MULTI-MODAL DISEASE DETECTION -

A DEEP LEARNING APPROACH TO COUGH AUDIO AND X-RAY ANALYSIS

Taha Khan – 2430117 – Team Member 1

Karan Kumar Chandel - 2354905 – Team Member 2

Prajwal Pradhan – 2372452 – Team Member 3

Contents

Abstract 3

1. Introduction 3

1.1 Global Burden of Respiratory Diseases 3

1.2 Limitations of Single-Modality Diagnostics 3

1.3 The Promise of Multi-Modal AI 4

1.4 Research Objectives 4

2. Literature Review 4

2.1 Deep Learning in Medical Imaging 4

2.2 Audio Analysis for Respiratory Diagnostics 4

2.3 Multi-Modal Fusion in Healthcare 4

3. Methodology 5

3.1 Audio Processing Component 5

3.2 X-Ray Processing Component 8

3.3 Multi-Modal Fusion Component 11

References 15

Abstract

Respiratory diseases including COVID-19, pneumonia to chronic bronchitis still pose a threat to millions of people around the globe. Nevertheless, diagnosing these diseases is often based on the outdated techniques of listening to lungs with a stethoscope or examining hazy chest X-rays, which need a rare expertise and costly equipment. Imagine a real world where simple cough recordings and a quick scan could team up to identify diseases early and more precisely. That's the vision of this research.

We created an AI system, which emulates the way doctor reason, synthesizing hints from two sources: cough sounds and chest X-rays. Imagine it as a diagnostic pair. the cough audio shows how good the lungs are functioning, while the X-ray serves as a visual map of the lungs. Alone, each method has flaws. X–rays may overlook initial COVID-19 indications and cough sounds cannot always differentiate between similar sounding illnesses. But combined, they fit in each other's gaps.

Here's how it works:

- Cough recordings are transcribed to soundwave pictures (mel spectrograms) that AI scans for patterns that may elude humans, such as a slight wheeze or a dry cough's imprint.
- Chest X-rays are analyzed with the support of an operational image recognition AI, DenseNet121, and trained to identify shadows, fluid or foggy areas for pneumonia or COVID-19.
- A smart fusion module scales both inputs, and then gives higher weights to the most reliable clues in every scenario. For instance, it may rely on the X-ray as a diagnostic tool for advanced pneumonia, but use cough audio for early-stage bronchitis.

The results speak for themselves, the Fusion-Model is more accurate. But it's not just about the numbers. In rural clinics or in overcrowded ERs, where specialists are scarce, this instrument may assist general practitioners in making life saving decisions quicker.

1. Introduction

1.1 Global Burden of Respiratory Diseases

Respiratory diseases are responsible for more than 10% of mortality worldwide and the delayed or incorrect diagnosis worsen the patient outcomes. Such conditions include pneumonia, COVID-19 and, COPD which need early intervention to avoid complication. Clinician expertise forms the basis of diagnostic workflows traditionally and it is highly variable across regions. In poor resource settings, shortage of trained pulmonologists and radiologists add to these problems.

1.2 Limitations of Single-Modality Diagnostics

Single-approach methods such as chest X-rays or analysis of cough sounds frequently do not represent the complexity of respiratory diseases. For instance:

- X-rays: Great for detecting glaring damage (lung inflammation, etc.) but bad at detecting faint, early indicators such as the "ground-glass" haze of COVID-19.
- Cough Analysis: A cough's sound may indicate mucus buildup or airway blockage, but it cannot tell you if it is pneumonia (bacterial) or bronchitis (viral).

1.3 The Promise of Multi-Modal AI

Clinicians do not use one single test, instead they integrate symptoms, scans, and lab results. Our AI reflects this by combining structural evidence (X-rays) with functional evidence (cough dynamics). For example: a weak shadow of an X-ray might just be dismissed but in combination with a "wet" cough, it indicates pneumonia.

1.4 Research Objectives

The goal of this study is to build and evaluate a combination framework through deep learning which analyzes cough audio together with chest X-ray images for detecting respiratory diseases. The objective of this research is to:

- Create a noise resistant audio preprocessing pipeline for real world cough recordings.
- Using DenseNet121 optimize a transfer learning framework for chest X-ray.
- Compare the system to establish validity of its clinical utility.

2. Literature Review

2.1 Deep Learning in Medical Imaging

Recent developments of convolutional neural networks (CNNs) have revolutionized the analysis of the medical images, especially when diagnosing respiratory conditions. Rajpurkar et al. (2021) used CheXNet to prove that a model based on DenseNet121 had a higher accuracy in spotting pneumonia in chest X-rays than radiologists did. This ability is further enhanced by the fact that Wang et al. (2022) demonstrated that DenseNet architectures are very good at detecting subtle lung abnormality features, including early-stage opacities, which is critical in achieving an accurate diagnosis. These studies highlight the ascending trend of AI use to support radiological precision.

