CIS 522 – Final Project – Technical Report

Pneumonia Fighter

April 2022

Team Members:

- Yuxin Kan; irenekxx; Email: `irenekxx@seas.upenn.edu`
- Peihan Li; perhanli; Email: `peihanli@seas.upenn.edu`
- Changqi Xiao; chthon; Email: `chthon@seas.upenn.edu`

Abstract

We apply many models in this project to learn from the chest x-ray images of selected patients, varying from naive models to complicated models to diagnose pneumonia. In this process, we implement k-nearest neighbor(kNN), logistic regression, basic CNN, AlexNet, VGG19, and ResNet, respectively. We find that naive models like logistic regression and kNN cannot fully extract features from images of chest X-ray as the dataset is imbalanced. 74 percent of the images are labeled as pneumonia, leading to high false positive rate (FPR) in the confusion matrix of our naive models. When implementing transfer learning, we tried various methods to overcome the imbalance of our dataset. Data augmentation balanced the dataset to some extent. The regularization methods like dropout and early stopping also prevented the dataset from overfitting. To improve the generalizability and efficiency of our model, we attempted to freeze different numbers of layers of the pretrained model. By using AlexNet and ResNet, we ultimately got over 90 percent test accuracy and obtained a good F1-score of about 0.92. According to the statistics of CDC, there is an increasing number of people dying from pneumonia over the past few years. One of the key concerns is over the initial diagnosis of pneumonia that is essential to preventing the spread of this infectious disease. Our model enables rapid initial diagnosis of pneumonia, indicating the scientific and societal relevance of our research.

1 Introduction

Pneumonia is an infection that inflames the air sacs in one or both lung(s), causing the air sacs (alveoli) to fill up with fluid or pus. This infection can make it hard for the oxygen you breathe in to get into your bloodstream. According to the CDC, during 2018, 1.5 million people in the United States were diagnosed

with pneumonia in an emergency department. More than 40,000 people died from this disease the same year in the United States.

Obviously, Pneumonia can affect people of all ages and is an unignorable public health issue worldwide. Finding a more efficient way to diagnose pneumonia turns out to be the first and most crucial step in restricting the spread of the disease. Two major methods currently used to detect this infection are chest X-rays and blood tests. In this project, we focus on the former to train data for a more efficient diagnosis of pneumonia. To be more specific, we hope to identify a comprehensive chest x-ray deep-learning model that assists the diagnostic accuracy in the preliminary stage.

Data we use for this project are 5,856 chest X-rays, obtained from the retrospective cohorts of pediatric patients in Guangzhou Women and Children's Medical Center, China, with an age range of 1 to 5. The dataset is quite imbalanced: around 74% images are labeled as Pneumonia.

To achieve high test accuracy, we applied a wide range of models, varying from naive models (e.g.: logistic regression) to more complicated models (e.g.: AlexNet and ResNet). When implementing transfer learning, we tried to overcome the imbalance of our dataset through data augmentation. We also tried unfreezing certain layers of pretrained network during training process. We made great efforts to tune the parameter and ultimately reached an accuracy of over 90% with AlexNet and Resnet.

2 Related Work

There are many methods on Kaggle and some of them also get an accuracy on test data. By referring to their work, we collect different methods from naive to complex, reproduce some of them, combine the advantages of them and build our own model.

To overcome the imbalance of the model, we use data augmentation methods like random cropping and random flipping to balance the data. Besides data augmentation, we refer to the work by Mateusz Buda[1]. We tried multiple ways to overcome this problem and finally come to the best method.

By combining and implementing different CNN models and trying different ways of balancing the datasets, our model can outperform most of current models.

3 Dataset and Features

This dataset is aimed to precisely identify medical diagnoses and treatable diseases, specifically pneumonia, by imaged-based learning from chest X-rays.

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care.

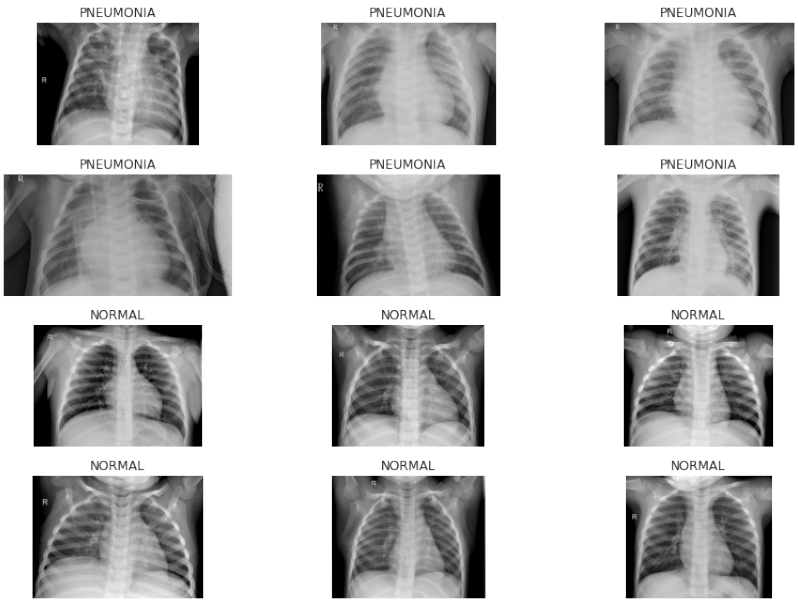

Figure 1: Samples

A total of 5,856 chest X-ray images from children are collected and labeled, including 3,883 characterized as depicting pneumonia (2,538 bacterial and 1,345 viral) and 1,349 normal. The datasets are organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal).

