Spring
2024

AI-Enhanced Radiology: Enhancing Lung Disease Diagnosis

LEVERAGING DEEP LEARNING TO OPTIMIZE DIAGNOSTIC PRECISION AND EFFICIENCY

SEAS 6402 – DATA ANALYTICS CAPSTONE

JONGKYU LEE – G33845168
NISHANTH NANDAKUMAR – G27281814
SAI KIRAN REDDY VELLANKI – G28981115
HARSHITA MADHUKAR BHARADWAJ – G38876302

May 7, 2024

Abstract

The application of artificial intelligence (AI) in medical diagnostics presents a promising avenue for enhancing the accuracy and efficiency of interpreting chest X-rays, which are pivotal in diagnosing complex lung diseases. Despite their critical role, these imaging processes are challenged by high variability and the subtleties of image interpretation, which can significantly impact patient outcomes. This study explores the utility of deep learning models, specifically VGG19 and ResNet50, applied to the NIH chest X-ray dataset to improve diagnostic processes. Our objectives were to assess the diagnostic accuracy these models can achieve, understand their impact on the efficiency of diagnostic processes, and identify challenges in integrating AI into clinical settings. The results indicated that while the ResNet50 model achieved the highest diagnostic accuracy of 73.42%, the performance of VGG19 models was less consistent, highlighting the complexity of the task. Ensemble methods improved robustness but did not surpass the top-performing individual models. Although AI models enhance certain aspects of diagnostic efficiency, they are not yet capable of replacing human radiologists, serving instead as a supplementary tool to aid them. This study underlines the necessity for further research in optimizing these technologies for clinical integration, focusing on improving model accuracy and developing effective strategies for their implementation in real-world settings.

Table of Contents

Table of Figures

Table of Tables

Introduction

Background

Despite significant advancements in medical imaging technology, the diagnosis of complex lung conditions from chest X-rays presents substantial challenges. These imaging procedures are critical due to their accessibility and cost-effectiveness, yet their interpretation is fraught with difficulties. Radiologists face the demanding task of discerning subtle variations in images under significant time pressures and high caseloads, where misdiagnosis or delays can severely impact patient outcomes.

Chest X-rays exhibit high variability in diagnostic interpretation, influenced by the complexity of images and the varying levels of observer experience. This variability, combined with the limitations of the human eye and the subjective nature of image assessment, underscores the need for enhanced diagnostic processes. The integration of AI, particularly deep learning models like VGG19 and ResNet50, offers promising solutions to augment accuracy and efficiency in detecting lung diseases, aiming to revolutionize medical diagnostics by reducing human error and inconsistency.

Objective

The objective of this project is to explore how effectively deep learning models, specifically VGG19 and ResNet50, can enhance the diagnostic precision and efficiency of identifying lung diseases in chest X-ray imaging. This involves leveraging recent technological advancements where AI and image recognition technologies have started to play a transformative role in improving the efficiency and accuracy of disease diagnoses.

Research Question

This study aims to answer several critical questions about the implementation of AI in medical diagnostics: What level of diagnostic accuracy can deep learning models like VGG19 and ResNet50 achieve when applied to the NIH chest X-ray dataset? Additionally, how do these models impact the efficiency of diagnostic processes in terms of time and resource utilization compared to traditional methods? Lastly, what are the key challenges associated with integrating these AI models into real-world clinical settings, and what potential solutions can be proposed to facilitate their seamless adoption?

Structure of the Report

This report is structured to provide a comprehensive analysis of the research topic. Following this introduction, the methodology section will detail the data acquisition process and the statistical techniques employed. Subsequent sections will present the data analysis, machine learning models, results, and a discussion of the findings in relation to the research objectives. Finally, the report will conclude with a summary of the findings and their implications.

Methodology

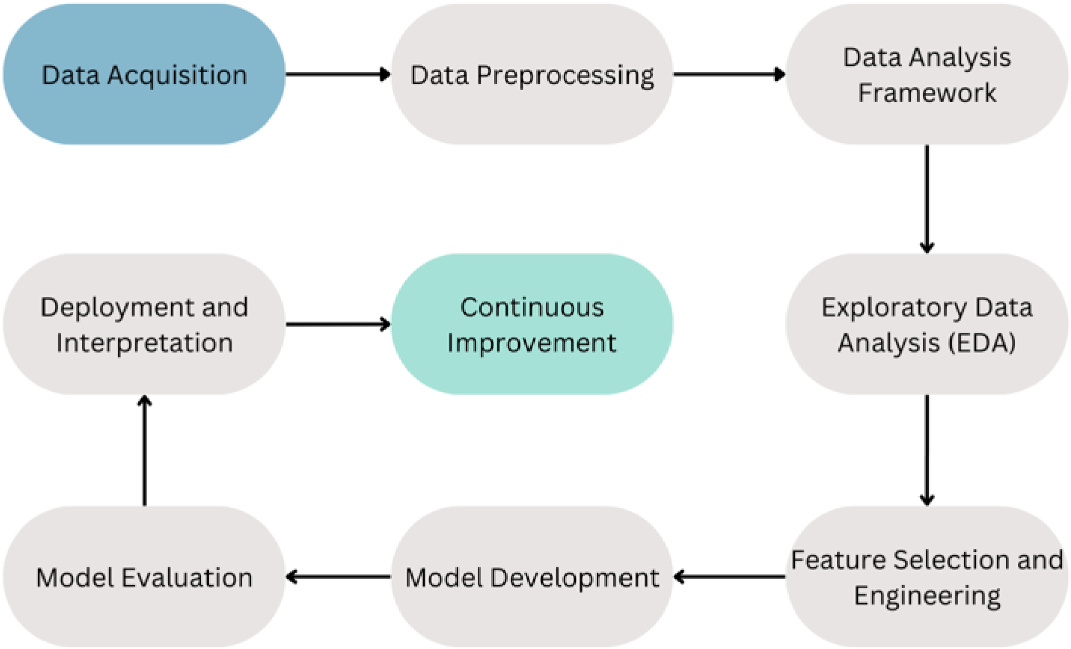


Figure 1: Data Science Process Flowchart

Figure 1 presents the comprehensive data science methodology framework utilized in this project. The process is initiated with data acquisition, where raw data is gathered, followed by data preprocessing to cleanse and prepare the data for analysis. Upon these foundations, a robust data analysis framework is constructed, enabling thorough exploratory data analysis (EDA) and feature selection, which feed into the model development phase. Post-development, the model is rigorously evaluated to assess its performance. The final stages involve deploying the model and interpreting its outputs to inform decision-making. Continuous Improvement is positioned as a future endeavor, highlighting areas for potential enhancements and advancements beyond the current scope of the project. This reflects the project's forward-looking approach, anticipating the integration of new insights and technological advancements to refine and evolve the model over time.

Data Acquisition

Chest X-Ray Data

The dataset for this study is derived from the substantial repository of chest X-rays available at the National Institutes of Health Clinical Center (National Institutes of Health (NIH), 2017). The Chest X-ray dataset, comprises 112,120 frontal-view X-ray images from 30,805 unique patients. It includes a diverse array of text-mined labels for fourteen common thoracic pathologies such as Atelectasis, Consolidation, Infiltration, and Pneumothorax, among others. Each image in the dataset can have multiple labels corresponding to the various diseases identified.

The selection of this dataset was motivated by its scale and diversity, which are critical for developing a robust deep learning model. It is approximately 27 times larger than the previous largest publicly available dataset, (OpenI), which contained only 4,143 frontal view images. This significant increase in data volume provides a more comprehensive representation of real patient population distributions and the challenges faced in realistic clinical settings.

Labels for the dataset were extracted using advanced natural language processing techniques from the associated radiological reports, ensuring high relevance and accuracy of the disease annotations. This dataset not only reflects the complexity of clinical diagnosis for chest X-rays, which is often considered more challenging than chest CT imaging but also addresses the limitations noted in earlier studies where the scarcity of data severely restricted the performance capabilities of deep neural networks.