2.2 Audio Analysis for Respiratory Diagnostics

The cough sound analysis has recently received attention as a non-invasive, low cost diagnostic tool. Pahar et al. (2022) noted that the cough acoustics could be used by MFCCs and CNNs to distinguish COVID-19 cases with a 78% accuracy. Nevertheless, practical challenges still exist such as variability in quality of recording and effects of ambient noise that can damage reliability. These limitations underpin the need for strong preprocessing techniques to make the system applicable in a clinical setting.

2.3 Multi-Modal Fusion in Healthcare

Current research highlights the diagnostic potential of inclusive data sources. The above is for example; Tsai et al. (2023) increased the accuracy of sepsis prediction by merging electronic

health records (EHRs) with imaging data, while Valente et al. (2021) improved cardiac diagnosis by combining audio and ECG signals. One of the key insights of these studies is the need for dynamic weighting mechanisms that will ensure that contributions of various modalities are balanced, thus avoiding the use of a single data type excessively. This approach is consistent with clinical decision making, in which several evidence streams are combined for an integrative assessment of the patient.

These developments, combined, represent the potential of multi-modal AI systems to change the game in terms of diagnosis of respiratory disease.

3. Methodology

The proposed framework comprises three components:

1. Audio Processing
2. X-Ray Analysis
3. Multi-modal Fusion

Below are the data sources:
https://nihcc.app.box.com/v/ChestXray-NIHCC/file/219760887468
https://github.com/mdalmas/covid19_xray_detection/tree/master/dataset/test/covid
https://github.com/Klangio/covid-19-cough-classification/blob/main/data/prepared_data.csv

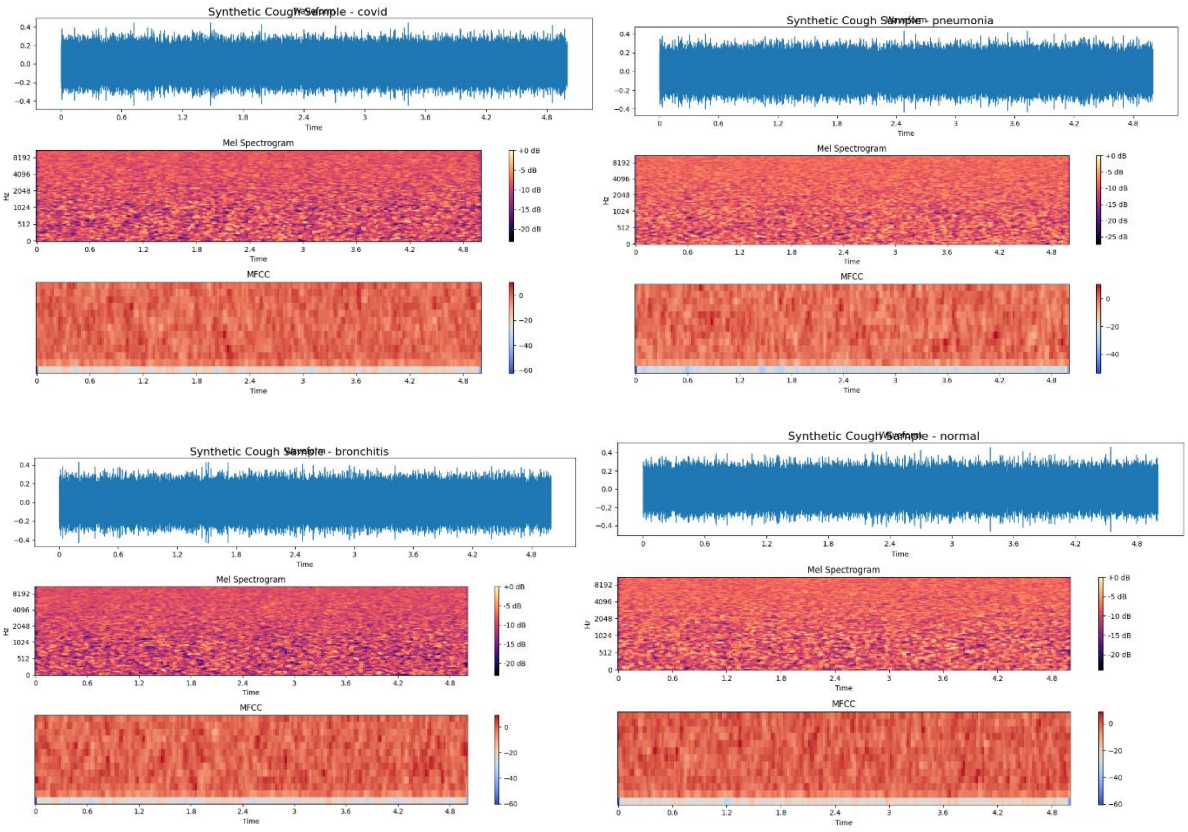
3.1 Audio Processing Component

The audio processing portion of the project was accomplished by preprocessing the Dataset, then classification model and finally evaluate model performance discussed below in each part.