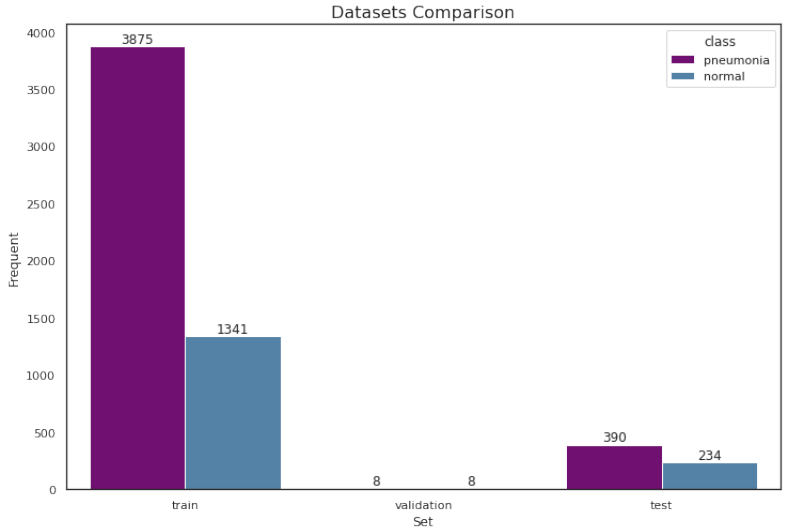

Figure 2: Datasets Comparison

Specifically, there are 5216, 16, 624 images in our training set, validation set, and test set respectively. As for our training set, the ratio of pneumonia samples to normal samples is about 2.89. One of the challenges of this dataset is class imbalance. Around 74% of images are labeled as Pneumonia and 26% are normal.

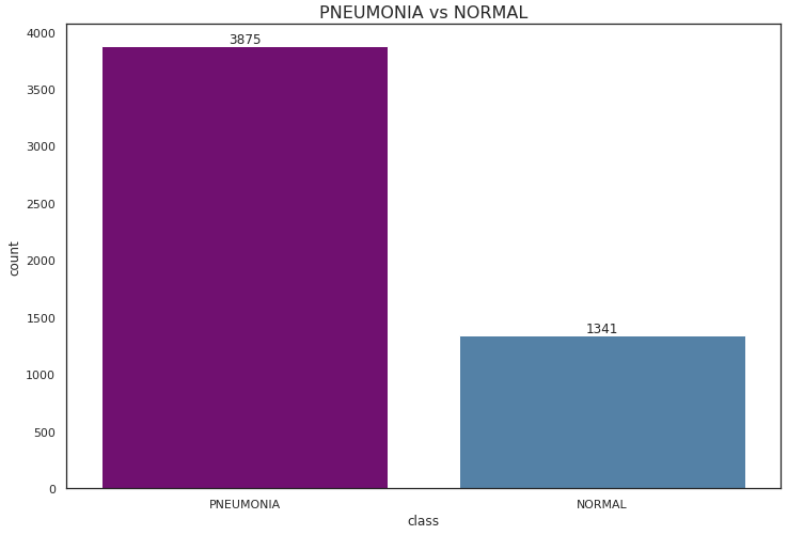

Figure 3: Pneumonia vs Normal

Another difficulty of handling the data is that image sizes are completely different between normal and pneumonia cases, so we have to cut the image reasonably to make them the same size without losing information.

Features like potential pathology on a tissue map that are hard for eyes to tell should be effectively identified by our models to make a referral decision with performance comparable to human experts, enabling timely diagnosis of irreversible severe lung loss.

4 Methodology

In our experiments, we mainly applied a general deep learning framework, especially convolutional neural network, to address this classification problem, as the development of CNN layers has allowed for significant gains in the ability to classify images and detect objects in a picture. To define a more composite representation by combining convolutional neural network with the basic framework, we also tried several non-deep learning models as benchmarks (kNN and logistic model). Moreover, to improve the performance and address lack of data, we utilize advanced techniques like transfer learning and fine-tuning.

Experimental results demonstrate that the proposed models substantially outperform baselines. Now let us analyze the reasons as follows.

As mentioned above, we employed kNN and logistic model as benchmarks. The k-nearest neighbors algorithm (kNN) is a simple non-parametric supervised learning method to solve classification problems. kNN algorithm is faster than deep learning models as it doesn't require training steps and does not derive any discriminative function from the training data. However, compared to deep learning, it does not work well with large and high-dimensional dataset. We chose k to be 10 as it performed best.

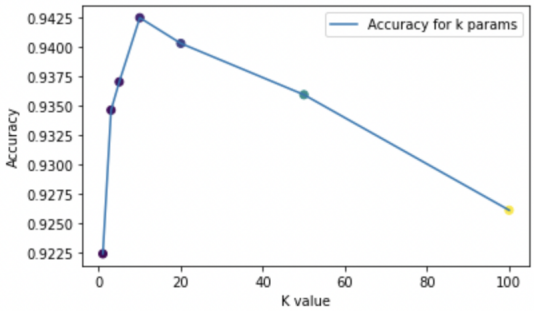

Figure 4: Accuracy for k parameters

Similarly, logistic model is a statistical model using logistic regression. It is suitable for classification problems with two possible outcomes. However, the major limitation of logistic regression is the assumption of linearity between the dependent variable and the independent variables. In real-world scenarios, linearly separable data is rarely found.

Convolutional neural network performs much better than non-DL models. Multiple processing layers to which image analysis filters are applied to convolve multiple filters across the image and produce a feature map that is used as input to the following layer. This architecture makes it possible to process images in the form of pixels as input and to give the desired classification as outputs. The first convolutional layer of a CNN is essentially a standard image filter with a ReLu activation function. Its goal is to take a raw image and extract basic geometric features. At the end of first layer, we applied a downsampling strategy like max pooling. The second convolutional layer accepts the features extracted by the first as its input, which allows it to combine these basic shapes into more complex features.

