APS360 Final Report

X-ray Diagnosis on Bacterial and Viral Pneumonia

Team 43
Weixuan Sun 1003803048
Jiawen Li 1002904253
Qiaoyi Yan 1003130774
Yiran Qiu 1004172290

Word counts: 2262
Penalty: 0%

Google Colab Link

https://colab.research.google.com/drive/1k8NYtR_vONczE_S6njCgt6RGb_SrJkVc?usp=sharing

I. Introduction

Aside from childbirth, pneumonia is the leading cause of hospital admission. It results in approximately 1 million adult hospital admissions annually – and 50,000 deaths [1]. It's now also one of the common symptoms of COVID-19 [2]. With our healthcare system heavily impacted by the current COVID-19 situation, the team sees an opportunity to help relieve the workload and stress for doctors and radiologists, and we came up with this project to develop an ML model that reads chest x-rays and diagnoses the presence of different types of pneumonia. This could potentially allow patients to self-diagnose and doctors to speed up their early-stage pneumonia x-ray diagnosis process.

Pneumonia can be caused by bacteria, fungi, and viruses. In this project, we will be mainly exploring bacterial and viral pneumonia, since fungal pneumonia is less common [1].

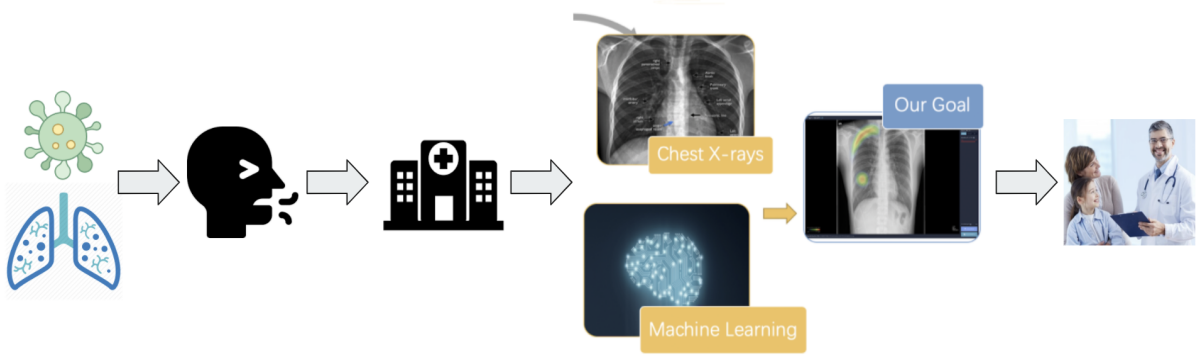


Figure 1. Create automated pneumonia detection tool

II. Background & Related Work

There have been multiple pieces of research that look into the clinical interpretability of machine learning techniques. In 2018, a paper published on Cell Press discussed the development of an AI system using transfer learning to classify medical images and demonstrated its applicability for diagnosis on pneumonia X-ray images [3]. The paper also concluded their tool could potentially be employed in a wide range of medical diagnosis [3].

Some companies also saw the bright prospect of using AI & deep learning for medical diagnosis. Lunit is a medical AI software company that provides AI-powered solutions for cancer diagnostics [4]. As shown in Figure 2, their commercial software, Lunit Insight CXR, detects the abnormality on a chest x-ray. Lunit claimed their product achieved a 97% ~ 99%

accuracy in ROC AUC, and are currently preparing for regulatory approval in various markets worldwide [5].

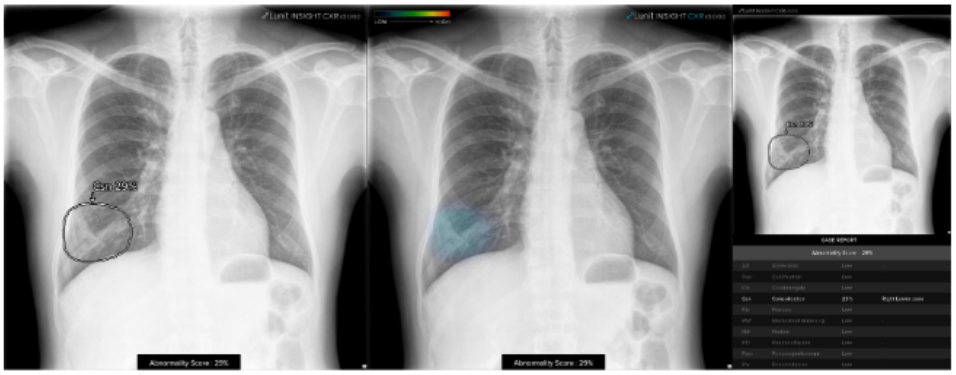


Figure 2. Lunit Insight CXR detects pneumonia in the chest x-ray image

III. Data Processing

The team used 5856 chest x-ray images provided by Kaggle [6], available at https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. The images are organized in three folders (training, validation, and testing), and have two subfolders for each folder (normal and pneumonia). Originally, there are only 16 images in total in the validation folder, which is not enough. The team will re-split the images as 70% training data, 15% validation data, and 15% testing data. Table 1 shows the folder structure after splitting.