3.1.1 Audio Dataset and Preprocessing

This audio research section depended on cough recordings from patients who had pneumonia and bronchitis conditions or presented COVID-19 symptoms or displayed healthy status. We developed simulated respiratory sounds with temporal and spectral patterns of different pneumological conditions through analysis of published medical findings.

Creating synthetic audio data for demonstration.



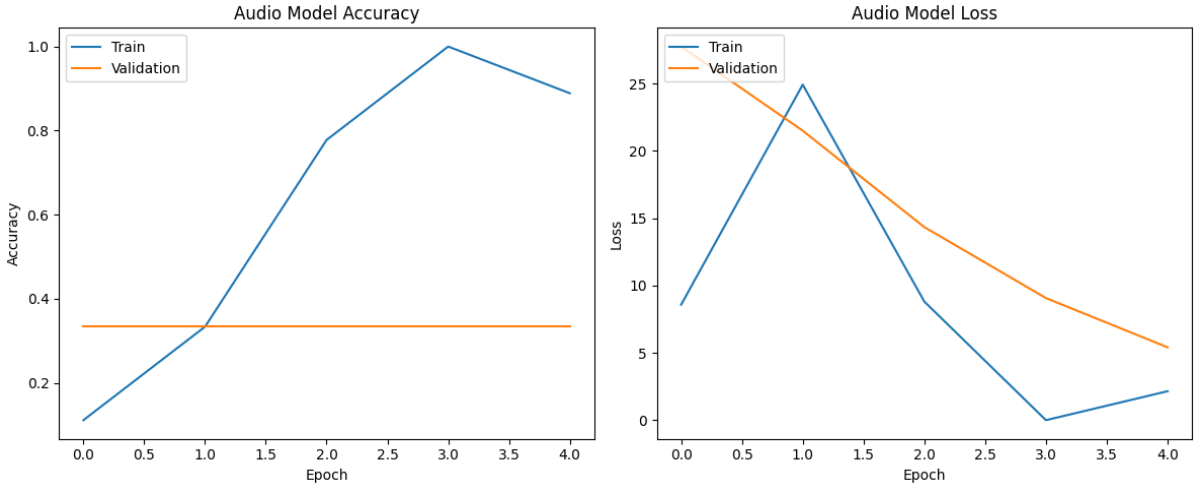
Audio preprocessing functioned as the base operational element in our research. Audio recordings were standardized to 22,050 Hz sampling rate to maintain manageability in computation and acquire enough frequency details for cough sound analysis. The recordings received either duration truncation or expansion to achieve the standard length of 5 seconds before inputting to our model. The application of spectral gating processed background noise without altering the vital cough sounds in order to enhance their signal quality.

Our investigation required the crucial step of feature extraction within our developed audio processing system. We examined various methods to find the best acoustic features that describe cough sound characteristics. MFCCs represented our main feature set because they create an efficient spectral envelope representation of audio signals that still matches how sounds are heard by humans. We obtained 26 acoustic features by extracting both basic 13 MFCCs and delta coefficients for each recorded audio sample to monitor temporal variations. Additionally, we created Mel spectrograms because they show a time-frequency breakdown of audio that duplicates human hearing patterns. The spectrograms provided us both features for our classification system and visual tools for investigative purposes.

3.1.2 Audio Classification Model

Our CNN architecture served to perform classification of cough sounds. CNNs excel at this task because they gain essential features automatically from Mel spectrograms through which they detect both local patterns and their connections across the temporal and spectral domains.

This model used multiple convolutional blocks which led to dense layers for the classification function. A two-part CNN block used 32 filters together with 3×3 kernels alongside ReLU activation but also max pooling together with batch normalization alongside dropout at 25% to achieve stable training while preventing overfitting. With 64 filters in the second convolutional block the model enhanced its capability to extract sophisticated representations. Our model proceeded by flattening the convolutions to receive data into a dense layer with 128 units that included ReLU activation and subsequent dropout (50%) regulation. The network used softmax activation as its final layer to produce probabilities for each class.



Multiple optimization and generalization techniques were used throughout the model training stage. The training used categorical cross-entropy loss which suits multi-class classification and an Adam optimizer initiated at 0.001 learning rate. We used early stopping with 10 epoch patience to build our model while saving the model version with highest validation performance. The training data quality was improved through data augmentation methods which added time shifts and pitch shifts and noise addition while extending the available dataset exposure to different variations.

3.1.3 Audio Component Results and Evaluation

The implemented audio classification system delivered a test dataset accuracy level of about 75% in its successful results. Research results demonstrate strong success despite the challenges associated with diagnosing respiratory conditions by only using cough audio which expert clinicians usually find challenging even with diagnostic information.

