Classifying Pneumonia from X-ray Images

PneumoniaHacker

Zehang Chen[a], Anni Pan[b], Mary Qin[b]

[a] *Department of Surgery, Cardiovascular Division, The Hospital of the University of Pennsylvania, Philadelphia, PA 19104*

[b] *University of Pennsylvania, Philadelphia, PA 19104*

P R O J E C T I N F O

Keywords:
Pneumonia
Chest X-ray Images
Deep Learning
Transfer Learning

C O N T E N T S

A B S T R A C T

Introduction Since the pandemic declaration from the World Health Organization in March 2020, COVID-19 has infected over 150 million people worldwide, causing the death of 90 million. Surging patient numbers levy a heavy burden on the healthcare systems, leading to physician burnout, delayed diagnosis, and treatment, which impedes response fighting the spread. Pneumonia is a common symptom amongst COVID-19 patients. Chest X-ray, the standard exam to diagnose pneumonia, dramatically assists in screening and treating COVID-19 infection. We aim to develop an algorithm to assist the classification of X-ray images in accelerating the classification of X-ray images to accelerate diagnosis, alleviating physician workload.

Methods We used a dataset from the Kaggle competition containing 5,968 labeled chest X-ray images and implemented a non-deep learning baseline model (multiple logistic classifier from `sklearn`). We then trained two basic deep learning models: a feed-forward neural network and a convolutional neural network. Subsequently, we applied a transfer learning model based on `ResNet18`, followed by a novel architecture model with a custom activation function, `ALReLU`. Model performances were assessed using testing accuracy and confusion matrix. Training curves and network filters were also plotted for model interpretation.

Results Non-deep learning classifier established a baseline testing accuracy of 61.538%. The basic deep learning model improved the performance (testing accuracies: feed-forward neural network, 65.7%; convolutional neural network, 72.3%). The novel architecture model further increased the testing accuracy to 81.571%. `ResNet18`-based transfer learning model achieved the highest accuracy (98.435%).

Conclusion We successfully developed a deep learning network for classifying chest X-rays into normal, viral, and bacterial pneumonia and expect the algorithm to be helpful in clinical use with physician supervision. We believe our algorithm can increase the speed and accuracy of pneumonia diagnosis.

1. Introduction

Since the World Health Organization announced the first cases in China, COVID-19 has spread around the globe. As one of the countries severely impacted by the disease, the United States has reported over 32 million cases to date, suffering over 570,000 tragical deaths [1]. Despite this pandemic initiated over last year, there are still considerable numbers of new cases reported daily [2]. The overwhelming number of patients levy significant stress on the healthcare system, which may lead to physician burnout and consequentially result in decreased quality of care, patient safety, physician turnover, and patient satisfaction [3–7]. Additionally, COVID-19 is highly contagious. Delayed diagnosis and treatment are likely to result in a broader spread and lead to more severe or even fatal clinical events.

One of the prominent conditions of COVID-19 patients is pneumonia. Chest X-rays, a standard procedure to diagnose pneumonia, are also frequently used as a standard test for suspected patients. In order to reduce physician workload and assist in accelerating diagnosis, we aim to develop an algorithm that assists in classifying chest X-rays into normal, viral pneumonia and bacterial pneumonia.

2. Dataset

We used the dataset provided from a Kaggle competition for training and testing the algorithm [8]. The dataset contains 5,908 labeled chest X-ray images that were pre-classified into training (N=5,284) and testing set (N=624). Training set images were classified into normal (N=1,342) and pneumonia (N=3,944), with pneumonia patients further categorized into stress-smoking induced (N=2), viral (N=1,407), and bacterial (N=2,535) based on cause. Testing set images were labeled similarly (normal, N=234; pneumonia, N=390 [viral, N=148; bacterial, N=242]). As cases labeled as stress-smoking induced pneumonia is scarce and only present in the training set, they were excluded from this study. Samples of representative images from each remaining group were shown in Figure 1. For convenience, numerical coding was used for groups: normal patients were labeled 0, while patients with viral and bacterial pneumonia were labeled 1 and 2, respectively.

3. Methods

3.1. Baseline: non-deep learning classifier

To establish a baseline for all deep learning classifiers, we first implemented a non-deep learning classifier, namely the multiple logistic classifier from `sklearn` [9]. We used its performance as a reference for evaluation. To better understand the classifier, we also plotted the resulting filters after training. The confusion matrix, as well as accuracy, were calculated for evaluation. It is worth noting that the baseline model took extensive time to train. Hence, the trained classifier was saved to expedite reproduction of the test accuracy calculation.

3.2. Basic deep-learning model: feedforward neural network

We implemented a basic feed-forward neural network with three fully connected linear layers for the basic deep model. In the for-

Email addresses: chenze@pennmedicine.upenn.edu (Z. Chen), annipan@seas.upenn.edu (A. Pan), qinyh@seas.upenn.edu (M. Qin).
All authors equally contributed.

