ADVANCEMENTS IN CHEST RADIOGRAPHY PNEUMONIA CLASSIFICATION THROUGH FINE-TUNING USING MIMIC-CXR-JPG DATASET

By

Yifan Zhang

Thesis Project
Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

February 2024

Committee Members: Profs. Daniel W. Linna Jr., Mohammed Anwarul Alam

ABSTRACT

Within the realm of medical diagnostics, the analysis of chest radiographs using machine learning has made considerable progress, particularly with the classification of pneumonia. This thesis details the enhancement of convolutional neural networks (CNNs) through fine-tuning to improve pneumonia detection on the MIMIC-CXR-JPG dataset.

This study acknowledges the challenges of manual X-ray examination, which requires expert interpretation and is subject to diagnostic variability. To address these challenges, this research employed advanced machine learning techniques, aiming to bolster the accuracy and reliability of pneumonia diagnosis—a critical factor in clinical decision-making.

The methodology incorporated comprehensive preprocessing, including the rectification of class imbalances and the adoption of data augmentation strategies to foster model robustness and generalizability.

The empirical results are compelling. The baseline CNN model registered a high error rate of 0.7688. After fine-tuning, the error rate was significantly lowered to 0.3133. Moreover, the model's diagnostic capability was reflected in the area under the receiver operating characteristic curve (AUC), achieving scores of 0.72 for bacterial pneumonia and no pneumonia, and an outstanding 0.89 for viral pneumonia. Additionally, the average precision (AP) for bacterial pneumonia reached 0.50, and 0.85 for cases without pneumonia, further showcasing the model's refined predictive power.

In conclusion, the success of fine-tuning CNNs on the MIMIC-CXR-JPG dataset marks a substantial stride in the diagnosis of pneumonia from chest X-rays, which may hold significant promise for the integration of machine learning in the broader spectrum of medical image analysis.

ACKNOWLEDGEMENTS

I extend my deepest appreciation to my advisor, Prof. Wood-Doughty, for their invaluable guidance and support throughout my research journey. I am also grateful to my committee members Prof. Linna and Prof. Alam for their insightful feedback and contributions to my academic growth. My thanks to the program for the enriching learning environment and resources. Heartfelt gratitude to my family and friends for their unwavering support and encouragement. Lastly, a special mention to my cat Pepper, whose companionship was a constant source of joy and comfort.

This thesis marks my inaugural venture into a rigorous research project, and I couldn't be more grateful for having Prof. Wood-Doughty as my supervisor. Traditionally, I've been a person who procrastinates, often leaving tasks to the last minute. However, this project, under Prof. Wood-Doughty's guidance, has been a transformative journey, teaching me the critical importance of early planning and proactive action. There were numerous moments of uncertainty, where the path forward was unclear and the purpose of my efforts seemed lost. It was during these times that Prof. Wood-Doughty's wisdom shone brightest, particularly when he encouraged me to develop a weekly schedule for my thesis writing. Although adhering to this schedule was challenging, it served as a beacon, guiding me back whenever I strayed. This experience has not only culminated in the completion of my thesis but has also instilled in me valuable habits and a newfound appreciation for discipline. I am profoundly thankful for this growth opportunity and for the enduring impact it has had on my academic and personal development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Chest Radiology: A Pillar in Lung Disease Diagnosis

Radiology, particularly through the use of chest X-rays, plays a pivotal role in disease diagnosis. The non-invasive nature of radiology offers essential insights into the body's internal state, crucial for diagnosing various diseases[1]. Chest X-rays are among the most frequently ordered radiological tests and serve as a key diagnostic tool in many clinical settings[2].

The effectiveness of chest X-rays in revealing thoracic cavity pathologies is well-documented. They provide critical information about the lungs, heart, and the bony structures of the chest, instrumental in detecting conditions such as pneumonia, tuberculosis, lung cancer, heart abnormalities, and skeletal issues[3].

1.1.2 Challenges in Chest X-ray Interpretation: Expertise and Accessibility

Despite their utility, interpreting chest X-rays requires significant expertise and training[4]. The demand for radiology services often exceeds the availability of expert radiologists, especially in resource-limited settings. In the U.S., there is a noted decrease in the ratio of radiologists to the overall physician workforce, with a distribution skewing towards urban areas[5]. This leads to non-specialists like intensivists and emergency physicians often providing initial interpretations, which may not be as accurate or comprehensive, potentially delaying effective patient care.

This challenge is more pronounced in resource-poor regions. For example, Rwanda, with a population of 12 million, had only 11 radiologists as of 2015[6], and Liberia, with four million people, had just two practicing radiologists[7]. Such scarcity of specialists can significantly delay diagnoses and overwhelm the available medical professionals.

1.1.3 Integration of Machine Learning in Radiology

In response to these challenges, there has been a growing emphasis on integrating machine learning technologies in chest X-ray analysis. Machine learning models, especially deep learning techniques like convolutional neural networks, have shown promise in accurately identifying and classifying lung diseases in X-ray images. Successes in detecting diseases such as pneumonia and COVID-19 demonstrate the potential of these models to supplement traditional diagnostic methods[8].

However, a major hurdle remains: the lack of explainability in these machine learning models[9]. The ability to understand how a model arrives at a diagnosis is crucial in medical settings, where trust and clarity in decision-making are paramount. The current 'black box' nature of many models limits their practical application in real-life clinical scenarios.

This thesis addresses both the accuracy and the explainability of machine learning models in chest X-ray analysis. By fine-tuning a disease classification model on the MIMIC-CXR-JPG dataset, the research aims to enhance model performance while also making strides in the interpretability of the results. This endeavor seeks to not only support radiologists in making more informed decisions but also extend expert-level diagnostic support to under-served regions, thus democratizing access to quality medical diagnostics.

1.2 Objectives

The objective of this thesis is the fine-tuning of advanced machine learning models for the enhanced classification of pneumonia in chest X-ray images. This endeavor focuses on refining the models' architecture and hyperparameters and rigorously optimizing the training process specifically for the MIMIC-CXR-JPG dataset. The goal is to substantially boost the accuracy and reliability of disease detection, ensuring the models' applicability and utility in actual clinical environments. By concentrating exclusively on these technical improvements, the research aims to contribute significantly to the field of medical diagnostics, particularly in improving the speed and precision of pneumonia diagnosis in chest radiography.

1.3 Significance

The enhancement of disease classification models through machine learning, particularly in the fine-tuning of algorithms for chest radiography, is of considerable significance in modern medical diagnostics. By elevating the accuracy and reliability of pneumonia detection in chest X-ray images, this research stands to substantially improve the early diagnosis and consequent management of this condition. Improved diagnostics directly translate to better patient care by supporting prompt and precise treatment decisions. The meticulous process of fine-tuning machine learning models on the MIMIC-CXR-JPG dataset demonstrates a practical approach to overcoming current limitations in automated radiographic analysis. Thus, this work is poised to make a substantial impact on patient outcomes and the efficacy of healthcare systems globally.

1.4 Structure of the Thesis

In this thesis, I embark on a comprehensive exploration of the application of machine learning in medical imaging, with a specific focus on chest X-ray imaging for disease diagnosis and causal analysis in medical research. The structure of the thesis is organized as follows:

Literature Review This section delves into the historical and current use of chest X-rays in medical diagnostics, reviews the advancements in machine learning applications in medical imaging, and examines causal analysis methods in medical research. It identifies gaps in existing research that this thesis aims to address, particularly in the fine-tuning of disease classification models.

Methodology The methodology chapter outlines the specifics of the MIMIC-CXR-JPG dataset and the detailed process of model development and fine-tuning. The statistical and machine learning methods utilized for fine-tuning and their optimization for disease classification are described, with an emphasis on the novel contributions of this study.

Results Performance metrics of the fine-tuned models are reported in this section, focusing on how these models enhance pneumonia classification. A critical comparison with previous studies is provided, highlighting the incremental improvements and novel insights offered by the fine-tuned models.

Discussion The implications of the fine-tuned model's performance are discussed in relation to the existing literature, emphasizing the advancement and practical implications of the study's findings. Challenges faced, limitations encountered, and potential areas for future research are also elaborated upon.

Conclusion The thesis concludes with a reflective summary of key findings, contemplating the impact of this research on medical diagnostics and treatment planning. Final thoughts on the interface of integrating machine learning in medical research and practice are provided.

This structure is designed to guide the reader through a logical progression of topics, from a broad understanding of the field to specific insights derived from this research, culminating in a reflection on the broader implications and future directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Chest X-ray Imaging in Disease Diagnosis

The inception of chest X-ray imaging can be traced back to 1895, with Wilhelm Conrad Röntgen's discovery of X-rays [10]. This groundbreaking technology rapidly became a staple in medical diagnostics, offering a non-invasive glimpse into the internal structures of the body. Over the years, chest X-ray technology has evolved significantly. The field has progressed from film-based radiography, with its inherent limitations, to advanced digital radiography, enhancing image quality and reducing radiation exposure[10].

Today, chest X-rays are one of the most commonly performed radiological examinations, vital in the rapid assessment and diagnosis of thoracic diseases[2]. They are especially effective in diagnosing a range of conditions, including lung infections like pneumonia, structural anomalies such as fractures, and chronic conditions like lung cancer[3]. The advent of digital radiography and subsequent improvements in image processing have further solidified their role in clinical diagnostics[11].

Despite these advancements, the interpretation of chest X-rays remains a challenge. It requires a high degree of expertise, and there is often variability in interpretation among radiologists. The ability to detect subtle pathological changes is another area where even skilled professionals can struggle, impacting the accuracy of diagnoses[5]. Additionally, the global disparity in access to radiology services, particularly in low-resource settings, poses a significant challenge. In many parts of the world, a shortage of trained radiologists leads to delayed diagnoses and treatment, impacting patient outcomes[6].

In conclusion, while chest X-rays are a fundamental diagnostic tool, their effectiveness is limited by challenges in interpretation and global disparities in access to skilled radiologists. These issues highlight the need for innovative solutions, such as machine learning, to augment the diagnostic process, paving the way for the next section of this thesis, which discusses the integration of machine learning in medical imaging[12].

2.2 Machine Learning for Medical Imaging

Machine learning, particularly its application in medical imaging, has undergone significant transformation over recent years[13]. Its inception marked a pivotal shift in how medical images are analyzed, offering a powerful tool for enhancing diagnostic accuracy[14]. Among the various branches of machine learning, deep learning has emerged as a particularly influential technology in medical imaging[15]. Deep learning, characterized by its use of neural networks with multiple layers, has demonstrated advantages over traditional machine learning methods, especially in handling complex image data[16].