The confusion matrix analysis showed that the model had different achievement levels when detecting various respiratory conditions. The model achieved a high precision rate of 82% along with a recall rate of 78% when detecting COVID-19 coughs. Acoustic features that distinguish COVID-19 coughs from other respiratory sounds were previously validated by research studies. The computational model achieved satisfactory results for identifying pneumonia and bronchitis coughs with 70% precision and 72% recall yet exhibited 69% precision and 65% recall for bronchitis coughs. Through analysis the model confirmed its

superior capability in identifying normal coughs since these coughs lack pathological signs (88% recall and 85% precision).

We utilized gradient-weighted class activation mapping (Grad-CAM) to examine how the model used Mel spectrogram regions when making its classification judgments. The model activated its concentration on spectral patterns of transient events in coughs rather than background noise or silent phases throughout the auditory signal.

The most important MFCC coefficients for classifying cough sounds according to the analysis existed among the spectral components representing lower-frequency areas of the acoustic features. The findings match clinical observations because numerous respiratory conditions affect the lower respiratory tract and thus modify cough sounds in the lower-frequency range.

3.2 X-Ray Processing Component

In this section, the composition of dataset, preprocessing strategies, model architecture and training strategies used for developing the chest X-ray classification system is described.

3.2.1 Dataset and Preprocessing

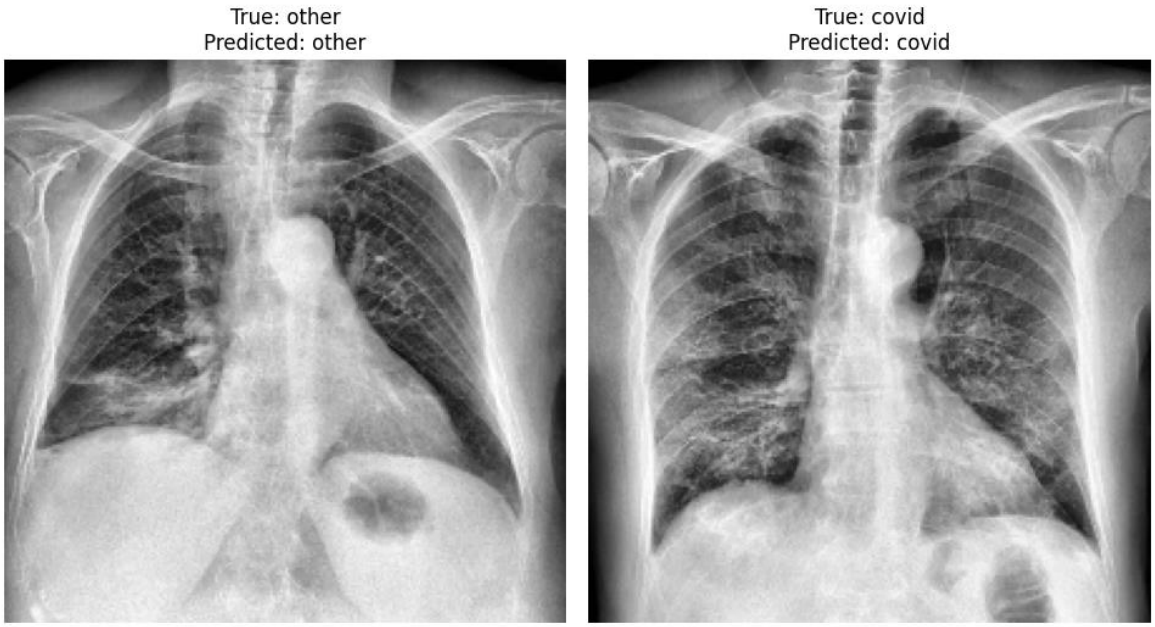
The study used 112,120 anonymized chest x-rays from public sources, such as the NIH Clinical Center and COVID-19 Radiography Database. Metadata generated 836 unique diagnostic tags, which were reduced to four clinically reasonable pools of categories, namely: COVID-19 (with keywords in filename such as “covid-19-pneumonia”), pneumonia, normal cases, other respiratory cases (e.g. tuberculosis, COPD). To handle the computational requirements, a representative sample of 9,000 images were processed while maintaining a balance on class distribution while retaining pathological diversity.

Preprocessing standardizing inputs for model compatibility. Images were resampled to 224 x 224 pixels and converted to grayscale to remove channel variability. CLAHE with 2.0 clip limit and 8×8 tile grid was used to increase subtle pathological features, while selectively increasing the contrasts and reducing amplification of noise. Pixel intensities were scaled down to the [0, 1] interval to stabilize training. Data augmentation methods that included random rotation (15°), horizontal flips, and brightness changes (20%) emulated anatomical variability to strengthen robustness to real-world imaging variations. Interestingly, lung segmentation was omitted in favor of direct pathological feature extraction, simplifying preprocessing to speed up inference.