Finally, we utilized advanced techniques including transfer learning and fine-tuning to improve the performance. Transfer learning has proven to be a highly effective technique, particularly when faced with domains with limited data. Rather than training a completely blank network, by using a feed-forward approach to fix the weights in the lower levels already optimized to recognize the

structures found in images in general and retraining the weights of the upper levels with back propagation, the model can recognize the distinguishing features of a specific category of images, much faster and with significantly fewer training examples and less computational power.

We mainly used transfer learning by importing and comparing three pre-trained models: AlexNet, VGG, and ResNet. In particular, we fine-tune these pre-trained models on our datsets by replacing and retraining the parameters of the output, fully-connected layer of the pre-trained model, while freezing the other layers. By replacing the last layer, we can do binary classification with two outputs.

AlexNet, VGG and ResNet are very popular CNN models. AlexNet and ResNet have similar amount of parameters. However, the main base element of ResNet is the residual block. In other words, when going deeper, ResNet requires a lot of computations which means more training time and energy required. Compared to them, VGG not only has a higher number of parameters but also has a decreased accuracy. It takes more time to train a VGG with reduced accuracy.

Apart from architecture each model applies, for fine-tuning, other parameters are Adam optimizer, cross-entropy loss as our loss function, initial learning rate of 1e-5 and early stopping. The regularization used in AlexNet is L2 with a weight decay of 1e-5. According to analysis above, VGG requires more time to train so we assigned it epoches of 20, larger than the other two models.

5 Results

To evaluate different models, accuracy was measured by dividing the number of correctly labeled images by the total number of test images. Overall, kNN can only achieved an accuracy of 0.77 with precision of 0.73, recall of 0.99 and F-score of 0.99. Logistic Regression has similar performance of 0.78. Our vanilla CNN model yields an accuracy of 0.78 but obviously performs much better according to its confusion matrix. However, the ratio of false positivity is still pretty high. By employing a transfer learning framework and fine-tuning of pre-trained models, our experiments demonstrated compelling performance. Both AlexNet and ResNet achieved high accuracy of 0.90, while VGG also reached accuracy of 0.87. The detailed visualizations of the training process and results are presented below.

6 Discussion

In our project, we aimed to design and train a model to maximize the accuracy of distinguish pneumonia samples from normal samples.
First, we were provided three datasets (train, val, test) by Kaggle. Considering the size of these datasets, we did some data augmentation, such as rotation, flipping.

Next, we trained our non-DL (kNN, Logistic Regression), our base-DL(Basic CNN), and our advanced models (finetuned AlexNet, VGG19, Resnet) on the training set and the validation set, and tested our models on the test set. To improve the performance of our models, we have tried different finetuning methods (finetune different layers), learning rates and regularization methods (dropout, early stopping).
Finally, we compared the performance of all those models and came out with our best model.

6.1 Findings

We have done some experiments on our models, and here are the results.

6.1.1 kNN

From Figure 4 above, we can see that when k = 10, the accuracy is the highest, so we choose k = 10. We tested our model with this k on the test set.

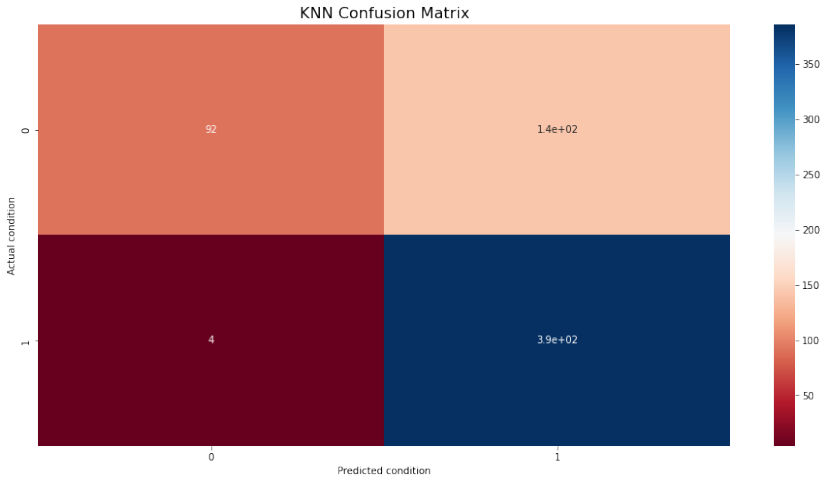

Figure 5: kNN Confusion Matrix on the Test Set

And here are some metrics of this model:

- Accuracy: 0.77
- Precision: 0.73
- Recall: 0.99
- F-Score: 0.84

From the metrics above, we can know that this model has a high false positive rate, and the accuracy and the precision is not ideal.

