**APS360 Applied Fundamentals of Machine Learning**

**Project Proposal**

**Group members: Charles Huang, Emma Zhao, Hannah Zhang, Tim Guo**

**1 Introduction**

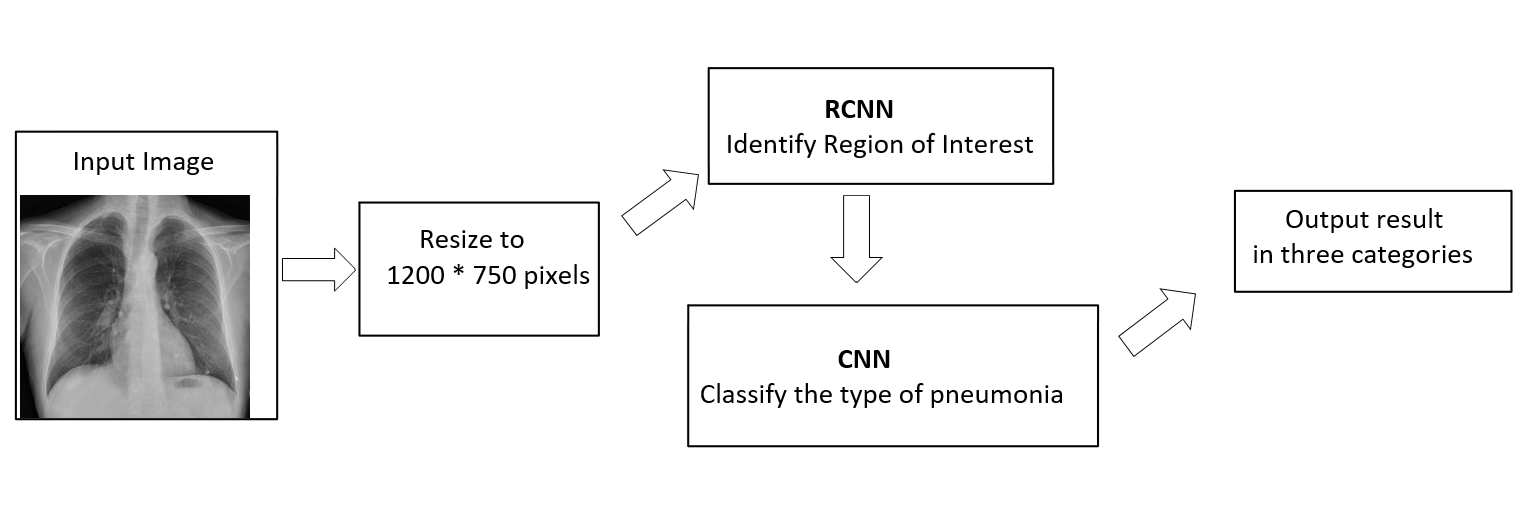
Over 100 million people get infected with pneumonia globally each year, and more than 1 million children under 5 years old are killed by pneumonia annually.[1][2] In some regions, the high death rate of pneumonia can be caused due to the shortage of medical resources and healthcare workers[3]. So, an inexpensive method to diagnose pneumonia effectively is needed.

Chest X-rays (CXRs) are widely applied for pneumonia diagnosis due to their low cost and efficiency.[4] However, the process of reading CXRs is systematic and time-consuming. It takes an average of 1-2 minutes for a well-trained radiologist, which is a rare and expensive resource for some regions, to read one CXR, while a hospital can generate hundreds of CXRs per 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 architecture for pneumonia image classification tasks.

**2 Illustration & Figure**

****

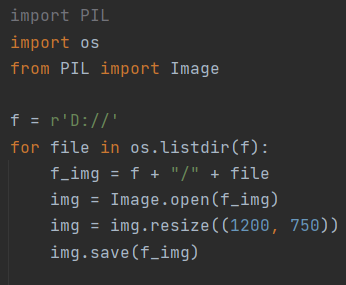
**3 Background & Related Work**

Due to the systemic and time-consuming process of reading a CRX, 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] is able to 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 one to five years old from Guangzhou Women and Children’s Medical Center, Guangzhou. Low-quality pictures such as the ones with low contrast have already been removed from the dataset to ensure better training results.

However, these images do not have identical sizes and we proposed to use simple code in python to resize all images into 1200\*750 pixels. The code demonstrated below simply iterates through all images in the folder and resize every single one into the size of 1200\*750 pixels. The resized dataset can be used as inputs to our CNN. By running the following code in a Python shell, it will resize all images within the designated folder f which will contain all datasets.



Dataset: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

**5 Architecture**

The architecture we decided to use is Fast RCNN combined with CNN. Since the dataset of patients is a chest x-ray, we need to focus the analysis on the lung part on the image using Fast RCNN. Then the ROI images are passed through CNN for classification as we did in labs. To use the architecture, data images are normalized and passed in a Fast RCNN model. There can be outputs after the Fast RCNN model to check the accuracy of ROI. Then we can read the training result from CNN.