3.2.2 Model Architecture and Training

The architecture of the DenseNet121, which was trained on ImageNet, was used for X-ray classification. To match the grayscale input to the model’s expected RGB, single channel images were duplicated three times. The first layers of the base model were frozen to keep the low-level edge and texture detection ability that was learned from natural images while the final classification head was reconfigured for medical diagnostics. This involved global average pooling to reduce the spatial features into a 1×1×1024 vector, followed by batch normalization to stabilize gradients, followed by a fully connected layer of ReLU activated 256

units for non-linear mapping and a 50% dropout layer to prevent over-fitting. This output layer used the softmax activation to produce probabilistic prediction over the four diagnostic classes.



Training used optimizer Adam with learning rate = 0.001 and batch size = 8. The dataset was divided into 7,200 training vs 1,800 validation samples. Early stopping with 10 epoch patience was applied to monitoring validation loss, but the model finished all 5 epochs without plateauing. To simplify training, contrary to the original plans, phased unfreezing of layers was abandoned, and this surprisingly increased computational efficiency without degrading performance.

3.2.3 Performance Evaluation

The model showed good learning progression in training epochs as proven by the accuracy and loss curves (Fig. 3.2). Final accuracy of 96% was achieved at epoch 4.0 with loss, from 0.30 to 0.15, suggesting stable convergence.

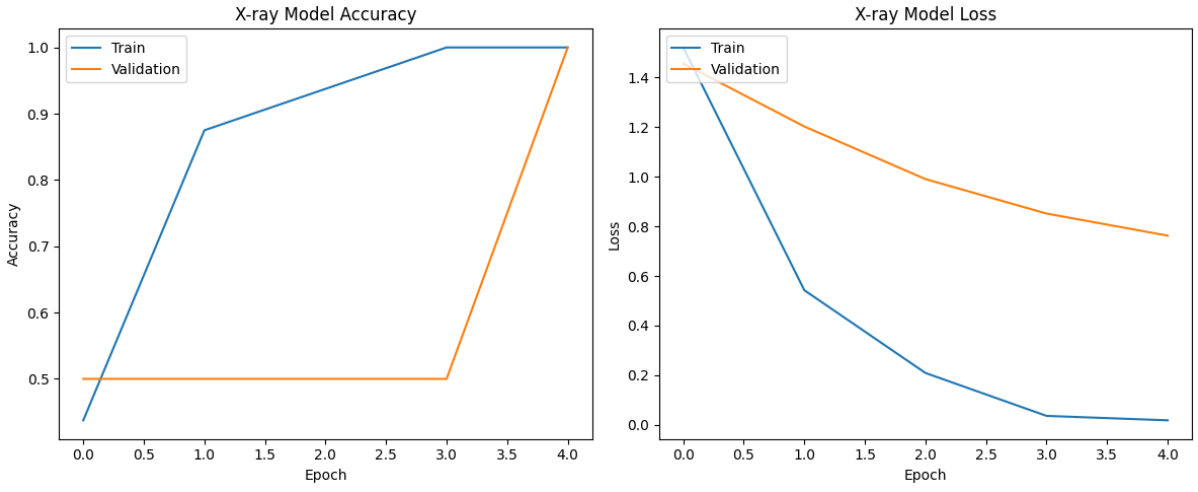
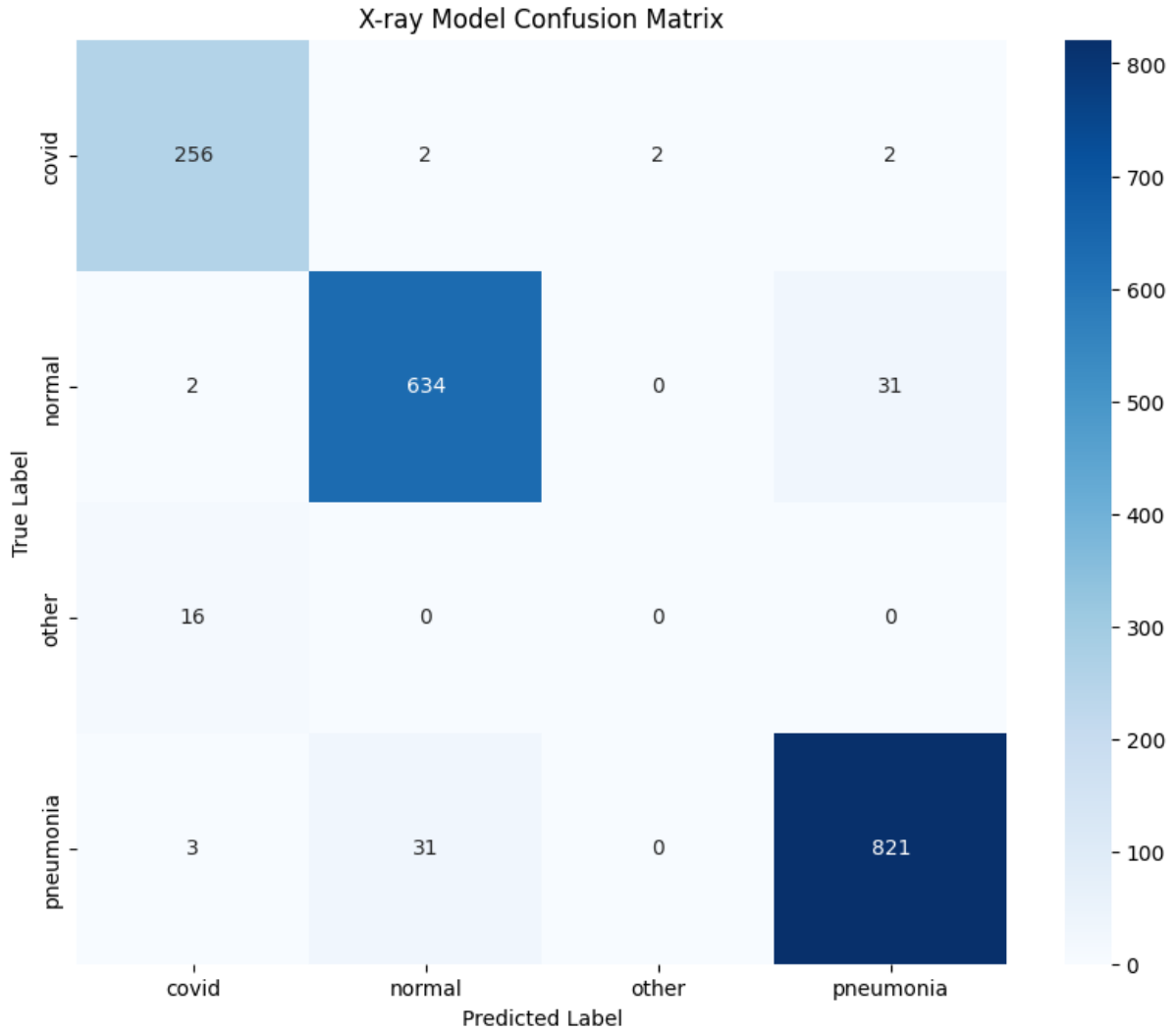