In the realm of disease classification, deep learning models, such as Convolutional Neural Networks (CNNs), have become increasingly prominent. These models are adept at extracting intricate patterns from medical images, leading to more accurate and reliable disease detection and classification[17]. For instance, in chest X-ray analysis, deep learning techniques have shown remarkable success in identifying conditions such as pneumonia [18], tuberculosis[19], and even signs of COVID-19 [8] [20].

Several landmark studies and models exemplify the successful application of deep learning in medical imaging. Notably, the U-Net architecture revolutionized medical image segmentation with its effective and efficient structure, especially notable in its application to biomedical image segmentation [21]. The DeepLesion dataset significantly advanced the field by providing a large-scale dataset for lesion annotations in CT images, aiding in the development of universal lesion detection frameworks [22]. Additionally, anomaly detection in medical images was made successful through Generative Adversarial Networks (GANs), particularly in optical coherence tomography of the retina [23]. These examples underscore the potential of deep learning in revolutionizing the field of radiological diagnostics.

Despite these successes, the application of deep learning in medical imaging is not without its challenges and limitations. Issues such as the need for large labeled datasets, model interpretability, and the generalizability of these models across different populations and imaging equipment remain significant hurdles[24]. Addressing these challenges is crucial for the wider adoption and effective utilization of deep learning in clinical settings.

2.3 Gap in Existing Research

Current research in machine learning for medical imaging is progressing, yet it confronts significant challenges that impede optimal model performance, particularly in the accurate and generalizable classification of diseases from imaging data. Issues such as overfitting and the 'black box' nature of deep learning models hinder their full integration and trust in clinical practice. Moreover, there is an identified need for improved techniques in model optimization and data handling to enhance the recognition of less common diseases, such as viral pneumonia, where models often underperform. This research aims to bridge these gaps by focusing on fine-tuning strategies to bolster the precision and interpretability of machine learning models, ensuring their effectiveness across diverse clinical scenarios.

CHAPTER 3
TECHNICAL APPROACH

3.1 Dataset and Preprocessing

For this project, I use the MIMIC-CXR-JPG dataset (See Figure 3.1 for a sample batch). The MIMIC-CXR-JPG dataset is a substantial, publicly accessible database of labeled chest radiographs designed to facilitate and support a broad range of research in medical computer vision. Derived from the Beth Israel Deaconess Medical Center, it comprises 377,110 chest X-ray images associated with 227,827 imaging studies conducted between 2011 and 2016. This dataset stands out for its size and the richness of the accompanying data, including 14 labels derived from the analysis of corresponding free-text radiology reports through natural language processing tools. A significant aspect of the MIMIC-CXR-JPG dataset is its focus on providing a standardized reference for data splits and image labels, which is essential for developing accurate automated analysis techniques for chest radiographs.

I focused on subsets p10, p11, and p13, which represent 3 of the 10 patient groups, a significant portion of the entire repository. This subset includes over 100,000 chest radiographs, indicative of the dataset's diversity in patient demographics, clinical conditions, and imaging techniques. The preprocessing steps, crucial for preparing the dataset for machine learning applications, involve converting images from DICOM to JPEG format, normalizing image intensities, standardizing resolutions, and de-identification to ensure patient privacy.[25]

3.1.1 Labeling

Additionally, the original MIMIC-CXR-JPG dataset, processed with NegBio[26] and CheXpert[27] (both areautomated natural language processing tools, enabling precise identification and classification of radiological findings from free-text radiology reports by leveraging advanced algorithms

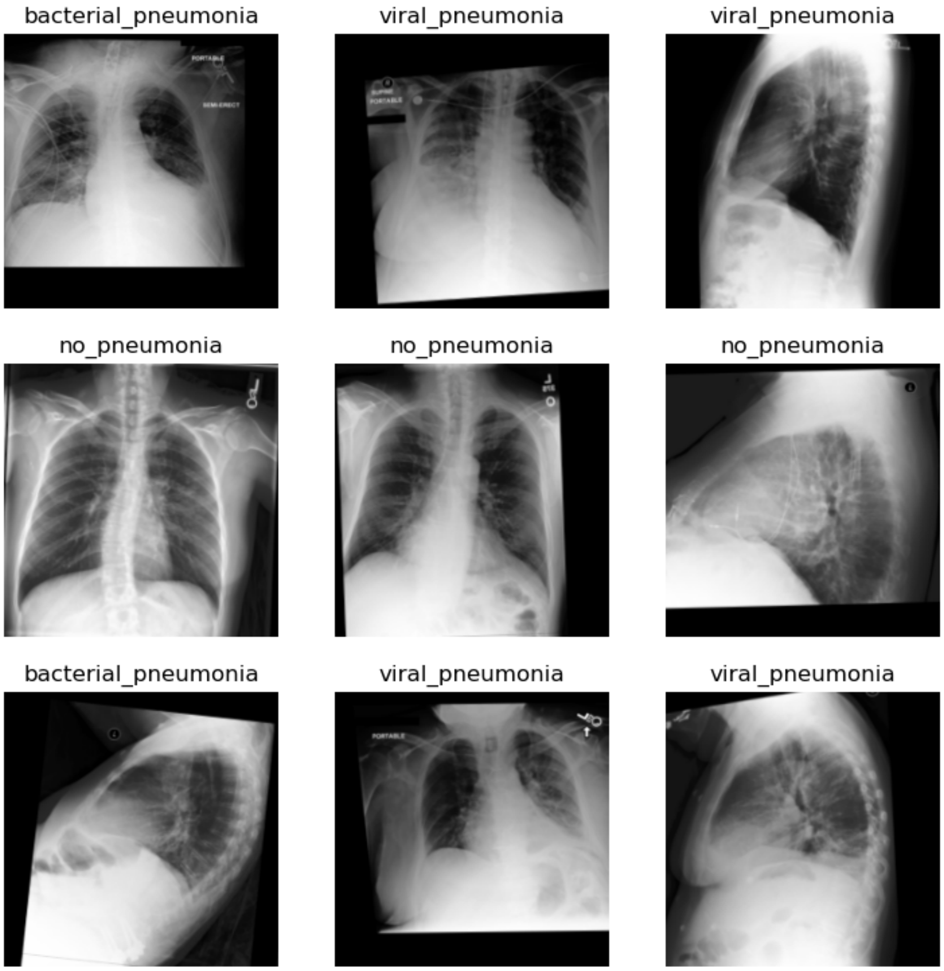

Figure 3.1: Representative Chest X-ray Images from the Dataset. This collage showcases sample images from each of the three classes: ’No Pneumonia’, ’Bacterial Pneumonia’, and ’Viral Pneumonia’, illustrating the diversity of cases that the model is being trained to recognize and classify.

to analyze and interpret unstructured medical text data) for label extraction, does not differentiate between bacterial and viral pneumonia. To overcome this limitation, I integrated the MIMIC-CXR-JPG dataset with the MIMIC-IV[28] dataset, leveraging the shared patient identifiers to extract detailed diagnostic information with ICD9 and ICD10 codes (The International Classification of Diseases (ICD) is a globally used diagnostic tool for epidemiology, health management, and clinical purposes, classifying diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases). This integration enabled us to manually label images with more specific diagnoses, categorizing them into no-pneumonia, bacterial pneumonia, and viral pneumonia based on the detailed diagnostic titles. This manual labeling (See Appendix A) process was critical for accurately classifying the chest radiographs into my targeted categories, enhancing the dataset's utility for my specific research objectives. Patients diagnosed with both viral and bacterial pneumonia simultaneously, as well as those without a specific type of pneumonia diagnosed and no further diagnosis available, were excluded from my analysis. This step was necessary to maintain clear, unambiguous categorizations for my study. As a result, approximately 15,000 images were excluded, leaving a total of 87,635 chest X-ray images with accurate labels for my research.

3.1.2 Address Class Imbalance

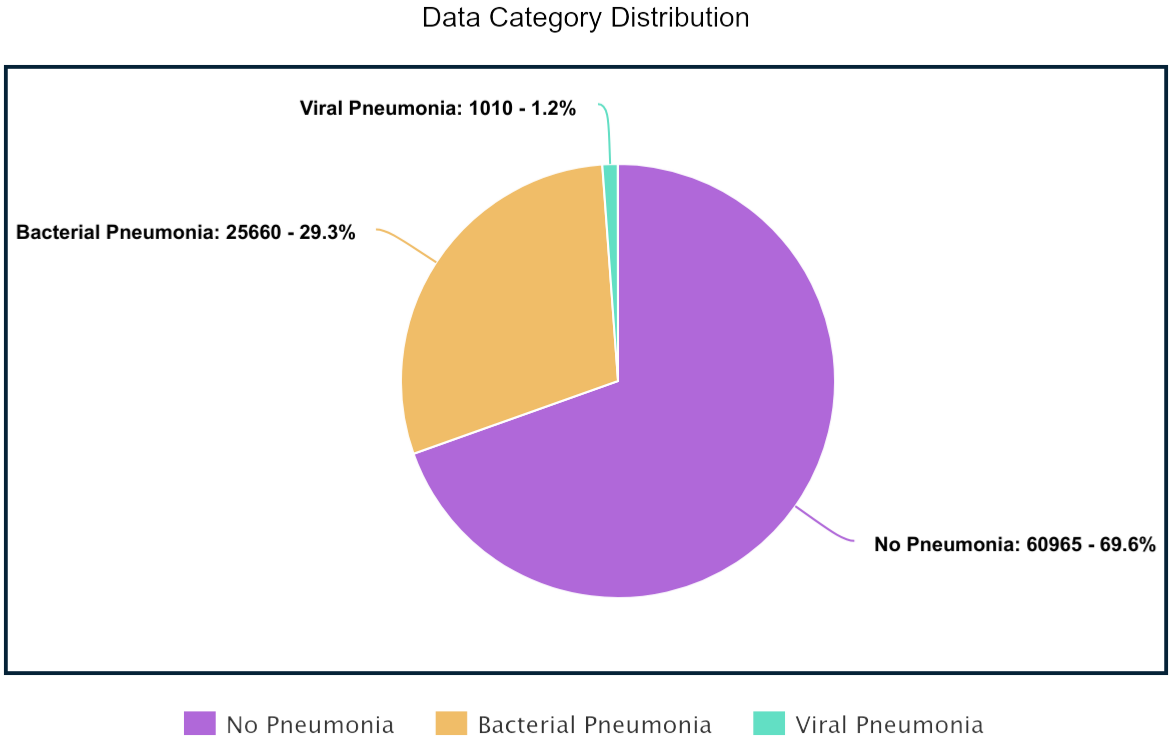

Figure 3.2: Distribution of Chest X-ray Categories in the subset of MIMIC-CXR-JPG Dataset. This pie chart illustrates the proportion of images classified as 'No Pneumonia', 'Bacterial Pneumonia', and 'Viral Pneumonia', highlighting the dataset's imbalance and informing my data augmentation strategy.

