**APS360 Applied Fundamentals of Machine Learning**

**Final Report - Part A**

**Total Word Count: 2494 excluding figures, captions, and references. No Penalty.**

**1 Introduction**

Over 100 million people are infected with pneumonia globally every year, and more than 1 million children under 5 years old are killed by pneumonia annually [1][2]. In developing countries, the high death rate of pneumonia can be caused due to the shortage of medical resources and healthcare workers [3].

To detect such a common disease, chest x-rays (CXRs) are widely performed for pneumonia diagnosis due to the low cost and efficiency [4]. However, the process of reading CXRs is systematic and time-consuming, in which it takes an average of 1-2 minutes for a well-trained radiologist to read one CXR, while hundreds of CXRs are generated from a normal hospital every day [5].

Since the CXR reading process is repetitive and time-consuming, using a well-trained machine learning model to read and classify the types of pneumonia can significantly improve efficiency and minimize labour overhead. Notably, machine learning methods like deep neural networks have proven their potential for different classification tasks including image classification [6].

Therefore, the goal of this project is to design a neural network model that can classify between normal patients and pneumonia-infected patients based on the input chest x-ray.

**2 Illustration & Figure**

The following diagram is the illustration of the model we used. The input images are resized to 500\*400 pixels and are inputted into a convolutional neural network to classify for the output. The output will be classified into two categories which are the normal and pneumonia-infected.

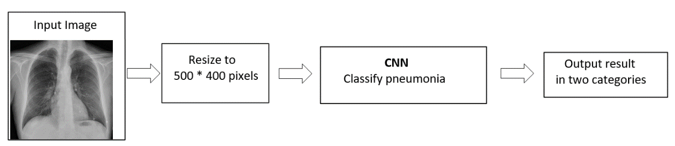
****

Figure 1 – Illustration of the neural network model we used

**3 Background & Related Work**

Since reading CRXs is systemic and time-consuming, deep learning is introduced to assist with faster diagnosis. Currently, deep learning and neural networks are widely applied in computer-aided diagnostics of diseases such as pneumonia, cardiomegaly and lung cancer. Existing work such as the “Automatic estimation of heart boundaries and cardiothoracic ratio from chest x-ray images” calculates the cardiothoracic ratio automatically which is an indicator that assists with cardiomegaly diagnosis and takes a significant amount of time to calculate normally [7]. Other work such as Jaiswal et al [8] can identify pneumonia and its potential cause using Fast RCNN with a training loss of 0.22.

**4 Data Processing**

The dataset is chosen among the chest x-ray images from retrospective cohorts of pediatric patients of age 1-5 from Guangzhou Women and Children’s Medical Center, Guangzhou [9]. The existing dataset has already removed low-quality pictures such as the ones with low contrast to ensure better training results.

The original raw dataset folder contains all 4887 images including chest x-rays of normal and pneumonia-infected patients. To categorize all images into a 70-15-15 random split into training, validation, and testing dataset along with its labels: normal and pneumonia, a Python script is created to perform such an operation [Appendix A]. The code simply opens the dataset folder and shuffles all images randomly using numpy.random.shuffle(). Then, it splits the dataset into 70-15-15 using numpy.split() and stores each set of image data into its corresponding folder. During the storing process, each image will be stored in the folder named after its label based on its image name which contains the label. All the images had different sizes and were resized to 500\*400 pixels using the resize() function. This dimension ensures that Google Colabotory does not run out of RAM.

However, after training with the primary model, the accuracy remains constant at 66% every epoch. We discovered that the number of the images labelled as pneumonia is twice as much as the other label, which caused an unbalance in the dataset. The model simply guessed pneumonia for all the training images and its accuracy remained the same at around 66%. Therefore, we used the os.remove() function call from the os library to selectively remove images labelled pneumonia to balance out the dataset. The following table shows the final distribution of each data set.

