APS360 PROJECT PROGRESS REPORT

Haoo Wang
Student# 1006669384
hbo.wang@mail.utoronto.ca

Ariana Lin
Student# 1007113053
ariana.lin@mail.utoronto.ca

Rosalind Wang
Student# 1006930519
rosalind.wang@mail.utoronto.ca

Muchen Liu
Student# 1006732145
muchen.liu@mail.utoronto.ca

1 PROJECT DESCRIPTION

Pneumonia, as it has been discovered by human beings, has been affecting millions of people every year, and brought about four million people to death(Ruuskanen et al., 2011). In many developing countries, pneumonia is still the major cause of death among old and young people. The goal that our team wants to achieve is to develop a machine learning model to diagnose pneumonia is that by training a model to detect pneumonia through X-ray images could improve the efficiency and the accuracy of diagnosis compared to doctors checking it themselves. Currently, our model is forged with a convolutional neural network to extract the features from the image imputed, after the feature extraction and flattening, the output would be connected to an artificial neural network which contains two fully connected layers for classification.

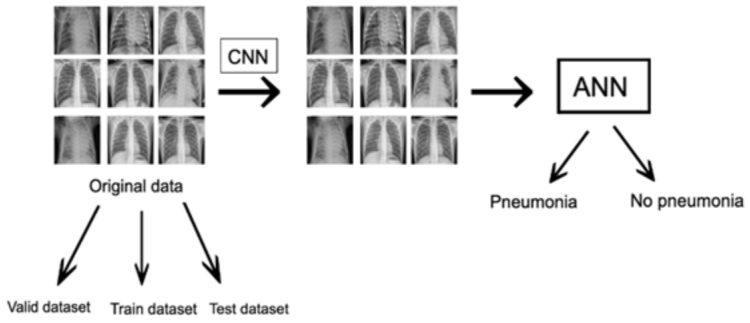

Figure 1: Structure of the model (CNN + ANN)

2 INDIVIDUAL CONTRIBUTIONS AND RESPONSIBILITIES

2.1 PROJECT MANAGEMENT

The team chose to discuss through WeChat as the communication method. The group unanimously agreed to set Tuesday at 7:00 pm Toronto time as the weekly meeting time. This way we can discuss if and how the new things we learned in Monday's lecture can be applied to our projects and have the opportunity to ask the TA relevant questions in Wednesday's tutorial. If there are special circumstances (travel arrangements, midterms, etc.) that prevent a group member from attending a meeting, then the rest of the group needs to be notified at least 24 hours in advance so that they can decide whether to change the meeting time or to have one of the participants relay the content of the meeting to the absent group member after the meeting normally arranged.

Figure 2: Overview of project management for the progress report

We use the Status Sheet to keep track of the project progress. It does not contain details such as project duration and task relationships but focuses more on the project status and the process of completion. The project status sheet also includes the person who performed each task, so the project leader can have a better understanding of the progress of each team member and know who is responsible for problems when they occur.

Tasks	Assigned To	Priority Level	Progress Rate	Status
Data Loading	All	*****	100%	Done
Data Processing&Cleaning	A&R	*****	100%	Done
Baseline model	H&R	****	100%	Done
Classifier	M&A	***	100%	Done
Primary model	All	****	100%	Done
Improve&Tune	W&R	*****	30%	In progress
Sanity Check	M&A	**	0%	Planned
Final Testing	All	*	0%	Planned

Figure 3: Status Sheet

There will be a strict upper limit of 9 pages for the main text of the initial submission, with unlimited additional pages for citations.

2.2 Responsibility and Accomplishment

According to the requirements of the progress report. So far, the team should have obtained all or most of the data. This means the tasks of data cleaning, constructing the baseline model and the initial feature engineering module and the classifier, providing the preliminary results of the baseline and primary model.

Table 1: **The tasks division and status**

Tasks	Responsible member	Status
Data cleaning	Ariana Lin,Rosalind Wang	June 24
Baseline model	Haobo Wang,Rosalind Wang	June 28
Classifier	Muchen Liu,Ariana Lin	July 7
Primary model	All team members	July 10

Ariana Lin and Rosalind Wang were responsible for data processing and cleaning. They need to load the medical imaging file and normalize the data. Although the data is automatically split into training, validation, and test sets, they split the data by the proportion of 0.7, 0.15 and 0.15. Haobo Wang and Rosalind Wang were responsible for making decisions on the baseline model and generating initial results. Muchen Liu and Ariana Lin needed to develop modules that take in the feature engineering module's consolidated features. The whole team needed to compare the performance according to different models of deep machine learning and according to hyper parameters. In the future steps, the team plans to improve the model by tuning the values of different hyper parameters (convolutional layers, number of epochs etc.), respectively. We will test the model and print all the relative results on Google Colab to keep track of how the modification of the model influences the result. The whole team will contribute equally and pass the final result.