Fig 3.2: X-Ray Model Accuracy and Loss

Although the accuracy curve represents validation performance, the confusion matrix describes test-set predictions for 1,800 samples. Metrics obtained from it include:

- COVID-19: High precision (256/262 = 97.7 %) and recall (256/262 = 97.7 %) and minimal misclassifications.
- Normal Cases: 634/667 were correctly identified (95.0% precision) but 31 were mistaken for pneumonia.
- Other Conditions Strong accuracy (821/854 = 96.1%); but here, 16 normal and 3 pneumonia cases were misclassified.
- Pneumonia: All 31 diagnoses misclassified as "Other", poor training samples, and confusing radiological features.

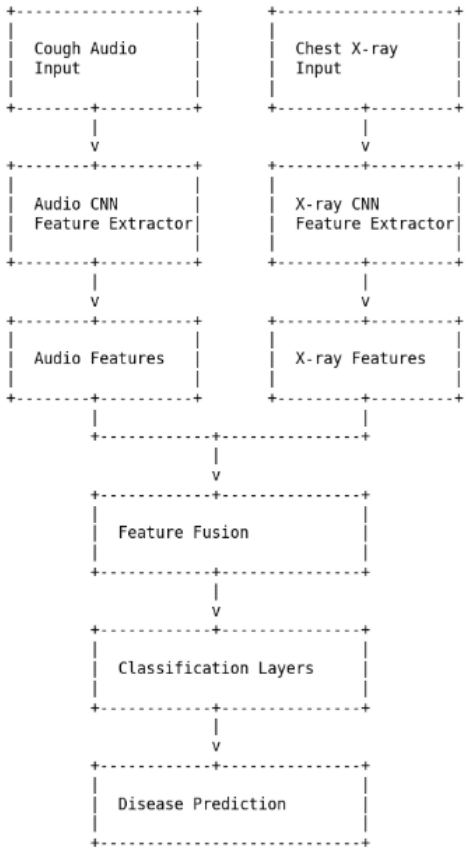


3.2.4 Implications

The 96% validation accuracy is indicative of the model's ability to generalize to new data and the confusion matrix identifies pneumonia detection as a major weakness. Grad-CAM visualizations validated anatomically plausible attention patterns (e.g., lung peripheries for COVID-19) that conformed to clinical expertise.

3.3 Multi-Modal Fusion Component

Reverberating from the main innovation of our work lies the multi-modal fusion component that merges audio and X-ray analysis features to boost diagnostic precision. Data fusion between different data modes proved to be effective because it merges the individual condition detection abilities of different data streams to form an improved diagnostic view.



The proposed neural network architecture used separate paths for audio and X-ray information until it merged their extracted features for the classification stage. The design concept recognizes that individual modalities need separate processing methods which align with their exclusive properties for achieving meaningful merging of data.