- Training (4100 images)
  - Normal (1109 images)
  - Pneumonia (2991 images)
- Validation (878 images)
  - Normal (237 images)
  - Pneumonia (641 images)
- Testing (878 images)
  - Normal (237 images)
  - Pneumonia (641 images)

Table 1. Folder hierarchy after reorganizing

Under the pneumonia subfolder, bacterial and viral pneumonia are labeled in the image's filename, as shown in Figure 3, which means we have to create two subfolders to segregate bacterial and viral pneumonia before training.

person100_bacteria_475.jpeg (75.58 KB)

person10_virus_35.jpeg (139.58 KB)

Figure 3. Bacteria and virus labels on the name of the images

The team created 3 sets of data to accommodate different approaches to this diagnosis project, either a multiclass (3 classes) classification or two binary classifications (Normal vs Pneumonia and Bacteria vs Virus), as shown in Figure 4.

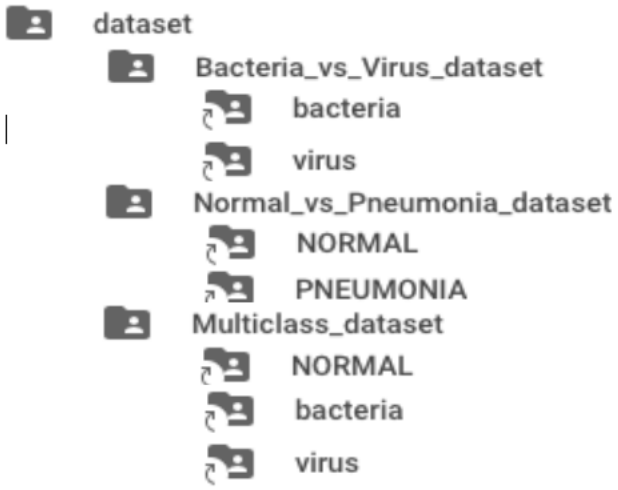


Figure 4. Different dataset folders for different classification tasks

The x-ray images come in both RGB and Grayscale. We changed all images to Grayscale and then resized them to the same dimension (1050 x 760). Three of the preprocessed images are shown in Figure 5.

Figure 5. Preprocessed images' example

We noticed that the original dataset is imbalanced and considered collecting more normal or viral pneumonia images. However, such medical images are expensive to acquire, therefore we decided to use a generative adversarial network to create artificial training data. In Figure 6, the generated viral pneumonia image's lung cavity is more blurry than the generated normal image.

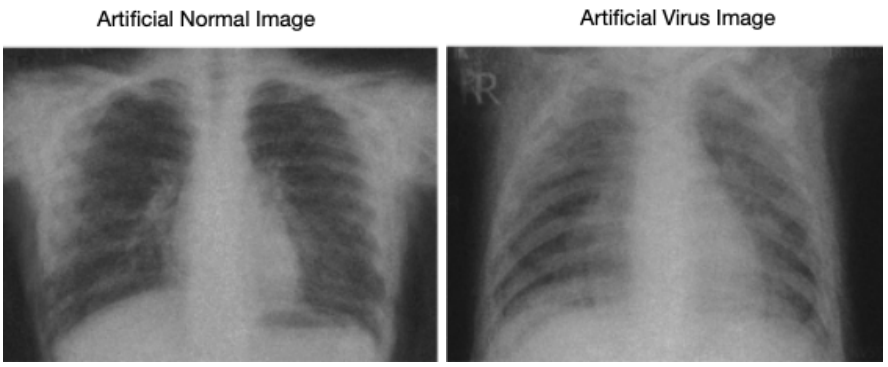


Figure 6. Examples of GAN output

We generated 96 artificial normal images and 128 artificial viral pneumonia images as extra training data. The team also acquired 20 images from Google for additional final testing.

IV. Baseline Model

In the baseline model, we treated the problem as a multi-class image classification problem, and attain reasonable performance with a Support Vector Machine (SVM). A visualization of the data flow is shown in Figure 7.

