DEEP LEARNING DRIVEN PNEUMONIA SCREENING FROM X-RAY IMAGES: PERFORMANCE ANALYSIS OF CNN AND VIT ARCHITECTURES

Srikar Gowrishetty[1], Venkata Jyothi Priya Mulaka[1], Eli Antoine[2]

[1] *Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL USA*
[2] *J.Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL USA*

PRESENT ADDRESS:
Srikar Gowrishetty, Graduate
Applied Data Science Program
Engineering Education Department
University of Florida
Gainesville , FL 32608

Venkata Jyothi Priya Mulaka, Graduate
Applied Data Science Program
Engineering Education Department
University of Florida
Gainesville , FL 32608

Eli Antoine, Graduate
Biomedical Engineering Program
University of Florida
Gainesville, FL 32608

SHORT TITLE:
Pneumonia Detection in X-rays using Deep Learning Models

This work was supported by:
Professor Kuang Gong
Assistant Professor
Department of Biomedical Engineering
University of Florida
Gainesville, FL, 32608

Counts:
Abstract – 157 words
Manuscript – 2831 words
Figures – 12

SUMBITTED TO: *Department of Biomedical Engineering*

ABSTRACT

Pneumonia remains a leading cause of respiratory failure and mortality worldwide, necessitating rapid and accurate diagnostic tools. While chest X-rays are the standard imaging modality for diagnosis, interpretation is subject to inter-observer variability and radiologist fatigue. This study evaluates the efficacy of deep learning models in automating pneumonia screening. We implemented and compared two distinct architectures: a Convolutional Neural Network (ResNet-50) and a Vision Transformer (ViT-B/16). The models were trained and tested on the Kaggle Chest X-Ray Images (Pneumonia) dataset consisting of 5,856 images. Our results indicate that the Vision Transformer outperforms the CNN architecture, achieving a test accuracy of 90% and a weighted precision of 91%, compared to ResNet-50's accuracy of 84% and weighted precision of 87%. Specifically, the ViT model demonstrated superior balance between sensitivity and specificity, significantly reducing false alarms in normal cases. This report discusses the implementation details, performance metrics, and the potential clinical implications of deploying Transformer-based models for medical image analysis.

KEYWORDS: Pneumonia screening, Deep learning, Convolutional neural networks (CNN), Vision Transformers (ViT), Medical imaging, Chest X-ray, Computer-aided diagnosis

1. Introduction

1.1 Clinical background

Pneumonia is an infection of the lung parenchyma in which the alveoli and surrounding tissues become inflamed and filled with fluid or pus.[7] This process alters gas exchange and can lead to hypoxia, respiratory distress, and systemic complications such as sepsis. Once a virus or bacterium enters the respiratory tract of an infected person, it causes the body to produce an inflammatory response and release macrophages and cytokines to combat the pathogen. Continued activation of the inflammation pathway can result in blood vessels and capillaries being more permeable resulting in pneumonia pathology as fluid is able to more easily enter the alveoli which are usually filled with air. Common pathogens include bacterial pathogens (e.g., *Streptococcus pneumoniae*) and viral infections (e.g., influenza, SARS-CoV-2), as well as aspiration of gastric contents in vulnerable patients.[1]

Symptoms typically include cough, fever, dyspnea, chest pain, and fatigue, but clinical presentation can be non-specific, especially in children, older adults, and immunocompromised patients. Early and accurate detection is therefore critical for triage and timely initiation of antibiotics or antiviral therapy, oxygen support, and hospitalization when indicated.[1]

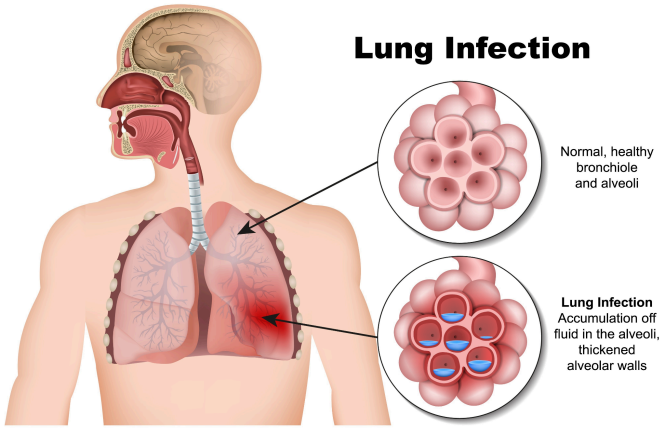


Figure: "Diagram of pneumonia pathology"

1.2 Role of chest X-ray imaging

Chest radiography is the first-line imaging modality for suspected pneumonia.[1] X-rays use high-energy photons to generate two-dimensional projections of the thorax, where bones, soft tissues, and air-filled spaces attenuate radiation to different degrees. Bones appear white because they are very dense and contain mainly calcium ions which have a high atomic number; meaning the x-ray radiation is encountering more electrons relative to the other structures in the body. This results in the bone attenuating/absorbing the x-ray radiation. Comparatively, soft tissues appear gray since they attenuate the x-ray radiation a lot less than bones do. Soft tissue is less dense and it is mainly made up of carbon, hydrogen, oxygen, and nitrogen which have low atomic numbers; meaning that the x-ray radiation is interacting with less electrons compared to bones. In a normal chest X-ray, the lungs appear relatively dark, with clear visualization of ribs, heart borders, and diaphragm. In pneumonia, affected lung regions typically show white, cloudy

opacities (consolidations or infiltrates) in areas that should otherwise be radiolucent.[1]

Chest X-rays are inexpensive, fast, and widely available, so they are routinely used in emergency departments and inpatient settings to guide decisions on admission, oxygen therapy, and escalation of care. However, interpretation can be challenging when findings are subtle, diffuse, or confounded by overlapping structures (e.g., heart, ribs, mediastinum).[1] This is especially apparent in diseases that have similar pathology to pneumonia and generate x-ray images that look very similar. One example is pulmonary edema which generates x-ray images with white opacities that resemble that of pneumonia x-ray images. One distinguishing feature is that the white opacities in pulmonary edema images usually form an outline similar to that of a bat.