Figure 1: Sample images from dataset

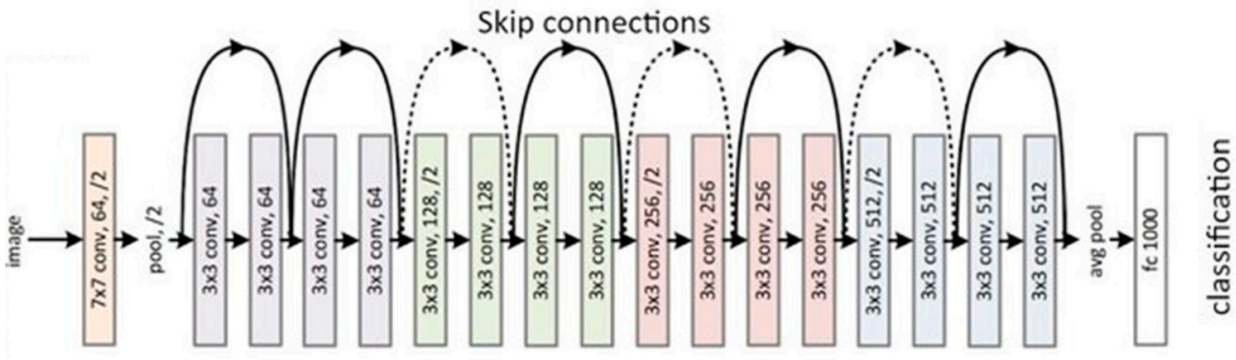

Figure 2: The architecture of `ResNet18`

ward function, we applied `ReLU` regulation after each layer. We used a cross-entropy loss and SGD optimizer with the epoch number of 20.

We split the training data into training (N=4,484) and validation (N=800) portion and kept the same test dataset for comparing performance with other models. The loss and accuracy from training and validation datasets were recorded and plotted during the training step. The test accuracy and confusion matrix were calculated for evaluation.

3.3. Basic deep-learning model: convolutional neural network

We created a convolutional neural network with the following architecture:

- Convolution (input=1, output=4, kernel_size=3)
- Convolution (input=4, output=8, kernel_size=3)
- Pool Layer
- Fully Connected Layer (127 008, 128)
- Fully Connected layer (128, 3)

The parameters were chosen based on consideration of both performance and random access memory (RAM) limitations. Cross-entropy loss and SGD optimizer were used in the model.

Similar to the feedforward neural network, we split training data into training and validation portions. We recorded the loss and accuracy from training and validation datasets and then calculated the test accuracy and confusion matrix for evaluation. Finally, to enhance the interpretability of the neural network, we plotted the feature maps after training.

3.4. Advanced model: `ResNet18`

We applied transfer learning with the pre-trained `ResNet18` model in Pytorch as our advanced model [10]. The model works by training the layers to fit a residual mapping. The model structure and the dimensions of each layer are shown in Figure 2 [11].

We used batch gradient descent with a batch size of 16 and chose Cross entropy loss and SGD optimizer for training. Training the model takes about 3 hours, so we decided to save the model and retrieve them for further analysis. We recorded the loss and accuracy from training and test sets at every epoch and then calculated the test accuracy and confusion matrix for evaluation.

To visually understand the components of our network which classifies the images, we used Grad-CAM [12]. It identifies target concepts with crucial gradients flowing into the final convolutional layer and produces a coarse localization map highlighting corresponding regions in the images that are related to the concept [13].

3.5. Novel architecture model

We implemented a convolutional neural network with a custom activation function, `ALReLU` [14]. The `ALReLU` is a modification of `ReLU` and is mathematically defined as follows:

$$f(x) = \begin{cases} x & \forall x > 0 \\ |\alpha x| & \forall x \leq 0 \end{cases}, \text{where } \alpha = 0.01$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} f(x) = \begin{cases} 1 & \forall x > 0 \\ -\alpha & \forall x \leq 0 \end{cases}, \text{where } \alpha = 0.01$$

The model structure, as suggested Mastromichalaki et al., is as below:

- Five convolutional layers, the first of which has a kernel size of 5×5 and the remaining 3×3. Their out channel sizes are 32, 64,