By leveraging this enriched dataset, the project aims to enhance the capabilities of computer-aided detection and diagnosis (CAD) systems in real-world medical scenarios, tackling the variability and diagnostic challenges inherent in chest X-ray analysis.

Data Preprocessing and Augmentation

Once the raw data was acquired, the next critical phase was to preprocess the data to make it suitable for analysis.

Key preprocessing steps included:

Matching Images with Disease Labels: The dataset comprised raw images identified by numbers, accompanied by an Excel file detailing the corresponding diseases. The first step was to align each image with its correct disease labels to establish a cohesive dataset.

Data Cleaning: This step involved removing any corrupted or irrelevant images, standardizing the image formats, and handling any missing data to ensure a clean and consistent dataset for the subsequent analysis.

Resizing and Normalization: All images were standardized to a uniform size (224x224 pixels) and normalized to align pixel values within a consistent range. This step ensured that the images were compatible with deep learning models requiring fixed input sizes.

Data Augmentation

Augmentation Pipeline: The augmentation pipeline was designed to enrich the dataset through transformations like horizontal flipping, brightness and contrast adjustments, shifting, scaling, rotating, applying random gamma, adding Gaussian noise, and padding. These transformations increased the diversity of training samples, enhancing the model's generalization abilities.

Imbalance in Class Distribution: The dataset initially suffered from class imbalance, with some classes having fewer samples. Data augmentation generated synthetic samples to balance the class representation, aiming for approximately 10,000 images per class.

Limited Size of Original Dataset: The original dataset for some classes were insufficient for robust model training. Data augmentation expanded the dataset by generating additional samples from the existing data for these classes.

Sampling Images to Reduce Class Imbalance: For classes with fewer than 10,000 images, synthetic samples were generated to augment the data as mentioned above, while for classes with more than 10,000 images, random sampling was used to reduce the count. This approach ensured balanced representation across classes.

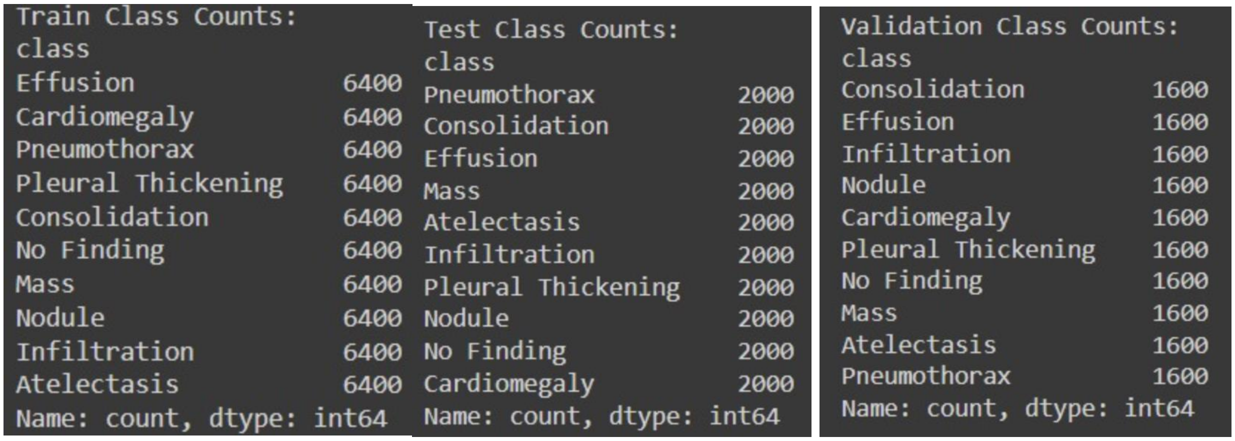
Checking Image Counts After Sampling: The script then verified the count of images in each directory, ensuring balanced class representation suitable for training.

Improving Model Generalization: The augmentation pipeline included various transformations to help the model learn invariant features, enhancing generalization and reducing overfitting.

Enhancing Model Robustness to Variations in Imaging Conditions: Augmentation introduced variations in the training data to mimic real-world scenarios, improving robustness to patient positioning, lighting conditions, and image quality.

Dataset Split

The dataset was then split as follows: 64% for training (6,400 images), 20% for testing (2,000 images), and 16% for validation (1,600 images). This balanced split ensured ample data for training while reserving adequate samples for testing and validation.



```
Train Class Counts:
class
Effusion              6400
Cardiomegaly          6400
Pneumothorax          6400
Pleural Thickening    6400
Consolidation         6400
No Finding            6400
Mass                  6400
Nodule                6400
Infiltration          6400
Atelectasis           6400
Name: count, dtype: int64
```

```
Test Class Counts:
class
Pneumothorax          2000
Consolidation         2000
Effusion              2000
Mass                  2000
Atelectasis           2000
Infiltration          2000
Pleural Thickening    2000
Nodule                2000
No Finding            2000
Cardiomegaly          2000
Name: count, dtype: int64
```

```
Validation Class Counts:
class
Consolidation         1600
Effusion              1600
Infiltration          1600
Nodule                1600
Cardiomegaly          1600
Pleural Thickening    1600
No Finding            1600
Mass                  1600
Atelectasis           1600
Pneumothorax          1600
Name: count, dtype: int64
```

Table 1: Class count for Train, Test and Validation sets

Data Analysis

Exploratory Data Analysis

The initial plot below, illustrates the distribution of pathologies by count, revealing 'Infiltration' as the most prevalent pathology, closely followed by 'Nodule' and 'Atelectasis,' with similar occurrence rates. 'Mass' and other pathologies follow suit.

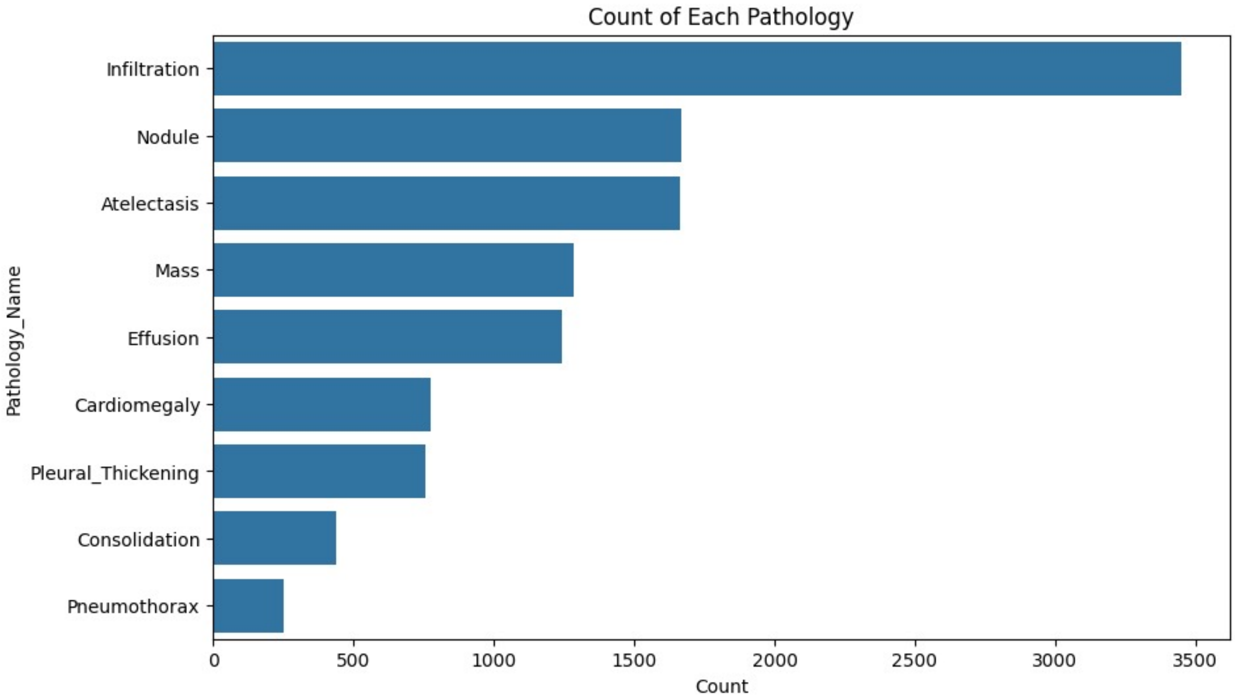


Figure 2: Count of Each Pathology

In the subsequent graph depicting pathology counts by gender, a clear trend emerges: males lead across all pathologies except for 'Cardiomegaly' and 'Pneumothorax.'