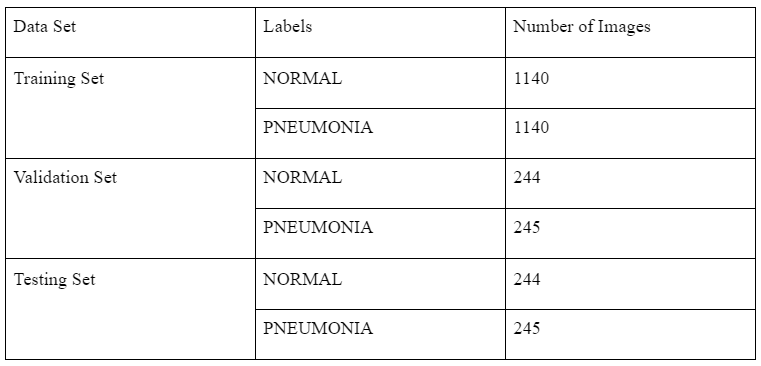


Table 1 - Final Data Distribution among Three Sets

**5 Architecture**

Initially, our team decided to use a CNN model to detect the Chest X-rays. The model takes in a 500x400 image input and passes it through a convolutional layer with a 5x5 filter and ten output channels. Then the model would perform a 2x2 max pooling with the processed data. A similar convolution process is performed with a 3x3 filter and twenty output channels with another 2x2 max-pooling layer. Afterward, there are three fully connected layers (FC layers) with 128 hidden units and 64 hidden units respectively in between (Figure 2).