6.1.2 Logistic Regression

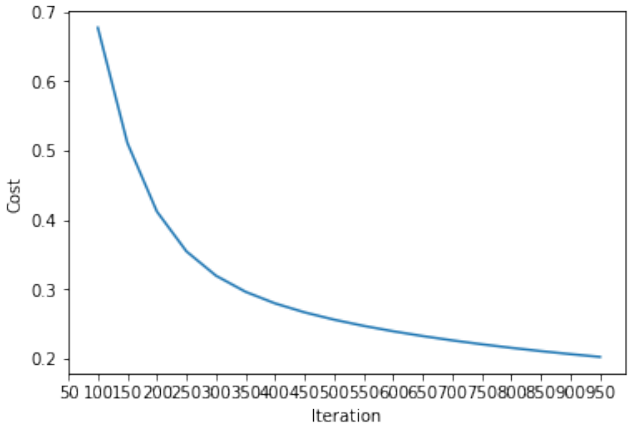

Figure 6: Accuracy of Logistic Regression

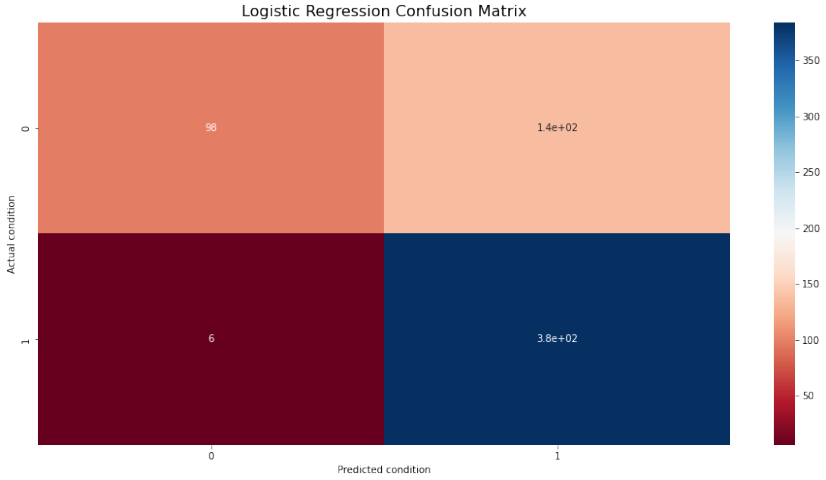

Figure 7: Logistic Regression Confusion Matrix on the Test Set

And here are some metrics of this model:

- Accuracy: 0.78
- Precision: 0.74
- Recall: 0.98
- F-Score: 0.84

From the metrics above, we can know that this model has similar performance with kNN.

Now, we can conclude that the capability of these simple models is not enough for our image classification problem.

6.1.3 Basic CNN

The basic CNN is composed of two convolutional layers and two fully connected layers as well as max-pooling layers and ReLU activation layers. The detailed architecture is as followed:

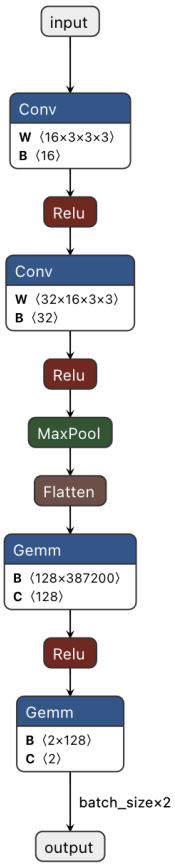

Figure 8: Basic CNN Architecture

And we set the learning rate $= 10^{-5}$, the training epoch $= 20$, the optimizer as Adam. Through training, we have loss and accuracy as followed:

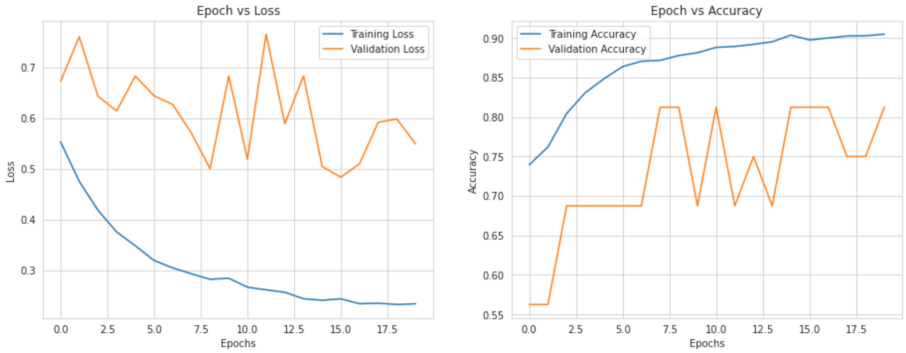

Figure 9: Accuracy and Loss of Basic CNN

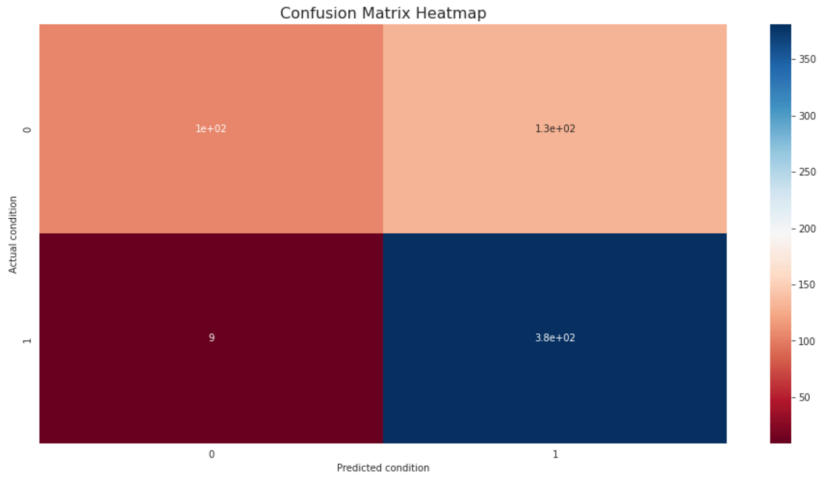

Figure 10: Basic CNN Confusion Matrix on the Test Set

And here are some metrics of this model:

- Accuracy: 0.78
- Precision: 0.75
- Recall: 0.98
- F-Score: 0.86

From the metrics above, we can know that this model has similar performance with the above two no-DL models. So next we need to increase the complexity of the model.