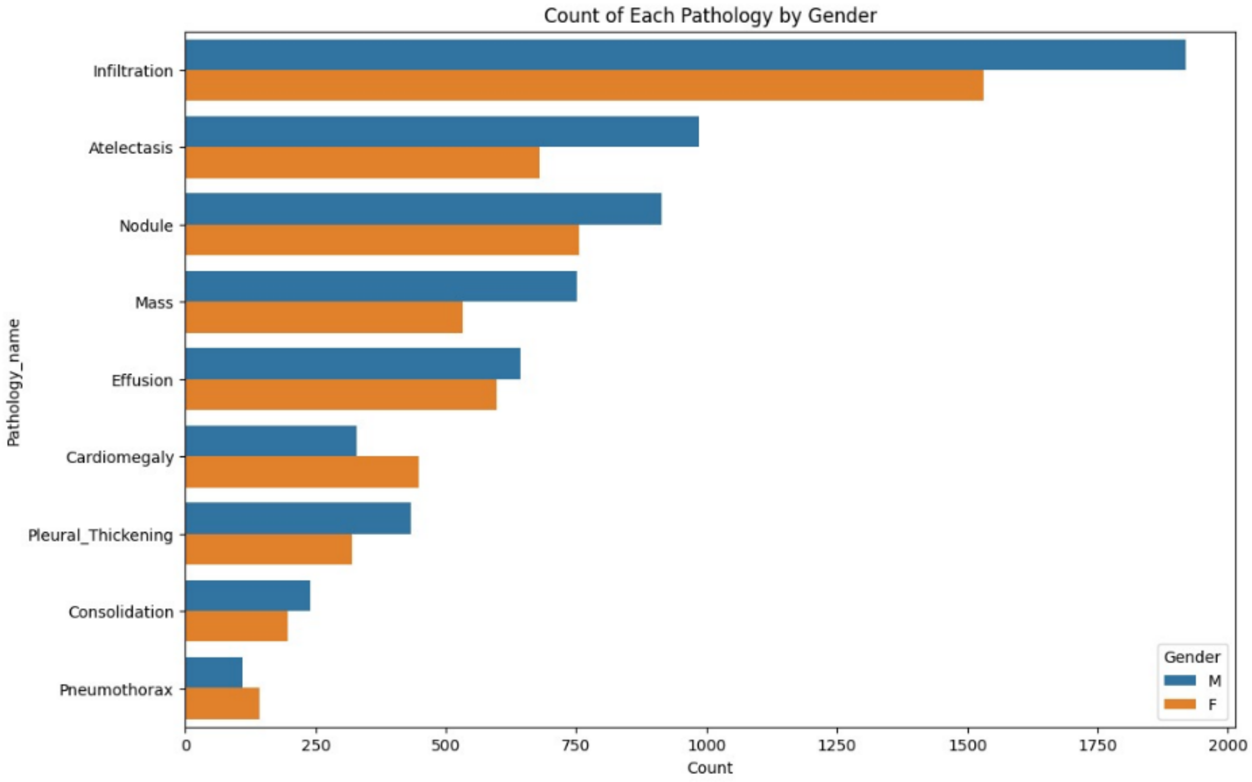


Figure 2: Count of Each Pathology by Gender

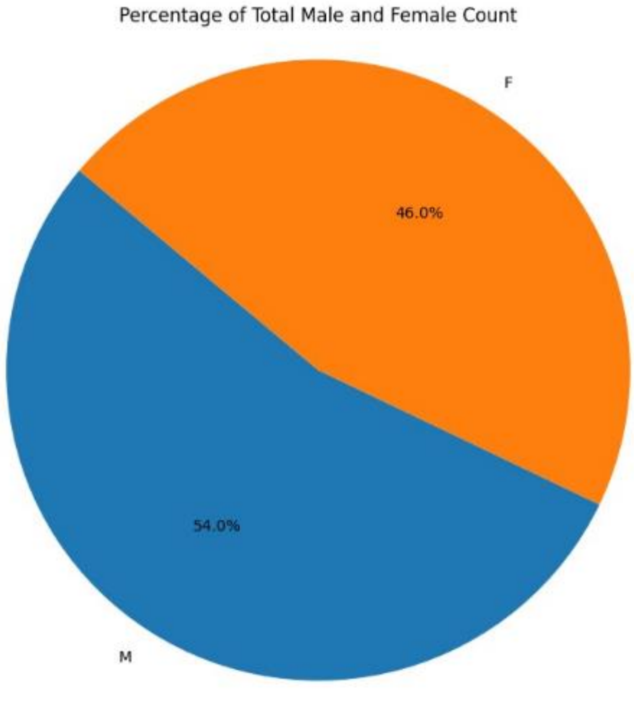


Figure 3: Percentage Count of Males vs Females

Considering the overall dataset, males represent 54% while females constitute 46% as shown in Figure 3.

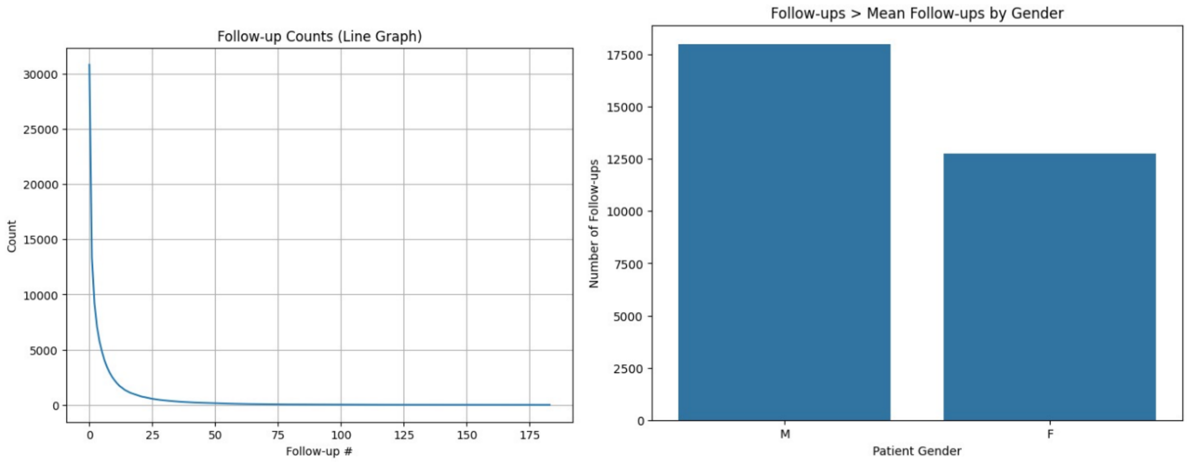


Figure 4: Follow up Count and Mean follow-ups by Gender

The plots above delves into patient follow-up counts, showcasing a notable trend; the majority of patients lack follow-up appointments. Additionally, it's discernible that males tend to have more follow-ups compared to females, as evidenced by gender distribution.