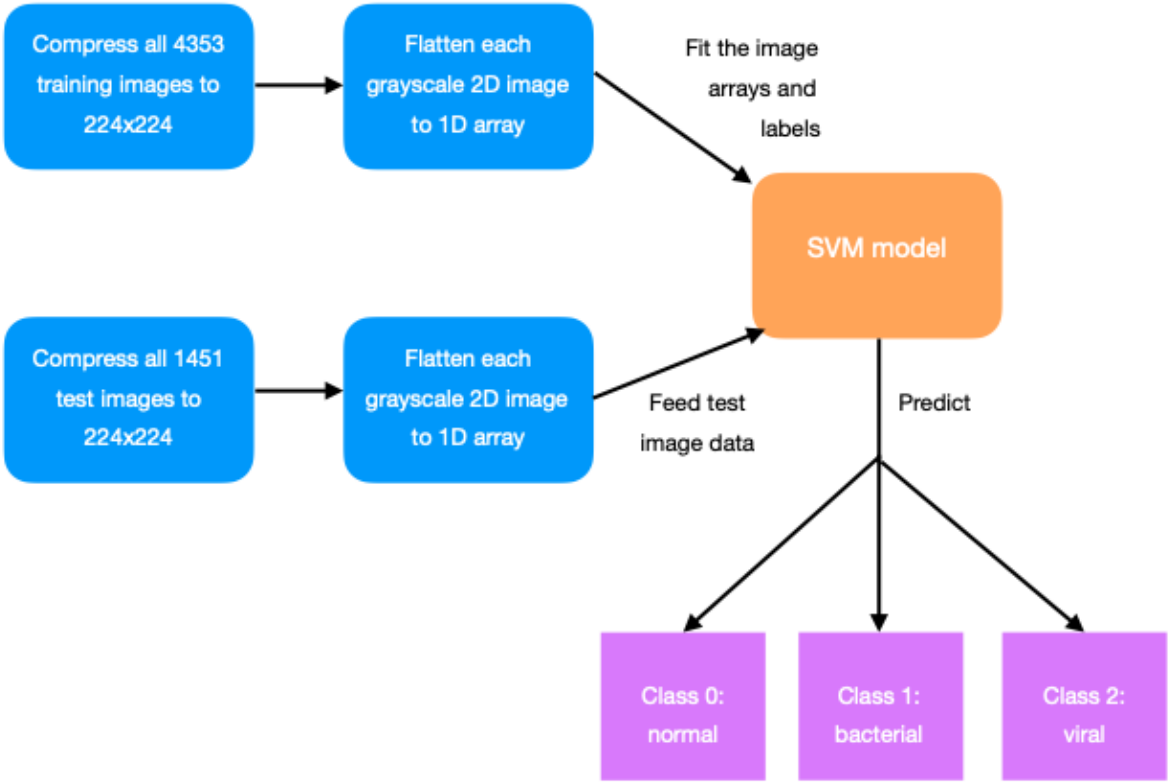


Figure 7. A visualization of the SVM baseline model

```
SVC model:  SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma=0.001, kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)

Classification report:
              precision    recall  f1-score   support

           0       0.88      0.90      0.89       380
           1       0.76      0.87      0.81       691
           2       0.69      0.51      0.59       380

    accuracy                           0.78      1451
   macro avg       0.78      0.76      0.76      1451
weighted avg       0.78      0.78      0.77      1451
```

Figure 8. SVM model hyperparameters and classifications

The model performs best in the normal class and worst in the virus class. The recall for viral class is only 51%, meaning the model misclassifies 49% of the virus cases. The overall test accuracy across all 3 classes is 78% (Figure 8).

V. Model Architecture

The team created two prototypes, “CNN from scratch” and transfer learning. We trained and tuned the two models separately on the same training and validation dataset and after comparing the two models, we decided to move forward with the “CNN from scratch” model since it achieved a higher validation accuracy.

The team divided the classification into two substages. The first substage is used to classify if a patient is normal or has pneumonia. The second substage is to classify which type of pneumonia a patient has, either bacterial or viral.

The architecture of the Normal vs Pneumonia classifier consists of 10 layers, where the detailed information is shown in Figure 9. The first convolutional layer creates 10 channels and the second layer generates 20 11*20 feature maps, which are then fed into a two-layer fully connected network. To avoid overfitting, the team also applied four 20% dropout layers on each of the convolutional and linear layers. ReLU activation function was applied. We used a batch size of 64, a learning rate of 0.0006, and trained the model for 12 epochs.