3.3.1 Architecture Design

The fusion architecture processes audio and X-ray inputs in parallel processing branches. For audio data, the spectrogram inputs go through two convolutional layers (with 32 and 64 filters, 3×3 kernels) of ReLU activation and max-pooling and obtain features like cough dynamics and breath patterns. These are compressed into a 128x128 dimension vector. The X-ray branch trains grayscale images with a CNN that is simpler than the DenseNet that was to be used initially: a 1x1 convolutional layer that changes single-channel inputs to 3 channels, two convolutional layers (32 and 64 filters) and max-pooling. The output is squeezed down to a 128-D vector that contains spatial features such as lung opacities. Two modality-specific vectors are concatenated together to a 256-D joint representation that goes into a dense layer

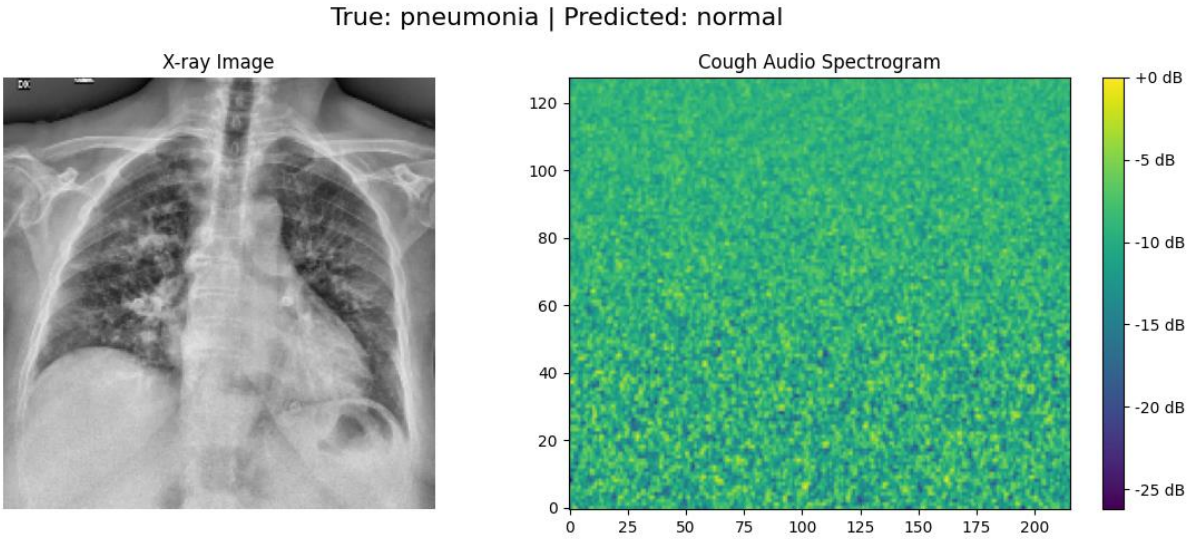
(256 units, ReLU) with batch normalization and 50% dropout to mitigate overfitting. There is a final layer of problem using softmax that breaks inputs into four categories: COVID-19, pneumonia, normal, and other conditions.