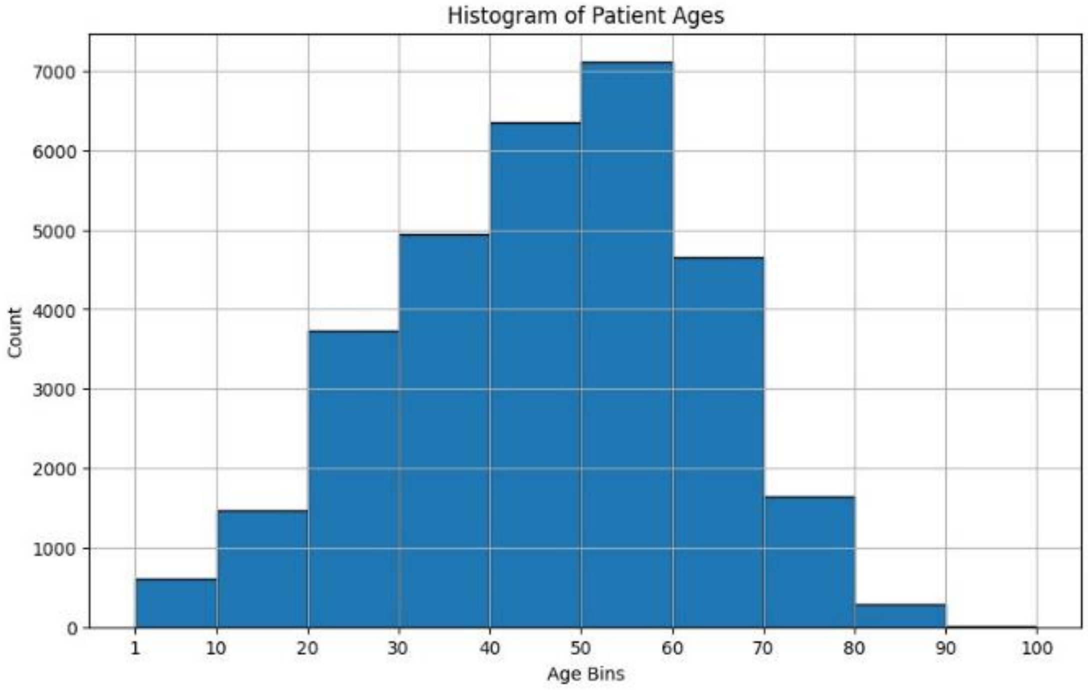


Figure 5: Histogram of Patients Age

Turning to patient age distributions, the highest concentration falls within the age range of 50 to 60 years, closely followed by 40 to 50 years, and subsequently 30 to 40 years

Machine Learning Models

Overview

In our project, we utilized pretrained models VGG19 and ResNet50 to enhance the diagnostic accuracy and efficiency of identifying lung diseases from chest X-ray images. These models, renowned for their deep learning capabilities, have been trained on extensive datasets, allowing them to effectively recognize and classify intricate image patterns. To further boost the performance and robustness of our diagnostic system, we implemented an ensemble method that combines the strengths of both VGG19 and ResNet50. This approach aimed to leverage the complementary features captured by each model, potentially leading to improved diagnostic outcomes and reduced errors compared to using each model individually.

VGG 19 Model

The VGG19 model is a convolutional neural network architecture known for its depth and simplicity. Originally developed for large-scale image recognition, VGG19 consists of 19 layers with trainable parameters, including 16 convolutional layers and 3 fully connected layers (Boesch, 2021).

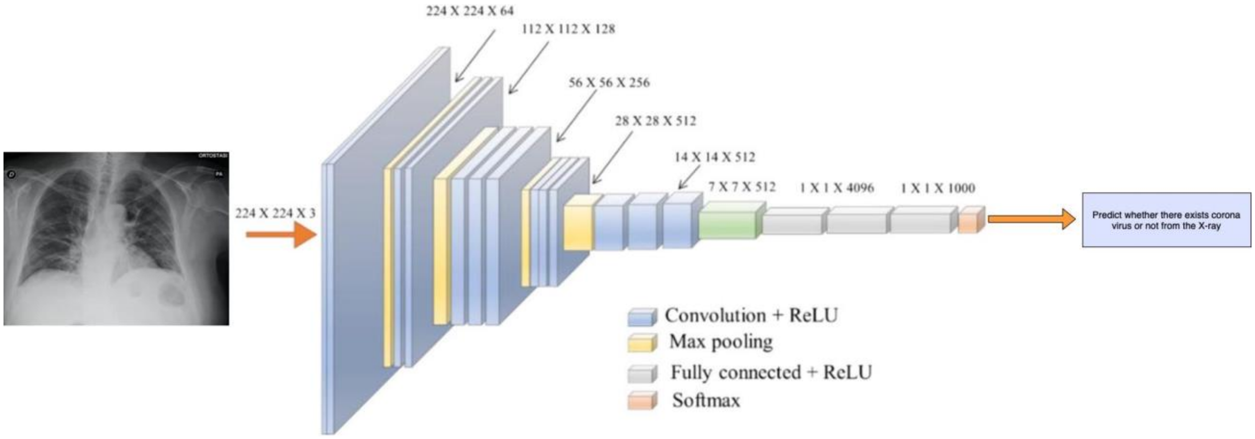


Figure 5: VGG 19 Model Architecture (Patel, 2020)

Key Features of VGG19:

- **Depth:** Comprising 19 layers, VGG19 utilizes multiple convolutional layers with small (3x3) filters, which enables the capture of complex features at various levels.
- **ReLU Activation:** Each convolutional layer is followed by a Rectified Linear Unit (ReLU) to introduce non-linearity, enhancing the network's ability to learn diverse features.
- **Pooling Layers:** Max pooling is performed after several convolutional layers to reduce spatial dimensions, thus decreasing the parameter count and computational complexity.

- **Fully Connected Layers:** Three dense layers towards the end of the network increase its decision-making capability.
- **Softmax Output:** The final layer uses a softmax function to classify input images into multiple categories, providing probabilities for each category.

Feature Selection and Engineering for VGG19 Model

The development of VGG19 Model 1 employed strategic modifications to optimize it for the task of classifying lung diseases from chest X-rays. Key to this process was the utilization of a pre-trained VGG19 model, which has been extensively trained on the ImageNet dataset. To tailor it for our specific use, the top layers of the network were removed in Model 3, allowing for the customization of the network's architecture to better suit our data's characteristics without overfitting. The model was adapted to handle input images resized to 224x224 pixels, maintaining the standard input format for deep learning models in image recognition.

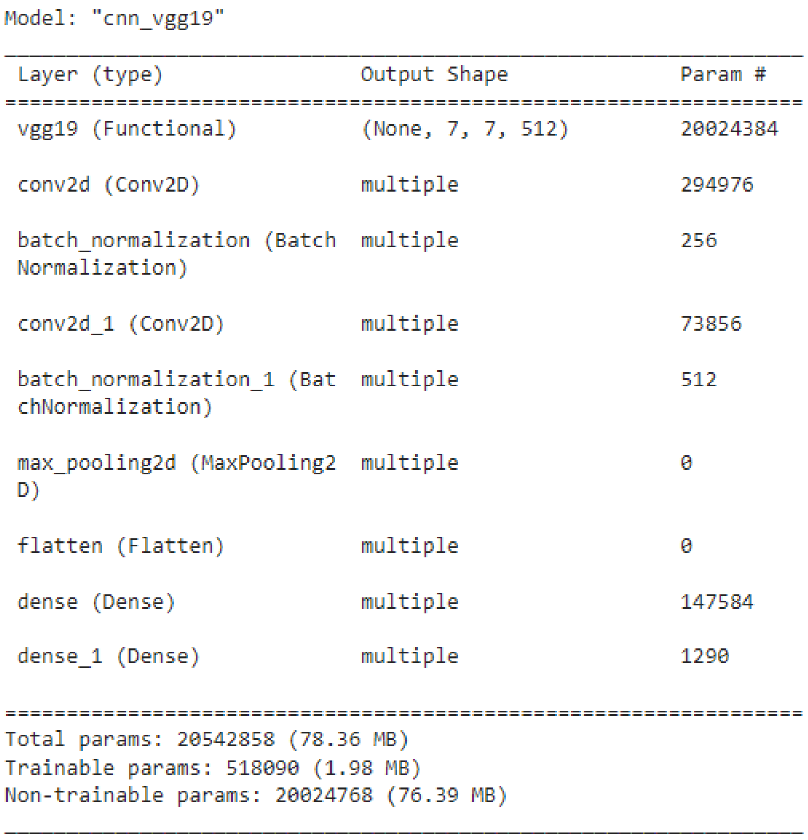
Further refinement was achieved by integrating additional convolutional layers with 64 and 128 filters, respectively. These layers were augmented with batch normalization, which stabilizes the learning process by normalizing the input layer by re-centering and re-scaling. This helps in speeding up the training process and reducing the sensitivity to network initialization. Additionally, Rectified Linear Unit (ReLU) activation functions were used to introduce non-linearity, enhancing the model’s ability to learn complex patterns in the data. The use of max pooling after the second convolutional layer effectively reduced the spatial dimensions of the feature maps, thus decreasing the number of parameters and computational complexity.