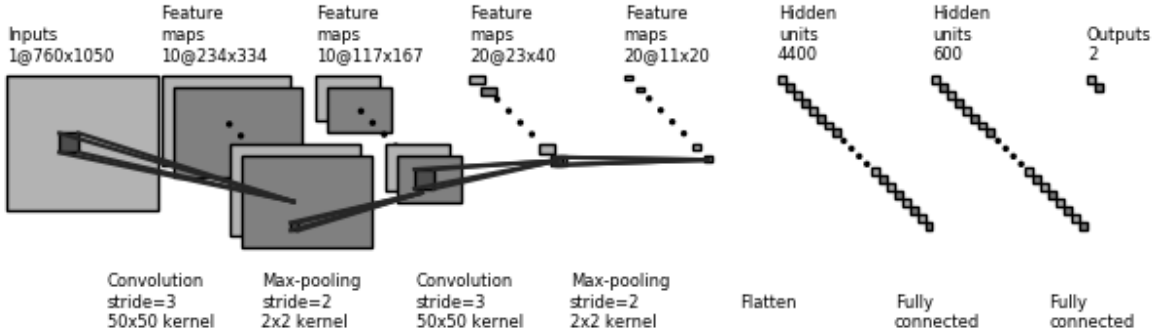


Figure 9. Normal vs Pneumonia binary classifier CNN model

```
NP model:
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
           Dropout-1       [-1, 1, 760, 1050]               0
            Conv2d-2       [-1, 10, 237, 334]          25,010
         MaxPool2d-3       [-1, 10, 118, 167]               0
           Dropout-4       [-1, 10, 118, 167]               0
            Conv2d-5         [-1, 20, 23, 40]         500,020
         MaxPool2d-6         [-1, 20, 11, 20]               0
           Dropout-7               [-1, 4400]               0
            Linear-8                [-1, 600]       2,640,600
           Dropout-9                [-1, 600]               0
           Linear-10                  [-1, 2]           1,202
================================================================
Total params: 3,166,832
Trainable params: 3,166,832
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 3.04
Forward/backward pass size (MB): 15.35
Params size (MB): 12.08
Estimated Total Size (MB): 30.48
----------------------------------------------------------------
```

Figure 10. A total of 10 layers and 3,166,832 training parameters

Similar to the NP classifier, the architecture of the Bacteria vs Virus classifier also contains 10 layers, with details shown in Figure 11. The first convolutional layer creates 8 channels and the second layer generates 15 11x20 feature maps. We use a batch size of 64, a learning rate of 0.0003, and trained for 13 epochs.

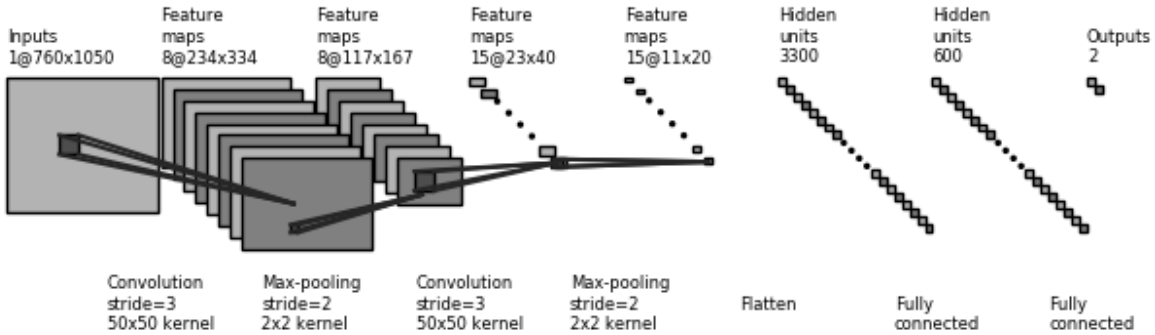


Figure 11. Bacterial vs Viral Pneumonia binary classifier CNN model

```
BV model:
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
           Dropout-1        [-1, 1, 760, 1050]               0
            Conv2d-2         [-1, 8, 237, 334]          20,008
         MaxPool2d-3         [-1, 8, 118, 167]               0
           Dropout-4         [-1, 8, 118, 167]               0
            Conv2d-5          [-1, 15, 23, 40]         300,015
         MaxPool2d-6          [-1, 15, 11, 20]               0
           Dropout-7                [-1, 3300]               0
            Linear-8                 [-1, 600]       1,980,600
           Dropout-9                 [-1, 600]               0
           Linear-10                   [-1, 2]           1,202
================================================================
Total params: 2,301,825
Trainable params: 2,301,825
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 3.04
Forward/backward pass size (MB): 13.49
Params size (MB): 8.78
Estimated Total Size (MB): 25.31
----------------------------------------------------------------
```

Figure 12. BV model has a total of 10 layers and 2,301,825 training parameters.

VI. Results and Evaluation

Quantitative Results

- **Best Normal VS Pneumonia Model:**
  - Learning Rate = 0.006, Batch size = 64, Epochs = 12
  - Final Training Accuracy: 0.9822
  - Final Validation Accuracy: 0.9272

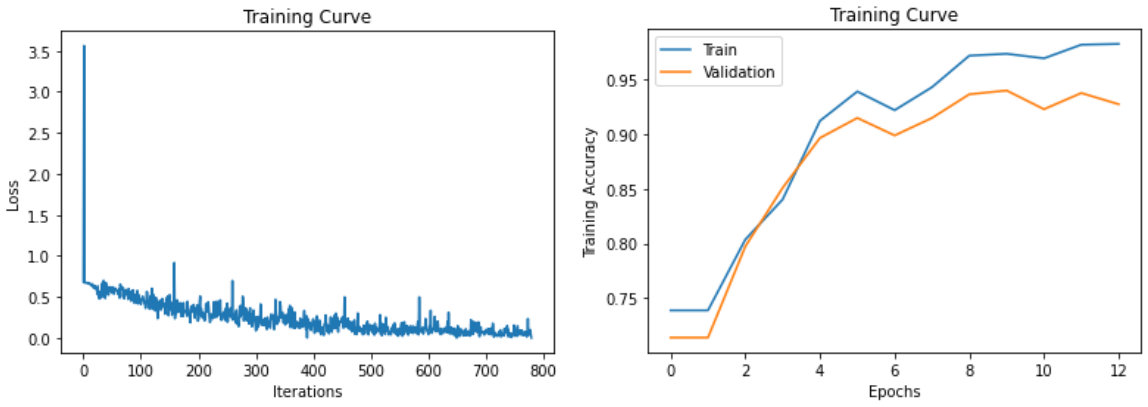


Figure 13. Training curves for best NP model on iterations and epochs

Normal vs Pneumonia Model	
Final Testing Accuracy	**0.928**
True Positive* Rate	0.967
True Negative Rate	0.841
False Positive Rate	0.153
False Negative Rate	0.037
*Treating Pneumonia as Positive	

