Tiffany Lin

Github: https://github.com/Luwuinnz/CECS456_DLProject

456 Deep Learning Project Report

1. Introduction

Pneumonia is a lung infection that causes inflammation of the alveoli, leading to difficulty breathing, coughing, and feverish symptoms. Its severity ranges from mild to life-threatening. However, prompt and effective diagnosis of pneumonia dramatically increases patient survival. Chest x-rays are the primary tool in diagnosing this disease, but manual interpretation is prone to human error, especially in high-volume clinical environments. Deep learning has strong potential for diagnostic accuracy and reducing costs.

The objective of the project is to design, train, and evaluate deep learning models given a dataset. Training and evaluating by the Chest X-Rays (pneumonia) dataset, I implemented two different models: a custom CNN and a ResNet50 transfer learning model to compare the validity of my model's performance. Performance was assessed using confusion and classification matrices, and accuracy and loss graphs.

2. Dataset and Related Work

The *Chest X-Rays (pneumonia)* dataset from Kaggle contained a total of 5,863 labeled chest X-ray images divided into two categories: normal and pneumonia. Manual intervention was not needed since the dataset is already split into train and test sets, with 5,216 training images and 624 test images. However, the dataset is highly imbalanced, with significantly more pneumonia cases than normal cases. This imbalance heavily influenced model behavior, which is discussed in the analysis section.

3. Methodology

Model 1: Custom CNN

The custom CNN model architecture consists of 32, 64, and 128 convolution filters, each followed by maxpooling. After flattening the feature maps, they were passed into a 128 unit dense layer and dropout 0.5. The output node comprised one sigmoid node for binary prediction. With the Adam optimizer, the learning rate was 1e-4, and the training was for 10 epochs.

Model 2: ResNet50 Transfer Learning

The ResNet50 architecture had ImageNet pretrained weights, provided by TensorFlow Keras Applications. ResNet was originally introduced by He et al. (2015) in their paper *'Deep Residual Learning for Image Recognition'*. The base model froze all the convolutional layers to preserve pretrained feature representations. Adding a custom classification head, it consisted of globalaveragepooling, a dense layer with 256 units with dropout 0.5, and a sigmoid output layer. Model implementation included ImageNet preprocessing, data augmentation, earlystopping when model shows no improvement, and reduceLRonplateau callbacks.

A. Experimental Setup

To improve model generalization, running the formatting.py code augmented the original Kaggle training dataset that both models used. Testing and validation set untouched by augmentation.

For consistency purposes, all experiments used images of size 224x224; normalized to 0 -1 range; a batch size 32; 1e-4Adam optimizer. The test set was reserved strictly for final evaluation and was not used during training. CNN model was calibrated by experimenting with threshold sweeps from 0.3 to 0.7; dropout value from 0.3 to 0.5, and epochs 8-15. The CNN model best performed with threshold sweep at 0.5; dropout at 0.5, and epoch at 10.

B. Measurement

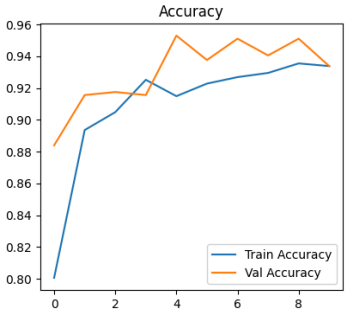
The measurements used are accuracy/loss graphs, classification matrix: precisions, recall, F1-score and confusion matrix. They captured different aspects of model behavior–uniquely in a class imbalance since there were more pneumonia cases than normal in the dataset. Measurements were computed after training and validation to ensure consistency across two models. The models produced probability outputs that were converted into binary predictions.

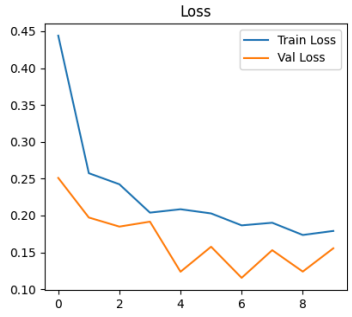
Scikit-learn's classification report and confusion matrix generate the evaluation metrics. Matplotlib and seaborn were used to plot confusion matrices and training curves to visualize the model's performance.