Model Development for VGG 19 – Model 1

In constructing the VGG19 Model 1, significant emphasis was placed on designing a robust classification head. The network architecture was structured to flatten the 3D feature maps into 1D feature vectors, which were then processed through dense layers. A dense layer with 128 units was utilized for intermediate feature processing, which was essential for transforming the learned features into a format suitable for classification. The culmination of the model architecture was a final dense layer equipped with a softmax activation function, used for outputting the probabilities for each class, providing a quantifiable measure of model confidence across the multiple classes.

The model was compiled with Adam optimizer, chosen for its efficiency in handling large datasets and its effectiveness in converging quickly. Adam adjusts the learning rate throughout training, which provides a more refined approach to reaching the global minimum. The loss function was set to categorical crossentropy, helping to measure the loss between the predicted values and the actual values.

For training, the model underwent 10 epochs using both training and validation data generators, enabling the effective measurement of the model's performance and generalization capabilities over time. Metrics focused primarily on accuracy, which is critical for evaluating the model's success in correctly classifying the diagnostic images.



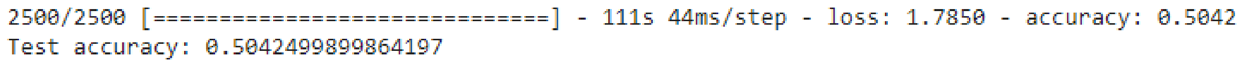
```
Model: "cnn_vgg19"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 vgg19 (Functional)          (None, 7, 7, 512)         20024384

 conv2d (Conv2D)             multiple                  294976

 batch_normalization (Batch  multiple                  256
 Normalization)

 conv2d_1 (Conv2D)           multiple                  73856

 batch_normalization_1 (Bat  multiple                  512
 chNormalization)

 max_pooling2d (MaxPooling2  multiple                  0
 D)

 flatten (Flatten)           multiple                  0

 dense (Dense)               multiple                  147584

 dense_1 (Dense)             multiple                  1290

=================================================================
Total params: 20542858 (78.36 MB)
Trainable params: 518090 (1.98 MB)
Non-trainable params: 20024768 (76.39 MB)
_________________________________________________________________
```

Figure 6: Model 1 Configuration

Model Evaluation for VGG 19 – Model 1

The performance of the VGG19 Model 1 was thoroughly assessed using a comprehensive testing protocol on the dataset. The evaluation focused on measuring both the loss and the accuracy to determine the model's effectiveness in classifying various thoracic pathologies.

During the final testing phase, the model achieved a test accuracy of 50.42% and a loss of 1.7850. These results were obtained over a total of 2,580 evaluation steps, with the process taking approximately 111 seconds, averaging 44 milliseconds per step. This level of accuracy, while indicative of the model's potential to identify correct diagnoses from X-ray images, also highlights areas for improvement. The test loss score suggests that there are substantial mismatches between the predicted outcomes and the actual labels, which could be attributed to various factors including the model's architecture, the complexity of the task, or limitations in the training dataset.



```
2500/2500 [==============================] - 111s 44ms/step - loss: 1.7850 - accuracy: 0.5042
Test accuracy: 0.5042499899864197
```

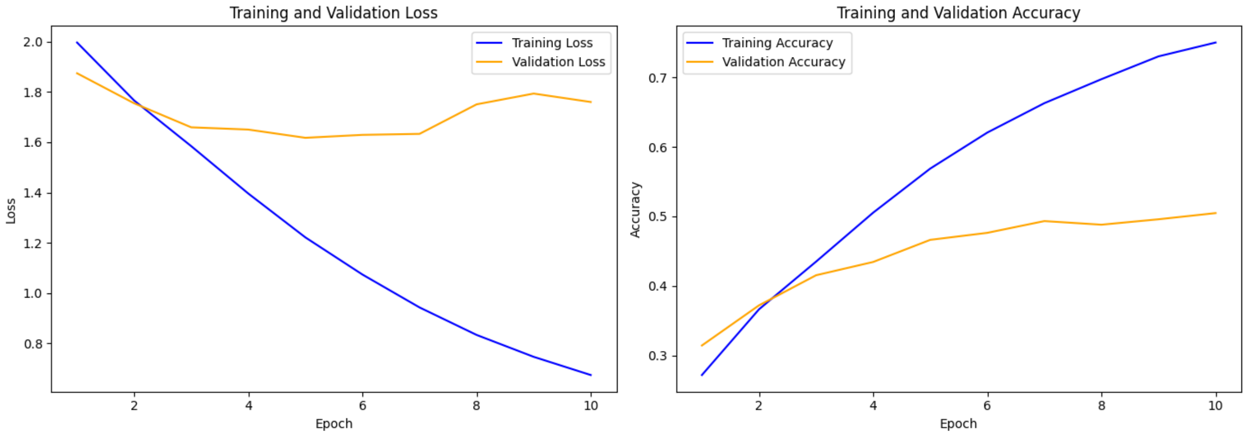


Figure 7: Training and Validation Loss and Accuracy graph – Model 1

Model Development for VGG 19 – Model 2

Building on the foundation set by VGG19 Model 1, the second iteration of the VGG19 architecture incorporates several strategic enhancements to further refine its performance for lung disease diagnosis from chest X-rays. Model 2 retains the core structure of the pre-trained VGG19 base, adapted by removing the top layers and adding custom layers suited for the specific task. This model continues to handle 224x224 RGB images, and the configuration includes additional convolutional layers and batch normalization to stabilize learning, similar to the first model.

