CS598 Deep Learning for Healthcare: Project Proposal

Ujjal Saha
ujjals2@illinois.edu

Ashish Pradhan
apradh6@illinois.edu

Dilip Ravindran
dilipr2@illinois.edu

Faizan Khan
fkhan71@illinois.edu

1. INTRODUCTION

In the course CS598 Deep learning for Healthcare we learnt how Data Scientists provide effective solutions to various healthcare related problems using Deep Learning models which is backed by state-of-the-art AI (Artificial Intelligence) techniques. During the course we not only learned how to work with structured data as well as unstructured data, but went on to implement supervised learning as well as unsupervised learning such as clustering, regression, classification, prediction etc. Focusing on various types of Neural Networks and its implementations, their extensive usage in the medical field gave us great exposure to the real world examples that used CNN (Convolutional Neural Networks), RNN (Recurrent Neural Network), GNN (Graph Neural Networks) etc. We submitted assignments in the form of programming exercises on some benchmark datasets, primarily using PyTorch in Python. This is the stage where we completed the theoretical knowledge of the course and we will now be doing a Project that will solve a real world problem using the techniques that were taught throughout the course. The projects were mentioned in the Project requirements documents [1] to choose from.

2. MOTIVATION

It is critical to accurately classify positive Covid-19 cases because of its comparatively high R0 factor and mortality rate compared to other widespread airborne disease like the common flu. Not just that, classifying positive cases quickly is of paramount importance in order to stem the spread, as the most important form of defense against the virus is to socially isolate positive cases. As the virus primarily attacks the lungs and hampers pulmonary functions, leading to inflammation and even respiratory failure, Chest Xray images can be the only form of data which researchers can actually visualize and therefore diagnose with.

2.1 Problem Statement

With the COVID-19 pandemic still around and with the chest x-ray data set availability we are trying to do a classification modeling where we can classify a new chest x-ray data as that from a covid-19 patient or not. We will build a trained data set from the covid-19 chest x-rays and then using CNN we will build the model. There are already quite a few covid-19 predictions model algorithms present, and we will consider those as our comparison model. We will also do a survey where we will present our model and a comparison of the existing model that we found on our research as well as literature survey. As many approaches are going to improve the accuracy, we are reproducing the results of the existing prediction models and try to find a scope where we can contribute to further improvements either in accuracy metrics or in resource utilization.

2.2 Care for the Problem

The current testing process is mostly done using the saliva or nasal swab sample that takes many hours or even days for the results. Faster diagnosis can be achieved using image classification of Xray images as it takes mins to scan them and few more mins to get the end image. So, for hospital patients with pneumonia this can be a faster way of testing or this can even serve as an eligibility criterion for further testing for confirmation of COVID. Moreover, this is a fresh research field where better approaches and algorithms are being developed, so our project could also contribute to the research with metrics data and comparison charts.

3. LITERATURE SURVEY

We have done considerable research about how using CNN detection of COVID-19 can be achieved using chest Xray dataset [3] [4] [5] [6] [7] [8]. We have also reviewed various studies on CNN models, the various layers that were used, forward passes, filters etc. Various data source providers that were referred in the project requirement document were studied and discussed about what will probably be the best fit for our project. Based on our research and literature survey (references are provided in reference section), few models have already been proposed with good accuracy. We will reproduce those models and aim for improving the metrics or utilization of computational resources.

4. DATASET

Since chest X-ray data for COVID-19 is limited, data will have to be collected from multiple sources. Following is the plan for collecting data for the project from publicly available repositories.

1) Collect confirmed COVID-19 Xray cases from multiple sources [9] and [10]. There are 514 such images.
2) Collect chest X-rays that have been tagged as non-COVID from NIH chest X ray dataset [11]. Select 2000 images.

Thus, the entire dataset would be ~2500 images, combining the COVID and non-COVID cases. The proposed dataset partitioning is as follows – 70% training, 20% validation and 10% test. Images would be downsized to 224 x 224 pixels to bring in uniformity.