**6 Baseline Model**

Our baseline model is random forest which had a substantial impact on medical image computing over the last decade [9]. The essence of this method is to build multiple trees in randomly selected substances of the feature space [10]. Using random forest, some samples are randomly selected from a dataset, then the algorithm will construct decision trees based on the selected samples and the related labels. Therefore, to apply random forest, we need to randomly select a set of data, and normalize the samples in order to use the algorithm. The algorithm will then output the result of training.

**7 Ethical Considerations**

The ethical issues with our task are grouped to two categories: data sourcing, and clinical development. Firstly, when collecting data, patients with different genders or different races may have different types of symptoms. If data is not comprehensive, the result of the training may be biased. If the dataset is not accurate, the training result will not be trustworthy when it is applied to real situations. When the technique is applied in clinics, the inaccurate results provided to customers will be even more problems to the doctors. They still need to explain to patients about the disease in detail, and even need to explain the wrong result from the program.

**8 Project Plan**

*8.1 Task Summarization*

|  |  |  |
| --- | --- | --- |
| Name | Proposal  (Completed) | Project  (Refer to the timeline below) |
| Charles Huang | -Background & Related Work  -Data Processing | 1a, 2a, 3, 4b, 5, 6 |
| Emma Zhao | -Architecture  -Baseline Model  -Ethical Consideration | 1b, 2b, 3, 4a, 5, 6 |
| Hannah Zhang | -Project Plan  -Risk Register | 1a, 1c, 2a, 3, 4b, 5, 6 |
| Tim Guo | -Introduction  -Illustration & Figure | 1b, 2b, 3, 4a, 5, 6 |

*8.2 Timeline*

We decided to divide the project into several milestones in order to precisely track our completion. The following list outlined our plan on this project.

1. Milestone 01 (deadline: Feb 24th, 2021)
   1. Data cleaning / processing
   2. Baseline Model Implementation
   3. Model Training and Verification (including unit testing)
2. Milestone 02 (deadline: March 4th, 2021, prone to change base on the progress report deadline)
   1. Implementation for the neural network (NN)
   2. NN training and verification (including unit testing)
3. Progress Report (deadline: TBD, approximately on week 8)
   1. Works will be divided evenly among group members once the rubric is out
      1. All the tasks below will be separated to smaller part according to the sections individual team member is responsible for
   2. Report read through
4. Milestone 03 (deadline: March 18th, 2021)
   1. Incorporate the suggestions from the progress meeting with TA to improve our model
   2. The second round of NN training and verification
5. Presentation (deadline: March 25th, 2021)
   1. Works will be divided evenly among group members once the rubric is out
      1. All the tasks below will be separated to smaller part according to the sections individual team member is responsible for
   2. Make slides and write scripts
   3. Presentation run through
6. Project Report (deadline: March 25th, 2021)
   1. Works will be divided evenly among group members once the rubric is out
      1. All the tasks below will be separated to smaller part according to the sections individual team member is responsible for
   2. Documentation of individual team members’ contribution

*8.3 Meeting Schedule*

Our team will be holding meetings on a weekly basis, with constant group messaging in case any emergency happens. The meeting time will be either Wednesday 6:30 - 8:00 pm, or Thursday 9:00 - 11:00 am. In the meeting, we will go over the obstacles we face and decide whether to reach out to the TA for help. WeChat and Zoom will be our primary communication method, we have also collected the members’ emails and phone numbers as back up. During the project period our team have agreed on the following Code of Conduct:

1. Member needs to reply to any work-related message sent as an acknowledgement
2. Due to the timezone difference within the team, we understand that there might be a delayed response. Hence, we have all agreed on sending a daily work report every night summarizing the current issues and the modifications.

**9 Risk Register**

*9.1 Failure in Data*

1. Different resolutions might interfere the result
   1. We need to go into the dataset and manually delete the data set with lower image quality
2. Lack of dataset
   1. To solve this we will need to find similar datasets online through different competition sites or Github resources.

*9.2 Failure in model*

1. The created model is not able to classify different kinds of chest diseases
   1. We will need to research for a better model or alter the hyperparameters to find the appropriate solutions
2. Training time longer than expectation

*9.3 Teamwork and collaboration*

1. Time zone difference among team members
   1. We have members currently living in both the EST timezone and CST. We will solve any potential issues through thorough communication, and strictly follow the code of conduct

**10 Link to Colab**

<https://colab.research.google.com/drive/1RgAl8L2E9QQ3LaMLcqeeXKABJZ2KwcZ5?usp=sharing>

**References**

[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] Sciencedirect.com. 2021. *Random Forest - an overview | ScienceDirect Topics*. [online] Available at: <https://www.sciencedirect.com/topics/engineering/random-forest> [Accessed 11 February 2021].

[10] Web.archive.org. 2021. *Wayback Machine*. [online] Available at: <https://web.archive.org/web/20160417030218/http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf> [Accessed 11 February 2021].