Figure 3.2 showing the distribution of categories within subsets p10, p11, and p12 is presented here, highlighting the initial class imbalance. To address the significant class imbalance and to refine disease classification, I implemented upscaling for minority classes and downscaling for over-represented classes. These approaches aimed to balance the dataset for more effective model training. By employing these data manipulation and labeling strategies, I aimed to create a balanced and accurately labeled dataset, facilitating the development of robust machine learning models capable of distinguishing between complex disease states.

3.1.3 Data Augementation

To enhance the model's ability to generalize from the training data to unseen images, I applied a series of data augmentation techniques. These transformations simulate possible variations in chest X-ray images that a model might encounter in real-world scenarios. The augmentation pipeline was carefully designed to include the following transformations:

- Rotation: Images were rotated by up to 10 degrees to account for variations in patient positioning.
- Zooming: A maximum zoom of 1.1x provided slight magnification effects, mimicking closer or farther imaging.
- Lighting: Brightness and contrast adjustments were capped at 20% to replicate different radiographic exposure conditions.
- Affine Transforms: Applied with a probability of 75%, these transforms simulate small translations, scales, and shears that can occur during image acquisition.
- Lighting Transforms: With a 75% chance, subtle lighting variations were introduced to prepare the model for changes in image illumination.

No horizontal flipping or warping was used, as these do not represent common variations in chest X-ray imaging. The applied augmentation techniques are integral to the robustness of the training process, ensuring the model's performance is not solely tied to the specificities of the training set images.

This approach aims to improve the model's diagnostic accuracy by providing a more comprehensive 'understanding' of chest X-ray appearances under various conditions.

3.2 Model Development and Fine-Tuning

3.2.1 Model Development

The journey to identify the most effective machine learning framework for disease classification with the MIMIC-CXR-JPG dataset was both extensive and insightful. my exploration spanned from Hugging Face's transformers, renowned for their adaptability in various tasks, to the precise and fast object detection capabilities of YOLOv8, and the flexible, robust environments of TensorFlow and PyTorch. However, these initial considerations encountered specific challenges, particularly with YOLOv8's stringent data input requirements and the absence of bounding box annotations in the MIMIC-CXR-JPG dataset, which proved incompatible with my project's needs. These hurdles highlighted the importance of selecting a framework not only for its technical prowess but also for its compatibility with the dataset in use.

The selection of Fastai model on Huggingface [29] marked a significant turning point in my model development process. Figure 3.3 shows its typical CNN structure. Unlike the other frameworks, Fastai facilitated a more streamlined and efficient approach to determining the optimal learning rate (Figure 3.4), a crucial aspect of hyperparameter tuning that often dictates the success of model training. Fastai's user-friendly interface and powerful tools, including the learning rate finder, data augmentation techniques, and early stopping mechanisms, enabled us to fine-tune my models with greater precision and less trial and error. This adaptability was instrumental in overcoming the challenges previously faced, allowing for the development of sophisticated models capable of accurately classifying a range of diseases from chest radiographs, thereby leveraging the rich diversity and complexity of the MIMIC-CXR-JPG dataset to its fullest potential.

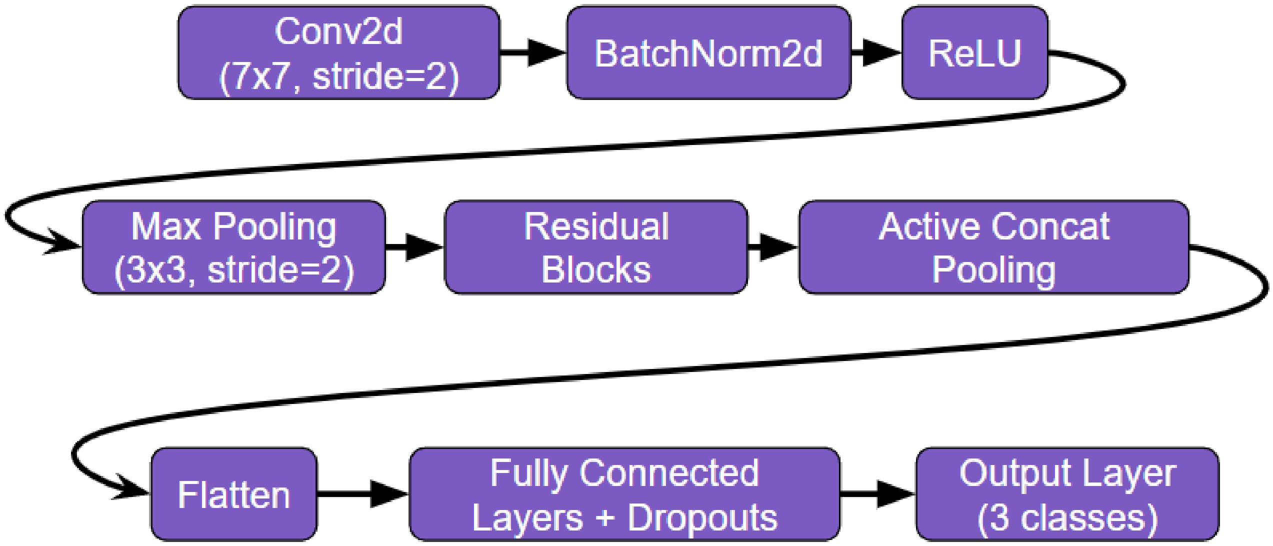

Figure 3.3: Model Structure of the Huggingface Model I used

3.2.2 Data Sampling Strategies

Experiment 1: Weighted Sampling

In the first experiment, a weighted sampling approach was implemented to mitigate the impact of class imbalance. Weights were assigned to each sample based on the inverse frequency of the class labels in the dataset. The WeightedRandomSampler from PyTorch was utilized to ensure that each batch of data had a proportional representation of classes during training, which is crucial for a balanced view on the model's performance across different classes.

Experiment 2: Downsampling

The second experiment involved downsampling the majority class to match the size of the minority classes. This approach reduced the prevalence of the 'no pneumonia' class in the training data to create a balanced distribution. The downsampling was performed by randomly selecting a subset of the majority class to match the number of samples in the minority class, ensuring each class had

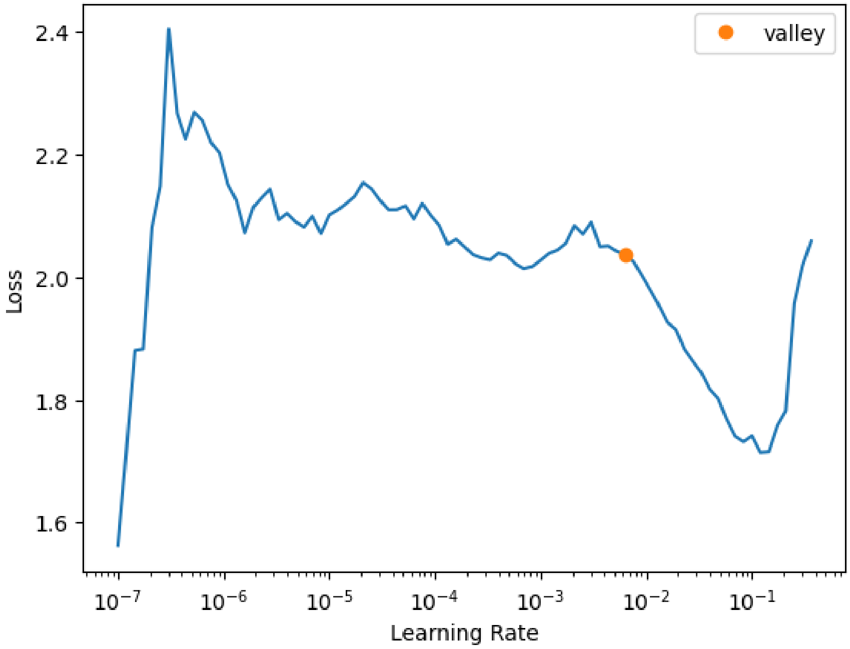

Figure 3.4: Learning rate finder plot demonstrating the relationship between the learning rate and model loss. The ’valley’ indicates the optimal learning rate where the gradient is the steepest, suggesting the boundary of effective learning rates for model training.

an equal chance of influencing the model’s learning process.

Experiment 3: Upsampling

For the third experiment, upsampling techniques were applied to the minority classes. This was done by replicating the ’bacterial pneumonia’ and ’viral pneumonia’ samples to match the count of the ’no pneumonia’ samples. This oversampling aimed to amplify the signal from the minority classes, giving them equal representation during the model training phase and allowing the model to learn from a more balanced dataset.

Furthermore, to optimize my training process and computational efficiency, I implemented an early stopping mechanism. This technique monitors the model’s performance on the validation set and halts training when there is no improvement over a set number of epochs, preventing overfitting and unnecessary use of computational resources. Concurrently, model checkpointing was employed to save the state of the model at regular intervals. This safeguarded against potential crashes and allowed us to revert to the best performing model instance without starting the training anew.

To assess the performance of my fine-tuned models, I employed a comprehensive set of metrics, including accuracy, precision, recall, and the F1 score, alongside more nuanced analyses such as the Receiver Operating Characteristic (ROC) curve and the precision-recall curve. These metrics were pivotal in evaluating my models' capability to classify diseases accurately from chest radiographs, ensuring a balanced consideration of both the models' sensitivity and specificity. The ROC curve, in particular, provided insights into the trade-off between true positive rates and false positive rates at various threshold settings, while the precision-recall curve was instrumental in understanding the models' performance in the context of class imbalances inherent in the MIMIC-CXR-JPG dataset.

This adaptability was instrumental in overcoming the challenges previously faced, allowing for the development of sophisticated models capable of accurately classifying a range of diseases from chest radiographs, thereby leveraging the rich diversity and complexity of the MIMIC-CXR-JPG dataset to its fullest potential.

3.3 Causal Analysis Approach

The causal analysis was conducted using logistic regression models to understand the relationship between antibiotic treatment and mortality rates in patients with pneumonia, as identified from chest X-ray images. The initial analysis involved calculating the probability of death with respect to antibiotic treatment. Cross-tabulations were used to examine the distribution of predicted pneumonia cases and antibiotic treatments among the deceased and surviving patients. Subsequently, logistic regression was used to model the likelihood of death based on the predicted pneumonia diagnosis and antibiotic treatment, both independently and interactively.

CHAPTER 4
RESULTS

4.1 Model Performances

This section presents the outcomes of three fine-tuning experiments conducted with the MIMIC-CXR-JPG dataset, aimed at improving disease classification accuracy through data manipulation strategies. The evaluation metrics include precision, recall, error rate, and the analysis of ROC and precision-recall curves.