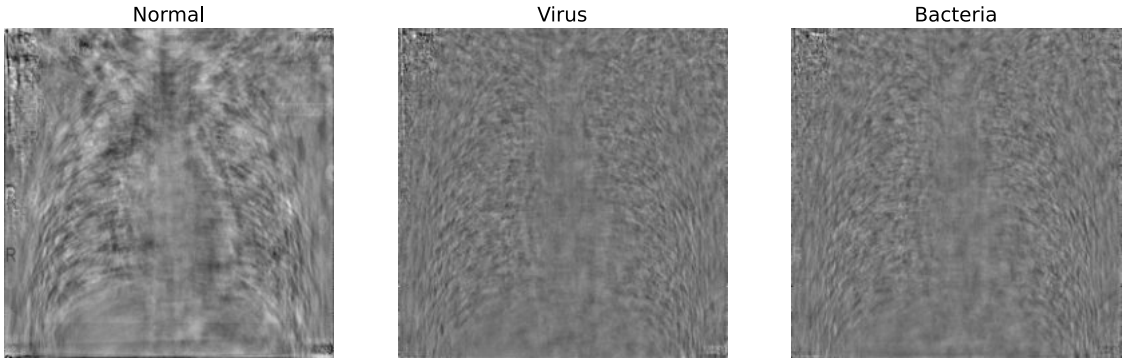

Figure 3: Filters for multiple logistic classifier

128, 256, and 512, respectively. Each convolutional layer is followed by a max-pooling layer, a batch normalization layer. A dropout layer with dropout rates being 0.1, 0.2, 0.3, 0.4, and 0.5, respectively and then an `ALReLU` layer.

- A global average pooling layer followed by `ALReLU`, batch normalization, and dropout with a rate of 0.3.
- A linear layer with 256 outputs, followed by an `ALReLU`, a layer norm, a dropout with a rate of 0.4, a linear layer with three outputs, and finally a `softmax` layer.

We split training data into training and validation portions similar to the feedforward neural network and the convolutional neural network. We used batch gradient descent with a batch size of 16 and chose Cross entropy loss and SGD optimizer for training. We recorded the loss and accuracy from training and validation datasets and then calculated the test accuracy and confusion matrix.

4. Results

4.1. Baseline: non-deep learning classifier

The non-deep learning classifier was the multiple logistic classifier from `sklearn`. After training, the filter used to predict all three groups were plotted and shown in Figure 3. The overall test accuracy was 61.538%, and the corresponding confusion matrix was shown in Figure 4. The accuracy varied significantly between groups. The accuracy for normal patients was only 35.470%, which was close to random guessing. The overall model accuracy for pneumonia patients was 77.179%, almost doubling the accuracy for normal patients. Model performance was suboptimal for viral pneumonia patients (57.432%). Patients with bacterial pneumonia had the highest accuracy (89.256%).

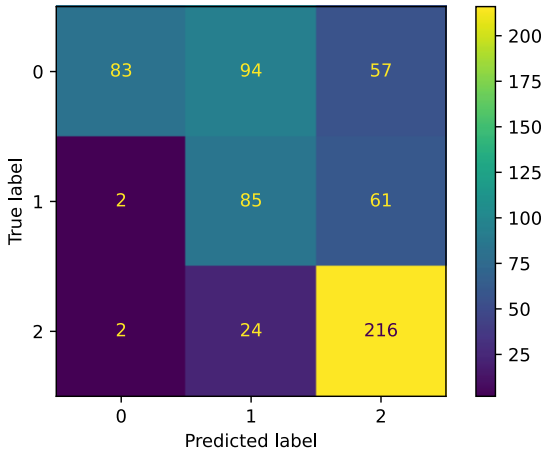

Figure 4: Confusion matrix for multiple logistic classifier
Labels: 0, normal; 1, viral pneumonia; 2, bacterial pneumonia.

4.2. Basic deep-learning model: feedforward neural network

We trained the feedforward neural network on training and validation datasets with 20 epochs. As shown in Figure 6, the training loss steadily decreased as epoch increased, while the validation loss first decreased then remained relatively constant. Similarly, the training accuracy increased to about 87%, while the validation accuracy was relatively constant at around 75% as plotted in Figure 7.

The feedforward neural network had an overall test accuracy of 65.705%, and the confusion matrix was shown in Figure 8. The accuracy varied between groups. The accuracy for normal patients was 37.179%, which was similar to random guessing. The overall model accuracy for pneumonia patients was 82.821%, which was more than double the normal cases. In particular, the accuracy was 71.622% for identifying viral pneumonia and 89.669% for bacterial pneumonia.