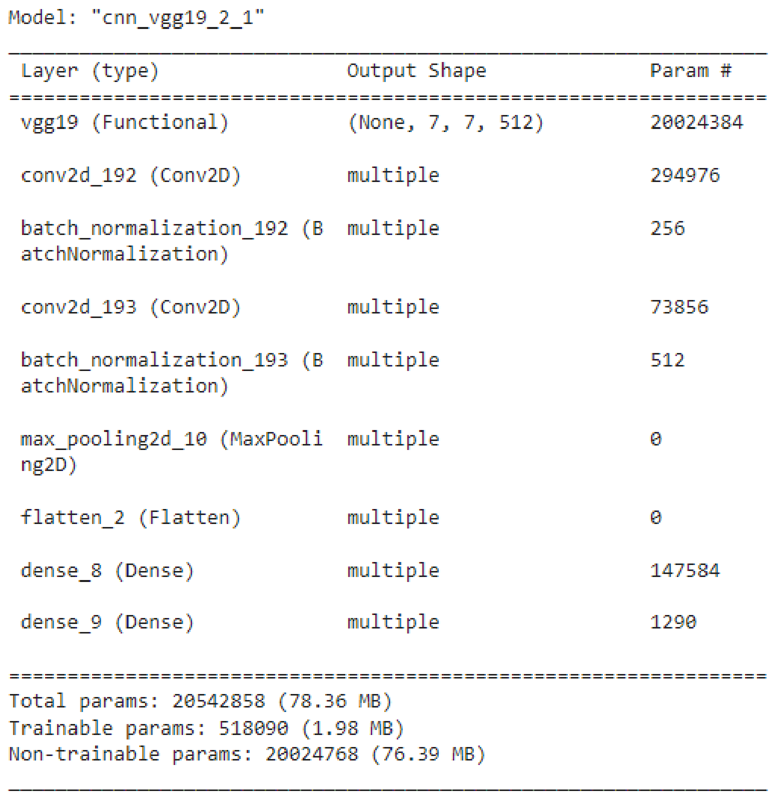
Key Features and Differences from Model 1:

Extended Training Regimen:

To enhance the model's ability to learn from a diverse set of imaging data, the training duration for Model 2 has been significantly extended to 50 epochs. This increase from the 10 epochs used in Model 1 allows for a more comprehensive learning process, taking advantage of a larger volume of data iterations to refine feature detection and classification accuracy.

Early Stopping Mechanism:

Model 2 introduces an early stopping mechanism with a patience setting of three epochs. This development means that the training will automatically cease if there is no improvement in the validation loss for three consecutive epochs. This approach is designed to provide a tighter control over convergence, preventing overfitting and ensuring that the model training stops at an optimal point for performance.

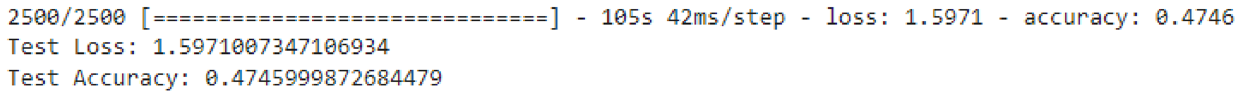


```
Model: "cnn_vgg19_2_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 vgg19 (Functional)          (None, 7, 7, 512)         20024384  
                                                                 
 conv2d_192 (Conv2D)         multiple                  294976    
                                                                 
 batch_normalization_192 (B  multiple                  256       
 atchNormalization)                                              
                                                                 
 conv2d_193 (Conv2D)         multiple                  73856     
                                                                 
 batch_normalization_193 (B  multiple                  512       
 atchNormalization)                                              
                                                                 
 max_pooling2d_10 (MaxPooli  multiple                  0         
 ng2D)                                                           
                                                                 
 flatten_2 (Flatten)         multiple                  0         
                                                                 
 dense_8 (Dense)             multiple                  147584    
                                                                 
 dense_9 (Dense)             multiple                  1290      
                                                                 
=================================================================
Total params: 20542858 (78.36 MB)
Trainable params: 518090 (1.98 MB)
Non-trainable params: 20024768 (76.39 MB)
_________________________________________________________________
```

Figure 8: Model 2 Configuration

Model Evaluation for VGG 19 – Model 2

VGG19 Model 2 achieved a test loss of 1.5971 and an accuracy of 47.46% over 2,500 evaluation steps. Despite the extended training up to 50 epochs and the incorporation of early stopping to optimize convergence, the model exhibited a lower accuracy compared to Model 1, which achieved a higher accuracy of 50.42%. This reduction underscores potential issues in model tuning or overfitting despite the more sophisticated training regimen. Further optimizations in model architecture and training strategies are necessary to improve its diagnostic performance and surpass the effectiveness of Model 1.



```
2500/2500 [==============================] - 105s 42ms/step - loss: 1.5971 - accuracy: 0.4746
Test Loss: 1.5971007347106934
Test Accuracy: 0.4745999872684479
```

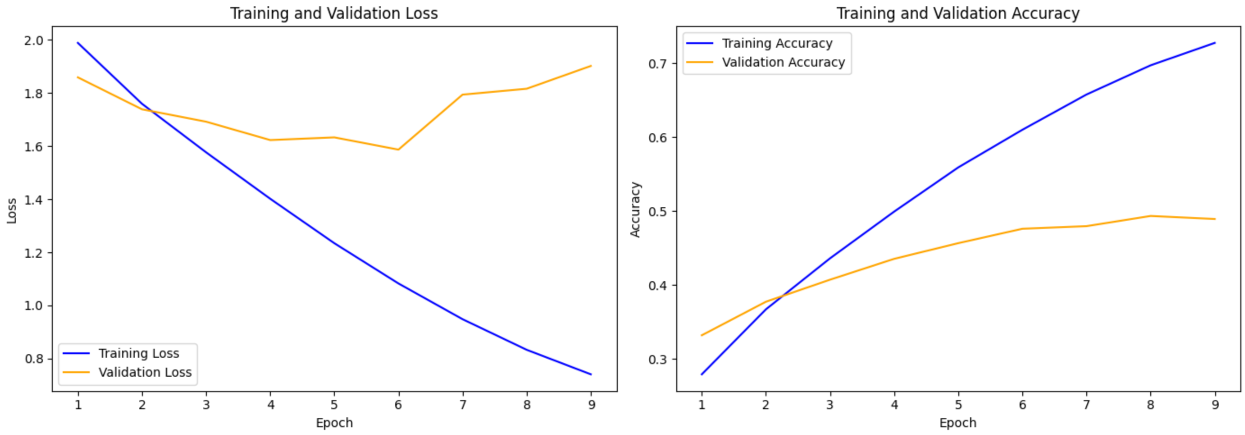


Figure 9: Training and Validation Loss and Accuracy - Model 2

Model Development for VGG 19 – Model 3

Building on the foundational aspects of the previous VGG19 models, Model 3 introduces changes to the training configuration to enhance its performance. A pivotal modification in this iteration is the adaptability in layer trainability. Unlike Model 1, which employed a fully frozen base VGG19 model to leverage pre-trained features without further adaptation, Model 3 strategically unfreezes the last two sets of convolutional layers (block4 and block5). This adjustment is aimed at fine-tuning the model on more complex and abstract features specifically relevant to the variations and complexities observed in lung disease imaging.

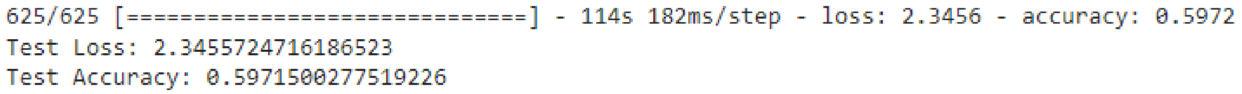
Key Features and Differences from Model 1 and 2:

1. **Layer Trainability:**
   - **Adjustable Training Layers:** The selective trainability of deeper layers in Model 3 allows for enhanced learning of nuanced features that are critical for accurate lung disease diagnosis, which was not possible with the completely frozen layers in Model 1.
2. **Impact on Diagnostic Accuracy:**
   - **Enhanced Specificity:** By enabling training on the latter stages of the network, Model 3 is tailored to better adapt to and learn from the specific characteristics of lung diseases visible in X-rays. This is expected to improve the model's diagnostic accuracy by refining its ability to detect subtle and intricate patterns that differentiate various thoracic conditions.

These targeted adjustments in Model 3 aim to strike a balance between harnessing the robust, generalizable features learned from large datasets (as with the original VGG19) and fine-tuning the model to enhance its specificity and effectiveness in medical imaging diagnostics. The approach underscores a shift towards a more dynamic adaptation of pre-trained networks, promising improvements in performance metrics and practical utility in clinical settings.