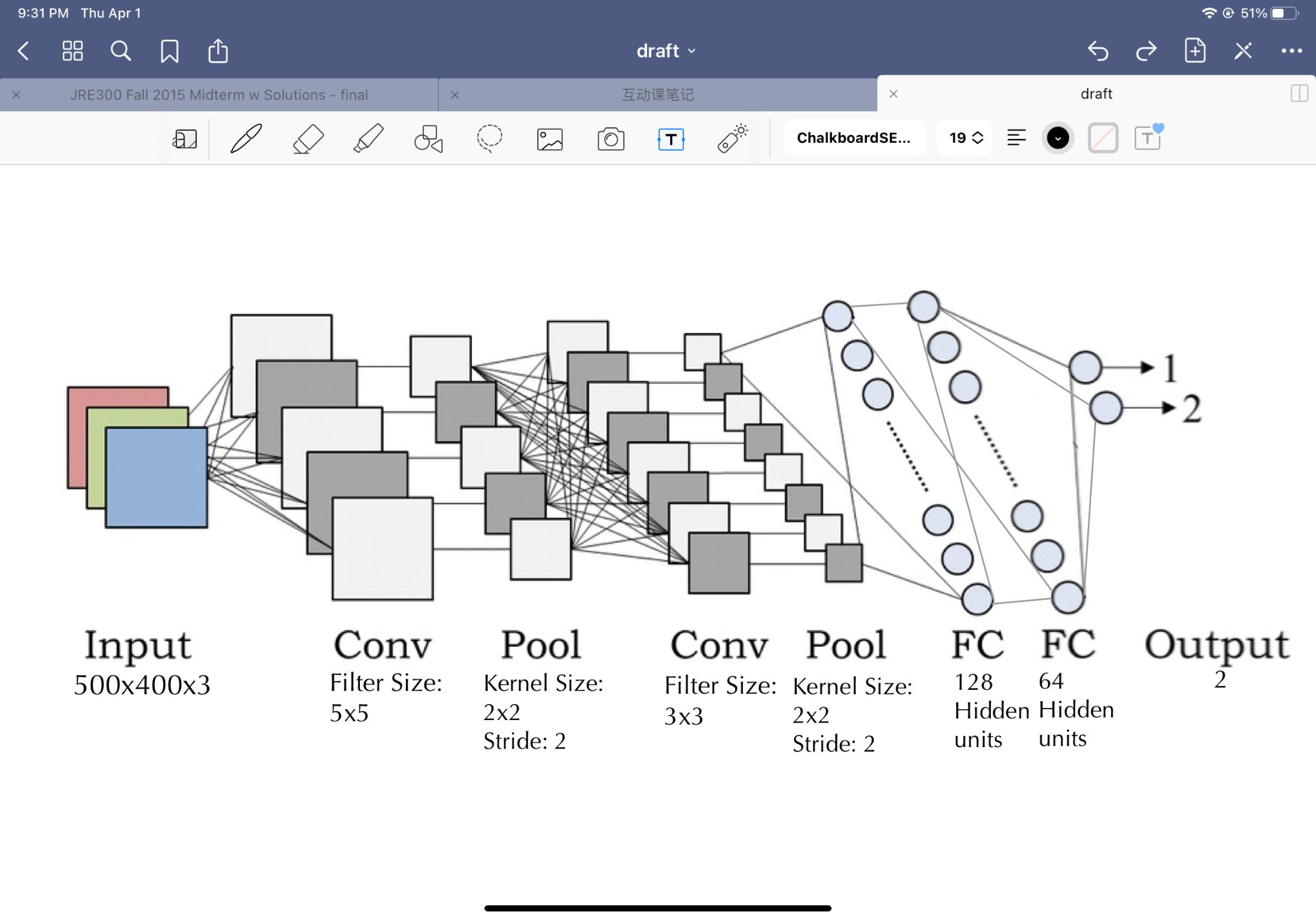


Figure 2 – CNN Architecture

However, we’ve encountered a GPU memory limitation with this model, which will be detailly discussed in section 12. With further research on the topic, our team decided to use the transfer learning method with AlexNet to reduce the computation power.

According to recent research, pre-trained models for the ImageNet database provide a performance boost in CXRs diagnosis [10]. Therefore, our team chose to use transfer learning with the AlexNet model. The pre-trained parameters in the AlexNet model are fed into our Alex\_NN() model that contains three FC layers. There are 2000 hidden units and 100 hidden units respectively in each FC layer (Figure 3). We choose to set up the FC layer with a larger number of hidden units following less hidden units to ensure our model captures essential information in the learning process. ReLu activation function is used between the FC layers.

Although the model performs as a binary classifier, we treat our model as a multi-classification model to acquire the classification probabilities. Since Pneumonia may lead to death, its diagnosis should be treated carefully. Our group decided that it is better to estimate a patient's likelihood of having Pneumonia rather than outputting the final result. Therefore, we set up two output units and used the softmax activation function in the final layers. Notice that we still make the final classification to find our model's accuracy throughout the training process.