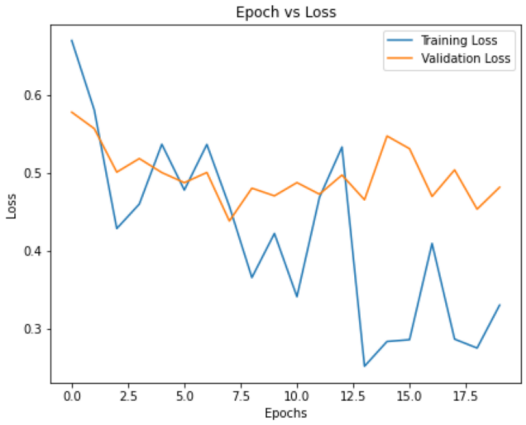

Figure 6: Loss for feedforward neural network

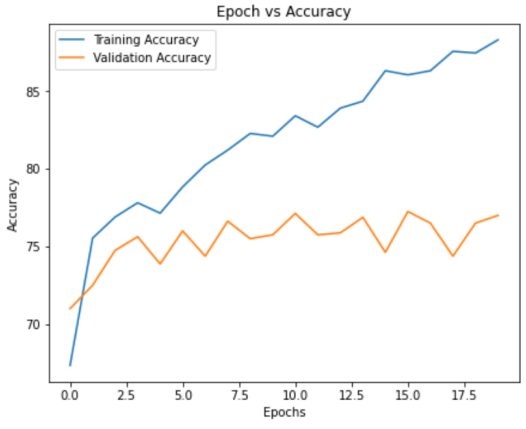

Figure 7: Accuracy for feedforward neural network

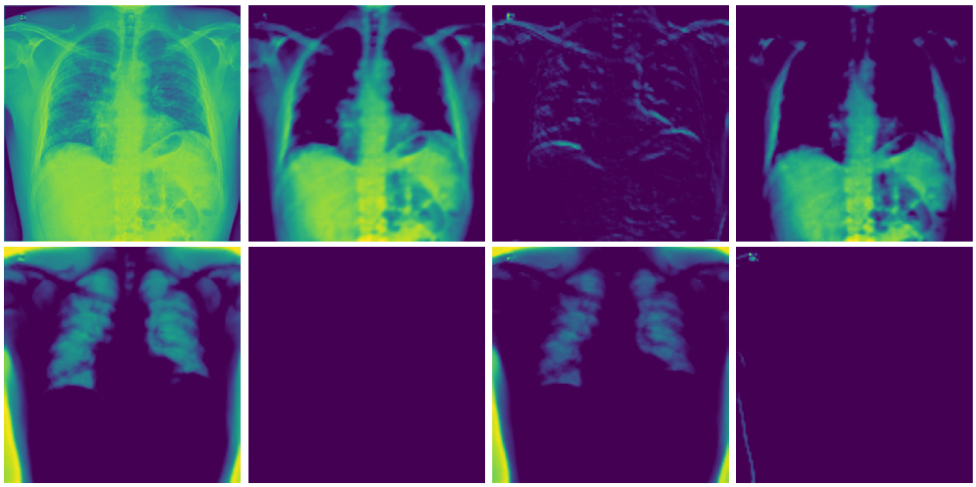

Figure 5: Convolutional neural network feature maps

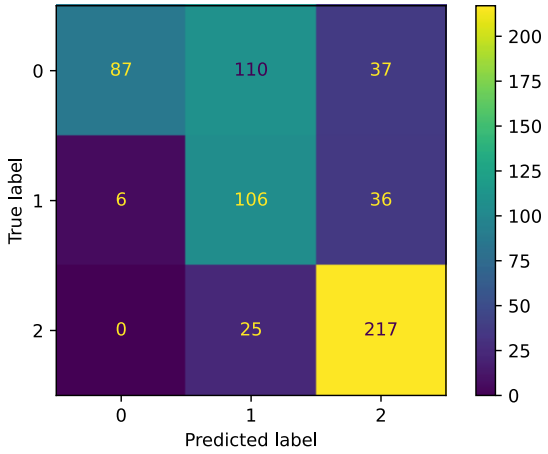

Figure 8: Confusion matrix for feedforward neural network
Labels: 0, normal; 1, viral pneumonia; 2, bacterial pneumonia.

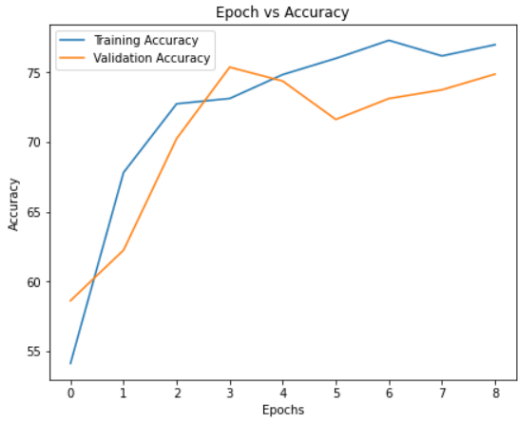

Figure 10: Accuracy for convolutional neural network

4.3. Basic deep-learning model: convolutional neural network

We trained the convolutional neural network on training and validation datasets with nine epochs. As shown in Figure 9, both training loss and validation loss decreased as the epoch increased. Similarly, the training and validation accuracy was plotted in Figure 10. Both accuracies significantly increased during the first few epochs, then steadily improved to about 75%.