1.3 Limitations of visual interpretation

Radiologists and clinicians face multiple challenges when diagnosing pneumonia from chest X-rays:

- **Subtle findings:** Faint opacities or early disease may be easily missed, especially in overloaded clinical settings.[1]
- **Mimicking conditions:** Pulmonary edema, atelectasis, hemorrhage, acute respiratory distress syndrome (ARDS), effusions, and some malignancies can look similar to pneumonia.[1]

- **Variability:** Differences in acquisition protocols, projection (AP vs. PA), exposure, and equipment across hospitals can lead to inconsistent image appearance.[1]
- **Inter-reader variability:** Different clinicians may disagree on the presence or severity of pneumonia, particularly in borderline cases.[1]

These factors motivate computer-aided diagnosis systems that can highlight abnormal regions and provide a second opinion to human readers.

1.4 Deep learning for chest X-ray analysis

Deep learning models, especially convolutional neural networks (CNNs), have demonstrated strong performance in detecting thoracic pathologies from chest radiographs. Prior work such as CheXNet has shown that CNNs can reach radiologist-level performance for pneumonia detection by learning complex texture and intensity patterns directly from large image datasets.[2]

More recently, Vision Transformers (ViTs) have emerged as an alternative architecture that treats an image as a sequence of patches and applies self-attention to model long-range dependencies. ViTs can capture global context more naturally than CNNs, which may be advantageous when opacities are diffuse or distributed across multiple lung zones.[5]

1.5 Project goals

In this project we aim to:

1. Implement a ResNet-50 CNN and a ViT-B/16 model for binary classification of chest X-rays into NORMAL vs PNEUMONIA.[1]
2. Investigate the impact of class imbalance, where pneumonia examples greatly outnumber normal images, on model behavior.
3. Apply balanced sampling (WeightedRandomSampler) and fine-tuning strategies to improve performance, especially on the NORMAL class.
4. Use Grad-CAM and ViT attention maps to visualize which image regions drive the predictions, and qualitatively compare how CNNs and transformers "look" at the X-ray.

2. Materials and Methods

2.1 Dataset Description

We use the Kaggle Chest X-Ray Pneumonia dataset. The dataset is organized into train, validation, and test folders under a root `chest_xray` directory and contains posterior–anterior chest radiographs of pediatric patients. The task is binary classification[3]:

- Label **0 – NORMAL**
- Label **1 – PNEUMONIA**

The dataset split used in this project is:

- **Training:** 1,341 NORMAL, 3,875 PNEUMONIA
- **Validation:** 8 NORMAL, 8 PNEUMONIA

- **Test:** 234 NORMAL, 390 PNEUMONIA (total 624 images)

This training split is strongly imbalanced, with approximately three times more pneumonia images than normal images.

2.2 Preprocessing and data augmentation

All images are converted to RGB and resized to 224 × 224 pixels. Different transforms are applied for training vs. validation/testing. The processing steps are implemented using *torchvision.transforms* and are consistent with the ImageNet normalization used by the pretrained backbones.[1]

Training transforms (imbalanced and balanced experiments):

- Resize or RandomResizedCrop to 224 × 224
- Random horizontal flip
- Random rotation (±10–15°)
- Convert to tensor
- Normalize using ImageNet mean and standard deviation:
    - mean = [0.485, 0.456, 0.406]
    - std = [0.229, 0.224, 0.225]

Validation and test transforms:

- Resize to 224 × 224
- Convert to tensor
- Normalize with the same ImageNet statistics

- No random augmentations applied

The batch size is 32, and all models are trained on GPU when CUDA is available.

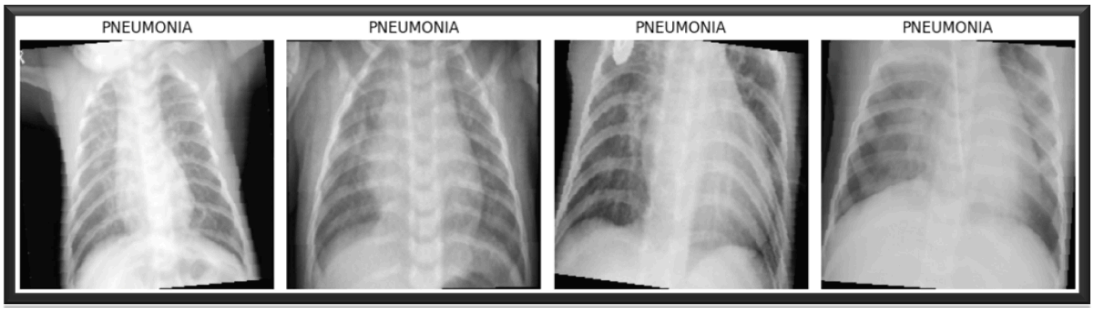


Figure: "Example image of the augmentations made to the images"

2.3 Handling class imbalance

To address the large imbalance in the training set (NORMAL: 1,341 vs PNEUMONIA: 3,875), we employ a WeightedRandomSampler during later experiments.

1. We compute the class counts from the training dataset:
    - `class_counts = [1341, 3875]` (NORMAL, PNEUMONIA).
2. We compute inverse-frequency class weights:
    - `class_weights = 1.0 / class_counts`.
3. Each training sample is assigned a weight equal to the weight of its class.
4. The `WeightedRandomSampler` uses these weights to draw samples with replacement, so each mini-batch becomes approximately class-balanced even though the underlying dataset is imbalanced.

This strategy forces the model to see NORMAL and PNEUMONIA images equally often during training and reduces the bias toward predicting "PNEUMONIA."[1]

2.4 ResNet-50 architecture and training

We use the ResNet-50 implementation from `torchvision.models` with ImageNet pretraining (`ResNet50_Weights.IMAGENET1K_V1`).[4]

- All layers are initially frozen (`requires_grad = False`).
- We unfreeze only `layer4` (the last residual block) for fine-tuning.
- The original 1,000-class fully connected (FC) layer is replaced by a new linear head with 2 output neurons for the binary task.