6.1.4 AlexNet

We built our model on the pretrained $AlexNet$[2], and we replaced the last fully connected layer with a 512×2 linear layer. We set the training epoch = 15, the optimizer as Adam. Here are some experiments result on this model by tuning hyperparameters.

learning rate = 10^{-5}, early-stopping

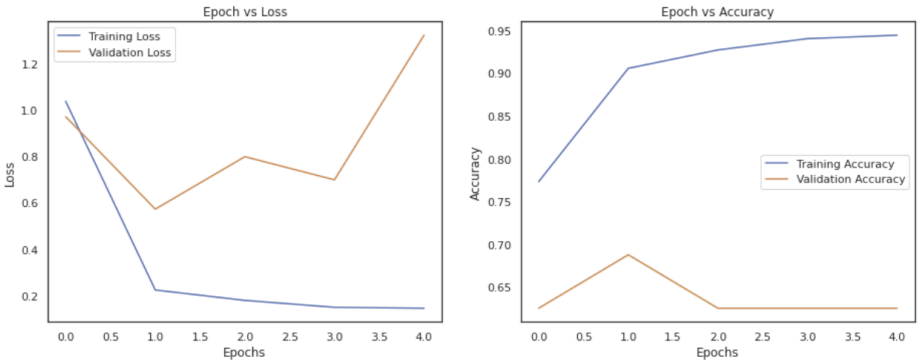

Figure 11: Accuracy and Loss of AlexNet (lr=1e-5, early-stopping)

Test Accuracy: 0.80

learning rate = 10^{-5}, learning rate scheduler, early-stopping

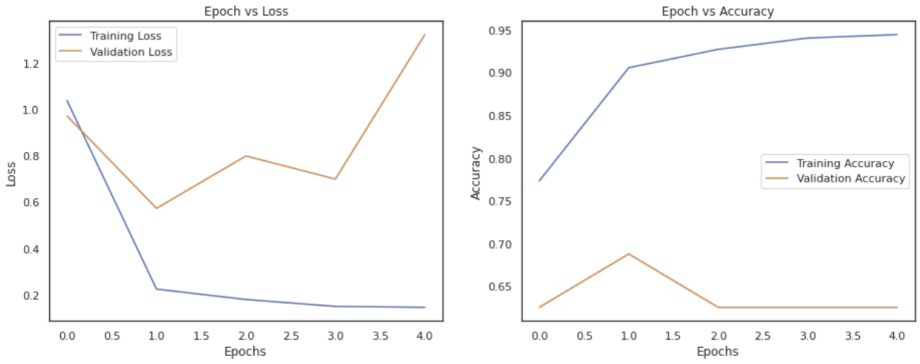

Figure 12: Accuracy and Loss of AlexNet (lr=1e-5, lr_scheduler ,early-stopping)

Test Accuracy: 0.89

learning rate = 10^{-4}

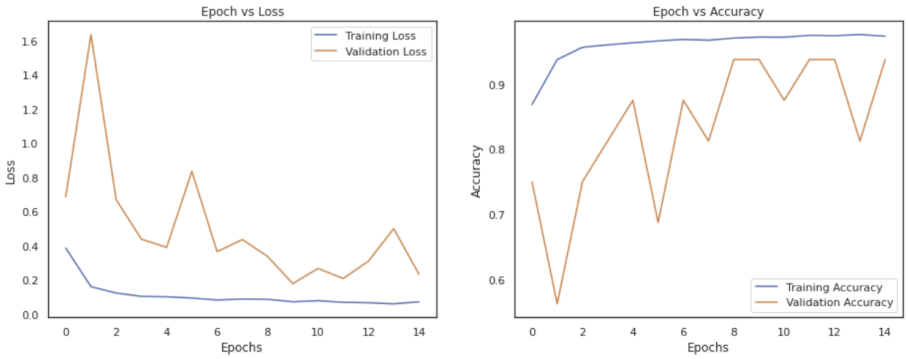

Figure 13: Accuracy and Loss (lr=1e-4) of AlexNet

Test Accuracy: 0.91

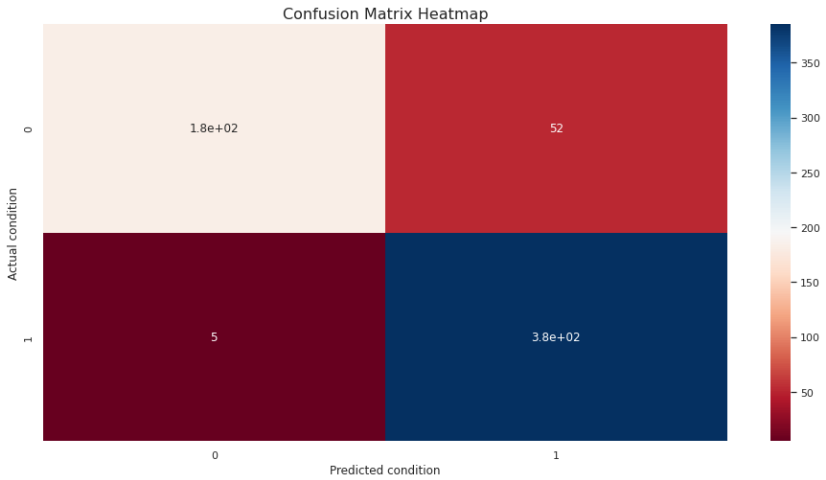

Figure 14: AlexNet Confusion Matrix on the Test Set

- Accuracy: 0.91
- Precision: 0.88
- Recall: 0.99
- F-Score: 0.93

From the figures above, we can conclude our AlexNet based model will have the best performance under a relatively big learning rate without early-stopping, which has an accuracy of more than 0.91.