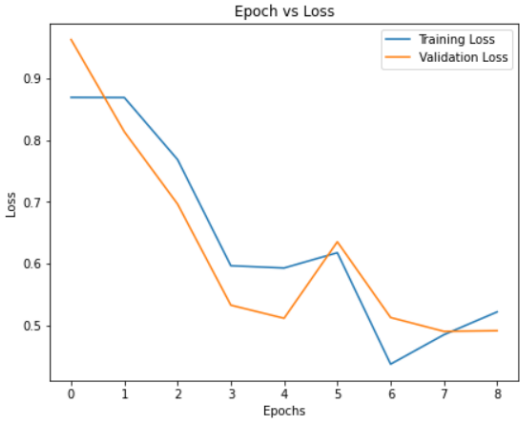

Figure 9: Loss for convolutional neural network

After training the feature maps were plotted in Figure 5. The convolutional network had an overall accuracy of 72.276%, and the confusion matrix was shown in Figure 11. The accuracy for each group varied. For normal patients, it had a mediocre accuracy of 62.821%. However, the accuracy for viral pneumonia patients was merely 44.594%, i.e., slightly better than guessing. On the conturary, the accuracy for bacterial pneumonia patients is 98.347%, which was almost perfect.

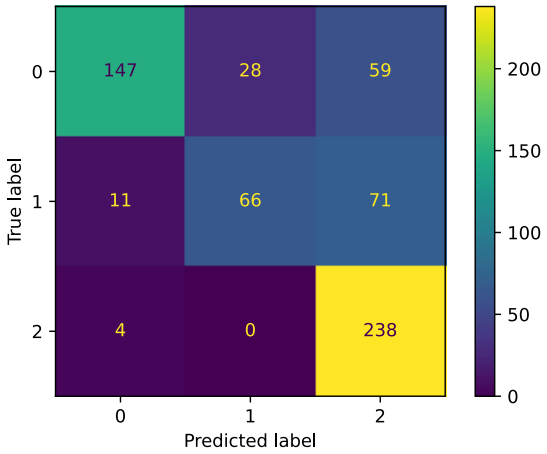

Figure 11: Confusion matrix for convolutional neural network
Labels: 0, normal; 1, viral pneumonia; 2, bacterial pneumonia.

4.4. Advanced model: ResNet18

For the advanced model using ResNet18, the training loss plummeted during the first two epochs and then steadily decreased (Figure 12). Likewise, slow yet steady increments followed the surge in the initial two epochs for the training accuracy (Figure 13). As shown in Figure 15, we took three images from each class, visualized the localization maps using Grad-CAM [12], and juxtaposed them with the original images. The model performed well in all groups with an overall testing accuracy of 98.435%. The confusion matrix is shown in Figure 14.

All three groups showed high accuracies, with 99.359% on normal patients, 97.647% on patients with viral pneumonia, and 98.718% on patients with bacterial pneumonia.

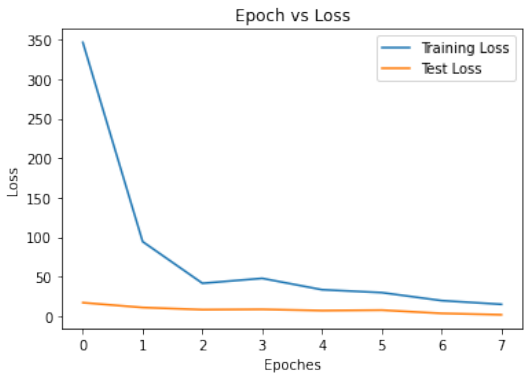

Figure 12: Loss for ResNet18

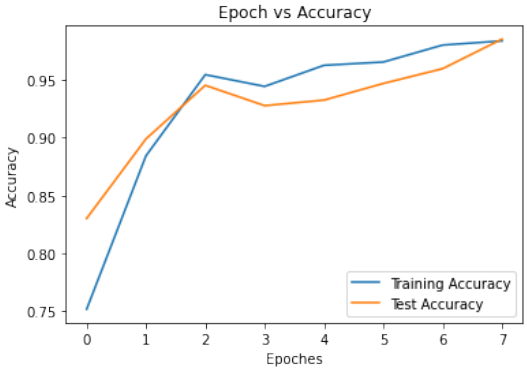

Figure 13: Accuracy for ResNet18

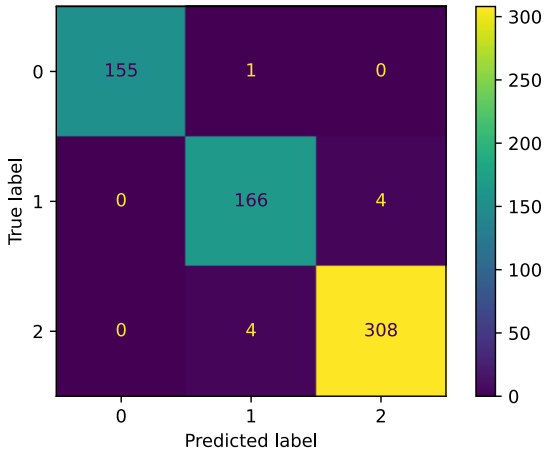

Figure 14: Confusion matrix for ResNet18
Labels: 0, normal; 1, viral pneumonia; 2, bacterial pneumonia.