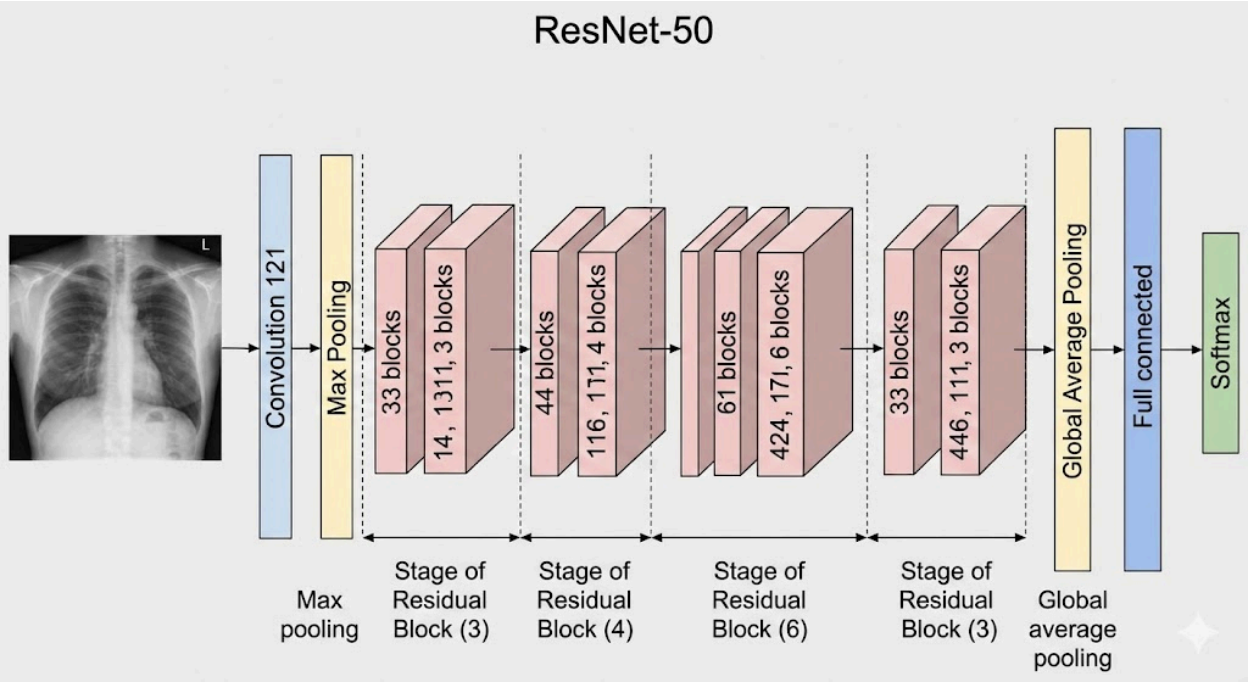


Figure: ResNet-50 Architecture for X-ray Images

The loss function is cross-entropy loss, and optimization uses Adam with two parameter groups:

- `layer4` parameters: learning rate `1e-5`
- new FC head: learning rate `1e-4`

2.4.1 Initial training (imbalanced)

In the first stage, we train ResNet-50 for 5 epochs using the original imbalanced training set and shuffled batches. We track training and validation accuracy across epochs and save the checkpoint with best validation accuracy (`resnet_best.pth`).

2.4.2 Balanced training

In the second stage, we retrain ResNet-50 using the WeightedRandomSampler in the training DataLoader, with the same architecture and optimizer settings, for 10 epochs. This setup improves exposure to NORMAL cases and aims to correct the earlier bias toward PNEUMONIA.

2.5 ViT-B/16 architecture and training

We use the Vision Transformer ViT-B/16 model from `torchvision.models` with ImageNet pretraining (`ViT_B_16_Weights.IMAGENET1K_V1`).[5]

- All parameters are first frozen.
- We unfreeze only the last encoder block (`vit.encoder.layers[-1]`) to adapt high-level features to pneumonia patterns.
- The original classification head is replaced with a 2-class linear head (`nn.Linear(num_ftrs, 2)`).

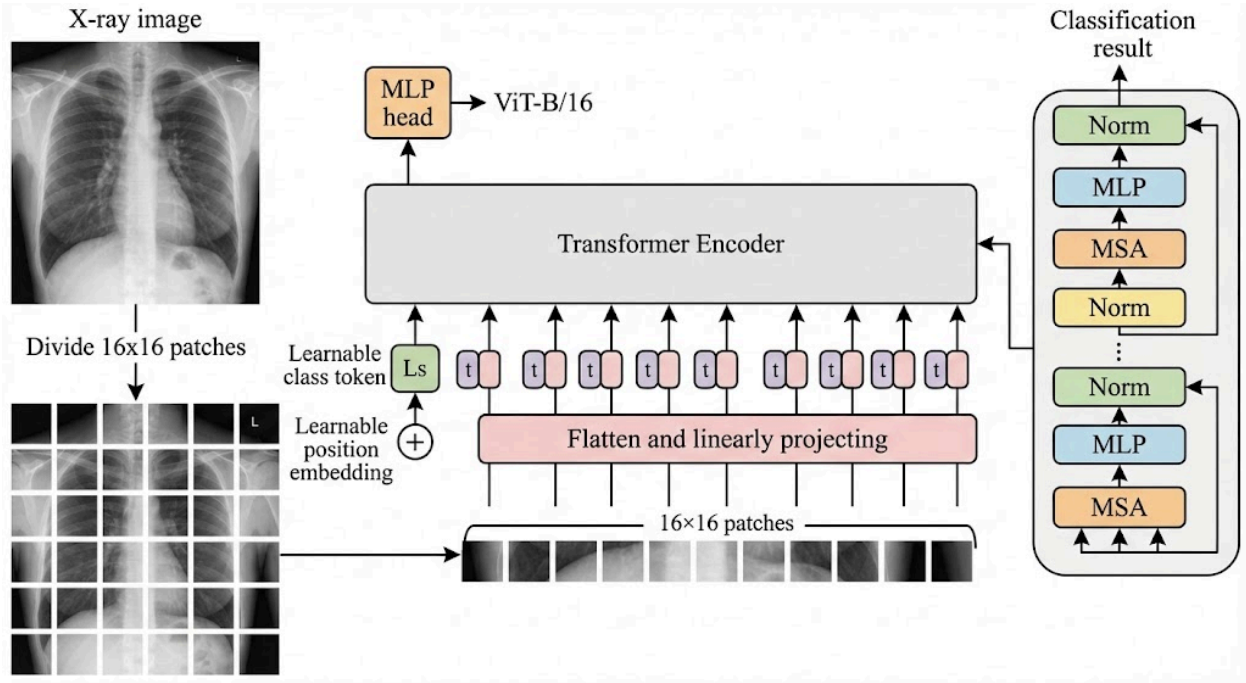


Figure: ViT-B/16 Architecture for X-ray Images

The loss function is again cross-entropy. We use Adam with two parameter groups:

- Last encoder block: learning rate `1e-5`
- New head: learning rate `1e-4`

2.5.1 Initial ViT training (imbalanced)

We train ViT-B/16 for 5 epochs on the imbalanced training set using the same basic augmentation pipeline and save the best validation checkpoint (*vit_best.pth*). Validation accuracy stabilizes around 0.75.