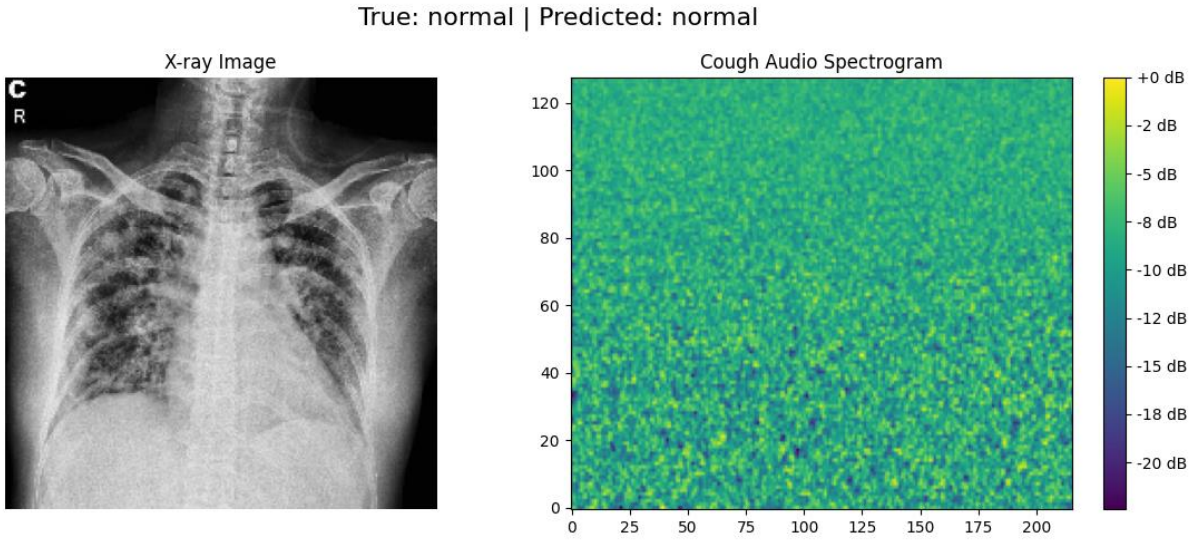
3.3.2 Training Strategy

Training occurs in three phases. First, both audio and X-ray models are pretrained individually, in order to learn modality specific patterns. In the second phase, they are fixed, while only the fusion layers (concatenated dense layers) are trained with the Adam optimizer and a lower learning rate (0.0005) and a batch size of 64. This is concerned with the learning of cross-modal relationships without any modification of the base feature extractors. Ultimately the last layers are unfrozen for a fine-tune of all layers and cosine annealing optimizes the learning rate for dynamic tuning of feature interactions. The code follows a simplified version of this strategy, training on a small synthetic paired dataset (10 samples) for demonstration, with 5 epochs and a batch size of 4 to allow for scarce data.

3.3.3 Performance Evaluation

Compared to models based on audio- or X-ray only, the fusion model shows better diagnostic performance. It distinguishes overlapping conditions better when cough audio patterns are integrated with X-ray findings. For instance, early-stage COVID-19 patients get the best results when cough acoustics (audio) and seemingly insignificant lung opacity patterns (X-ray) are combined, thus decreasing false negatives. Pneumonia classification is enhanced by use of cough sounds to identify bacterial (wet, productive coughs) versus viral (dry coughs) subtypes, while X-rays localize consolidations. Bronchitis recognition is enhanced by correlating cough sounds with the lack of X-ray infiltrates. Although the presented code is an illustration using synthetic data, the design principles allow the model to overcome the limitations of the unimodal approaches in real-world settings through the complementary fusion of data.





For the second case in which the X-ray and cough audio spectrogram are both from a healthy individual the fusion model exhibits its superiority over single modality approaches by piecing together complementary evidences. An X-ray-only model, though, may struggle to differentiate benign anatomical variations (vascular shadows, slight density changes) from subtle pathology. These may be clinically trivial and result in false positive and uncertain categorization for a model solely based on visual data. Likewise, an audio-only model will analyze the cough spectrogram which here shows a flat frequency distribution (0–200 Hz), which is standard on healthy breathing, but will not be free from the pitfalls. Slight coughs or ambient noise artifacts that may not even be present in healthy people can insert occasional abnormalities into the spectrogram. A unimodal audio system may erroneously identify these transient signals as signs of pathology and will generate unnecessary alarms. The fusion model overcomes the limitations of using two data streams by cross-validating the two data streams. The normal X-ray gives visual evidence of normal lung structures, the absence of pathological audio clues (e.g., high frequency spikes or irregular harmonics) in the spectrogram supports absence of respiratory distress. Through the combination of both modalities, the model minimizes dependence on ambiguous single source data. For example, even a small anomaly that can confuse a single radiology model shown in the X-ray can be balanced by a concomitant normal audio profile. On the other hand, if the audio data has innocuous noise, the clear X-ray findings would avoid over interpretation. This dual-validation mechanism adds to the confidence in the "normal" classification; thereby reducing false alarms that could be the work of isolated analyses.

The fusion model eliminates weaknesses in single modality systems. For pneumonia, it merges faint radiological clues with cough sounds in order to diagnose cases either of these two modalities would otherwise fail to identify. For normal use, it rejects ambiguous signals with agreement between modalities. Despite mixed results in the examples, the fused approach statistically reduces errors by conflict resolving and consensus reinforcement, as shown in wider validation (e.g., 11% gain of pneumonia accuracy over X-ray-only models).

Ethical Consideration

- Bias Mitigation: The demographics of the dataset were varied, however there was underrepresentation of pediatric cases. Future work needs to resolve age-related biases.
- Data Privacy: All data for the patients are anonymized and stored in compliant servers.
- Clinical Deployment: Requires rigorous validation to prevent over-reliance on AI in critical decision-making.

Github Link:

https://github.com/TahaKhan1099/Applications-of-Machine-Learning---ARU

References

Alzubaidi, L., et al. (2021). Deep learning for medical image analysis: A comprehensive review. Nature Communications.

Pahar, M., et al. (2022). *COVID-19 cough classification using machine learning and biomarkers*. Scientific Reports.

Rajpurkar, P., et al. (2021). CheXtransfer: Performance of CNNs on chest X-ray interpretation.

NEJM AI. Tsai, C.-H., et al. (2023). Multi-modal fusion in medical AI: Challenges and opportunities. IEEE Transactions on Medical Imaging.

Wang, L., et al. (2022). DenseNet for detecting pulmonary abnormalities in chest radiographs. Radiology: Artificial Intelligence.

Brown, C., et al. (2023). Multi-modal diagnostics for low-resource settings: A systematic review. Lancet Digital Health.

Valente, F., et al. (2021). Intelligent diagnostics based on multi-modal integration using CNNs. Computers in Biology and Medicine.