Model Evaluation for VGG 19 – Model 3

VGG19 Model 3, the most effective variant in our series, achieved a test accuracy of 59.72% with a loss of 2.3457 over 625 steps. Despite its improved performance, the model displayed a significant disparity between training accuracy and validation results, with validation accuracy exhibiting fluctuations and not paralleling the steady gains seen in training. This suggests potential overfitting to the training data, pointing to the need for enhanced regularization techniques and model adjustments to better generalize to unseen data.



```
625/625 [==============================] - 114s 182ms/step - loss: 2.3456 - accuracy: 0.5972
Test Loss: 2.3455724716186523
Test Accuracy: 0.5971500277519226
```

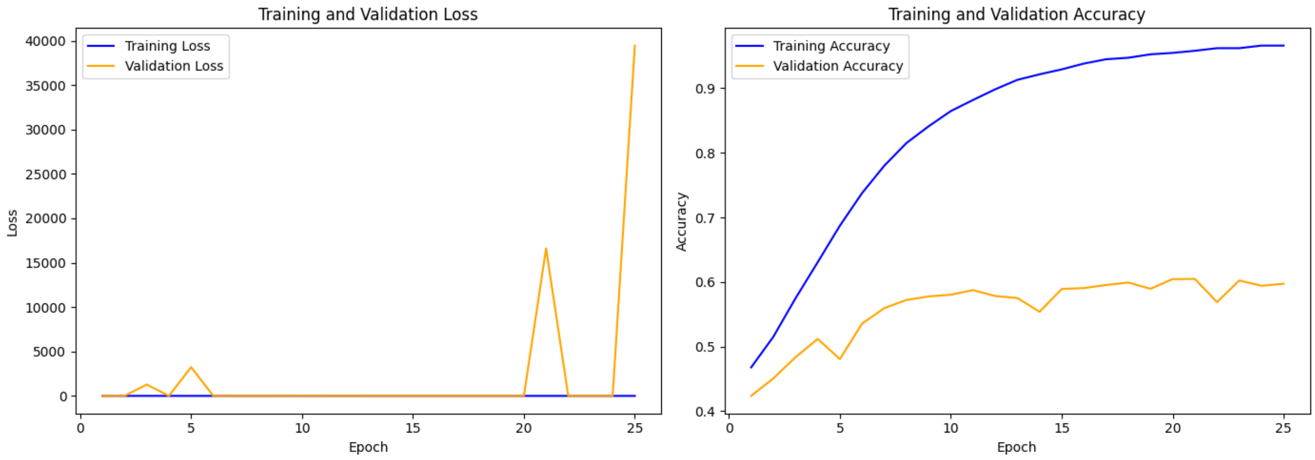


Figure 10: Training and Validation Loss and Accuracy - Model 3

ResNet 50 Model

ResNet models, short for Residual Networks, are a family of deep convolutional neural networks that were developed to address the vanishing gradient problem associated with training very deep architectures. Introduced by He et al. in their seminal 2015 paper, ResNet models utilize skip connections, or shortcuts, to jump over some layers. These connections allow the model to learn an identity function, ensuring that the deeper layers can perform at least as well as the shallower ones by directly carrying over activations from previous layers. This design facilitates the training of networks that are significantly deeper than those previously used, with models often exceeding hundreds of layers. ResNet has achieved remarkable success, notably winning the 2015 ImageNet Challenge in classification, detection, and localization tasks, and has since become a foundational architecture for many computer vision applications (He et al., 2015).

Feature Selection and Engineering for ResNet 50

The ResNet Model utilizes the robust framework of ResNet50, pre-trained on ImageNet, to adapt to the specific requirements of diagnosing lung diseases from chest X-ray images. The model harnesses the deep residual learning framework to facilitate the training of deeper networks, crucial for capturing intricate image details necessary for medical diagnosis. In this model, the original fully connected output layer of ResNet50 is replaced with a new linear layer designed to output predictions across 10 disease classes, tailoring the network's output to the specific classifications required for this application.

Model Development for ResNet 50 – Model 1

The development of ResNet Model 1 was focused on optimizing the model for high performance in medical imaging contexts. After replacing the output layer, the model was fine-tuned on the dataset of chest X-rays, using a GPU to handle the computational load effectively. An SGD optimizer with a momentum of 0.9 and a learning rate scheduler was employed to refine the training process, adjusting the learning rate by a factor of 0.1 every 7 epochs to prevent overfitting and enhance convergence. The model was trained in both training and validation phases to monitor and maximize accuracy while minimizing overfitting. This approach ensures that the model not only learns effectively from the training data but also performs well on unseen data, crucial for its deployment in real-world clinical settings. The use of cross-entropy loss helps in quantifying the difference between the predicted probabilities and the actual class labels, providing a reliable measure of model performance throughout the training.

Model Evaluation for ResNet 50 – Model 1

ResNet Model 1 achieved a test accuracy of 58.53% with a loss of 1.1981. This performance indicates a moderate level of diagnostic precision. The loss value suggests that while the model can categorize lung conditions with a reasonable degree of reliability, there remains significant room for improvement in minimizing errors and enhancing the accuracy of predictions.

```
Test Loss: 1.1981 Acc: 0.5853
```

Model Development for ResNet 50 – Model 2

ResNet Model 2 introduces changes in its development approach compared to Model 1, particularly in its training configuration. In Model 2, all layers of the ResNet50 architecture are unfrozen, allowing for complete trainability. This change enables the model to fine-tune all layers directly, rather than just adapting the final classification layers, which significantly enhances its ability to learn more detailed and complex features specific to lung diseases visible in chest X-rays.

Additionally, the optimizer used in Model 2 is AdamW, noted for incorporating weight decay, which helps in regularizing and preventing overfitting more effectively than the standard SGD used in Model 1. The learning rate scheduler is also adjusted to a ReduceLROnPlateau type, which reduces the learning rate when the validation loss plateaus, providing a more dynamic response to changes in training progress than the step-based reduction used previously. This nuanced control over the learning rate and layer trainability aims to improve the model's accuracy and adaptability to the intricate variations in medical imaging data.

Model Evaluation for ResNet 50 – Model 2

ResNet Model 2 achieved a test accuracy of 58.13% and a loss of 1.2124. This performance indicates a slight improvement in loss reduction compared to Model 1. The complete trainability of layers and the optimized learning rate adjustment strategy in Model 2 contributed to this enhanced precision.

```
Test Loss: 1.2124 Acc: 0.5813
```

Model Development for ResNet 50 – Model 3

ResNet Model 3 enhances adaptability by unfreezing all layers, allowing the entire network to learn from chest X-ray data. It employs the AdamW optimizer with weight decay for improved handling of overfitting and uses a ReduceLROnPlateau scheduler to adjust the learning rate based on validation performance, optimizing training effectiveness. The model is trained intensively with a focus on dynamically fine-tuning parameters to maximize accuracy and minimize loss.

Model Evaluation for ResNet 50 – Model 3

ResNet Model 3 demonstrated substantial progress in its testing phase, achieving a test accuracy of 64.89% with a loss of 1.0531. This performance indicates a notable improvement in the model's ability to accurately classify lung diseases from chest X-rays, reflecting the effectiveness of fully training all network layers and optimizing the training process through adaptive learning rate adjustments. The results show that the model is becoming increasingly reliable, offering more precise detection capabilities compared to its predecessors.

```
Test Loss: 1.0531 Acc: 0.6489
```

Model Development for ResNet 50 – Model 4

ResNet Model 4 introduces several critical enhancements over previous iterations. This model configuration involves unfreezing all layers to enable full trainability, allowing for deeper and more precise adaptations to the specific features of lung disease imaging in chest X-rays. To refine optimization, the model employs the AdamW optimizer with weight decay, which helps mitigate overfitting more effectively than standard optimizers. Additionally, Model 4 incorporates an advanced learning rate adjustment strategy with the ReduceLROnPlateau scheduler. This scheduler reduces the learning rate in response to plateaus in validation loss, facilitating more nuanced model training dynamics.