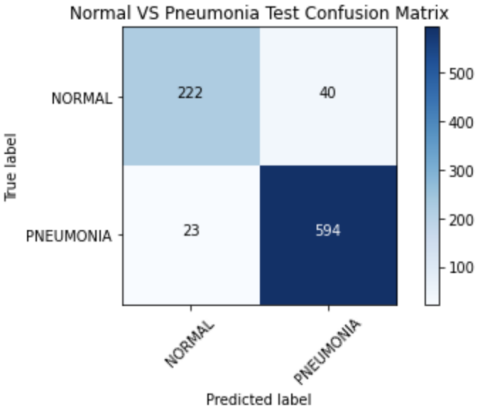


Figure 14. Final testing results and confusion matrix for best NP model

- **Best Bacterial VS Viral Pneumonia Model:**
  - Learning Rate = 0.003, Batch size = 64, Epochs = 13
  - Final Training Accuracy: 0.7993
  - Final Validation Accuracy: 0.7552

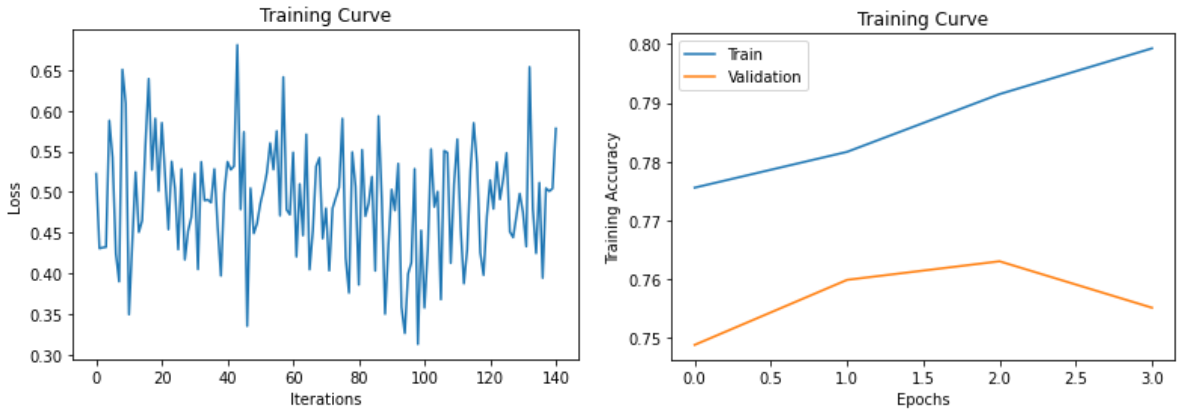


Figure 15. Training curves for best BV model on iterations and epochs

Bacterial vs Virus Model	
Final Testing Accuracy	**0.756**
True Positive* Rate	0.6
True Negative Rate	0.833
False Positive Rate	0.167
False Negative Rate	0.4
*Treating Virus Pneumonia as Positive	

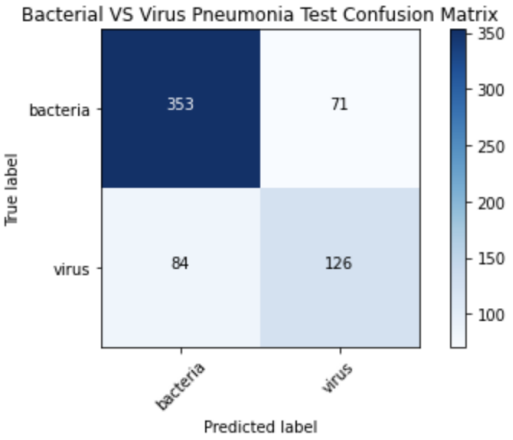


Figure 16. Final testing results and confusion matrix for best BV model

Qualitative Results

Below are a few qualitative examples of how our models perform on the testing dataset.

For Normal and Pneumonia classification:

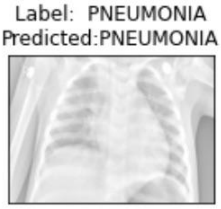
Sample pictures	True Label	Output distribution tensor	Classification result with probability
	Normal	tensor([0.0862, 0.9138])	The probability of this x-ray to be detected as Normal is 0.91, which is correct.
	Pneumonia	tensor([0.9033, 0.0967])	The probability of this x-ray to be detected as Normal is 0.90, which is correct.