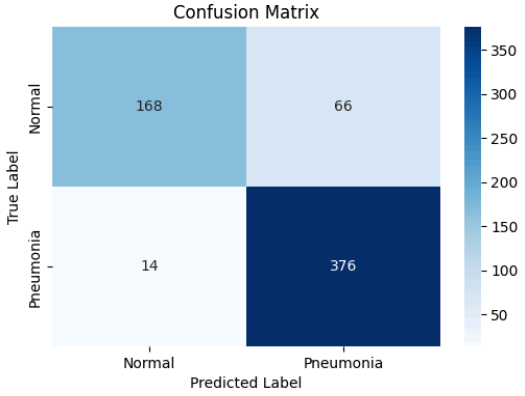
4. Results Analysis, Intuitions, Comparison

(i) The CNN model with dropout(0.4), epoch(10), and threshold(0.5) produced:

	Precision	Recall	F1 Score
Normal	0.92	0.72	0.81
Precision	0.85	0.96	0.90

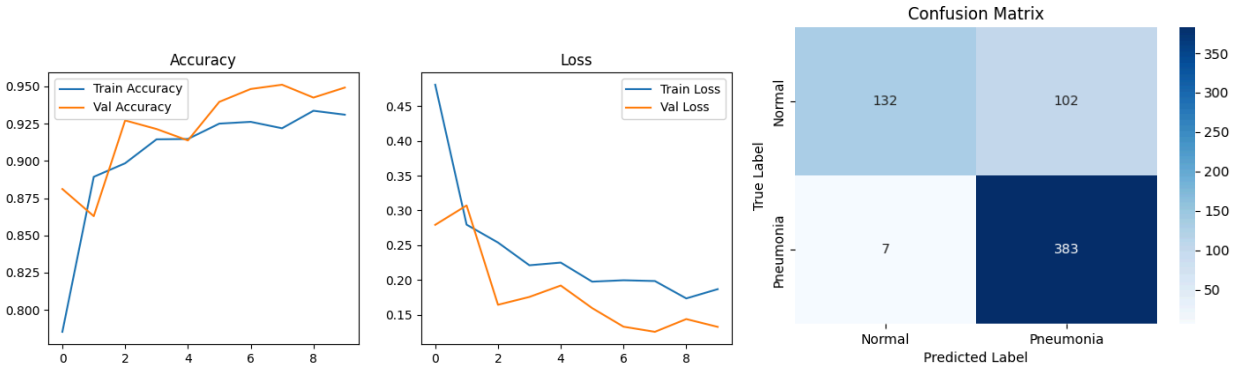






(ii) The CNN model with dropout(0.5), epoch(10), and threshold(0.5) produced:

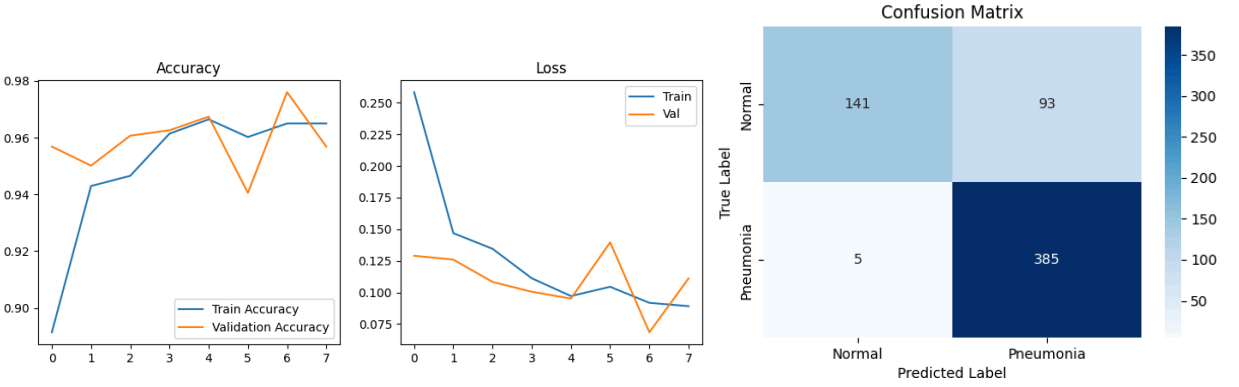
	Precision	Recall	F1 Score
Normal	0.95	0.56	0.71
Pneumonia	0.79	0.98	0.88