6.1.5 VGG19

We built our model on the pretrained $VGG19$[3]. We set the training epoch = 20, the optimizer as Adam. And we've tried several finetuning policies, including finetuning the whole classifier or only finetuning the last linear layer and we've experimented on different sets of learning rates and regularization. Here are some results.

finetune classifier, learning rate = 10^{-4}

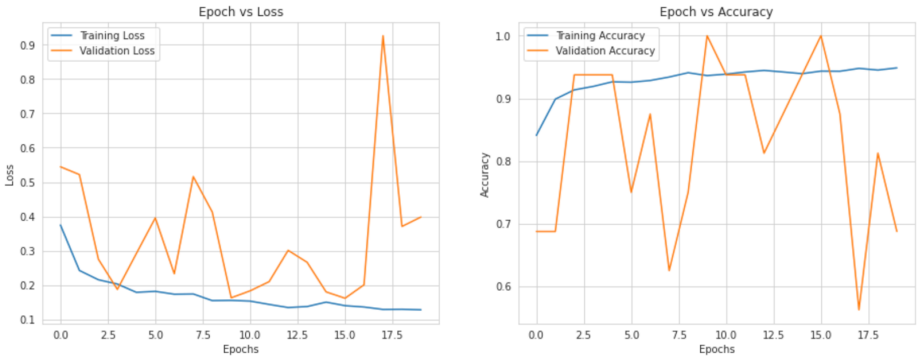

Figure 15: Accuracy and Loss of VGG19 (finetune classifier, lr=1e-4)

Test Accuracy: 0.79

finetune last layer, learning rate = 10^{-5}**, learning rate scheduler**

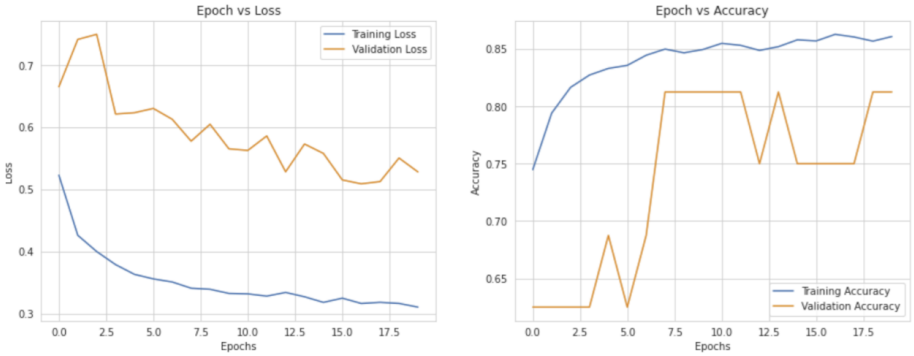

Figure 16: Accuracy and Loss of VGG19 (finetune last layer, lr=1e-5, lr_scheduler)

Test Accuracy: 0.82

finetune classifier, learning rate = 10^{-4}, learning rate scheduler, early-stopping

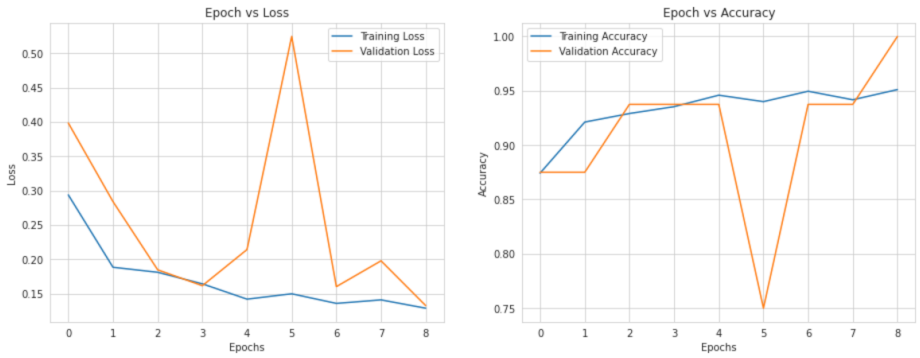

Figure 17: Accuracy and Loss of VGG19 (finetune classifier, lr=1e-4, lr_scheduler, early-stopping)

Test Accuracy: 0.87

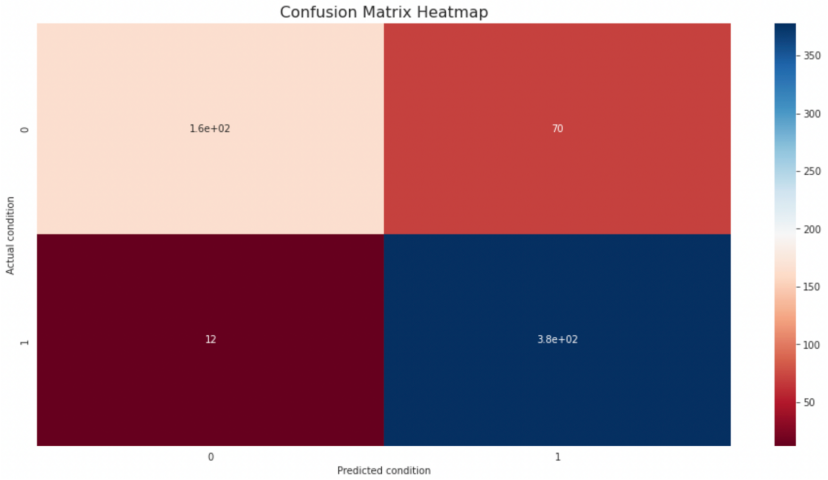

Figure 18: VGG19 Confusion Matrix on the Test Set

And here are some metrics of this model:

- Accuracy: 0.87
- Precision: 0.84
- Recall: 0.97
- F-Score: 0.80

From the figures above, we can conclude our VGG19 based model will have the best performance under a relative big learning rate with learning rate scheduler and early-stopping, which has an accuracy of 0.87. But compared to AlexNet, the pretrained VGG19 is harder to finetune.