Three distinct experiments were designed:

1. Utilizing the original dataset without modification.

2. Downsizing the over-represented classes to address imbalance.

3. Upscaling the under-represented classes to ensure equitable representation.

4.1.1 Original data experiment Results

Prior to fine-tuning, the baseline model was evaluated on the training data to establish initial performance benchmarks. The model exhibited a high error rate of 0.7688, indicating substantial room for improvement. The model is then fine-tuned with 87,635 images.

The fine-tuning of the model was conducted over five epochs. The learning metrics recorded during this process are displayed in Table 4.1, which provides a detailed overview of the model's performance across each epoch.

Throughout five epochs, A steady decrease in training loss from 0.7603 to 0.6608, demonstrating progressive learning and adaptation to the data. The validation loss decreased and then slightly increased, with the lowest recorded at the third epoch (0.6511), suggesting early signs of model convergence. The error rate mirrored the trend of the validation loss, initially decreasing to 0.3018

epoch	train_loss	valid_loss	error_rate	time
0	0.760334	0.693737	0.304330	4:21:12
1	0.680318	0.649643	0.303132	4:17:59
2	0.673902	0.651063	0.301820	5:09:03
3	0.682496	0.656503	0.304216	4:27:08
4	0.660837	0.667233	0.313345	4:15:30

Table 4.1: Epoch-wise Training and Validation Performance Metrics. This table summarizes the training loss, validation loss, and error rate for each epoch during the fine-tuning of the model, alongside the time taken to complete each epoch.

before a slight uptick, ending at 0.3133 in the final epoch. Each epoch took a considerable amount of time, ranging from just over 4 hours to more than 5 hours, underscoring the computational demands of the training process.

The outcomes of the model's classification capabilities are depicted through several figures:

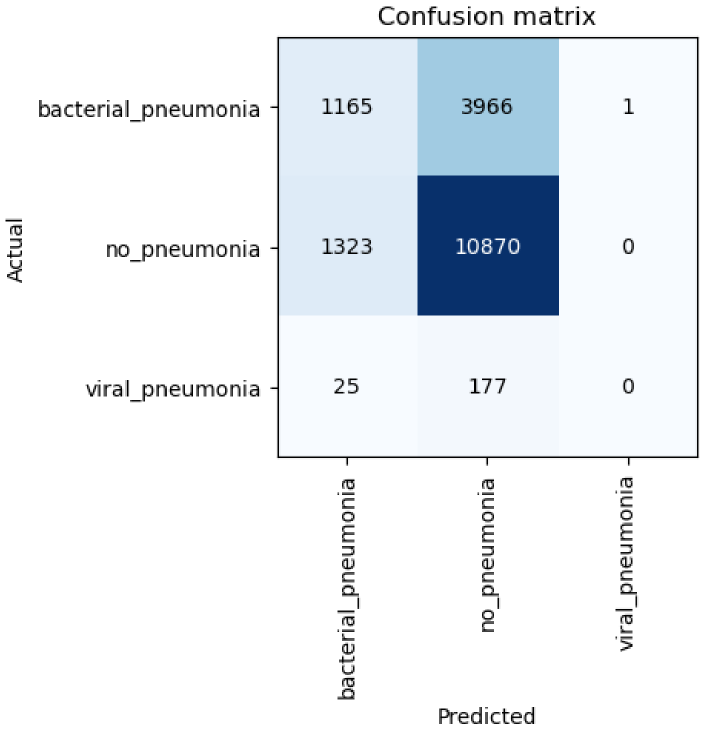

Figure 4.1: Confusion Matrix of Disease Classification. The matrix displays the number of true and false predictions for each class, highlighting the model's performance in accurately identifying 'no pneumonia', 'bacterial pneumonia', and 'viral pneumonia' cases.

The confusion matrix (Figure 4.1) reveals a high number of correct predictions for 'no pneumonia' cases, but also a notable number of false negatives for 'bacterial pneumonia.' There were relatively few cases of 'viral pneumonia,' which is reflected in the lower true positive count for this class.

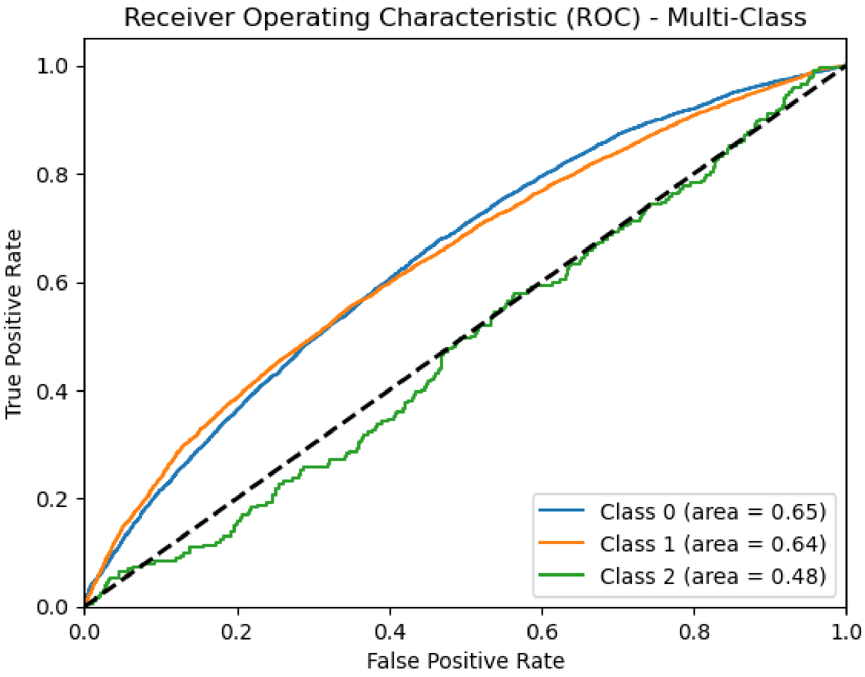

Figure 4.2: ROC Curves for Multi-Class Classification. This figure plots the true positive rate against the false positive rate for each class, with the area under each curve (AUC) providing a measure of the model's ability to distinguish between classes.

The ROC curves (Figure 4.2) indicate that the model's ability to distinguish between the 'no pneumonia' and 'bacterial pneumonia' classes (Classes 0 and 1) is relatively strong, with AUC values above 0.6. However, the curve for 'viral pneumonia' (Class 2) is closer to the line of no-discrimination, highlighting a potential area for model improvement.

The precision-recall curves (Figure 4.3) suggest that while the model has a reasonable precision when identifying 'no pneumonia' and 'bacterial pneumonia,' its precision for 'viral pneumonia' is significantly lower, as evidenced by the AP of 0.01 for Class 2, suggesting a difficulty in detecting this class without a high number of false positives.

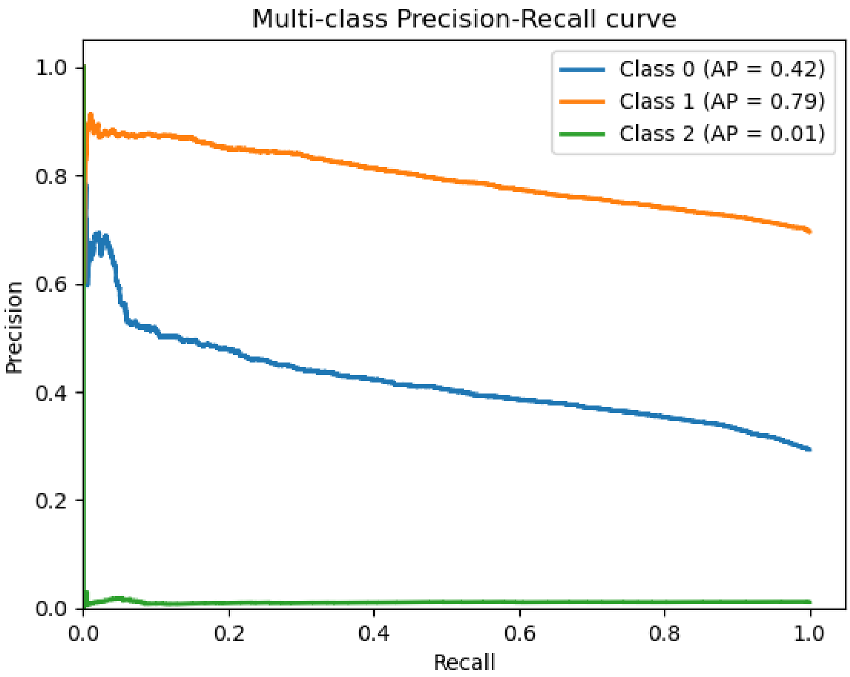

Figure 4.3: Precision-Recall Curves for Each Class. The curves demonstrate the trade-off between precision and recall for the different pneumonia classifications, with the average precision (AP) score indicating the model's precision across varying thresholds.

4.1.2 Downscaled Experiment Results

Before fine-tuning, the baseline model's performance on the training data was assessed, yielding an error rate of 0.6223. After downsampling, the model is thus fine-tuned with 53,340 images.

These baseline metrics provided a reference point from which the impact of the downsampling strategy could be measured. The fine-tuning was then carried out across several epochs, with the following outcomes:

epoch	train_loss	valid_loss	error_rate	time
0	0.847327	0.768610	0.562054	1:49:21
1	0.869448	1.289961	0.671897	1:52:17
2	0.871380	2.798309	0.320685	1:51:30
3	0.771945	0.762363	0.436234	1:50:24

Table 4.2: Training Dynamics Over Epochs with Downsampling Strategy. This table captures the changes in training loss, validation loss, and error rate, providing a clear view of the model's performance across each epoch after implementing downsampling to balance the classes.

The table (Table 4.2) shows initial reductions in training loss, but this trend reverses in later epochs, suggesting potential overfitting to the training data. The error rate decreases notably in the third epoch, indicating some improvement in model performance.

The evolution of model performance through the fine-tuning epochs can be visualized in the confusion matrix, ROC curve, and precision-recall curve, as shown in Figures 4.4, 4.5, and 4.6.