2.5.2 Balanced ViT training

Next, we retrain ViT-B/16 with the WeightedRandomSampler and 10 training epochs, improving exposure to NORMAL images and reducing pneumonia bias.

2.5.3 Final fine-tuning

For the final stage, we further fine-tune ViT-B/16 on the balanced setup with stronger augmentations (RandomResizedCrop, rotation, ColorJitter) to improve generalization. The best model from this phase is used for the final test evaluation reported in Section 3.4.

2.6 Evaluation metrics

For each trained model and configuration, we evaluate on the fixed test set (234 NORMAL, 390 PNEUMONIA). Using `scikit-learn`, we compute:

- **Confusion matrix**
- **Accuracy**
- Per-class **precision**, **recall**, **F1-score**
- **Macro averages** (unweighted mean across classes)
- **Weighted averages** (weighted by class support)

These metrics summarize both overall performance and how well each class is recognized.

2.7 Explainability: Grad-CAM and ViT attention

To interpret model predictions, we generate[1]:

- **Grad-CAM heatmaps for ResNet-50:**
  - Use gradients of the target class with respect to the last convolutional feature map.
  - Produce a heatmap over the X-ray that highlights regions contributing most to the prediction.

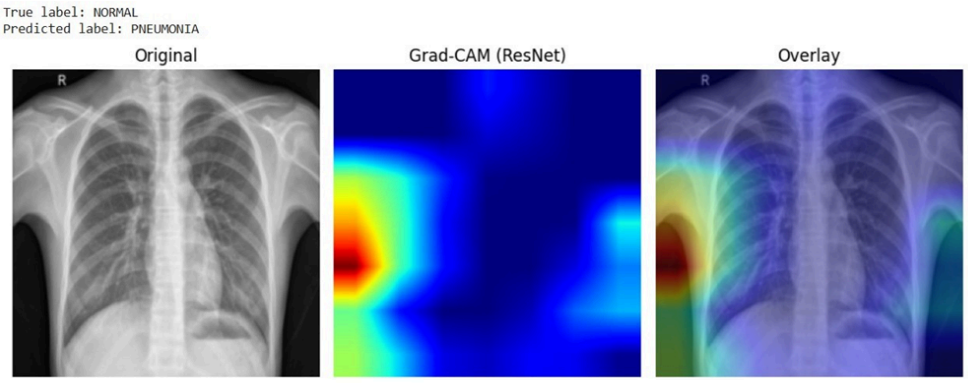


Figure 1: "Grad-CAM highlights localized texture interpreted as abnormal by ResNet-50

- **Attention maps for ViT-B/16:**
  - Visualize patch-wise attention weights aggregated across heads and layers.
  - Show how the transformer distributes focus across different lung regions.

These visualizations help assess whether the models attend to clinically meaningful structures (e.g., consolidated lung regions) or are distracted by artifacts and borders.

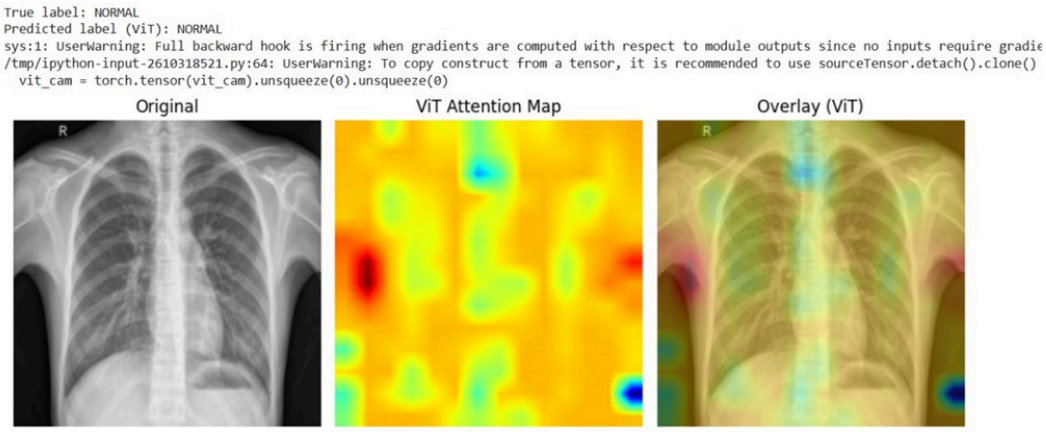


Figure 2: ViT attention remains globally distributed, leading to correct prediction."

3. Results

3.1 Baseline performance with imbalanced training

3.1.1 ResNet-50 (imbalanced)

Using the original imbalanced training set, ResNet-50 achieved the following test metrics:

- **Accuracy:** 0.80
- **NORMAL:** precision 0.97, recall 0.48, F1-score 0.64
- **PNEUMONIA:** precision 0.76, recall 0.99, F1-score 0.86
- **Macro average (P/R/F1):** 0.87 / 0.74 / 0.75
- **Weighted average (P/R/F1):** 0.84 / 0.80 / 0.78

The model was very sensitive to pneumonia (recall 0.99) but frequently misclassified NORMAL images as PNEUMONIA (recall 0.48), reflecting the training imbalance.