Table 2: **The distribution of work for the rest of the project**

Tasks	Task Description	Assigned To	Deadline
Final Model Iteration	Improve and tune classifier module	Haobo Wang,Rosalind Wang	July 30
Final Model Iteration	Sanity Check	Muchen,Ariana	August 1
Final Model Iteration	Final Testing	All team members	August 2
Final Model Iteration	Final Report Initiation	All team members	August 7
Project Presentation	PowerPoint,Recording	All team members	August 8
Project Final Report	Finalize Report	All team members	August 15

3 Notable Contribution

This part provides a detailed summary with results on this project of the team.

3.1 Data Processing

The data of chest X-Ray images of pneumonia is from patients aged 1 to 5 years from Guangzhou Women's and Children's Medical Center by accessing the public data platform, Kaggle, published by PAUL MOONEY in 2018(MOONEY, 2018). Although the data is already divided in testing, training, and validation including normal chest X-Ray images and pneumonia chest X-Ray images, the team reclassified the images for testing, training, and validation in a more proper proportion.

The chest X-Ray images from kaggle published by PAUL MOONEY are in JPEG format, which basically has no effect on our project compared with JPG format. There are 5856 images in the dataset, with 1583 normal and 4273 pneumonia. Since the inputs of the neural networks should be in the same size, the team resized images into the same size to easily get these images loaded and processed. As exceeded shrinking will lead to deformation of features and patterns inside the image, the team resized the original images with dimension of 1500 *1400 to 100*100.

After cleaning and resizing all the images to a considerably smaller size (100*100), the team decided to use 70 percent of each class to be in the training set, 15 percent to be in the validation set, and 15percent to be in the testing set. Therefore, there would be 1110 normal and 2990 pneumonia in the training dataset; 237 normal and 642 pneumonia in the validation dataset; and 236 normal and 641 pneumonia in the test dataset.

To ensure randomness in the splitting process, the team randomly selected the images that appear first from datasets. For example, the team used the first 1110 normal images from the datasets as our training dataset.

Since machine learning algorithms cannot operate on label data directly, the team converted all input variables and output variables to be numeric. The one-hot encoding would be applied to change the label 'normal' to 0, and 'pneumonia' to 1.

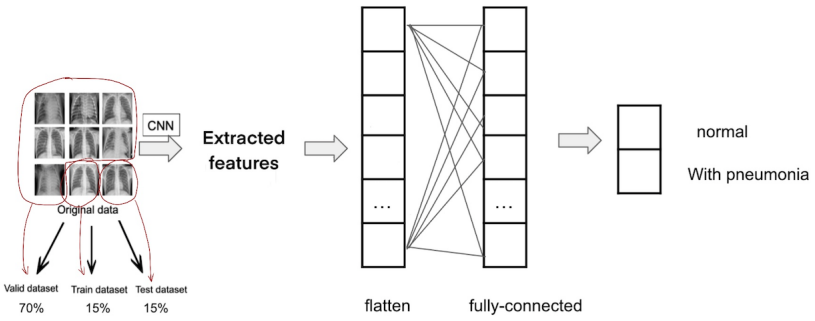

Figure 4: Detailed structure of the model(CNN + ANN)

3.2 Baseline Model

The baseline model is inspired and modified from the models in the tutorials and practical, which consists of one convolutional layer, with a conv2D and a max-pooling operation. ReLu activation function is used in the baseline model. Hyperparameters in the baseline model include kernel size, number of epochs, number of convolutional layers, steps per epoch, validation steps, etc. In this model, the result of the accuracy turns out to be approximately 0.3, which can be evidenced by the figure below. In terms of the quantitative and qualitative results, the learning curves produced by the model are shown in figure[6]. Based on the learning curves, the model's loss has a general decreasing trend, while the model's accuracy is relatively low. For the qualitative results, it can be observed that the model loss fluctuates at a higher value, which indicates that the model is not learning the data very well. The reason for this situation may be that the convolution layers and the epochs are not enough, leading to an occurrence of underfitting.

```
test accuracy: 0.2690992057323456
test loss: 9.802160263061523
```

Figure 5: Accuracy and loss report of the baseline model

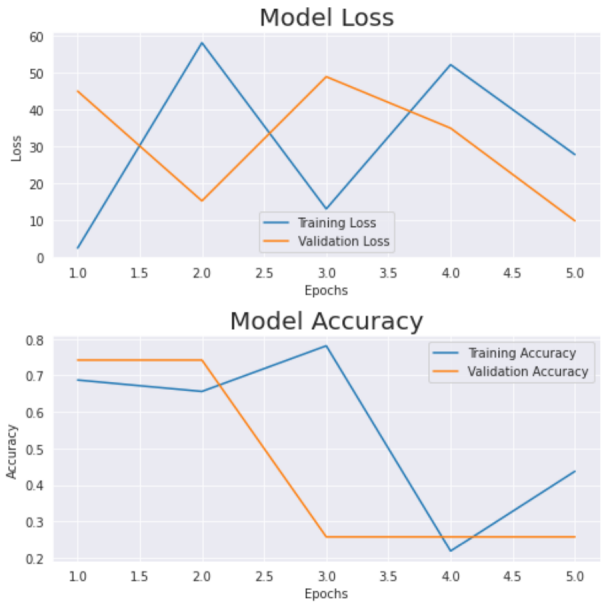

Figure 6: Learning curves of the baseline model

An unexpected challenge the team met during the training process is the learning curves keep fluctuating. From the figures, there are many dramatic increases and drops that occur in the learning curves. This challenge may have a negative impact on the analysis of the data and the reliability of the project. Overall, the complexity of the baseline model can be considered relatively simple and feasible and easy to be implemented.