Since data is collected from multiple sources, the related metadata is often in different formats and may not have a uniform pattern. Due to this, it may not be possible to do a lot of analysis based on demographics like age, gender, or country. During our initial research we found the images data can be combined from multiple sources as mentioned above, but during the project if we encountered any challenge that would restrict is combining any data, we will use only 1 data set and still try to achieve our project goal.

5. APPROACH

Our approach is primarily aimed at replicating existing research on classification of Xray images, focusing on the ones which have been applied on Covid Images. The data will be sourced from 2 different sources due to the limited number of images from a single source. Normal images will be sourced from the NIH Xray Data Repository. Extensive preprocessing might need to be done to standardize the images to maintain uniformity for the model to train on. We will aim to publish detailed comparison of at least 3 existing models which have performed well with high accuracy, precision, recall, AUC and F1 score. After successfully replicating the models and if we have time remaining, our intention is also to create a new model with a unique combination of various layers which might improve those metrics without compromising on computation speed. The pipeline for our project will consist of the following steps as shown in the below **Figure 1**:

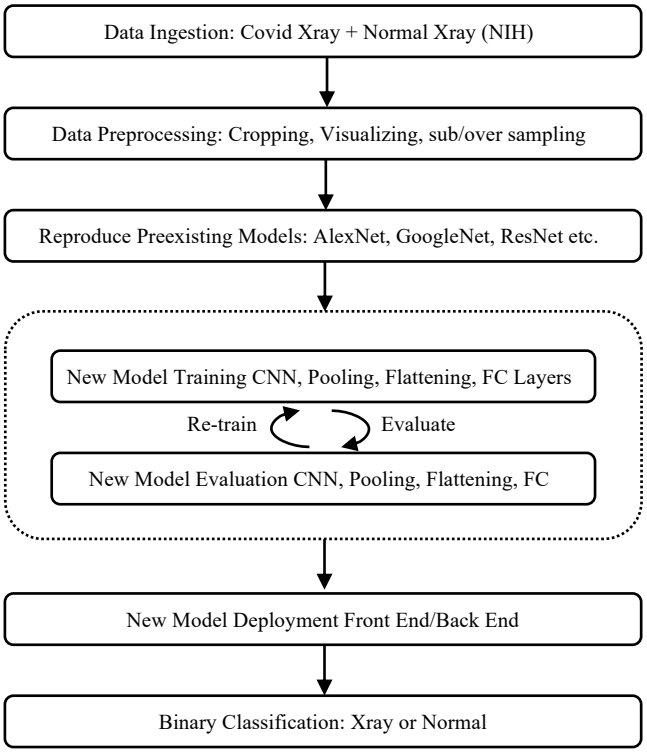

Figure 1: Proposed steps that will be followed as the solution for the problem during the project execution phase

6. EXPERIMENTAL SETUP

The hardware will primarily be laptops with inbuilt GPUs for higher computational power. If required, AWS resources (SageMaker, EC2 instances with GPUs) can be used to run the high computational modeling parts. Project will be done solely using the Python Language and will use Pytorch for deep learning modules. TensorFlow could also be used if needed for more effective visualizations. AWS and Google Cloud offers special student credit program to use their respective platform and we will try to use the University student account to take advantage of the credit program.

7. PROJECT TIMELINE

We will follow a Friday to Friday Sprint. We will call each Sprint with sprint number with 1st sprint as Sprint1 so on and so forth. Below **Table 1** shows tentative timelines that we are planning to follow for the project deliverables and completion.

Sprint1 04/02– 04/09	1) Dev/Test Environment Setup 2) Dataset analysis and fetch procedures 3) POC with CNN Modeling with the dataset 4) Additional Paper Readings 5) Draft Documentation
Sprint2 04/09– 04/16	1) Reproduce CNN Model for the existing 2-3 proposed models 2) Record the metrics and resource utilization 3) Draft Documentation
Sprint3 04/16– 04/23	1) Reproduce CNN Model for the standard 2-3 models (AlexNet, GoogleNet, RestNet etc.) 2) Record the metrics and resource utilization 3) Final Project Documentation
Sprint4 04/23– 04/30	1) Dev/Test – the new proposed model (provided previous sprints completion) 2) Fine tune layers, filters etc. to improve metrics or resource utilization 3) Final Test comparing all the models 4) Final Project Documentation
Sprint5 04/30– 05/07	1) Develop a front end for User Interaction with the prediction modeling and visualizations a. Interactive x-ray covid-19 prediction b. Display the prediction results c. Display comparison report with other models d. Display metrics and resource utilization In graphs and charts 2) Project Cleanup and Quality Procedures 3) Code Packaging and User Manual 4) Final Project Documentation
05/08– 05/09	1) Project Presentation Slide 2) Project Presentation Video 3) Project Submission