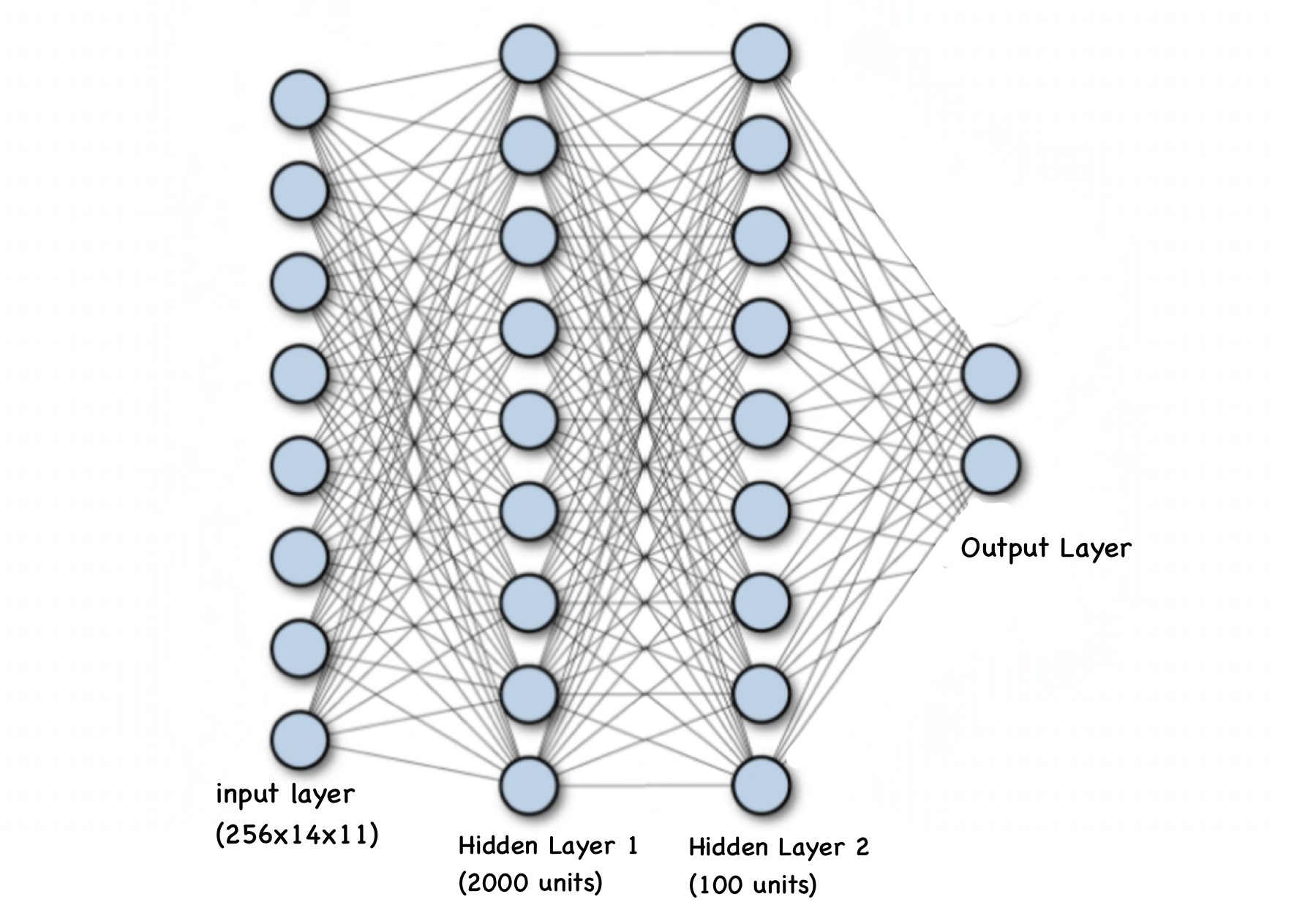


Figure 3 – Illustration of the neural network model we used

The model was trained using the Adam optimizer as it computes adaptive learning rates for each parameter and incorporates momentum in the calculation. The Adam optimizer is also computationally efficient and uses less memory [11].

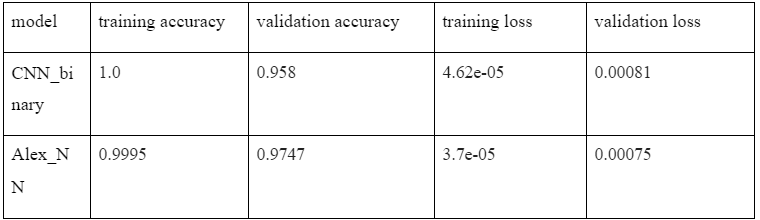
The following hyperparameters were used in training of our final model: num\_epoches = 48, learning\_rate = 5e-6, batch\_size = 128. They were tuned for acceptable accuracy.

**6 Baseline Model**

The baseline model we chose is Random Forest (RF) with 82% accuracy. This algorithm has contributed to image classification since it is capable of using categorical data sets, and not sensitive to over-fitting, and good at dealing with outliers in training data [5]. Therefore, we consider it would be a good model for image classification. The algorithm will construct decision trees based on the selected samples and the related labels, and remember the decision rules for each class. Thus, to use the RF, we need to randomly select a set of training data, and normalize the samples to call the algorithm. Once done, we can fit the training data in the model and predict the test result.

**7 Quantitative Results**

7.1 Comparing different models based on quantitative results

Table 2 - Training and validation results from different models

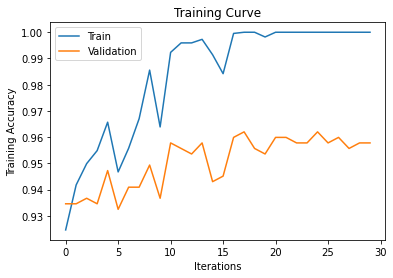
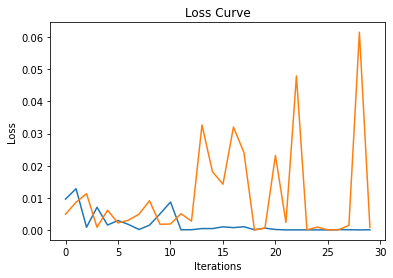