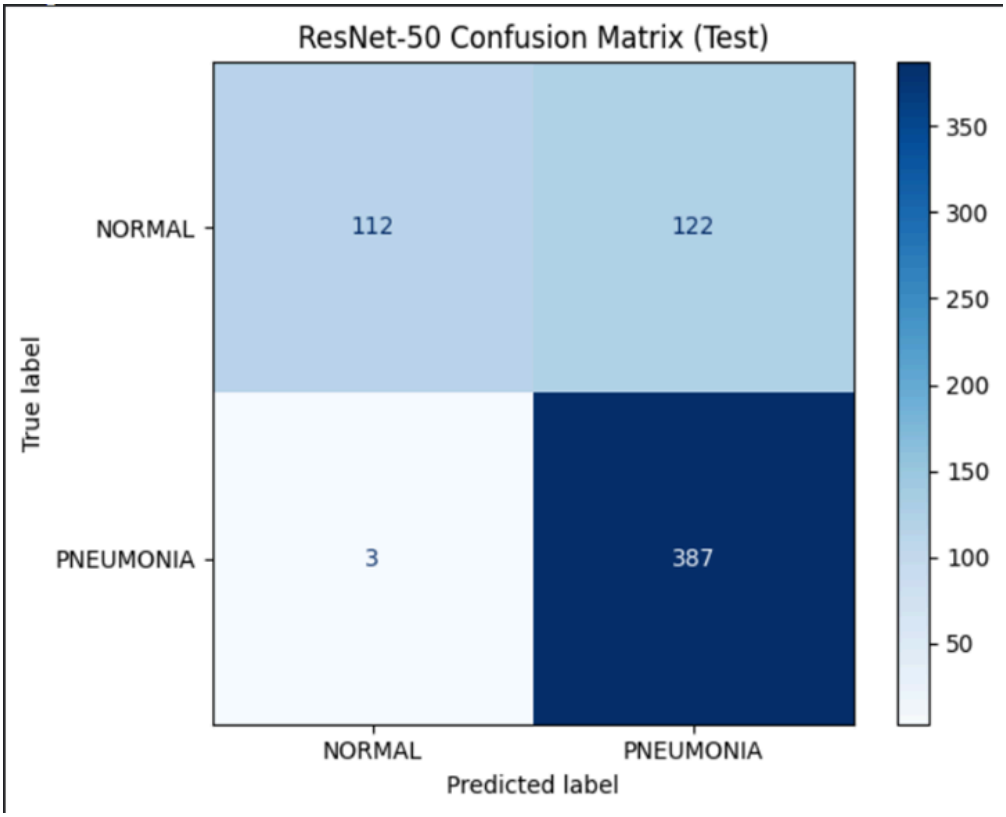


Figure 3: Confusion Matrix - ResNet50 (Imbalanced Dataset)

3.1.2 ViT-B/16 (imbalanced)

The baseline ViT-B/16 model showed slightly higher overall performance:

- **Accuracy:** 0.81
- **NORMAL:** precision 0.98, recall 0.51, F1-score 0.67
- **PNEUMONIA:** precision 0.77, recall 0.99, F1-score 0.87
- **Macro average (P/R/F1):** 0.88 / 0.75 / 0.77

- **Weighted average (P/R/F1):** 0.85 / 0.81 / 0.80

Like ResNet-50, ViT strongly favored pneumonia predictions but achieved slightly better NORMAL recall and overall accuracy.

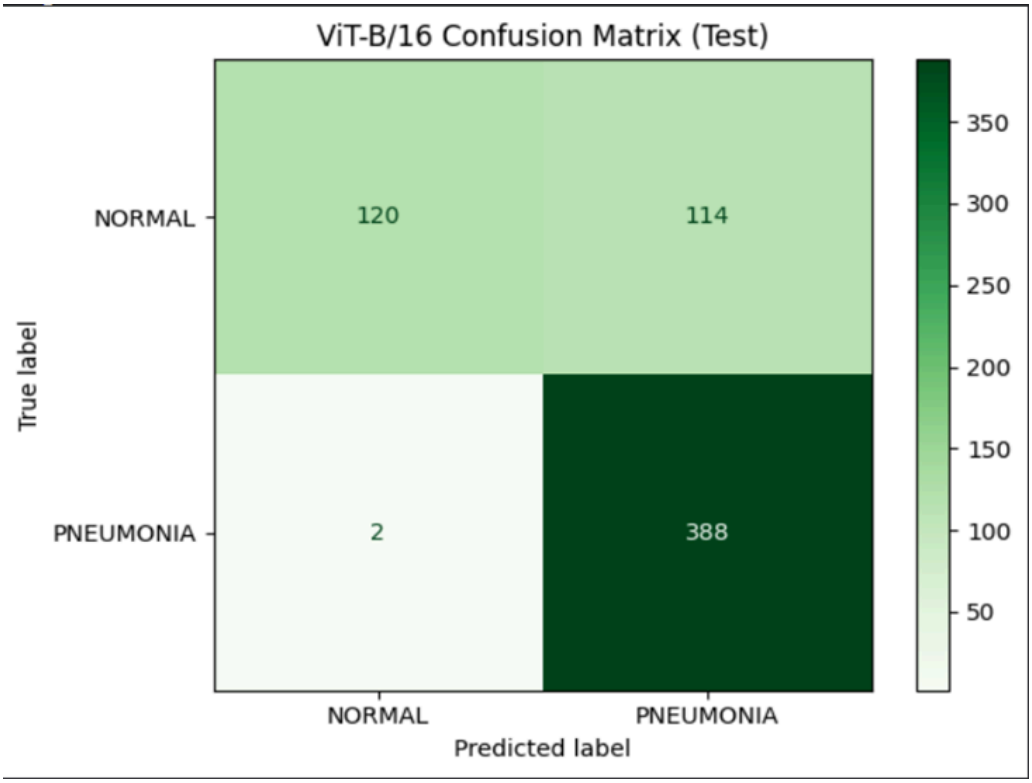


Figure 4: Confusion matrix – ViT-B/16 (Imbalanced Dataset)

3.2 Effect of balanced training

3.2.1 ResNet-50 (balanced)

After introducing the WeightedRandomSampler and retraining for 10 epochs, ResNet-50 performance improved:

- **Accuracy:** 0.85

- **NORMAL:** precision 0.98, recall 0.61, F1-score 0.75
- **PNEUMONIA:** precision 0.81, recall 0.99, F1-score 0.89
- **Macro average (P/R/F1):** 0.89 / 0.80 / 0.82
- **Weighted average (P/R/F1):** 0.87 / 0.85 / 0.84

Balanced training substantially improved NORMAL recall from 0.48 to 0.61 while maintaining very high pneumonia sensitivity (0.99).