Table 2. Qualitative x-ray examples of normal lungs and lungs with pneumonia

For Bacterial and Viral Pneumonia classification:

Sample pictures	True Label	Output distribution tensor	Classification result with probability
	Bacteria	tensor([0.3154, 0.6846])	The probability of this x-ray to be detected as bacterial pneumonia is 0.68, which is correct.
	Virus	tensor([0.7386, 0.2614])	The probability of this x-ray to be detected as viral pneumonia is 0.74, which is correct.

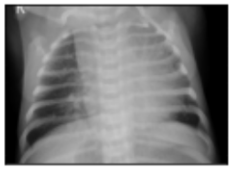
Label: virus Predicted:bacteria 	Virus	tensor([0.4629, 0.5371])	The probability of this x-ray to be detected as bacterial pneumonia is 0.54, which is incorrect.
Label: bacteria Predicted:virus 	Bacteria	tensor([0.5473, 0.4527])	The probability of this x-ray to be detected as viral pneumonia is 0.55, which is incorrect.

Table 3. Qualitative x-ray examples of normal lungs and lungs with pneumonia

Performance Analysis

Compared with diagnosing radiographs on normal or pneumonia lungs, our model has relatively low correctness on classifying bacterial or viral pneumonia x-rays. This is because it's challenging to differentiate between bacterial and viral pneumonia chest radiographs. Both radiographs have nodules and patchy areas of peribronchial ground-glass opacity and airspace consolidation which are resulted by lymphatic infiltrates in the alveolar septa [7, 8]. The x-ray for a normal lung cavity is clear as nothing is blocking the x-ray. However, the chest radiographs of both kinds of pneumonia have airspace opacity, lobar consolidation, or interstitial opacities which is comparatively more significant than the difference between bacterial and viral pneumonia [9]. The small difference between radiographs of bacterial and viral chest radiographs results in relatively low correctness on classifying (non linear classification).

VII. Evaluate models on new data

To better evaluate the performance of our models, we tested the model with chest x-rays images collected from Radiopaedia.org, a community where radiologists share real patients' x-rays with diagnosis results.

We selected chest x-rays that are within our testing criteria, and have resolutions greater than 1050*760. We also ensure that the radiologists are certain about their diagnosis on the x-rays to minimize the potential classification errors in the dataset.

For Normal and Pneumonia classification:
We tested our NP model on 23 images, with 8 normal chest x-rays and 15 pneumonia x-rays, and achieved an overall testing accuracy of 86.95%. Here are four qualitative examples.

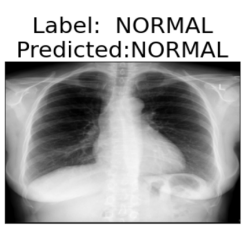
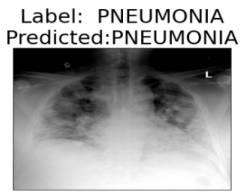
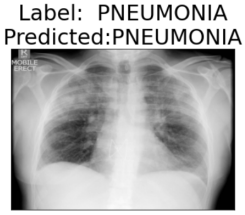
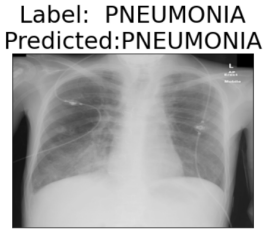
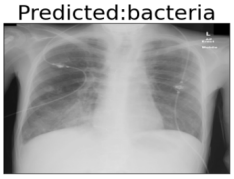
Sample pictures	True Label	Output distribution tensor	Classification result with probability
	Normal[10]	tensor([8.6470e-01, 1.3530e-01])	The probability of this x-ray to be detected as Normal is 0.8647, which is correct.
	Pneumonia[11]	tensor([7.9272e-08, 1.0000e+0])	The probability of this x-ray to be detected as Pneumonia is 1, our model is almost 100% certain this is a pneumonia x-ray.
	Pneumonia[12]	tensor([8.7590e-02, 9.1241e-01])	The probability of this x-ray to be detected as Pneumonia is 0.912, which is correct.
	Pneumonia[13]	tensor([3.2605e-05, 9.9997e-01])	The probability of this x-ray to be detected as Pneumonia is 0.999, which is correct.

Table 4. Qualitative x-ray examples of normal lungs and pneumonia lungs

For Bacterial and Viral Pneumonia classification:
We tested our BV model on 15 images, with 8 bacterial pneumonia x-rays and 7 viral pneumonia x-rays, and achieved an overall testing accuracy of 80.67%. Here are three qualitative examples.

Sample pictures	True Label	Output distribution tensor	Classification result with probability
	Bacteria[13]	tensor([0.8066, 0.1934])	The probability of this x-ray to be detected as bacteria pneumonia is 0.8066, which is correct.