Figure 4 - Loss and accuracy plot for CNN\_binary

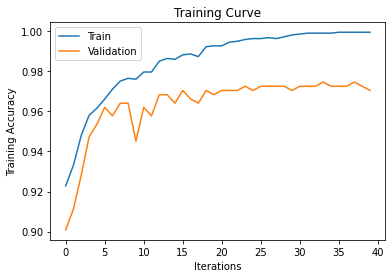
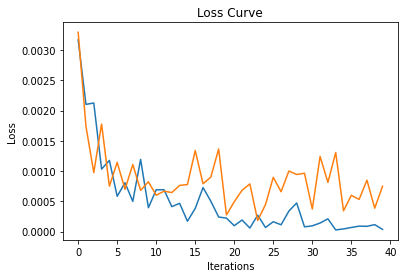


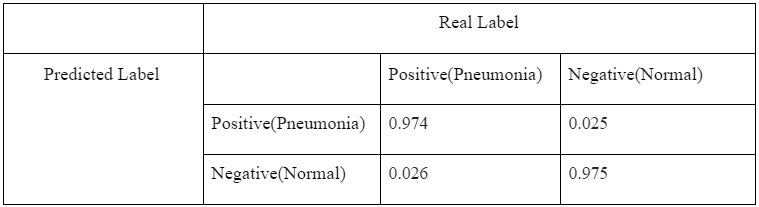
Figure 5 - Loss and accuracy plot for Alex\_NN

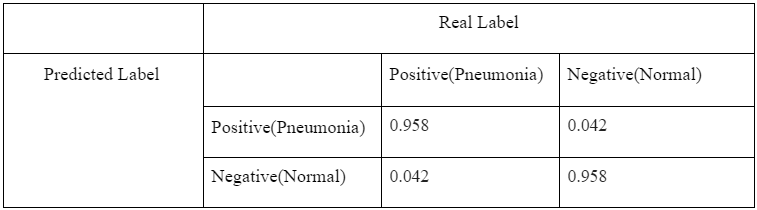
The table and figures above shows the final results from three different models after tuning parameters. The two models picked both have much higher accuracy than the baseline model.

The CNN\_binary model has a very high training accuracy and small training loss. But the validation accuracy is much smaller than training accuracy, and the validation loss shows a large fluctuation. So there may be a bit overfitting.

The Alex\_NN model has a trend of improving. The validation loss is decreasing and fluctuates in a small range, and validation accuracy is turning to stable. We picked the model at the 38th epoch out of the 40 epochs, since it has the highest validation accuracy.

7.2 Confusion matrix on the models

Table 3 - Confusion matrix for the Alex\_NN

Table 4 - Confusion matrix for the CNN\_binary

We care about the false negative rate. Since the false positive means the pneumonia is falsely labeled with normal, this situation will delay the treatment of the patients in real situations. Alex\_NN has a smaller false negative rate.

According to the analysis above, the final model we chose is the Alex\_NN with tuned parameters.

**8 Qualitative Results**

Comparing the model with quantitative results, we have picked Alex\_nn. With 97.5% validation accuracy, the model is expected to output accurate results for unseen samples. Our model will get the possibility of states and output the prediction with labels (e.g. “prediction for (patient) is normal”). For the representative samples below, all the 6 inputs get accurate results.

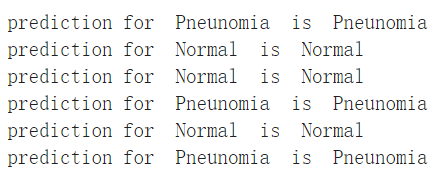
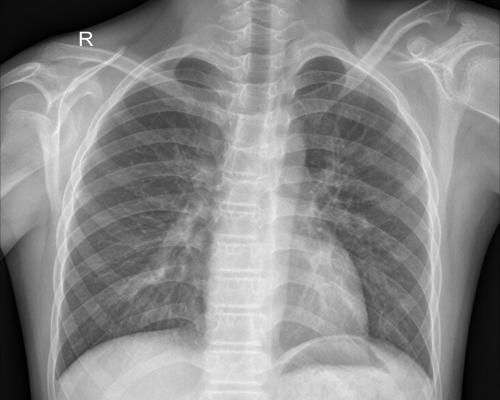


Figure 6 - Prediction output printed by our final model

**9 Evaluate Model on New Data**

9.1 Unseen Data

In **Section 4**,we have discussed the distribution of the data. The testing set remained unseen to our model, while the training set and validation set are used for training and tuning hyperparameters.