4.5. Novel architecture model

As shown in Figure 16, for the novel architecture model, the training loss fluctuated violently over the epoches, reaching the highest value in the fourth epoch and the lowest value in the last epoch. The fluctuation in the validation loss was much milder. There was minimal decrease in the validation loss in the end when compared to the initial value. As shown in Figure 17, both the training and the validation accuracy have an increasing trend over the epoches despite some fluctuations.

The novel architecture model achieved an accuracy of 81.571% on the test data, and the confusion matrix is shown in Figure 18.

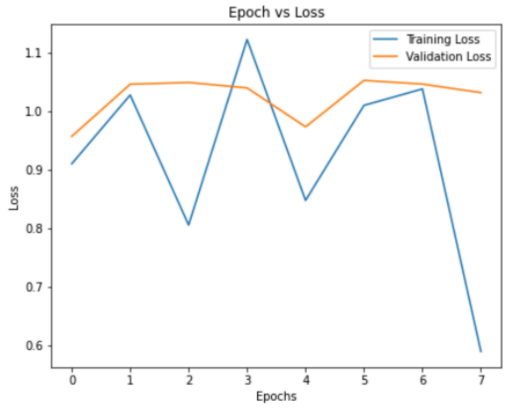

Figure 16: Loss for ALReLU neural network

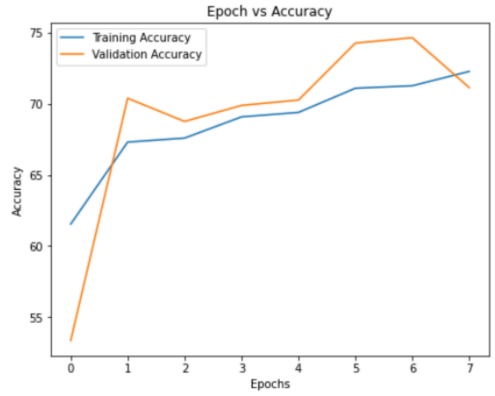

Figure 17: Accuracy for ALReLU neural network

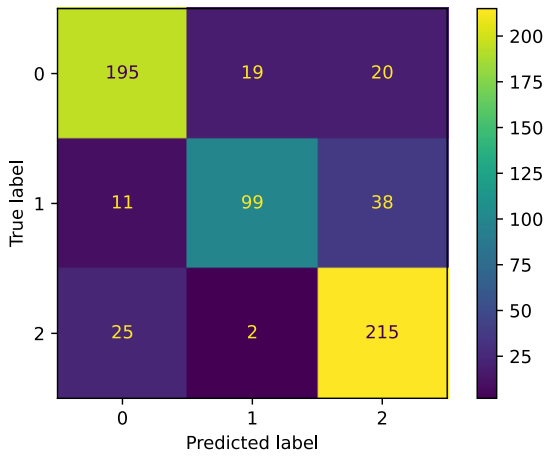

Figure 18: Confusion matrix for ALReLU neural network
Labels: 0, normal; 1, viral pneumonia; 2, bacterial pneumonia.