Label: virus Predicted:virus 	Virus[12]	tensor([0.2507, 0.7493])	The probability of this x-ray to be detected as viral pneumonia is 0.74, which is correct.
Label: virus Predicted:virus 	Virus[11]	tensor([0.3883, 0.6117])	The probability of this x-ray to be detected as viral pneumonia is 0.61, which is correct.(COVID-19)

Table 5. Qualitative x-ray examples of normal lungs and lungs with pneumonia

Our model also successfully identified the COVID-19 chest x-ray as viral pneumonia despite the model has not been trained on any COVID-19 dataset. Our research found that COVID-19 pneumonia and influenza virus pneumonia have a large amount of overlap, and no significant differences were detected [14]. Therefore our model was still able to predict correctly even though COVID-19 is a new type of virus.

VIII. Discussion

For the first substage of the classification model, our NP model can classify normal and pneumonia classes with a test accuracy of 92.8%. It also correctly predicts 4 out of 4 chest x-ray images when we test it on a new Normal vs Pneumonia dataset. This indicates that our NP model is well trained and can generalize well to unseen data.

As for the second substage of the classification model, the test accuracy of the BV model is 75.5%, which is lower than our expectation even though we know classifying different types of pneumonia is a much more challenging task. The confusion matrix implies that most classification errors are caused by the low recall of 60% of Virus class, even though it's already increased by 9% from the baseline model's Virus class recall (51%).

The result is unusual in a way that we didn't foresee earlier. In both the multi-class baseline model and our final model made up of 2 binary classifiers, the major source of classification errors is the Virus class. As we dive deeper into the dataset, we realize that the Virus dataset is small and not diverse, thus the BV model has not learned from sufficient Virus class examples. Specifically, in the dataset, each patient contributed 4 frontal chest x-ray images. Since patients did not change their postures while being x-rayed, the images from the same patient are almost identical. This also explains why the BV model fails to predict the x-ray image of a COVID-19 patient when we evaluate it on new data. What we have learned from the project is that the quality of the dataset is as important as the structure and parameters of the neural network.

Though the overall model performance is satisfactory, it is not reliable enough for real-life implementation. Ideally, there should be zero tolerance allowed for any of the misclassifications. One small error could heavily impact a patient or potentially make a life-or-death difference, as we will discuss more in the ethical considerations section.

IX. Ethical Considerations

It is difficult to obtain informed consent for secondary uses of patient data when the purposes of such uses are unknown at the time of data collection. Many patients are unaware of possible conflicts of interest regarding data sharing, especially for commercial uses. Patients' rights to be informed and to give consent before their data is shared may not be respected, infringing upon the fundamental human right to privacy [15].

All the chest X-ray images are from children in China and the ones characterized as pneumonia are either bacterial or viral. However, many other types of pneumonia exist. Therefore, our model may fail to generalize to adult patients or diagnose other pneumonia types.

Even if our model has high performance and generalizability, hospitals must understand the risks before adopting it. The diagnoses of the model must be confirmed by more experts and doctors before proceeding, since misclassified cases may result in medical malpractice [16].

X. Project Difficulty / Quality

There is a day-to-day discrepancy rate of 3-5% in radiology practice [17]. It is challenging for the neural network to correctly classify between different diseases, since all x-ray images are black and white, which means the model cannot classify based on colour, but other nuances and features instead. Our model performs relatively well in classifying pneumonia patients and people with healthy lungs, reaching a testing accuracy of 93%; however, the model only reached a 77% accuracy on classifying bacterial and viral pneumonia is because the difference between the chest radiographs of two kinds of pneumonia is small.

Both kinds of pneumonia have airspace consolidation and airspace opacity which shows similar signs on the radiograph, thereby hard to find patterns and to classify. The model should only be used on clinical trials if the accuracy rate reaches a certain level, potentially lower error rate than humans. The team tried to use a stacked convolutional autoencoder to denoise the chest x-rays, but it still didn't reach an accuracy that is higher than 95%. The other problem that the team encountered is the limited amount of x-rays and the imbalanced dataset. Since medical images are always rare and expensive to obtain, the team created a generative adversarial network to generate more x-ray images for testing. The team also trained and tuned different transfer learning models, such as Alexnet, Googlenet, and Resnet for comparison; however, those models did not outperform the binary classification CNN models the team had.

References

[1] By: Krystina Ostermeyer RN, "Bacterial vs Viral Pneumonia - Elite Learning", Elite Learning, 2020. [Online]. Available: https://www.elitecme.com/resource-center/nursing/bacterial-vs-viral-pneumonia/. [Accessed: 08- Aug- 2020].

[2] "Updates on Novel Corona Virus (COVID-19)", Who.int, 2020. [Online]. Available: https://www.who.int/maldives/news/detail/31-01-2020-updates-on-novel-corona-virus-(COVID-19)#:~:text=The%20virus%20can%20cause%20a,and%20deaths%20can%20occur. [Accessed: 08- Aug- 2020].