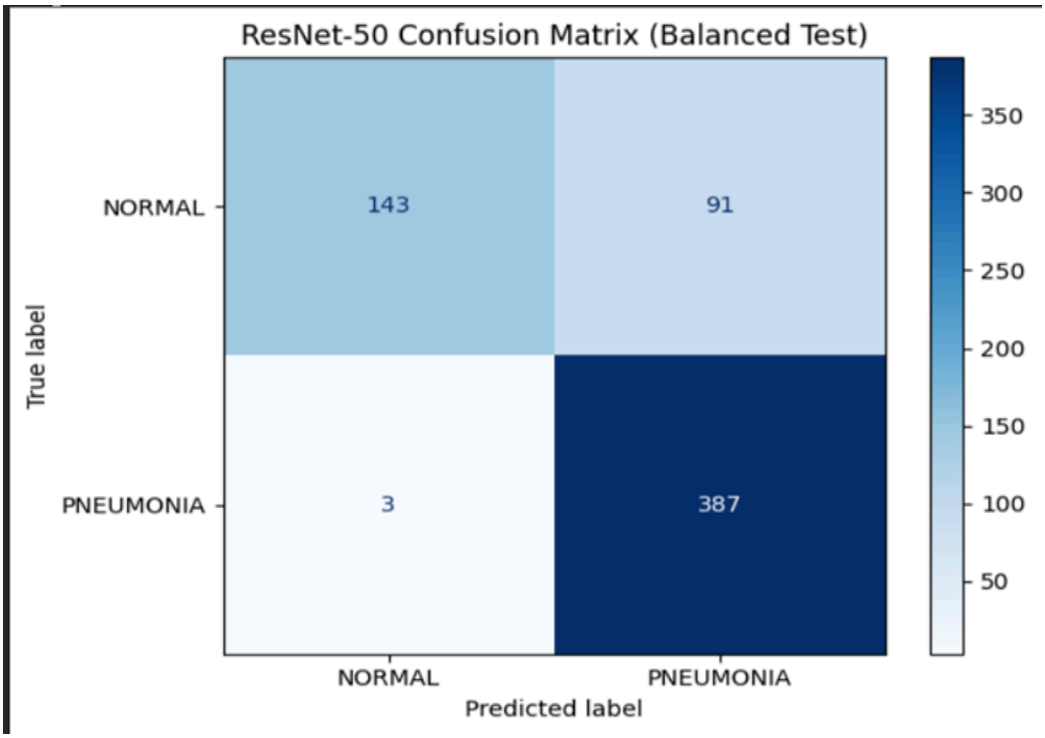


Figure 5. Confusion matrix – ResNet-50 (Balanced Dataset).

3.2.2 ViT-B/16 (balanced)

Balanced training also benefited ViT-B/16:

- **Accuracy:** 0.91
- **NORMAL:** precision 0.95, recall 0.79, F1-score 0.86
- **PNEUMONIA:** precision 0.89, recall 0.98, F1-score 0.93
- **Macro average (P/R/F1):** 0.92 / 0.88 / 0.90
- **Weighted average (P/R/F1):** 0.91 / 0.91 / 0.90

Compared to the imbalanced run, NORMAL recall improved from 0.51 to 0.79, and overall accuracy increased from 0.81 to 0.91, indicating a much more balanced classifier.

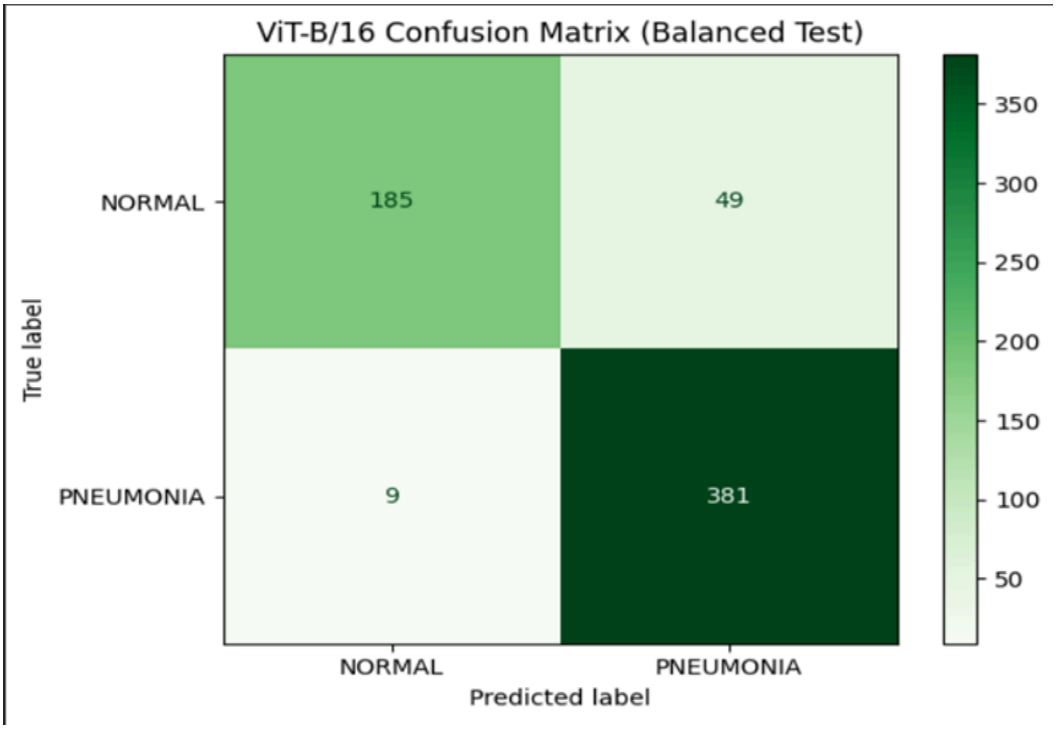


Figure 6. Confusion matrix – ViT-B/16 (balanced training).

3.3 Final fine-tuned ViT-B/16

With additional fine-tuning using stronger augmentation, the final ViT-B/16 model achieved:

- **Accuracy:** 0.92
- **NORMAL:** precision 0.96, recall 0.82, F1-score 0.89
- **PNEUMONIA:** precision 0.90, recall 0.98, F1-score 0.94
- **Macro average (P/R/F1):** 0.93 / 0.90 / 0.91
- **Weighted average (P/R/F1):** 0.92 / 0.92 / 0.92

This configuration provided the best overall performance and the most symmetric treatment of both classes, with only modest overfitting observed between training and validation curves.

