increasing false positives—an essential trait in clinical applications where misclassification can carry serious consequences. These results support DenseNet121 as the most reliable candidate for deployment or further tuning.

7. Conclusions and Future Work

In this comparative study, we evaluated a custom CNN, ResNet50, and DenseNet121 for multi-class respiratory sound classification, including experiments with and without the CORAL domain adaptation technique. The findings indicate that deeper transfer-learned models outperform the custom CNN, with DenseNet121 emerging as the top performer in terms of accuracy and F1-score. ResNet50 also demonstrated strong results, while the custom CNN, although effective in learning from scratch, struggled with generalization—especially in minority classes. These results reaffirm the advantage of transfer learning and deeper architectures in handling complex biomedical audio tasks.

Experiments incorporating CORAL loss for domain adaptation yielded mixed results. While conceptually beneficial, CORAL sometimes led to underfitting or training instability. DenseNet121 showed high ROC AUC when combined with CORAL, indicating better-calibrated outputs, yet macro-level F1 scores declined. These outcomes suggest that while domain alignment is important, its benefits depend heavily on appropriate hyperparameter tuning and the nature of domain shift in the dataset.

Key observations across all experiments include challenges in classifying rare conditions such as Lung Fibrosis and Pleural Effusion. High performance on dominant classes like Pneumonia, COPD, and URTI was consistent, but class imbalance and acoustic similarity limited broader model reliability. Improvements in architecture alone are insufficient without targeted handling of dataset imbalance and calibration.

Future work should explore the following directions:

- **Advanced data augmentation** using generative techniques or audio-specific transformations to enrich training diversity.
- **Improved handling of class imbalance**, through focal loss, oversampling, or synthetic data generation for rare classes.
- **Model calibration techniques**, such as temperature scaling or ensembles, to align predicted confidence with real-world reliability.
- **Stronger domain adaptation strategies**, including adversarial methods like DANN, for better generalization across recording settings.
- **Clinical integration**, via model compression for edge deployment, interpretability tools for user trust, and validation on prospectively collected real-world data.

In conclusion, the study highlights the promise of deep learning for automated auscultation, especially through DenseNet121, but also underscores the need for holistic model and data pipeline optimization for successful clinical adoption.

Acknowledgment: This project benefited from the use of ChatGPT (OpenAI) for assistance in writing, editing, and structuring sections under the authors' guidance.

References

1. Andrès, E., Gass, R., Charloux, A., Brandt, C., & Hentzler, A. (2018). Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0. Journal of Medicine and Life.

2. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine.

3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

4. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

5. Perna, D., & Tagarelli, A. (2019). Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) IEEE.

6. Piirilä, P., & Sovijärvi, A. R. (1995). Crackles: recording, analysis and clinical significance. European Respiratory Journal.

7. Chen, Z., Wang, H., Yeh, C.-H., & Liu, X. (2022). Classify respiratory abnormality in lung sounds using STFT and a fine-tuned ResNet18 network. In 2022 IEEE Biomedical Circuits and Systems Conference (BioCAS) , IEEE.

8. Rocha, B. M., Pessoa, D., Marques, A., Carvalho, P., & Paiva, R. P. (2020). Automatic classification of adventitious respiratory sounds: A (un)solved problem? Sensors.

9. Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. In European Conference on Computer Vision (ECCV) Workshops.

10. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering.

11. Piczak, K. J. (2015). ESC-50: Dataset for environmental sound classification. In Proceedings of the 23rd ACM International Conference on Multimedia.

12. Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Proceedings of the 32nd International Conference on Machine Learning.

****