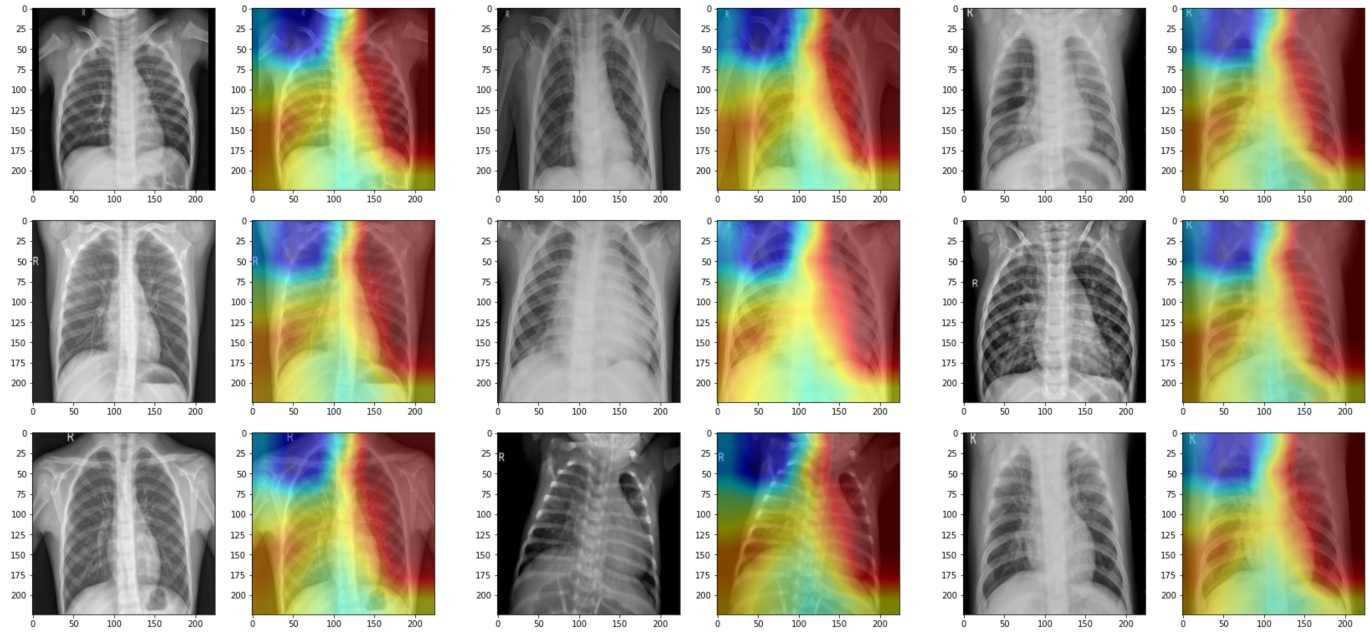

Figure 15: Localization maps
A *normal;* **B** *bacetiral pneumonia;* **C** *viral pneumonia.*

5. Model Evaluation and Interpretation

5.1. Baseline: non-deep learning classifier

The performance of the baseline non-deep learning classifier was mediocre (Figure 4). The classifier failed to classify normal patients into the correct category (35.470% accuracy). In real life, this will cause excessive reviewing or exams to be performed on normal patients. The performance was also unimpressive in identifying patients with viral pneumonia (57.432% accuracy) and suboptimal for bacterial infections (89.256% accuracy). Considering that we aim to develop an algorithm for preliminary screening for COVID-19 patients, who are likely to present with viral pneumonia, the multiple logistic classifier is not suitable for our purpose.

Upon reviewing the filters after training (Figure 3), it was evident that the filter corresponding to normal patniets presented a much more definite silhouette of the rib cages and lobes of the lung, while the borders for pneumonia patients were obscure. These differences were expected and align with typical clinical findings that opaque areas are commonly seen in X-rays for pneumonia patients. Pneumonia decreases the air-exchange capacity of the lung, resulting in lower air content, more similar to the surrounding tissues. Hence, the lung will be harder to identify on the image for pneumonia patients, resulting in less definite borders. Nonetheless, even patients with pneumonia may not be infected in all lobes, and the healthy lobes of an infected patient may confuse the classifier, leading to poor performance.

5.2. Basic deep-learning model: feedforward neural network

The feedforward neural network showed a moderate performance in the training step. As shown in Figure 6 and Figure 7, the training loss steadily decreased and the training accuracy increased, thus the model learned about pneumonia classification as epoch increased. However, the validation loss and accuracy stayed relatively constant, which indicates a potential problem of overfitting.

The test accuracy (65.705%) of the model was better than the non-deep benchmark, but still has areas to improve. In particular, it had an extremely low accuracy (37.179%) in identifying normal cases, but much higher accuracies in identifying pneumonia (71.622% for viral pneumonia and 89.669% for bacterial pneumonia). The model could thus effectively classify viral and bacterial cases but would usually misclassify normal cases as problematic ones. Considering our preliminary screening purpose, the model can effectively capture most pneumonia cases.

5.3. Basic deep-learning model: convolutional neural network

The convolutional neural network showed a better performance in the training step. As shown in Figure 9 and Figure 10, the loss for both training and validation datasets decreased, while the accuracy for both datasets increased. Thus the model improves and learns more about pneumonia classification as the epoch increases. Besides, it effectively addresses the overfitting problem in the feedforward neural network.

The test accuracy (72.276%) of the model was better than the non-deep benchmark, and slightly improved from the feedforward neural network. However, it had an imbalanced accuracy for the three groups. The model had an accuracy of 62.821% in identifying normal cases, 44.594% in identifying viral pneumonia, and 98.347% in identifying bacterial pneumonia. Thus it could effectively identify almost all bacterial cases but cannot capture the viral ones. This would be problematic for our purpose of identifying the viral cases.

As the generated feature maps shown in Figure 5, some filters focused on the spine, and some on the chest cavity. Because the model had a high accuracy for only bacterial pneumonia but not viral pneumonia, we believe the feature maps relate to most bacterial features but very few viral features.

