# Pneumonia Detection Using Vision Transformers (ViT) and ResNet (CNN)

The goal of this project was to develop a machine learning model capable of detecting **pneumonia from chest X-ray images**. To compare the performance of modern deep learning architectures, we implemented and trained two powerful models:

- **Vision Transformer (ViT)**
- **ResNet (Residual Convolutional Neural Network)**

Both models were trained to classify images into two categories: **normal** or **pneumonia**, but used different strategies for feature extraction — ViT uses attention over image patches, while ResNet leverages convolutional operations.

==========

## Required Libraries

Before running the project, ensure the following libraries are installed:

```
pip install torch torchvision torchaudio

pip install transformers

pip install matplotlib

pip install scikit-learn

pip install numpy

pip install opencv-python

pip install mlflow

pip install pyngrok
```

==========

## 1. Data Preprocessing

To make the data compatible with both models, we performed these steps:

- **Resizing:**
  All chest X-ray images were resized to **224x224** pixels.

- **Normalization:**
  Images were normalized using mean and standard deviation values (specific to each model).

- **Transformation Pipelines:**
  A combination of Resize, ToTensor, and Normalize functions from torchvision.transforms were used.

- **Dataset Structure:**
  The data was organized into train, val, and test folders and loaded using PyTorch's ImageFolder.

==========

## 2. Model Architectures

### A) ResNet (CNN)

ResNet is a powerful convolutional neural network known for its **skip connections** that help train very deep models efficiently.

Key layers used:

- Convolutional layers
- Batch normalization
- ReLU activation
- Skip (residual) connections
- Fully connected output layer with 2 classes

We used **ResNet-18** pre-trained on ImageNet, then fine-tuned it for pneumonia classification.

---

### B) Vision Transformer (ViT)

ViT treats an image as a sequence of patches and uses Transformer encoders to model global dependencies across the image.

Key steps:

- Divides image into patches (16x16)
- Flattens and embeds each patch
- Applies positional encoding
- Processes with Transformer blocks
- Outputs classification prediction

We used the **google/vit-base-patch16-224** model and modified it for binary classification

==========

# 5. Training Function

## 🔷 ResNet18 Version

- **Training Function:**
    - Uses a custom function `run_training()`.
    - Tracks:
        - Training loss
        - Validation loss
        - Accuracy per epoch
- **Prediction Method:**
    - Applies `sigmoid` activation to outputs.
    - Uses thresholding (> 0.5) to convert outputs to binary predictions.
- **MLflow Logging:**
    - Logs:
        - Training/validation metrics
        - Hyperparameters
        - Final model

---

## 🔶 Vision Transformer (ViT) Version

- **Training Setup:**
    - Loss function: `CrossEntropyLoss`
    - Optimizer: `AdamW`
    - Learning rate: `5e-5`
    - Epochs: 5
    - Device:
        - Uses `CUDA` if available
        - Falls back to CPU otherwise
- **Prediction Method:**
    - Uses `logits` from model output
    - Applies `argmax` to get predicted class

==========

## Model Performance Comparison

**ResNet18**

- Validation Accuracy: 93.97% — shows strong learning during training.
- Test Accuracy: 83.8% — good, but a drop compared to validation indicates potential overfitting or generalization issues.
- F1 Score: 0.883 — highlights a decent balance between precision and recall.
- Conclusion: Performs well for pneumonia detection but may benefit from further tuning or regularization to improve generalization.

**Vision Transformer (ViT)**

- Test Accuracy: 99% — excellent performance, significantly higher than ResNet18.
- Evaluation Strategy: Efficient inference using `torch.no_grad()` and `argmax`.
- Additional Metrics: Confusion matrix and classification report used for deeper insights.
- Conclusion: ViT demonstrated superior performance, likely due to its ability to model global relationships in image data, making it highly effective for medical image classification.

.