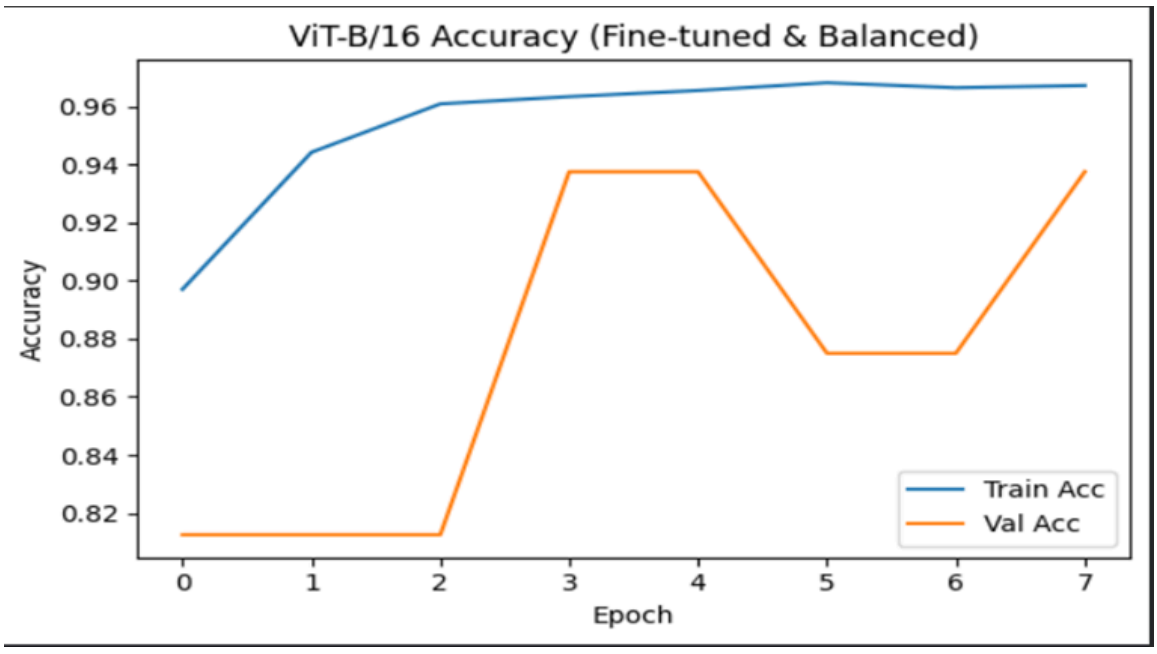
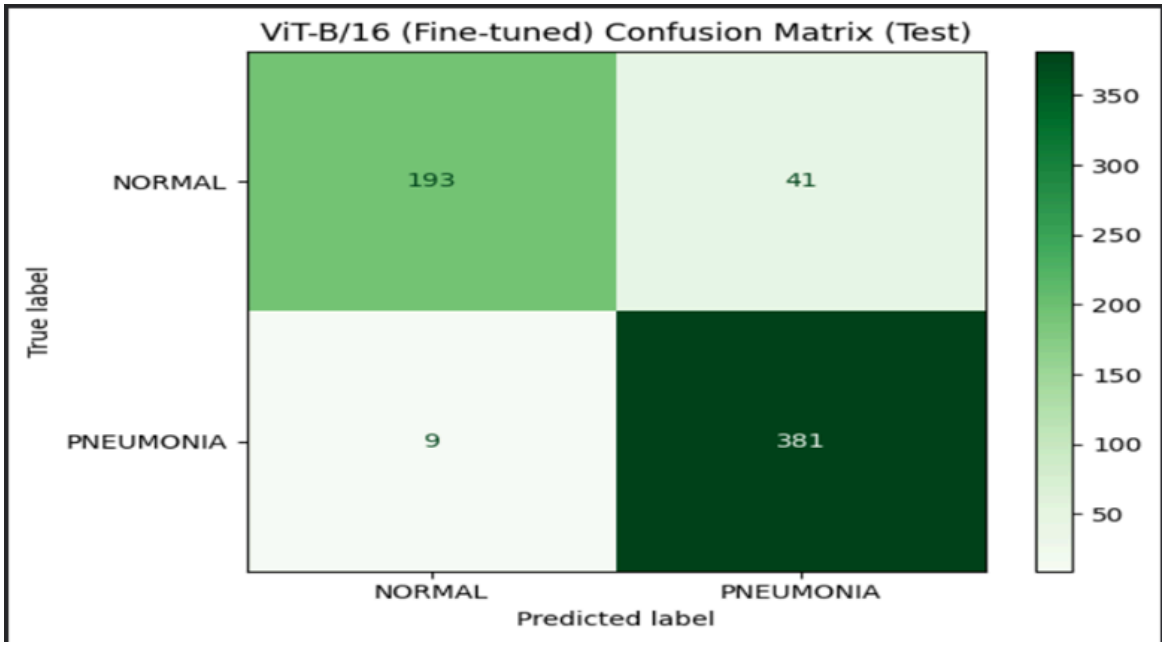


Figure 7. "Validation accuracy stabilizes near 0.92 after 10 epochs; prediction visualization confirms high confidence and correct labeling."



Example predictions: "Normal vs Pneumonia with predicted probabilities."

3.4 Qualitative explainability results

Representative Grad-CAM and attention map visualizations highlight qualitative differences between the CNN and ViT models.[1]

- **ResNet-50:**
    - In one example, the true label was NORMAL, but ResNet predicted PNEUMONIA.
    - The Grad-CAM heatmap showed intense activation over a localized shadow in the left lung, which corresponded to normal anatomical variation rather than pathology.
    - This suggests that ResNet can sometimes over-interpret local texture patterns as disease.
- **ViT-B/16:**

- In the same NORMAL case, ViT correctly predicted NORMAL.
- The attention map displayed distributed, patch-based activation across central lung fields and upper lobes, emphasizing global context rather than one isolated region.
- This behavior suggests better robustness to small artifacts and border effects.