5.4. Advanced model: `ResNet18`

As shown in Figure 13, the test accuracy of the model based on `ResNet18` was almost the same as the training accuracy at the end of training, indicating that our model addresses the problem of overfitting very well. With an accuracy of 99.359% on normal patients,

97.647% on patients with viral pneumonia, and 98.718% on patients with bacterial pneumonia, as shown in Figure 14, the advanced model does an excellent job in classifying X-ray images.

It was clear in Figure 15 that the most intense red color marked the left lobe of the lung. The higher region indicated that our model mainly focuses on the region mentioned above when classifying images, i.e., the lung instead of surroundings. Since pneumonia infects the lungs, it's reasonable that our model looks at this region when it classifies the images.

Compared to the left lobe, the right lobe had a relatively minor influence on the decision of our model. One possible reason would be that the information from the left lobe is already sufficient for our model to decide. However, it may also due to that our dataset mainly contained X-ray images of patients whose left lobe got infected, as either lobe may be infected pneumonia.

5.5. Novel architecture model

The performance of the novel architecture model was better than the non-deep learning baseline and both deep learning baselines, but not as good as the advanced model. As shown in Figure 18, the model achieves 83.333% accuracy on normal patients, 66.892% accuracy on patients with viral pneumonia, and 88.843% accuracy on patients with bacterial pneumonia. The largest source of error is classifying 25.676% of the viral pneumonia cases as bacterial pneumonia.

Considering our goal to screen for COVID-19 patients, who are likely to have viral pneumonia, our model's effectiveness would be limited due to the low accuracy on patients with viral pneumonia and the tendency of misclassifying viral pneumonia as bacterial pneumonia. Such misclassifications may be dangerous in real life, as it could miss out on COVID-19 patients.

6. Social Impact

With the highest accuracy of 98.435%, our model could help the doctors diagnose whether a patient has pneumonia and whether the infection was bacterial or viral. On occasions where the doctor and the model disagree, the doctor can double-check the X-ray image and decide. With supervision from physicians, we can further minimize misclassification and accelerate the diagnosis process.

Moreover, in the COVID-19 pandemic, our model could be used for pre-screening for patients with COVID-19. If the patient develops viral pneumonia, they should immediately get tested for COVID-19.

7. Conclusion

To classify the X-ray images into normal, viral, and bacterial pneumonia, we used the multiple logistic classifier, the feedforward neural network, the convolutional neural network, transfer learning with `ResNet18`, and a novel convolutional neural network with a custom activation function, `ALReLU`. Among these models, transfer learning with `ResNet18` achieved the highest test accuracy of 98.435%. We believe that the automated classification of pneumonia can increase the speed and accuracy of pneumonia diagnosis.

References

[1] Centers for Disease Control and Prevention. United states covid-19 cases and deaths by state, May 2021.

[2] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. The Lancet Infectious Diseases, 20(5):533–534, 2021/05/02 2020.

[3] Jenny Firth-Cozens and Joanne Greenhalgh. Doctors' perceptions of the links between stress and lowered clinical care. Social science & medicine, 44(7):1017–1022, 1997.

[4] Tait D Shanafelt, Katharine A Bradley, Joyce E Wipf, and Anthony L Back. Burnout and self-reported patient care in an internal medicine residency program. Annals of internal medicine, 136(5):358–367, 2002.

[5] Colin P West, Mashele M Huschka, Paul J Novotny, Jeff A Sloan, Joseph C Kolars, Thomas M Habermann, and Tait D Shanafelt. Association of perceived medical errors with resident distress and empathy: a prospective longitudinal study. Jama, 296(9):1071–1078, 2006.

[6] Eric S Williams, Linda Baier Manwell, Thomas R Konrad, and Mark Linzer. The relationship of organizational culture, stress, satisfaction, and burnout with physician-reported error and suboptimal patient care: results from the memo study. Health care management review, 32(3):203–212, 2007.

[7] Colin P West, Angelina D Tan, Thomas M Habermann, Jeff A Sloan, and Tait D Shanafelt. Association of resident fatigue and distress with perceived medical errors. Jama, 302(12):1294–1300, 2009.

[8] Praveen. Coronahack: Classify the x ray image which is having corona, May 2021.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[11] Tawsifur Rahman, Muhammad E. H. Chowdhury, Amith Khandakar, Khandaker R. Islam, Khandaker F. Islam, Zaid B. Mahbub, Muhammad A. Kadir, and Saad Kashem. Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray. Applied Sciences, 10(9), 2020.

[12] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. CoRR, abs/1610.02391, 2016.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[14] Stamatis Mastromichalakis. Alrelu: A different approach on leaky relu activation function to improve neural networks performance. CoRR, abs/2012.07564, 2020.