(a) (b) ©

Figure 7 - Example data images from (a) Training set; (b) Validation set; (c) Testing set.

The images in the testing set were randomly split during the data processing step, and were resized and transferred into AlexNet features by the AlexNet pre-trained model. These transferred AlexNet features would be fed into our trained Alex\_NN model to output the binary classification result. Finally, we would be using these outputs and compare them with the labels to measure the performance of our model.

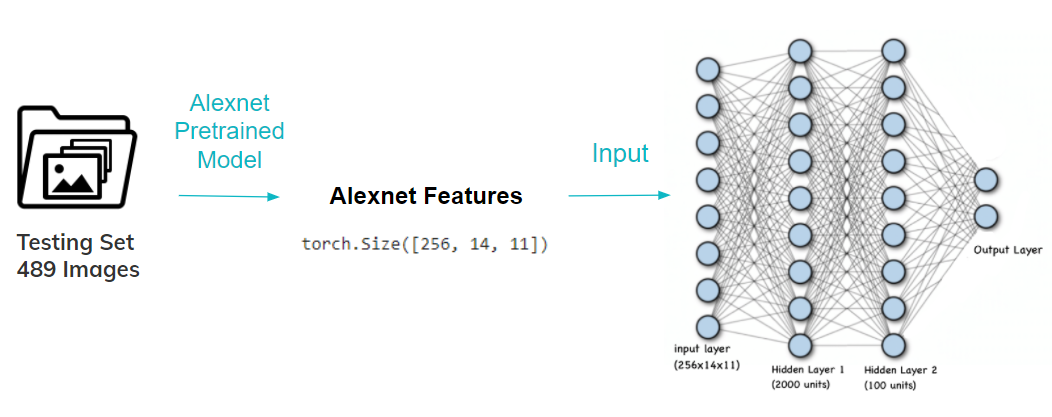


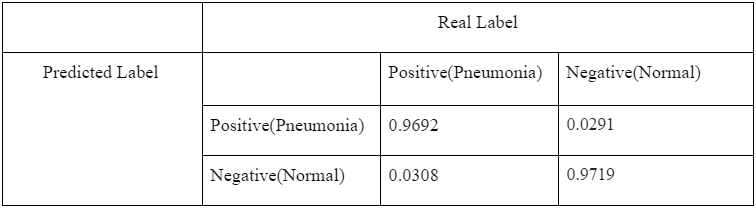
Figure 8 - Illustration of the model output classifications using the testing data

9.2 Expectation of Performance on the New Data

Our team would calculate the accuracy, false-positive and false-negative rate of the model by comparing its output with the testing set labels. Compared with validation results, we expect to get a lower accuracy and higher false rates on the unseen data since the validation data can be overfitted with hyperparameter tuning. However, our group still expects an accuracy greater than 96% on the unseen dataset because Figure 4 in **Section 7** shows that the model has validation accuracies that are mostly larger than 96% with little or no overfitting after 20 epochs. We also expected both false negative and positive rates to be below 4%.

9.3 Performance on the New Data

The Alex\_NN model achieves an accuracy of **97.06%** on the new data, which exceeds our expectations. The false rates are shown in the table below, which also meet the team’s expectations.

Table 5 - Confusion matrix for the Alex\_NN on the testing set

**10 Discussion**

10.1 Diagnostic Accuracy

The Alex\_NN model is performing well based on the observations from the previous section. It has an accuracy of 97.06%on the new data, meaning the model can diagnose 97.06% of the patients correctly.