[3] D. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning", Cell, vol. 172, no. 5, pp. 1122-1131.e9, 2018. Available: 10.1016/j.cell.2018.02.010 [Accessed: 08- Aug- 2020].

[4] "About us", Lunit.io, 2020. [Online]. Available: https://www.lunit.io/en/company/. [Accessed: 08- Aug- 2020].

[5] "AI Product", Lunit.io, 2020. [Online]. Available: https://www.lunit.io/en/product/insight_cxr3/. [Accessed: 08- Aug- 2020].

[6] "Chest X-Ray Images (Pneumonia)", Kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. [Accessed: 08- Aug- 2020].

[7] "Viral Pneumonia Imaging: Practice Essentials, Radiography, Computed Tomography", Emedicine.medscape.com, 2020. [Online]. Available: https://emedicine.medscape.com/article/360254-overview#a2. [Accessed: 08- Aug- 2020].

[8] "Typical Bacterial Pneumonia Imaging: Practice Essentials, Radiography, Computed Tomography", Emedicine.medscape.com, 2020. [Online]. Available: https://emedicine.medscape.com/article/360090-overview#a2. [Accessed: 08- Aug- 2020].

[9] R. (ACR), "Pneumonia | Lung inflammation - Diagnosis, Evaluation and Treatment", Radiologyinfo.org, 2020. [Online]. Available: https://www.radiologyinfo.org/en/info.cfm?pg=pneumonia. [Accessed: 08- Aug- 2020].

[10] H. Knipe, "Normal chest x-ray | Radiology Case | Radiopaedia.org", Radiopaedia.org, 2020. [Online]. Available: https://radiopaedia.org/cases/normal-chest-x-ray-1?lang=us. [Accessed: 08- Aug- 2020].

[11] O. Rizk, "COVID-19 pneumonia | Radiology Case | Radiopaedia.org", Radiopaedia.org, 2020. [Online]. Available: https://radiopaedia.org/cases/covid-19-pneumonia-139?lang=us. [Accessed: 08- Aug- 2020].

[12] C. Hacking, "Bilateral pneumonia in transplant lungs | Radiology Case | Radiopaedia.org", Radiopaedia.org, 2020. [Online]. Available: https://radiopaedia.org/cases/bilateral-pneumonia-in-transplant-lungs?lang=us. [Accessed: 08- Aug- 2020].

[13] C. Hacking, "Pneumonia in transplant lungs | Radiology Case | Radiopaedia.org", Radiopaedia.org, 2020. [Online]. Available: https://radiopaedia.org/cases/pneumonia-in-transplant-lungs?lang=us. [Accessed: 08- Aug- 2020].

[14] W. Palmer, "CT Reveals Differences Between COVID-19 and Influenza Pneumonias", Diagnostic Imaging, 2020. [Online]. Available: https://www.diagnosticimaging.com/view/ct-reveals-differences-between-covid-19-and-influenza-pneumonias. [Accessed: 08- Aug- 2020].

[15] C. Mouton Dorey, H. Baumann and N. Biller-Andorno, "Patient data and patient rights: Swiss healthcare stakeholders' ethical awareness regarding large patient data sets – a qualitative study", BMC Medical Ethics, vol. 19, no. 1, 2018. Available: 10.1186/s12910-018-0261-x [Accessed 02-Aug-2020].

[16] "What are the big ethical challenges for AI in healthcare? | Maneesh Juneja", Hall & Partners, 2020. [Online]. Available: https://www.hallandpartners.com/big-ethical-challenges-for-ai-in-healthcare. [Accessed: 02-Aug-2020].

[17] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?," *Insights into imaging*, Feb-2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5265198/. [Accessed: 02-Aug-2020].

Appendix: Individual Contributions Table

Tasks	Task Description	Weixuan Sun	Jiawen Li	Qiaoyi Yan	Yiran Qiu
Data preprocessing	Balance data, create data path and data loaders, save data to shared folder, resize data	100%	-	-	-
Building GAN	Building a GAN to generate more data	-	50%	-	50%
Building SVM baseline model	Building a baseline model to compare the neural network against	50%	50%	-	-
Building models	Building a CNN and a Transfer Learning model to classify the data	50%	50%	-	-
Model training	Training the CNN and Transfer Learning model	50%	20%	30%	-
Hyperparameter tuning	Tuning the model with different hyperparameter to achieve the best accuracy	10%	-	50%	40%
Model evaluation	Report the test accuracy and confusion matrices, and test the model with new data	10%	-	90%	-
Research and documentation	Search information that are helpful in coding or writing reports	25%	25%	25%	25%
Presentation	Preparing the slides, recordings and video editing	25%	25%	20%	30%
Tasks that are unable to complete	**Task Description**	**Weixuan Sun**	**Jiawen Li**	**Qiaoyi Yan**	**Yiran Qiu**
Building AutoEncoder	Building an AutoEncoder for denoising	40%	-	-	60%
Hyperparameters Tuning for Bacterial VS Virus Model	Tuning the model with different hyperparameter to achieve the best accuracy	30%	-	50%	20%