From the confusion matrix, the CNN model correctly identified 132 out of 234 normal cases and 383 out of 390 pneumonia cases. Although this model is very sensitive to pneumonia, with an accuracy of 83%, it still struggles to correctly identify normal images. False positives for pneumonia are moderately high at 102 and false negatives are low at 7, a desirable output from a diagnostic viewpoint.

(iii) The ResNet50 model produced these results:

	Precision	Recall	F1 Score
Normal	0.97	0.60	0.74
Pneumonia	0.81	0.99	0.89



The ResNet50 model demonstrates strong and stable performance, showing high overall accuracy and excellent sensitivity to pneumonia cases. The learning curves show no signs of overfitting, with training and validation accuracy/loss closely aligned throughout epochs. The performance is akin to CNN's performance, where the confusion matrix and class metrics indicate that the model identifies pneumonia reliably, and it also struggles with reliably identifying normal cases; however, it has a higher normal recall of 0.60. Compared to the custom CNN, ResNet50 delivers smoother learning and better generalization, confirming the advantages of deeper pretrained architectures in medical image classification.

Since the dataset is predominantly filled with pneumonia images, both models developed a predictive bias towards pneumonia. This explains the higher recall of pneumonia and the lower

recall of normal images. But this is a recognized limitation of medical datasets. This contributes to the higher pneumonia performance seen in both models, because pneumonia often produces clear, visual features such as opacities, whereas normal lungs vary more subtly.

Threshold sweeps demonstrated that adjusting the decision threshold changed performance balance without retraining the model. For example, increasing the threshold from 0.4 to 0.5 decreased normal recall dramatically–approximately 0.72 to 0.56 but slightly increased pneumonia recall by 0.01. This shows that the model is heavily biased because of the imbalance dataset. Both training and validation curves went down together, with the validation accuracy close to that of training accuracy. The learning curves do not show symptoms of overfitting, but both models exhibit a strong class bias toward pneumonia from dataset imbalance.

According to the accuracy/loss graphs, ResNet50 gave smoother learning curves, fewer fluctuations, and stronger generalization. As a transfer learning model, it was expected to show more meaningful improvement of normal class recall and balanced metrics compared to the CNN model because of its pretraining.

5. Conclusion

This project successfully implemented and evaluated two deep learning models for binary classification of healthy and pneumonia-infected chest X-rays. Both models achieved high pneumonia sensitivity, which is critical to clinical diagnostics and patient health. Due to dataset imbalance and difficulty in models to detect minute details, like differentiation in normal chest x-rays: the normal recall was lower. However, the ResNet50 model's recall was significantly better and had more stable generalization than the custom CNN. Threshold tuning demonstrated that model performance can be clinically calibrated without retraining. The CNN model provided an insightful baseline to larger and operationally deep learning models, while the ResNet50 model highlighted the realistic strengths in a clinically trained model. All in all, the project provided incredible insight into how deep learning models can be an effective tool to automate and supplement the diagnostic processes in the health industry. But like all models, issues grow from dataset imbalance and insufficient model calibration, which both heavily influence the classification of healthy normal chests to pneumonia-infected chests.

6. Contribution to Code

This project was individually conducted. All data preprocessing, CNN design, ResNet50 implementation, training, evaluation, tuning, visualizations, and report writing were completed solely by me.