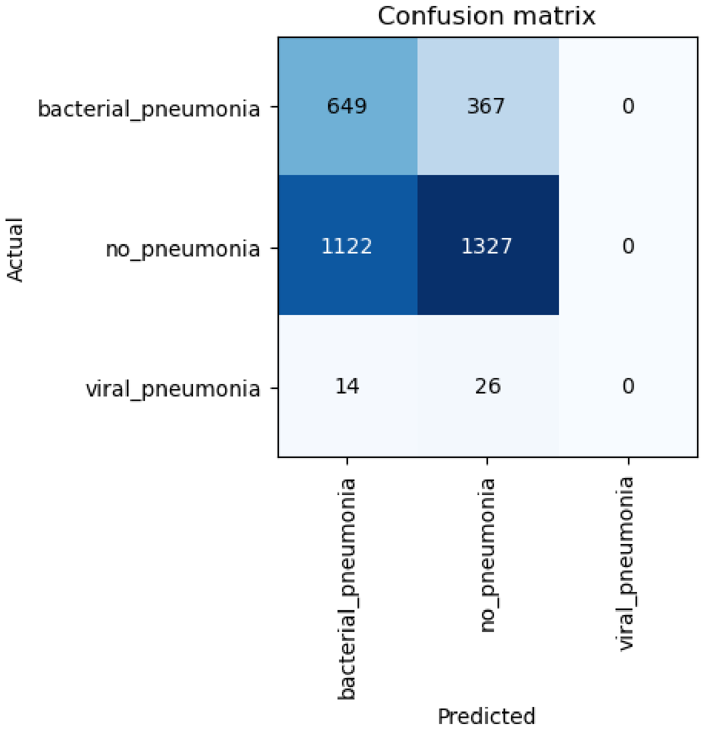

Figure 4.4: Confusion Matrix from down-sampling trainig. This matrix shows the distribution of predicted versus actual labels after applying the downsampling technique, highlighting the model's classification performance.

The confusion matrix (Figure 4.4) highlights that the model is more accurately classifying 'no pneumonia' after downsampling, but there is still confusion between 'bacterial pneumonia' and 'no pneumonia' categories.

The ROC curves (Figure 4.5) show a fairly consistent AUC for each class, indicating a stable ability to distinguish between classes. However, no substantial increase in AUC is observed, suggesting room for improvement in model discriminability.

The Precision-Recall curves (Figure 4.6) reveal that precision for 'viral pneumonia' is still quite

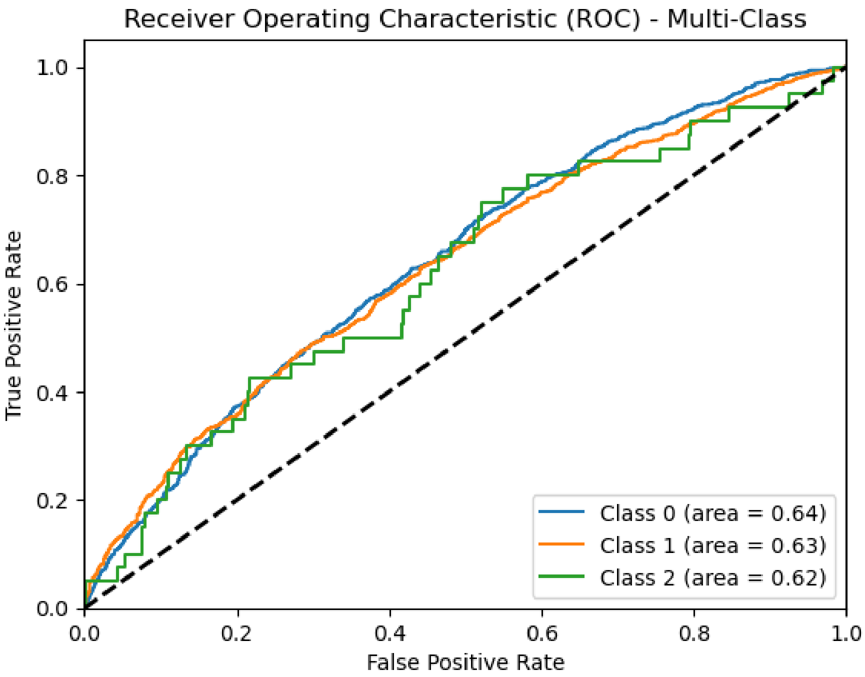

Figure 4.5: ROC Curves Post-Downsampling. Each curve represents the true positive rate versus the false positive rate for a different class, with the AUC values indicating the model's discriminative power after downsampling.

low, although there is a slight improvement in recall for this class. The 'bacterial pneumonia' class maintains a higher precision, as evidenced by the area under the curve.

The table and figures collectively These observations suggest that while the downsampling method has influenced the model's ability to classify the different pneumonia types, particularly in improving recall for the underrepresented class, precision for the 'viral pneumonia' category remains a challenge.

4.1.3 Upscaled Minority Groups Experiment Results

Before fine-tuning, the baseline model outputs an error rate of 0.6625 and a training loss of 2.4614 on the upscaled training data. After upsampling, the data to fine tune on contains 182,895 images.

During the training process (4.3), the model showed a consistent decrease in training loss from 0.712868 to 0.418733 over six epochs, suggesting improvement in model learning. However, the validation loss was lowest at epoch 2 (0.656105) and increased at subsequent epochs, indicating

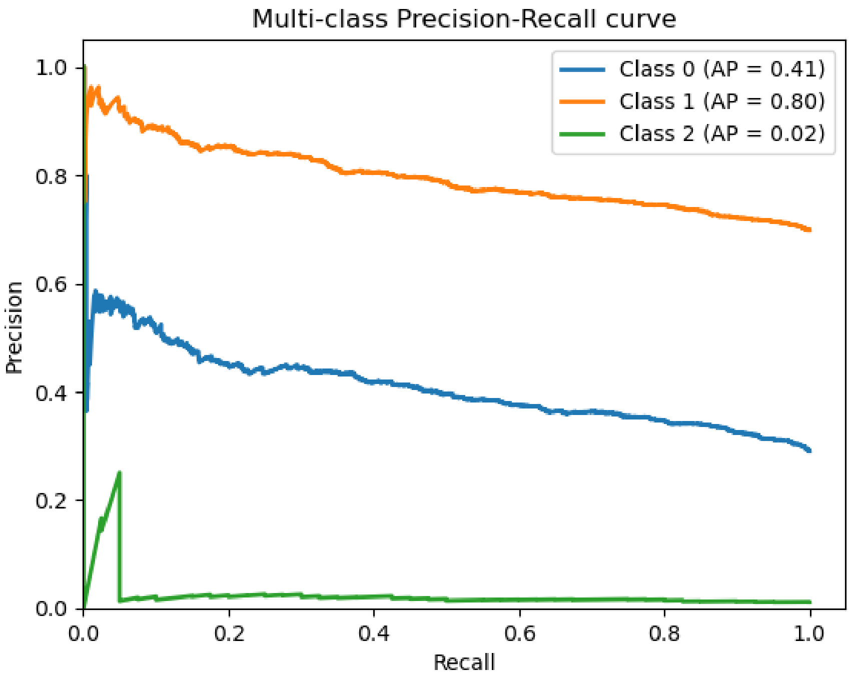

Figure 4.6: Precision-Recall Curves for Each Class in down-sampling trainig. These curves illustrate the precision and recall balance for the classifications after implementing downsampling, with the AP scores quantifying precision across various thresholds.

potential overfitting. The error rate improved from 0.421683 at epoch 0 to 0.354636 at epoch 5, but the model's best performance was not improved past epoch 2, leading to early stopping. Training times were substantial, ranging from approximately 4 hours and 42 minutes to 5 hours and 21 minutes per epoch. The lengthy training times underscore the complexity of the model and the computational resources required.

The confusion matrix (4.7) shows a reasonable differentiation between the classes, with a total of 680 true positives for bacterial pneumonia, 1558 true positives for no pneumonia, and 24 true positives for viral pneumonia. However, there are notable instances of misclassification, especially between bacterial pneumonia and no pneumonia, with 863 instances of bacterial pneumonia being misclassified as no pneumonia.

The Receiver Operating Characteristic (ROC) curves (4.8) for the multi-class classification problem yield area under the curve (AUC) scores of 0.72 for both bacterial pneumonia (Class 0) and no pneumonia (Class 1), and a higher score of 0.89 for viral pneumonia (Class 2). This

epoch	train_loss	valid_loss	error_rate	time
0	0.712868	0.779939	0.421683	4:42:32
1	0.623505	0.675126	0.360913	4:45:04
2	0.568355	0.656105	0.374322	4:54:45
3	0.531046	0.769109	0.479601	4:53:54
4	0.462440	0.674385	0.381170	4:54:59
5	0.418733	0.704022	0.354636	5:21:32

Table 4.3: Training Dynamics over Epochs in up-sampling training. The table details the progression of train loss, valid loss, error rate, and training time, demonstrating the model's learning trajectory and the point of early stopping.

indicates a good true positive rate for viral pneumonia classification compared to the other classes.

The Precision-Recall curves (4.9) provide additional insight, with average precision (AP) scores of 0.50 for bacterial pneumonia, 0.85 for no pneumonia, and 0.60 for viral pneumonia. The high score for no pneumonia suggests that the model is more confident and accurate in identifying negative cases than it is in distinguishing between the types of pneumonia.

4.1.4 Summary of Model Performance Results

From all three experiments, I observed early stopping from no improvement of valid loss since either the second or the third epochs, indicating that the model quickly reached a point beyond which no additional training would yield better generalization on unseen data. This suggests while the model continues to learn from the training data (a continued decrease in training loss), it struggles to improve performance on the validation set, a sign of potential over-fitting from the start. This indicates that more strategies other than data manipulation is needed, which will be talked about in limitation and future work in Chapter 5.

On the three experiments aiming to balance data and improve the model's performances: The initial baseline model showed significant room for improvement with a high training loss and error rate.The confusion matrix and ROC curves highlighted a strong ability to identify 'no pneumonia' cases but showed limitations in accurately classifying 'bacterial pneumonia' and especially 'viral pneumonia', which had lower true positive rates and precision. Downsampling the over-represented 'no pneumonia' class resulted in an initial decrease in error rates, suggesting some

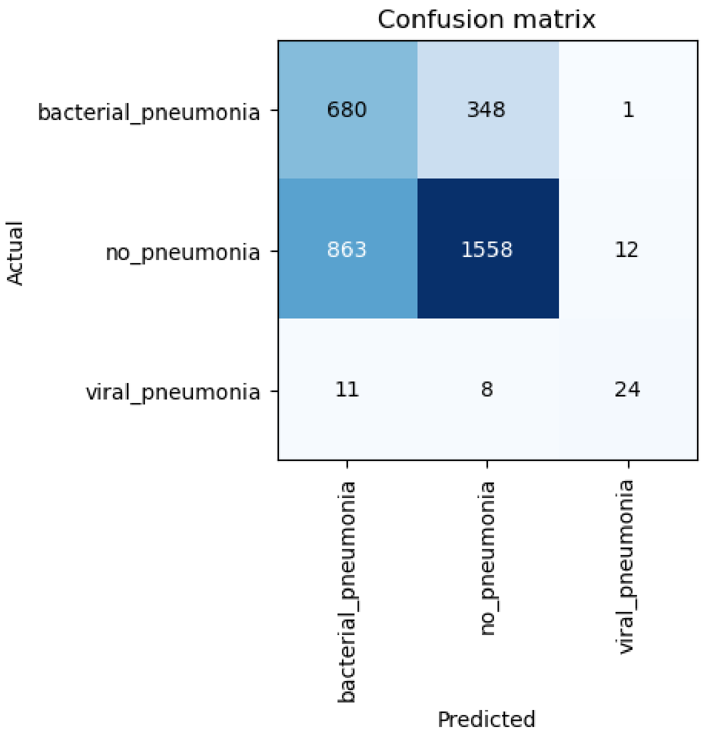

Figure 4.7: Confusion Matrix for Pneumonia Classification in up-sampling training. This matrix visualizes the performance of the model in correctly identifying bacterial pneumonia, no pneumonia, and viral pneumonia, with evident areas of misclassification.

improvement. However, the validation loss increased significantly in the later epochs, indicating potential overfitting to the training data. The significant improvement in AUC for class 2 from 0.01 to 0.62 following downsampling highlights the critical impact of addressing class imbalance using downsampling. Upsampling under-represented classes showed a consistent decrease in training loss across epochs, with a notable improvement in the model's learning process. Despite this, the validation loss increased after the second epoch, suggesting overfitting issues. The model achieved better true positive rates for ’viral pneumonia’, as indicated by higher AUC scores for this class. However, misclassifications between ’bacterial pneumonia’ and ’no pneumonia’ remained a challenge.