A significant innovation in Model 4 is the introduction of an early stopping mechanism during training. This feature sets a patience threshold that terminates training if no improvement is observed in the validation loss for a predefined number of epochs, preventing unnecessary computations and potential overfitting. The model is trained with a long horizon of up to 100 epochs to maximize learning potential, but with the safeguard of early stopping to ensure efficiency and model performance optimization.

Model Evaluation for ResNet 50 – Model 4

ResNet Model 4 achieved an impressive test accuracy of 73.95% with a test loss of 0.8806, marking it as the best performing model in our series. This significant improvement in both accuracy and loss indicates that the model's comprehensive training approach, which included full layer trainability and advanced optimization techniques, effectively captured the complexities of lung disease characteristics from chest X-rays. The combination of a dynamically adjusted learning rate and early stopping based on validation performance proved to be effective strategies for maximizing the model's diagnostic capabilities while ensuring computational efficiency.

```
Test Loss: 0.8806 Acc: 0.7395
```

Ensemble Model

Model Development for Ensemble – Averaging

The ensemble model integrates outputs from three distinct architectures: VGG19, VGG19 with transfer learning modifications, and ResNet50, to bolster the robustness and accuracy of disease prediction from chest X-rays. This strategy leverages the diverse strengths of each individual model to enhance diagnostic performance.

Ensemble Strategy:

- **Combination Technique:** The ensemble method utilizes averaging to combine the predictions from the three models. This approach harnesses the individual predictive capabilities and unique insights from each model, leading to a more accurate and reliable consensus prediction.

- **Prediction Process:** During the prediction phase, outputs from the VGG19, VGG19 with transfer learning, and ResNet50 models are first obtained independently. These predictions are then averaged to form a unified ensemble output, effectively increasing confidence in the diagnostic results and mitigating specific weaknesses that may be present in any single model.

Model Evaluation for Ensemble – Averaging

The ensemble model, combining VGG19, VGG19 with transfer learning, and ResNet50, achieved an overall accuracy of 66.19%. This approach effectively integrates the strengths of individual models to enhance diagnostic accuracy for lung diseases from chest X-rays. The ensemble's prediction performance, as assessed through precision, recall, and F1-score across various conditions, reflects a balanced improvement over single-model approaches, particularly in managing diverse types of lung pathologies with varying degrees of presentation complexity.

Model Development for Ensemble – Hard Voting

The hard voting ensemble model applies a straightforward yet effective voting strategy where predictions from the individual models—VGG19, VGG19 with transfer learning, and ResNet50—serve as votes for determining the final classification. Each model independently assesses the input chest X-ray image and casts a 'vote' for one of the possible disease categories. The final prediction for each image is determined by the majority of votes from the models. This method enhances the decision-making process by aggregating diverse analytical perspectives, which helps in reducing the likelihood of misdiagnosis due to model-specific biases or errors.

Model Evaluation for Ensemble – Hard Voting

The hard voting ensemble model achieved an accuracy of 62.50%, demonstrating a robust method of leveraging multiple predictions to enhance diagnostic performance. The precision, recall, and F1-scores for various disease conditions indicate that this ensemble method effectively balances model biases and errors, leading to improved diagnostic reliability across diverse lung pathologies. Although not the highest in overall accuracy, the hard voting strategy underscores its potential to provide consistent and dependable diagnostic outcomes when integrating insights from multiple models.

Results and Analysis

The individual performances of VGG19 models showed considerable variability, with accuracies ranging from 47.46% to 59.71%. This variability underscores the challenge of employing a single model type in the diagnostic interpretation of chest X-rays. The lowest test loss recorded among the VGG19 models was 1.597, suggesting some level of difficulty in achieving lower error rates.

In contrast, ResNet50 models demonstrated an overall higher performance compared to VGG19, with the fourth model achieving the highest accuracy of 73.42% and the lowest test loss of 0.9009 as highlighted in yellow in the table below among all tested models. The consistent improvement in both loss and accuracy across the ResNet50 models indicates a potentially better suitability for the complexities involved in interpreting chest X-rays.

The ensemble models, using both averaging and hard voting techniques, showed accuracies of 66.19% and 62.50%, respectively. Although these results are higher than most individual VGG19 models, they did not surpass the highest-performing individual ResNet50 model. This suggests that while the ensemble method adds robustness by reducing variability and potentially lowering the risk of overfitting, it may not always lead to superior predictive performance compared to the best individual models in specific settings.

Model	**Test Loss**	**Accuracy**
VGG 19		
Model 1	1.785	0.5042
Model 2	1.597	0.4746
Model 3	2.345	0.5971
RESNET 50		
Model 1	1.198	0.5853
Model 2	1.212	0.5813
Model 3	1.053	0.6489
Model 4	0.9009	0.7342
ENSEMBLE		
Model 1 – Averaging	-	0.6619
Model 1 – Hard Voting	-	0.6250

Table 2: Summary of Results of all Models

Conclusion

This study has critically examined the efficacy of utilizing advanced deep learning models, specifically VGG19 and ResNet50, to enhance the diagnostic accuracy and efficiency of lung disease identification from chest X-rays. Through the application of these models to the NIH chest X-ray dataset, the research successfully demonstrated that deep learning can achieve notable diagnostic accuracies, with the highest reaching 73.42% using the ResNet50 model. Despite this achievement, the maximum accuracy of around 70% suggests that these AI tools, in their current form, are supplementary and cannot replace radiologists. The technology serves to assist rather than supplant, highlighting the indispensable role of human expertise in medical diagnostics.

In terms of efficiency, the study explored the impact of AI on reducing time and resource utilization. Although AI models streamline some aspects of the diagnostic process, significant challenges persist in fully integrating these technologies into everyday clinical practice. Key issues include variability in model performance, the need for extensive data to train models effectively, and the integration of AI outputs into clinical workflows.

Future Work

Future research should aim to address several key areas to enhance the implementation of AI in medical diagnostics:

1. **Model Optimization:** Further refinement of the AI models to improve accuracy and reliability. This could involve experimenting with newer architectures, hybrid models, or advanced ensemble methods that might yield better performance than traditional or single-model approaches.
2. **Integration Strategies:** Developing more sophisticated integration strategies that allow AI tools to work seamlessly alongside radiologists. This includes creating interfaces that can efficiently present AI-generated insights in a way that complements the radiologist's workflow and decision-making process.
3. **Validation and Standardization:** Rigorous clinical validation of AI models is essential to ensure their reliability and effectiveness in diverse real-world settings. Additionally, standardizing the deployment of these models across different healthcare systems will be crucial to ensure consistent performance and trustworthiness.
4. **Ethical and Regulatory Considerations:** Addressing ethical concerns and regulatory requirements is vital for the adoption of AI in healthcare. Future work should focus on ensuring that AI applications uphold patient confidentiality, informed consent, and equitable healthcare delivery.
5. **Human-AI Collaboration:** Investigating the dynamics of human-AI collaboration to optimize both human expertise and AI capabilities. Studies could focus on defining best practices for

human oversight of AI diagnostics, including when and how human intervention should be prioritized.

By pursuing these areas, future research can pave the way for AI to become a more integral and effective component of medical diagnostics, ultimately leading to better patient outcomes and more efficient healthcare systems.

References

Boesch, G. (2021). *VGG Very Deep Convolutional Networks (VGGNet) - What you need to know*. [online] viso.ai. Available at: https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), pp.1904–1916. doi:https://doi.org/10.1109/tpami.2015.2389824.

Hoo-chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, Ronald M. Summers, Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, IEEE CVPR, pp. 2497-2506, 2016

National Institutes of Health (NIH). (2017). *NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community*. [online] Available at: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community.

Open-i: An open access biomedical search engine. https: //openi.nlm.nih.gov

Patel, K. (2020). *COVID -19 Detector with VGG-19 Convolutional Neural Network*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/python-based-project-covid-19-detector-with-vgg-19-convolutional-neural-network-f9602fc40b81.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald Summers, ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471, 2017.