6.1.6 ResNet

The difference between ResNet and other neural networks lies in the residual block. In traditional neural networks, each layer feeds into the next layer. In a network with residual blocks, each layer feeds into the next layer and directly into the layers about 2–3 hops away. The residual block looks like the following image.

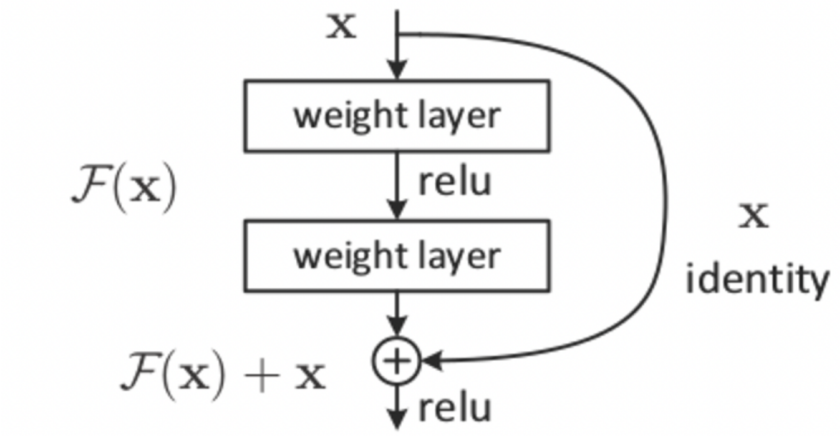

Figure 19: The structure of Resdual Block

ResNet with large number (even thousands) of layers can be trained easily without increasing the training error percentage and help in tackling the vanishing gradient problem using identity mapping.

We built our model on the pretrained $ResNet^{[4]}$, and we replaced the last fully connected layer with a 2048×2 linear layer. We set the learning rate $= 10^{-4}$, the training epoch $= 20$, the optimizer as Adam. And we applied L2 regularization on our model by setting weight_decay $= 1e - 4$. Here are some experiment results.

unfreeze last 3 layers of ResNet

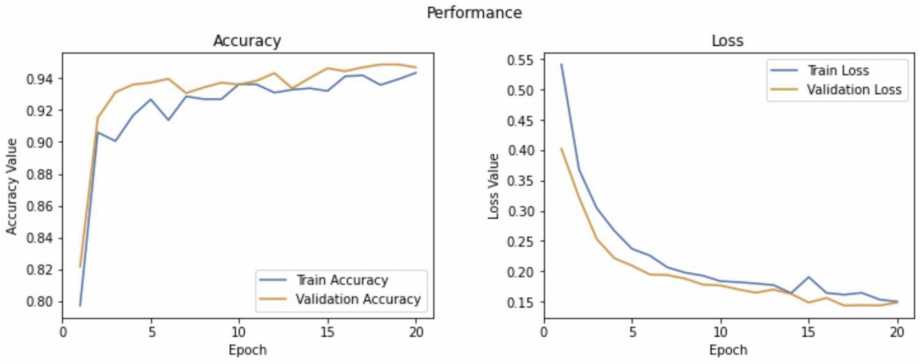

Figure 20: Accuracy and Loss of ResNet (unfreeze some layers)

Test Accuracy: 0.83

finetune last layer

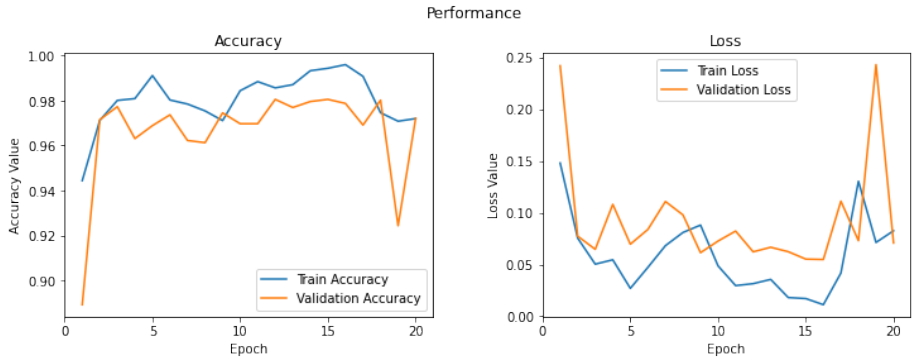

Figure 21: Accuracy and Loss of ResNet (finetune last layer)

Test Accuracy: 0.90

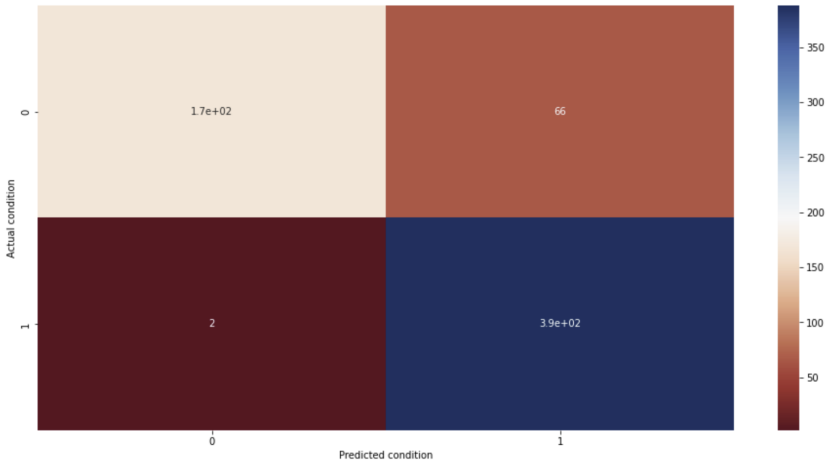

Figure 22: ResNet Confusion Matrix on the Test Set

And here are some metrics of this model:

- Accuracy: 0.90
- Precision: 0.99
- Recall: 0.85
- F-Score: 0.92

From the metrics above, we can know that this model has good performance on our problem as well. So we'll compare our models and figure our why we got these results.