These visualizations support the quantitative finding that the ViT model is less biased toward false pneumonia alarms while still capturing clinically relevant regions.

4. Discussion

4.1 Impact of class imbalance

Initial experiments clearly showed that training on an imbalanced dataset caused both ResNet-50 and ViT-B/16 to learn a strong preference for the PNEUMONIA class. NORMAL recall was around 0.48–0.51, meaning roughly half of normal chest X-rays were wrongly flagged as pneumonia. In a clinical setting, such behavior would lead to excessive false positives, unnecessary anxiety, and further testing.

The WeightedRandomSampler significantly mitigated this issue by enforcing approximately equal sampling of NORMAL and PNEUMONIA images. NORMAL recall rose to 0.61 for ResNet-50 and 0.79 for ViT-B/16, while pneumonia sensitivity remained extremely high ($\geq$0.98). This demonstrates that balanced sampling is a

simple yet powerful technique for improving fairness between classes without modifying loss functions or model architectures.

4.2 CNN vs. Vision Transformer behavior

Quantitatively, both networks can achieve strong pneumonia detection, but ViT-B/16 consistently outperforms ResNet-50 in terms of balanced performance:

- Higher test accuracy (0.91–0.92 vs. 0.85) after balancing.
- Better NORMAL recall (0.79–0.82 vs. 0.61).
- Higher macro F1-scores, indicating more symmetric treatment of classes.

Qualitatively, Grad-CAM and attention maps show architectural differences[2][4]:

- ResNet-50 often focuses on localized high-contrast regions, which may correspond to true consolidations but can also pick up on artifacts or normal structures.
- ViT-B/16 tends to distribute attention over broader lung regions, integrating global context and reducing the chance of over-reacting to a single suspicious patch.[5]

These observations suggest that transformer-based models may be particularly suitable for diffuse or subtle pathologies and for handling dataset variability in acquisition and anatomy.

4.3 Limitations

This project has several important limitations:

1. **Dataset size and composition:**
    - The validation set is extremely small (16 images), which can make early stopping and hyperparameter tuning unstable.
    - The dataset is pediatric and from a single source, which may limit generalizability to adults or images from other institutions.
2. **Binary task only:**
    - The classifier distinguishes only NORMAL vs PNEUMONIA. In practice, radiologists face a spectrum of diseases that can mimic pneumonia; a multi-label or multi-class setup would be more realistic.
3. **Limited hyperparameter search:**
    - Due to time and compute constraints, only a small set of learning rates and epochs were explored. More systematic tuning could further improve performance or reduce overfitting.
4. **No calibration analysis:**
    - We did not evaluate probability calibration or decision thresholds; for deployment, calibrated probabilities and clinically informed thresholds would be necessary.
5. **Single-image context:**
    - Models operate on single radiographs and do not incorporate clinical metadata (age, symptoms, lab values) that clinicians routinely use.

4.4 Potential clinical role

Despite these limitations, the final ViT-B/16 model demonstrates strong performance and good interpretability. In a real clinical workflow, such a system could:

- Serve as a triage tool, prioritizing suspicious studies for radiologist review.
- Provide visual explanations (heatmaps) to highlight areas of concern.
- Help standardize detection across different operators and busy clinical settings.

However, extensive external validation, calibration, and integration with hospital systems would be required before any clinical deployment.

5. Conclusions

In this project, we implemented and compared a ResNet-50 CNN and a ViT-B/16 Vision Transformer for automatic pneumonia detection from chest X-ray images. Both models were initialized with ImageNet weights and fine-tuned on the Kaggle Chest X-Ray Pneumonia dataset.

Key findings are:

1. Training on an imbalanced dataset produced high pneumonia recall but poor normal recall, leading to many false alarms.
2. Using a WeightedRandomSampler to balance class sampling during training significantly improved normal-class recall while preserving pneumonia sensitivity.
3. After balancing and fine-tuning, ViT-B/16 achieved the best overall performance with 0.92 accuracy and F1-scores of 0.89 (NORMAL) and 0.94 (PNEUMONIA), outperforming ResNet-50.
4. Grad-CAM and attention maps showed that ViT tends to use more global, distributed features, making it less sensitive to small benign shadows than ResNet-50.

These results suggest that transformer-based architectures, combined with careful handling of class imbalance and interpretability tools, are a promising direction for robust, trustworthy AI-assisted pneumonia screening from chest radiographs.

6. Future Scope

a. Expand to Multi-Class & Multi-Label Pneumonia Subtypes

Currently, the model only predicts *Normal vs Pneumonia*.
Future work can extend the classifier to:

- Bacterial vs Viral pneumonia
- Lobar vs multilobar infection
- Differentiating pneumonia from diseases that mimic it (ARDS, effusion, pulmonary edema)

This makes the model more clinically relevant and reduces misdiagnosis risk.

b. Expand to 3D CT Scan Models

Future work can extend the approach to:

- 3D volumetric CNNs
- Transformers for CT slices
- Early pneumonia detection before visible changes on X-ray

CT-based models achieve higher sensitivity and could complement X-ray models.

c. Build a Mobile or Web-Based Screening App

After optimization, the model could be converted into:

- A lightweight app for remote/rural diagnosis
- A telemedicine support tool
- A screening assistant during pandemics

This enhances accessibility in low-resource settings.

References

1. Caliman-Sturdza OA, Soldanescu I, Gheorghita RE. “SARS-CoV-2 Pneumonia: Advances in Diagnosis and Treatment.” *Microorganisms*. 2025.
2. Rajpurkar P, Irvin J, Zhu K, et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.” arXiv:1711.05225, 2017.
3. Mooney P. “Chest X-Ray Images (Pneumonia).” Kaggle, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.
4. He K, Zhang X, Ren S, Sun J. “Deep Residual Learning for Image Recognition.” *Proc. CVPR*, 2016.
5. Dosovitskiy A, Beyer L, Kolesnikov A, et al. “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.” *ICLR*, 2021.

6. Wayson MB, et al. “Suggested reference values for regional blood volumes in children and adolescents.” *Phys. Med. Biol*.(For using as the sample paper.)
7. National Heart, Lung and Blood Institute. (2022). *Pneumonia*. Www.nhlbi.nih.gov; NIH. https://www.nhlbi.nih.gov/health/pneumonia.