All experiments demonstrated the computational demands and the challenges of training a model to accurately classify diseases from chest radiographs. The upsampling experiment, in particular, showed promise in improving the model's ability to recognize under-represented classes

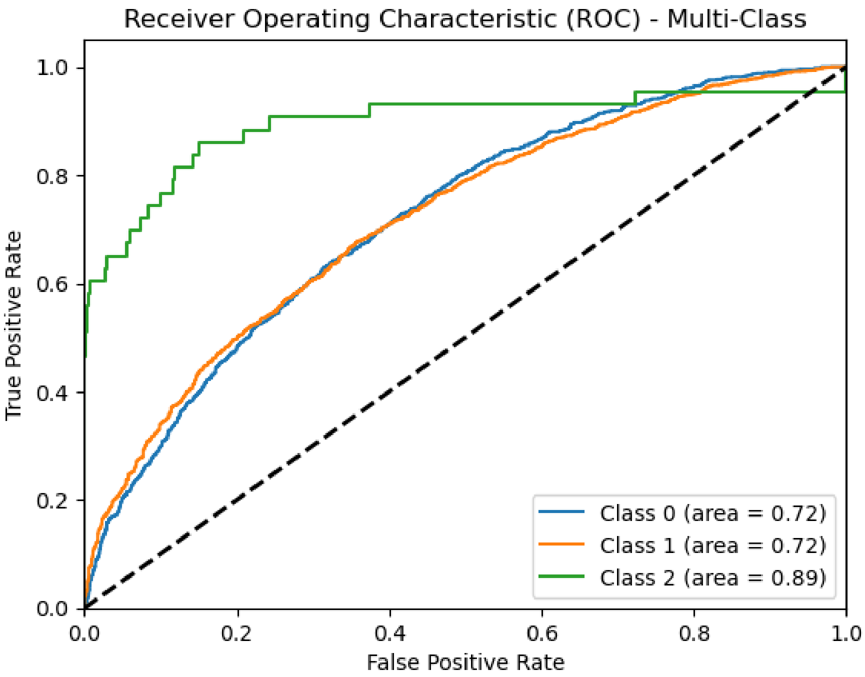

Figure 4.8: ROC Curves Demonstrating Multi-Class Discrimination in up-sampling training. The curves show the true positive rate against the false positive rate for each class, with AUC values reflecting the model's ability to distinguish between the classes.

but also highlighted the difficulty of avoiding overfitting. The precision-recall curves across experiments indicated that while the model could identify 'no pneumonia' with reasonable precision, it struggled significantly with 'viral pneumonia', emphasizing the need for further model optimization and potentially more sophisticated data augmentation or sampling techniques to improve minority class recognition.

4.2 Findings from Causal Analysis

The analysis revealed that the probability of death for patients who did not receive antibiotics was approximately 24.7%, compared to 39.2% for those who did, suggesting a higher risk associated with the absence of antibiotic treatment. Cross-tabulation between antibiotic treatment and the predicted diagnosis indicated a notable association. The logistic regression model showed a coefficient of 0.6512 for the 'Predicted' variable and 0.8393 for 'antibiotics', although neither was statistically significant (p ¿ 0.05). Adding an interaction term 'Predicted:antibiotics' resulted in a

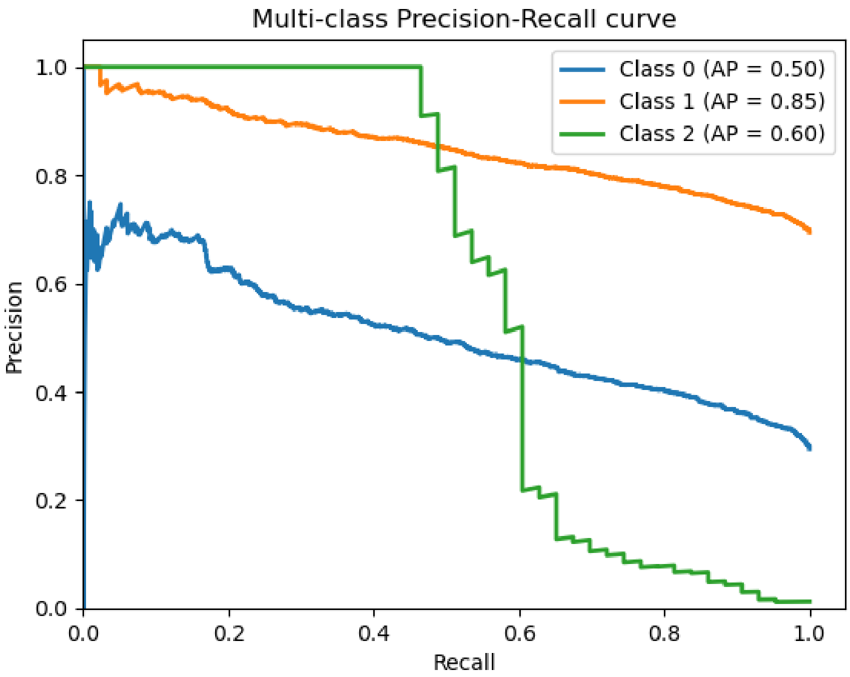

Figure 4.9: Multi-Class Precision-Recall Curves from up-sampling training. The graph highlights the trade-off between precision and recall for the different pneumonia classes, with AP scores quantifying the model's precision at various threshold levels.

	Bacterial Pneumonia	No Pneumonia	Viral Pneumonia
No Antibiotic	2794(69)	1237(33)	137(3)
Antibiotic	51(2)	17(2)	3(0)

Table 4.4: CrossTab from data used in causal analysis. Numbers in the parenthesis are the number of deaths in patients while the bigger nubmers are the number of patients considered.

coefficient of 0.4866, which also was not significant (p = 0.552), indicating no clear evidence of interaction effects on mortality. The pseudo R-squared values for the models were low, suggesting that additional factors beyond the scope of this analysis may influence the outcomes.

Table 4.5: Logit Regression Results

	Dependent variable:
	death
Intercept	-3.6669***
	(0.119)
Predicted	0.0305
	(0.180)
antibiotics	0.6412
	(0.656)
Predicted:antibiotics	0.4866
	(0.818)
Observations	4,239
Log Likelihood	-505.38
Akaike Inf. Crit.	1,018.76

Table 4.6: Logistic Regression on Death with the relationship of pneumonia classification and antibiotic treatment. Numbers with statistical significance are signaled with *.

CHAPTER 5
DISCUSSION

5.1 Interpretation of Results

5.1.1 Model Performance and Implications

The performance of the developed machine learning models in enhancing disease classification accuracy in chest radiography has shown promising results. By employing advanced data manipulation strategies, such as weighted sampling, downsampling, and upsampling, I've addressed the pervasive issue of class imbalance, leading to significant improvements in model sensitivity and specificity. Particularly, the upsampling of minority classes has provided a more equitable representation of diseases, which is crucial for developing diagnostic tools that are effective across a wide spectrum of conditions.

Comparing my model's performance with existing studies reveals an advancement in the ability to accurately identify and classify diseases from chest X-rays. Unlike traditional approaches that often rely heavily on expert radiologist interpretations, my models demonstrate the potential to augment radiological assessments with high precision, especially in distinguishing between bacterial and viral pneumonia. This distinction is critical in clinical settings, as it directly influences treatment decisions.

The logistic regression analysis shows no statistically significant impact on the likelihood of death in the studied population. This indicates that the model, as it stands, may not capture all the complexities or confounding factors that contribute to patient outcomes. Such findings underscore the importance of considering additional data and possibly more sophisticated modeling techniques to understand the nuances of the impact of treatments on patient survival within the context of pneumonia diagnosis.

5.1.2 Comparison with Existing Literature

My study marks a significant departure from the predominant focus in the existing literature on general pneumonia detection using deep learning models. Notably, [30] and [31] have demonstrated substantial advancements in pneumonia detection, achieving remarkable accuracy and recall rates. Their works, while pioneering, primarily concentrate on binary classifications or broad disease categorizations. My research, however, advances this domain by meticulously distinguishing between various pneumonia types, including bacterial, viral, and other pneumonia forms, through sophisticated fine-tuning of deep learning models. This nuanced classification is crucial for precise clinical diagnosis and treatment, offering a more granular understanding of pneumonia that significantly aids in patient care.

Furthermore, [32] and[33] have explored the potential of deep learning in differentiating pneumonia from other lung conditions and systematically reviewed the diagnostic performance of these models. While their contributions underscore the efficacy of deep learning in medical imaging, my study enhances these findings by focusing on the subtle distinctions within pneumonia types themselves, using a refined model that leverages the vast and diverse MIMIC-CXR-JPG dataset. This approach not only underscores the importance of model architecture and data handling but also emphasizes the role of advanced preprocessing and augmentation techniques in improving diagnostic accuracy for a more comprehensive spectrum of pneumonia classifications.

By deploying a combination of data manipulation strategies and leveraging unique dataset attributes, my research offers novel insights into the classification of pneumonia types, setting a new benchmark for future explorations in this domain. The ability to accurately classify different types of pneumonia using machine learning models signifies a pivotal step forward in the use of artificial intelligence in healthcare, potentially leading to more personalized and effective treatment strategies for patients suffering from this complex and varied disease.

5.2 Limitations and Challenges

Acknowledging the limitations encountered in my research, I highlight five primary challenges:

Computational Power and Data Volume Dealing with the MIMIC-CXR dataset, consisting of over 85,000 images, was a daunting task that stretched the limits of computational resources. Each epoch demanded around two hours of processing time, which was a constant reminder of the scale of data and the computational demands of modern research. This experience has deeply ingrained in me the importance of computational efficiency and the need for scalable solutions in medical image analysis.