3.3 Primary Model

The primary model that our group has built is constructed with the idea of combining convolutional layers that connect with fully connected layers. In the current stage, our primary model consists of four convolutional layers and two fully connected layers. The major task of convolutional layers is feature learning, by inputting images into the convolutional layer, some low-level features would be first extracted. After more convolutional layers are added, some high-level features would be learned by the model for future classification. The activation function we used for convolutional layers is ReLu since the calculation of it is simple and it could work well with a convolutional neural network. In our convolutional layers, a 3*3 kernel is used in all four layers, the first convolutional layer requires an input shape of (150*150*3) and it has 32 filters (which would give 32 outputs). Some padding was applied in order to keep the size of the output the same as the input. Most parameters for the second convolution layers are the same as the first layer, yet we want 64 outputs. Similarly, the third and fourth convolutional layers would generate 128 outputs. Also, 2*2 max pooling is applied after each of the convolutional layers, because it could efficiently reduce the spatial size of extracted features to reduce the time consumed in computing. After extracting features from the convolutional layers, the model would flatten the layer and move the outputs into fully connected layers for classification. The first fully connected layer has 64 output space, with a relu activation function and 12*12 kernel size. The last layer is also a fully connected one, yet with the softmax activation which would give us a result with a probability distribution.

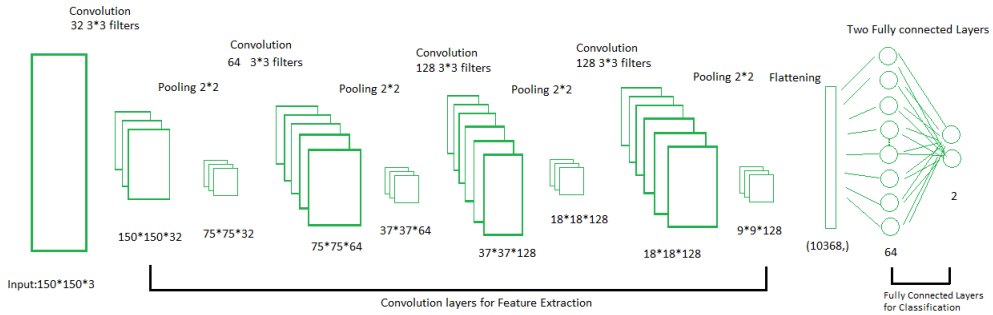

Figure 7: Architecture of the model

For quantitative data, through 10 epochs of training and validation, we have obtained training curves like the figure shown below.

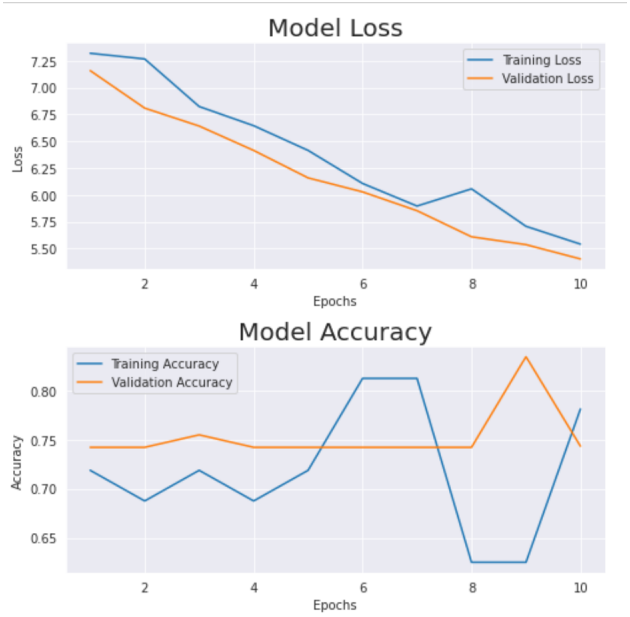

Figure 8: Learning curves of the primary model

A future challenge for us to solve is that we could see that the loss kept decreasing, yet the model accuracy is still quite unstable. Also, a test set is used, and the loss is 5.405 with an accuracy of 0.734. The interesting part about the result of this model is that although it is considerably fine, however, the loss and accuracy curve are not classical ideal training curves.This may be due to the immature hyperparameters and the small number of epochs, therefore, there still are improvements that could be addressed in our future steps.

REFERENCES

PAUL MOONEY. Chest x-ray images (pneumonia), 2018. URL `https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia`.

Olli Ruuskanen, Elina Lahti, Lance C Jennings, and David R Murdoch. Viral pneumonia. *The Lancet*, 377(9773):1264–1275, 2011.