However, misdiagnosing still exists. The false-positive rate and false-negative rate of the Alex\_NN model is 2.91% and 3.08%, meaning that 2.91% of the normal patients would be diagnosed with pneumonia, while 3.08% of the pneumonia-infected patients would be diagnosed as normal. The false-positive predictions can cause waste in medical resources by providing treatment on patients without pneumonia. And the false-negative predictions can cause more severe results, such as a delayed or unperformed treatment.

Despite these false rates, the Alex\_NN model still has a better diagnosing performance compared to human radiologists, who have an average diagnostic accuracy of 93.1% [12]. Therefore, our model is approximately 4% more accurate than professional human radiologists on diagnosing pneumonia with CXRs.

10.2 Time Efficiency

Our group recorded the time for the whole process of reading CXRs with our Alex\_NN model. The process includes loading data images, transferring images to AlexNet features, inputting features to the model, and using the model to output predictions. This whole diagnosis process took 65.1 seconds with an average of 0.133 seconds for each testing sample, significantly lower than the average reading time of 1-2 minutes by a radiologist. Therefore, using our model is much more time-efficient.

10.2 Conclusion

By having approximately 4% more accuracy and much less time spent than the human radiologists, we can conclude that our Alex\_NN model performs well. It exceeds the team’s expectation on meeting the project goal to effectively and efficiently classify pneumonia-infected patients based on the input chest x-ray.

**11 Ethical Consideration**

The ethical issues with the task are data sourcing and clinical development. For data sourcing, the X-ray images from patients may have representation and evaluation bias. Patients with different genders or different races may have different types of symptoms while the data for training and testing we got was only from Asian hospitals. Therefore, our model now fails to generalize to the users and may not be trustworthy in real situations. For clinical development, when applying the technique to a clinic, the wrong prediction may result in serious health issues. If a pneumonia patient is accidentally diagnosed as normal, the patient may delay the timing of treatment. Also, the doctors will need to explain to patients about the disease in detail, and the wrong result from the program will waste more time for doctors.

**12 Project Difficulty / Quality**

The project's overall difficulty is similar to what we expect at first, and the process is challenging yet manageable. Our team has gained a lot of hands-on experience and knowledge of machine learning through this project. The following is a summary of the problems we have encountered and their solutions.

***Unbalanced Dataset***

We realized that our data is unbalanced with 3500 pneumonia x-rays and 1500 normal x-rays. The unbalanced data cause the model to constantly classify the X-rays as "infected," which results in around 0.66 accuracies and a high false-negative rate. We balanced the data by manually excluding some pneumonia samples. Although this made our data size smaller, the CNN model's validation accuracy is still acceptable. Additionally, in our final model, the AlexNet pre-trained parameters incorporate data augmentation as well [13]. Although the data augmentation strategy is applied to the ImageNet data, it can effectively increase the amount of data and boost CXR diagnosis technology accuracy [10].

***GPU/CPU Memory***

The limitation of memory space is another obstacle in our project. Each x-ray image in the original dataset we obtained has a large resolution. Passing the images to the CNN requires a reasonably large memory on the GPU side. We often get the "CUDA/RAM out of memory" warning message, which forced us to reduce our batch size. However, as the batch size gets smaller, the model cannot capture informative gradient descents. Therefore, we changed our model to AlexNet transfer learning. The pre-trained data provided by AlexNet reduced the use of GPU memory and increased the accuracy of the model.

***High Similarity Between Bacteria and Viral Infection***

Another problem is that CXR readings can hardly distinguish between virus infection and bacterial infection. To verify this notion, our team contacted a professional radiologist for confirmation. We learned that CXRs reading could only be used to diagnose Pneumonia. To determine whether it is a virtual or bacterial infection, doctors often need further tests such as CT or blood tests. Our team decided to match our model's usability to reality, which is solely using CXR to diagnose the probability of having Pneumonia.

The final accuracy of our model performs better than our expectations. The decrease in the amount of data caused by balancing does not majorly affect our final model's accuracy. Our final model's result is still reasonable as it slightly outperforms the general diagnostic accuracy of CXR [12].