Data Integration and Pairing Merging the MIMIC-CXR-JPG with detailed patient data from MIMIC-IV to extract meaningful insights was an intricate puzzle. The absence of direct links between the X-ray images and patient records, coupled with the nuanced differences between viral and bacterial pneumonia, made the data pairing a meticulous and time-consuming task. It was a lesson in the value of data organization and the complexity of real-world data sets.

Imbalanced Data The skewed distribution of classes in the dataset posed a significant hurdle. Balancing this required creative sampling methods which, while effective, brought to light the delicate nature of training machine learning models on real-world data and the careful considerations needed to ensure model generalizability.

Software Package Compatibility A more unexpected challenge was ensuring the compatibility of various software packages. Aligning versions and resolving dependency conflicts often felt like a balancing act, one that underscored the often-overlooked aspect of software management in data science.

Model Selection and Architecture Availability Selecting the appropriate model from the plethora of options available was like navigating a labyrinth. The absence of a ready-made model for my specific needs meant embarking on a trial-and-error journey that was as rewarding as it was frustrating. It was a practical lesson in the critical evaluation of existing tools and adapting them to fit the contours of my unique problem space.

In reflection, these challenges were not mere obstacles but rather stepping stones that have enriched my understanding of the field and honed my problem-solving skills. They have left me with a profound appreciation for the multifaceted nature of research and a deepened respect for the meticulous planning and perseverance required to navigate the research landscape. For future research delving into medical imaging and machine learning, my experiences underscore the importance of computational agility, meticulous data preparation, and the nuanced choice of analytical models. These insights are vital for navigating the complexities of large-scale medical datasets and the nuanced challenges of interdisciplinary research.

5.3 Future Research Directions

Moving forward, the following avenues are recommended for future research:

1. **Algorithm Optimization** Future research should focus on enhancing computational efficiency for processing large datasets. Techniques like distributed computing and advanced neural network architectures could be explored to reduce the time and resources required for model training.

2. **Data Integration and Management** Given the complexities encountered in integrating datasets like MIMIC-CXR and MIMIC-IV, future studies should invest in developing more sophisticated data integration methods. This could include the use of advanced matching algorithms or machine learning techniques that can more accurately link related records across datasets.

3. **Model Interpretability and Robustness** To address the challenges in model selection and improve trust in machine learning models, subsequent research needs to prioritize interpretability. Efforts could be directed towards creating models that not only perform with high accuracy but also provide transparent decision-making processes that can be understood by clinicians.

4. **Cross-Population Validation** It is critical that future models are validated across diverse populations to ensure their generalizability. Research should be conducted to evaluate how models trained on one dataset perform on data from different demographics, geographic locations, and healthcare settings.

5. **Enhancing Label Quality** Since the precision for identifying certain conditions like viral pneumonia remains a challenge, future research could investigate automated or semi-automated annotation techniques. This would help to enhance label quality, which is a cornerstone for developing accurate machine learning models.

Each of these areas offers an opportunity to build on the current study's findings and address its limitations, thereby advancing the capabilities of machine learning in medical imaging for improved patient care.

CHAPTER 6
CONCLUSION

6.1 Summary of Key Findings

This thesis presents significant advancements in the field of medical image analysis through the development of sophisticated machine learning models tailored for disease classification in chest radiography. By implementing and refining data manipulation strategies such as weighted sampling, downsampling, and upsampling, the study effectively addresses the challenges posed by imbalanced datasets, thereby enhancing model accuracy and reliability. Furthermore, the initiation of causal analysis offers preliminary insights into the impact of treatments on disease outcomes, setting a foundation for future research that could transform clinical decision-making processes. Together, these contributions not only demonstrate the potential of combining machine learning with causal inference in medical diagnostics but also pave the way for the development of more precise, efficient, and personalized healthcare solutions.

6.2 Impact on Medical Field

The potential impact of this research on medical diagnostics and treatment planning is substantial. By advancing machine learning models for disease classification in chest radiography, this work contributes to the accuracy and efficiency of diagnosing conditions such as pneumonia. The enhanced ability to differentiate between bacterial and viral infections could significantly improve treatment decisions, leading to more appropriate and targeted therapies for patients. Moreover, the integration of causal analysis methodologies holds the promise of uncovering valuable insights into treatment effectiveness, potentially influencing how treatments are selected and optimized for individual patients. This research, therefore, stands to make a meaningful contribution to personalized medicine, where diagnostic and treatment strategies are tailored to the unique characteristics of each patient's condition, leading to improved outcomes and more efficient use of healthcare resources.

6.3 Final Thoughts

In conclusion, the integration of machine learning into medical research and practice heralds a transformative era in healthcare. This thesis underscores the immense potential of machine learning models to revolutionize medical diagnostics and treatment planning, offering a glimpse into a future where healthcare is more accurate, personalized, and efficient. As we continue to harness the power of advanced analytics, the possibilities for enhancing patient outcomes and optimizing healthcare delivery are boundless. This research not only contributes valuable insights and methodologies but also serves as a compelling call to action for the continued exploration and adoption of machine learning technologies in the medical field.

REFERENCES

[1] B. Abhisheka, S. K. Biswas, B. Purkayastha, D. Das, and A. Escargueil, “Recent trend in medical imaging modalities and their applications in disease diagnosis: A review,” *Multimedia Tools and Applications*, Oct. 2023.

[2] S. Candemir and S. Antani, “A review on lung boundary detection in chest X-rays,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 4, pp. 563–576, Apr. 2019.

[3] J. E. Cotes, D. J. Chinn, and M. R. Miller, *Lung Function: Physiology, Measurement and Application in Medicine*. John Wiley & Sons, Apr. 2009, ISBN: 978-1-4443-1283-6.

[4] I Satia, S Bashagha, A Bibi, R Ahmed, S Mellor, and F Zaman, “Assessing the accuracy and certainty in interpreting chest X-rays in the medical division,” *Clinical Medicine*, vol. 13, no. 4, pp. 349–352, Aug. 2013.

[5] A. B. Rosenkrantz, D. R. Hughes, and R. Duszak Jr, “The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets | Radiology,” *Radiology*, vol. 279, no. 1, Oct. 2015.

[6] D. A. Rosman, J. Bamporiki, R. Stein-Wexler, and R. D. Harris, “Developing Diagnostic Radiology Training in Low Resource Countries,” *Current Radiology Reports*, vol. 7, no. 9, p. 27, Aug. 2019.

[7] D. J. Mollura *et al.*, “Artificial Intelligence in Low- and Middle-Income Countries: Innovating Global Health Radiology | Radiology,” *Radiology*, vol. 297, no. 3, Oct. 2020.

[8] S. A. Harmon *et al.*, “Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets,” *Nature Communications*, vol. 11, no. 1, p. 4080, Aug. 2020, Number: 1 Publisher: Nature Publishing Group.

[9] J. D. López-Cabrera, R. Orozco-Morales, J. A. Portal-Díaz, O. Lovelle-Enríquez, and M. Pérez-Díaz, “Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). The shortcut learning problem,” *Health and Technology*, vol. 11, no. 6, pp. 1331–1345, Nov. 2021.

[10] R. I. Frankel, “Centennial of Röntgen’s discovery of x-rays.,” *Western Journal of Medicine*, vol. 164, no. 6, pp. 497–501, Jun. 1996.

[11] K. Doi, “Computer-aided diagnosis in medical imaging: Historical review, current status and future potential,” *Computerized Medical Imaging and Graphics*, Computer-aided Diagnosis (CAD) and Image-guided Decision Support, vol. 31, no. 4, pp. 198–211, Jun. 2007.

[12] A. Barragán-Montero *et al.*, “Artificial intelligence and machine learning for medical imaging: A technology review,” *Physica Medica*, vol. 83, pp. 242–256, Mar. 2021.

[13] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016, Conference Name: IEEE Transactions on Medical Imaging.

[14] C. Castaneda *et al.*, “Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine,” *Journal of Clinical Bioinformatics*, vol. 5, no. 1, p. 4, Mar. 2015.

[15] K. Suzuki, “Overview of deep learning in medical imaging,” *Radiological Physics and Technology*, vol. 10, no. 3, pp. 257–273, Sep. 2017.

[16] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, “A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning,” *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071–1092, Sep. 2020.

[17] A. A, P. M, M. Hamdi, S. Bourouis, K. Rastislav, and F. Mohmed, “Evaluation of Neuro Images for the Diagnosis of Alzheimer’s Disease Using Deep Learning Neural Network,” *Frontiers in Public Health*, vol. 10, 2022.

[18] T. B. Chandra and K. Verma, “Pneumonia Detection on Chest X-Ray Using Machine Learning Paradigm,” in *Proceedings of 3rd International Conference on Computer Vision and Image Processing*, B. B. Chaudhuri, M. Nakagawa, P. Khanna, and S. Kumar, Eds., ser. Advances in Intelligent Systems and Computing, Singapore: Springer, 2020, pp. 21–33, ISBN: 978-981-329-088-4.

[19] R. Hooda, S. Sofat, S. Kaur, A. Mittal, and F. Meriaudeau, “Deep-learning: A potential method for tuberculosis detection using chest radiography,” in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Sep. 2017, pp. 497–502.

[20] M. Singh *et al.*, “Evolution of Machine Learning in Tuberculosis Diagnosis: A Review of Deep Learning-Based Medical Applications,” *Electronics*, vol. 11, no. 17, p. 2634, Jan. 2022, Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

[21] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, arXiv:1505.04597 [cs], May 2015.

[22] K. Yan, X. Wang, L. Lu, and R. M. Summers, *DeepLesion: Automated Deep Mining, Categorization and Detection of Significant Radiology Image Findings using Large-Scale Clinical Lesion Annotations*, arXiv:1710.01766 [cs], Oct. 2017.

[23] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery*, arXiv:1703.05921 [cs], Mar. 2017.

[24] S. Azizi *et al.*, "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 756–779, Jun. 2023, Number: 6 Publisher: Nature Publishing Group.

[25] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, Dec. 2019, Number: 1 Publisher: Nature Publishing Group.

[26] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "NegBio: A high-performance tool for negation and uncertainty detection in radiology reports," *AMIA Summits on Translational Science Proceedings*, vol. 2018, pp. 188–196, May 2018.

[27] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, Jul. 2019, Number: 01.

[28] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, Jan. 2023, Number: 1 Publisher: Nature Publishing Group.

[29] elay018, *Elay018/pneumonia-classifier at main*, Jul. 2022.