6.1.7 Comparison and Implication

In summary, we have these results on the test set: By comparing these metrics,

Model	kNN	Logistic Regression	Basic CNN	AlexNet	VGG19	ResNet
Accuracy	0.77	0.78	0.78	0.91	0.87	0.91
Precision	0.73	0.74	0.75	0.88	0.84	0.85
Recall	0.99	0.98	0.98	0.99	0.97	0.99
F-Score	0.84	0.84	0.86	0.93	0.80	0.92

Table 1: Comparison of Metrics between Models

we can have these conclusions:

- The capability of our non-DL models is not enough to handle our problem.
- A basic four-layer CNN will underfit our problem.

- The preset parameters of VGG19 does not fit our dataset very well.
- AlexNet and ResNet have similar performance. But during the experiments, we found that the ResNet based models were more stable and could perform similarly under different randomly divided datasets.

We would recommend ResNet based models for this problem instead of CNN based model.

In general, ResNet is an ensemble of many relatively shallow residual blocks, and there exists a 'Skip Connection' (identity mapping) between the input and the output of each residual block, which just adds the output from the previous layer to the layer ahead. These shortcuts make a very deep network possible, and such a deep ResNet structure is complex enough to learn the distribution of our data. Through tuning and some regularization methods, this ResNet model can have very good and stable performance on our problem.

6.2 Limitations and Ethical Considerations

The main limitation of our model is that it can only discriminate between diseased and non-diseased individuals, i.e. our model is a binary classifier. In this case our model provides limited information. In a practical application scenario, it is often necessary to know not only whether a patient is infected with pneumonia, but also a lot of other information, such as the severity of the disease and the specific subtype of pneumonia. Therefore, only the binary results are of limited use in practical applications. Other limitations include the difficulty of collecting unbiased data, privacy concerns.

In term of ethical issues, our model might be biased because of the imbalance and bias of our current data. This might have negative implications if it's applied to real-world scenarios.

One the one hand, collecting chest xray images of patients infected with pneumonia is more difficult than collecting those of normal people, and if one person needs to do a chest x-ray, it's more likely that his pneumonia has reached the mid-stage. Thus, our model might not be able to detect early stages of pneumonia and might tend to give negative (not illed) results for unknown symptoms or marginalized groups, causing higher false negative rate, and this can cause physical and financial damage to patients, can also cause reputational and financial damage to the hospitals.

On the other hand, there may also be privacy concerns with collecting the patient's chest x-ray, where our model might be used in unintended, negative ways. Body data is very sensitive information, so improper data collection measures or data utilization methods of our model can lead to the public's private information being used for advertising, intimidation, and other inappropriate situations.

In term of ethical issues, our model might be biased because of the imbalance and bias of our current data. If it's applied to real-world.

6.3 Future Research Directions

To overcome the limitations of our model, the first issue we need to focus on is about data. We can try model data augmentation methods, such as resize, Gaussian blur, sharpness adjustmentm to balance our data. And considering our small dataset, a better choice may be to fuse the training and validation sets and increase the utilization of the data by cross-validation. If we need to collect more data, an important step in this process lies in the proper selection by computer vision specialists together with radiologists and medical physicists of an adequate training set. Special care must be taken to select the training set minimizes the potential for biases in the resulting models. Other improvements can be achieved by selecting other pre-trained models, adjusting the structure of the fine-tuned models, and tuning the hyperparameters.

Since our model is trained on the chest xray images, it can be used to diagnose other conditions, such as cancer, infection or air collecting in the space around a lung, by applying transfer learning.

7 Social Impact

Pneumonia is an infection that inflames the air sacs in one or both lungs. The infection causes the lungs' air sacs to become inflamed and fill up with fluid or pus. That can make it hard for the oxygen you breathe in to get into your bloodstream. According to the CDC, In the United States, 1.5 million people were diagnosed with pneumonia in an emergency department during 2018. More than 40,000 people died from the disease that year in the United States. Obviously, pneumonia can do harm to the world and it is a must to restrict the spread of it. So it is crucial to find a way to diagnose pneumonia efficiently. Currently, there are two main methods to detect the bacteria. They are chest X-rays and blood tests. In this project, we will focus on how to diagnose pneumonia from chest X-rays of patients. Precisely diagnosing pneumonia is the premise of preventing its spread, which has a profound social impact.

8 Conclusions

In this project, we trained various models to diagnose pneumonia based on the chest X-rays of patients. Models we apply vary from naive models (e.g.: logistic regression) to complicated models (e.g.: AlexNet and ResNet). For kNN, logistic regression, and basic CNN, the accuracy of predicting pneumonia is under 80 percent. This suggests that these models do not improve much as compared to prior diagnosis methods. As can be observed from the confusion matrix, the number of false positives (FP) is high for these models due to the imbalance of our training data. More complicated models do learn the features of x-ray images though. Specifically, the accuracy of VGG19 is about 87 percent. The accuracy of both AlexNet and ResNet is 91%. The number of FP decreases substantially after we implement data augmentation and regularization.

9 References

[1] Buda, M., Maki, A.& Mazurowski, M. (2017). A systematic study of the class imbalance problem in convolutional neural networks.
[2] Krizhevsky, A., Sutskever, I. & Hinton, E.G. (2012). ImageNet Classification with Deep Convolutional Neural Networks.
[3] Simonyan, K.& Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.
[4] He, K., Zhang, X., Ren, S. & Sun, J.(2016). Deep Residual Learning for Image Recognition.