**Reference**

[1] B. Dadonaite and M. Roser, “Pneumonia,” *Our World in Data*, 04-Nov-2018. [Online]. Available: https://ourworldindata.org/pneumonia#burden-of-pneumonia. [Accessed: 11-Feb-2021].

[2] “Global estimate of the incidence of clinical pneumonia among children under five years of age,” *World Health Organization*, 04-Mar-2011. [Online]. Available: https://www.who.int/bulletin/volumes/82/12/rudan1204abstract/en/. [Accessed: 11-Feb-2021].

[3] S. Naicker , J. Plange-Rhule, R. C. Tutt, and J. B. Eastwood, “Shortage of healthcare workers in developing countries--Africa,” *Ethnicity & disease*. [Online]. Available: https://www.ethndis.org/priorsuparchives/ethn-19-01s1-60.pdf [Accessed: 11-Feb-2021].

[4] Apps.who.int. 2021. [online] Available at: <https://apps.who.int/iris/bitstream/handle/10665/66956/WHO\_V\_and\_B\_01.35.pdf> [Accessed 11 February 2021].

[5] Medium. 2021. *Can Machine Learning Read Chest X-rays like Radiologists?*. [online] Available at: <https://towardsdatascience.com/can-machine-learning-read-chest-x-rays-like-radiologists-part-1-7182cf4b87ff> [Accessed 11 February 2021].

[6] Yadav, S. and Jadhav, S., 2021. *Deep convolutional neural network based medical image classification for disease diagnosis*.

[7] Ahmed H. Dallal, Chirag Agarwal, Mohammad R. Arbabshirani, Aalpen Patel, Gregory Moore, "Automatic estimation of heart boundaries and cardiothoracic ratio from chest x-ray images," Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis, 101340K (3 March 2017); <https://doi.org/10.1117/12.2254136>

[8] Amit Kumar Jaiswal, Prayag Tiwari, Sachin Kumar, Deepak Gupta, Ashish Khanna, Joel J.P.C. Rodrigues, Identifying pneumonia in chest X-rays: A deep learning approach, Measurement, Volume 145, 2019, Pages 511-518, ISSN 0263-2241, <https://doi.org/10.1016/j.measurement.2019.05.076>.

[9] P. Mooney, “Chest x-ray Images (Pneumonia),” 24-Mar-2018. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. [Accessed: 07-Apr-2021].

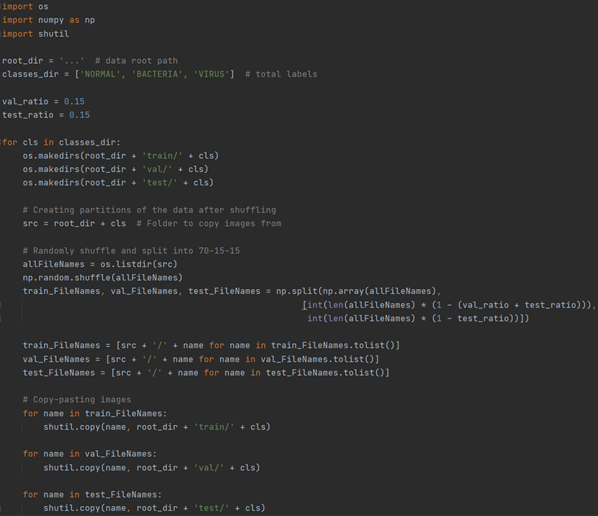
[10] Ke, Alexander & Ellsworth, William & Banerjee, Oishi & Ng, Andrew & Rajpurkar, Pranav. (2021). CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation.

[11] N. Kapoor, “Optimizers,” *Medium*, 19-Nov-2020. [Online]. Available: https://towardsdatascience.com/optimizers-88694509311c. [Accessed: 07-Apr-2021].

[12] Taghizadieh, Ali & Ala, Alireza & Rahmani, Farzad & Nadi, Akbar. (2015). Diagnostic Accuracy of Chest x-Ray and Ultrasonography in Detection of Community Acquired Pneumonia; a Brief Report. Emergency (Tehran, Iran). 3.

[13] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.

**Appendix A - Python Script of Splitting and Labelling Data**

****