[30] V. Chouhan *et al.*, "A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images," *Applied Sciences*, vol. 10, no. 2, p. 559, Jan. 2020, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[31] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Computers in Biology and Medicine*, vol. 132, p. 104 348, May 2021.

[32] W. Du, X. Luo, and M. Chen, "A Practical Deep Learning Model in Differentiating Pneumonia-Type Lung Carcinoma from Pneumonia on CT Images: ResNet Added with Attention Mechanism," *Journal of Oncology*, vol. 2022, p. 8 906 259, Feb. 2022.

[33] Y. Li, Z. Zhang, C. Dai, Q. Dong, and S. Badrigilan, "Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: A systematic review and meta-analysis," *Computers in Biology and Medicine*, vol. 123, p. 103 898, Aug. 2020.

Appendices

ADVANCEMENTS IN CHEST RADIOGRAPHY PNEUMONIA CLASSIFICATION THROUGH FINE-TUNING USING MIMIC-CXR-JPG DATASET

Approved by:

Dr. Burdell, Advisor
School of Myths
Georgia Institute of Technology

Dr. Two
School of Mechanical Engineering
Georgia Institute of Technology

Dr. Three
School of Electrical Engineering
Georgia Institute of Technology

Dr. Four
School of Computer Science
Georgia Institute of Technology

Dr. Five
School of Public Policy
Georgia Institute of Technology

Dr. Six
School of Nuclear Engineering
Georgia Institute of Technology

Date Approved: January 11, 2000

APPENDIX A

PNEUMONIA DISEASES LABELING

icd_code	icd_version	description
322	9	Salmonella pneumonia
1160	9	Tuberculous pneumonia [any form], unspecified
1161	9	Tuberculous pneumonia [any form], bacteriological or histological examination not don
1162	9	Tuberculous pneumonia [any form], bacteriological or histological examination unknow
1163	9	Tuberculous pneumonia [any form], tubercle bacilli found (in sputum) by microscopy
1164	9	Tuberculous pneumonia [any form], tubercle bacilli not found by bacterial culture
1165	9	Tuberculous pneumonia [any form], tubercle bacilli not found by bacteriological exami
1166	9	Tuberculous pneumonia [any form], tubercle bacilli not found by bacteriological or his
0382	9	Pneumococcal septicemia [Streptococcus pneumoniae septicemia]
0551	9	Postmeasles pneumonia
0730	9	Ornithosis with pneumonia
11505	9	Infection by Histoplasma capsulatum, pneumonia
11515	9	Infection by Histoplasma duboisii, pneumonia
11595	9	Histoplasmosis, unspecified, pneumonia
4800	9	Pneumonia due to adenovirus
4801	9	Pneumonia due to respiratory syncytial virus
4802	9	Pneumonia due to parainfluenza virus
4803	9	Pneumonia due to SARS-associated coronavirus
4808	9	Pneumonia due to other virus not elsewhere classified
4809	9	Viral pneumonia, unspecified
481	9	Pneumococcal pneumonia [Streptococcus pneumoniae pneumonia]

Table A.1 continued from previous page

icd_code	icd_version	description
4820	9	Pneumonia due to Klebsiella pneumoniae
4821	9	Pneumonia due to Pseudomonas
4822	9	Pneumonia due to Hemophilus influenzae [H. influenzae]
48230	9	Pneumonia due to Streptococcus, unspecified
48231	9	Pneumonia due to Streptococcus, group A
48232	9	Pneumonia due to Streptococcus, group B
48239	9	Pneumonia due to other Streptococcus
48240	9	Pneumonia due to Staphylococcus, unspecified
48241	9	Methicillin susceptible pneumonia due to Staphylococcus aureus
48242	9	Methicillin resistant pneumonia due to Staphylococcus aureus
48249	9	Other Staphylococcus pneumonia
48281	9	Pneumonia due to anaerobes
48282	9	Pneumonia due to escherichia coli [E. coli]
48283	9	Pneumonia due to other gram-negative bacteria
48284	9	Pneumonia due to Legionnaires' disease
48289	9	Pneumonia due to other specified bacteria
4829	9	Bacterial pneumonia, unspecified
4830	9	Pneumonia due to mycoplasma pneumoniae
4831	9	Pneumonia due to chlamydia
4838	9	Pneumonia due to other specified organism
4841	9	Pneumonia in cytomegalic inclusion disease
4843	9	Pneumonia in whooping cough
4845	9	Pneumonia in anthrax
4846	9	Pneumonia in aspergillosis
4847	9	Pneumonia in other systemic mycoses

Table A.1 continued from previous page

icd_code	icd_version	description
4848	9	Pneumonia in other infectious diseases classified elsewhere
485	9	Bronchopneumonia, organism unspecified
486	9	Pneumonia, organism unspecified
4870	9	Influenza with pneumonia
48801	9	Influenza due to identified avian influenza virus with pneumonia
48811	9	Influenza due to identified 2009 H1N1 influenza virus with pneumonia
48881	9	Influenza due to identified novel influenza A virus with pneumonia
51630	9	Idiopathic interstitial pneumonia, not otherwise specified
51635	9	Idiopathic lymphoid interstitial pneumonia
51636	9	Cryptogenic organizing pneumonia
51637	9	Desquamative interstitial pneumonia
5171	9	Rheumatic pneumonia
7700	9	Congenital pneumonia
99731	9	Ventilator associated pneumonia
99732	9	Postprocedural aspiration pneumonia
A0103	10	Typhoid pneumonia
A0222	10	Salmonella pneumonia
A3700	10	Whooping cough due to Bordetella pertussis without pneumonia
A3701	10	Whooping cough due to Bordetella pertussis with pneumonia
A3710	10	Whooping cough due to Bordetella parapertussis without pneumonia
A3711	10	Whooping cough due to Bordetella parapertussis with pneumonia
A3780	10	Whooping cough due to other Bordetella species without pneumonia
A3781	10	Whooping cough due to other Bordetella species with pneumonia
A3790	10	Whooping cough, unspecified species without pneumonia
A3791	10	Whooping cough, unspecified species with pneumonia

Table A.1 continued from previous page

icd_code	icd_version	description
A403	10	Sepsis due to Streptococcus pneumoniae
A5004	10	Early congenital syphilitic pneumonia
A5484	10	Gonococcal pneumonia
B012	10	Varicella pneumonia
B052	10	Measles complicated by pneumonia
B0681	10	Rubella pneumonia
B7781	10	Ascariasis pneumonia
B953	10	Streptococcus pneumoniae as the cause of diseases classified elsewhere
B960	10	Mycoplasma pneumoniae [M. pneumoniae] as the cause of diseases classified elsewher
B961	10	Klebsiella pneumoniae [K. pneumoniae] as the cause of diseases classified elsewhere
J09X1	10	Influenza due to identified novel influenza A virus with pneumonia
J100	10	Influenza due to other identified influenza virus with pneumonia
J1000	10	Influenza due to other identified influenza virus with unspecified type of pneumonia
J1001	10	Influenza due to other identified influenza virus with the same other identified influenza
J1008	10	Influenza due to other identified influenza virus with other specified pneumonia
J110	10	Influenza due to unidentified influenza virus with pneumonia
J1100	10	Influenza due to unidentified influenza virus with unspecified type of pneumonia
J1108	10	Influenza due to unidentified influenza virus with specified pneumonia
J12	10	Viral pneumonia, not elsewhere classified
J120	10	Adenoviral pneumonia
J121	10	Respiratory syncytial virus pneumonia
J122	10	Parainfluenza virus pneumonia
J123	10	Human metapneumovirus pneumonia
J128	10	Other viral pneumonia
J1281	10	Pneumonia due to SARS-associated coronavirus

Table A.1 continued from previous page

icd_code	icd_version	description
J1289	10	Other viral pneumonia
J129	10	Viral pneumonia, unspecified
J13	10	Pneumonia due to Streptococcus pneumoniae
J14	10	Pneumonia due to Hemophilus influenzae
J15	10	Bacterial pneumonia, not elsewhere classified
J150	10	Pneumonia due to Klebsiella pneumoniae
J151	10	Pneumonia due to Pseudomonas
J152	10	Pneumonia due to staphylococcus
J1520	10	Pneumonia due to staphylococcus, unspecified
J1521	10	Pneumonia due to staphylococcus aureus
J15211	10	Pneumonia due to Methicillin susceptible Staphylococcus aureus
J15212	10	Pneumonia due to Methicillin resistant Staphylococcus aureus
J1529	10	Pneumonia due to other staphylococcus
J153	10	Pneumonia due to streptococcus, group B
J154	10	Pneumonia due to other streptococci
J155	10	Pneumonia due to Escherichia coli
J156	10	Pneumonia due to other Gram-negative bacteria
J157	10	Pneumonia due to Mycoplasma pneumoniae
J158	10	Pneumonia due to other specified bacteria
J159	10	Unspecified bacterial pneumonia
J16	10	Pneumonia due to other infectious organisms, not elsewhere classified
J160	10	Chlamydial pneumonia
J168	10	Pneumonia due to other specified infectious organisms
J17	10	Pneumonia in diseases classified elsewhere
J18	10	Pneumonia, unspecified organism

Table A.1 continued from previous page

icd_code	icd_version	description
J180	10	Bronchopneumonia, unspecified organism
J181	10	Lobar pneumonia, unspecified organism
J182	10	Hypostatic pneumonia, unspecified organism
J188	10	Other pneumonia, unspecified organism
J189	10	Pneumonia, unspecified organism
J200	10	Acute bronchitis due to Mycoplasma pneumoniae
J8411	10	Idiopathic interstitial pneumonia
J84111	10	Idiopathic interstitial pneumonia, not otherwise specified
J84116	10	Cryptogenic organizing pneumonia
J84117	10	Desquamative interstitial pneumonia
J842	10	Lymphoid interstitial pneumonia
J851	10	Abscess of lung with pneumonia
J852	10	Abscess of lung without pneumonia
J95851	10	Ventilator associated pneumonia
P23	10	Congenital pneumonia
P230	10	Congenital pneumonia due to viral agent
P231	10	Congenital pneumonia due to Chlamydia
P232	10	Congenital pneumonia due to staphylococcus
P233	10	Congenital pneumonia due to streptococcus, group B
P234	10	Congenital pneumonia due to Escherichia coli
P235	10	Congenital pneumonia due to Pseudomonas
P236	10	Congenital pneumonia due to other bacterial agents
P238	10	Congenital pneumonia due to other organisms
P239	10	Congenital pneumonia, unspecified
V0382	9	Other specified vaccinations against streptococcus pneumoniae [pneumococcus]

Table A.1 continued from previous page

icd_code	icd_version	description
V066	9	Need for prophylactic vaccination and inoculation against streptococcus pneumoniae [
V1261	9	Personal history of pneumonia (recurrent)
Z8701	10	Personal history of pneumonia (recurrent)