Table 1: Project task breakdown in iteration where each iteration is a 7 day sprint except the last sprint with only 2 days

8. ACKNOWLEDGMENTS

Our thanks to Prof. Jimeng Sun for conducting the course and sharing his knowledges and making us learn the subjects through his video lectures, weekly reflection submissions, programming assignment submissions and the project work. Our thanks to all the course TAs for their amazing support throughout the program so far without which this course would have been more challenging.

Also, our thanks to ACM SIGCHI for allowing us to modify templates [2] they had developed.

9. REFERENCES

[1] Sun, Jimeng, Prof University of Illinois – Urbana Champaign, CS598 Projects: Deep Learning for Healthcare,

[2] ACM SIG PROCEEDINGS template. http://www.acm.org/sigs/pubs/proceed/template.html

[3] Geert Litjens, Thijs Kooi, etc A survey on deep learning in medical image analysis, Medical Image Analysis, Vol -22, Dec 2017, Pages 60-88, https://doi.org/10.1016/j.media.2017.07.005

[4] Prottoy Saha, Muhammed Sheikh Sadi, etc, EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers, Informatics in Medicine Unlocked, Vol 22, 2021, 100505, https://doi.org/10.1016/j.imu.2020.100505

[5] Morteza Heidari, Bin Zhenga, etc, Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms, International Journal of Medical Informatics Vol 144, Dec 2020, 104284, https://doi.org/10.1016/j.ijmedinf.2020.104284

[6] Baltruschat, I.M., Nickisch, H., Grass, M. et al. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Sci Rep 9, 6381 (2019), https://doi.org/10.1038/s41598-019-42294-8

[7] Rahib H. Abiyev and Mohammad Khaleel Sallam Ma'aitahcorresponding, Deep Convolutional Neural Networks for Chest Diseases Detection, Published online 2018 Aug 1, PMCID: PMC6093039, doi: 10.1155/2018/4168538

[8] Boran Sekeroglu and Ilker Ozsahin, Detection of COVID-19 from Chest X-Ray Images Using Convolutional Neural Networks, PubMed, Sep 18 2020, PMID: 32948098 https://doi.org/10.1177/2472630320958376

[9] Dataset contains 542 frontal chest X-ray images from 262 people from 26 countries. This also contains clinical attributes about survival, ICU stay, intubation events, blood tests, location, as well as freeform clinical notes for each image/case. This dataset can be accessed at https://github.com/ieee8023/

[10] Dataset contains COVID-19, no-findings and pneumonia images. This dataset is combination of the data from multiple sources. This contains 127 COVID-19 images collected from COVID-Chest-xray-dataset and Normal and pneumonia images are collected from ChestX-ray8 database. This can be accessed at https://github.com/muhammedtalo/ COVID-19COVID-19 data.

[11] NIH Chest xray Dataset, https://nihcc.app.box.com/v/ChestXray-NIHCC

About the authors:

Ujjal Saha is a graduate student from University of Illinois – Urbana Champaign pursuing MCS-DS and CS598 is his 7th course in the program. By profession Mr. Saha is a Sr. Software Engineer working for a Fortune 500 software company with 14+ years of experience in software engineering.

Ashish Pradhan is a graduate student from University of Illinois – Urbana Champaign pursuing MCS-DS and CS598 is his 6th course in the program. By profession Mr. Pradhan is Data Analyst working for a Non Profit Foundation with 5+ years of experience in software engineering.

Dilip Ravindran is a graduate student from University of Illinois – Urbana Champaign pursuing MCS-DS and CS598 is his 3rd course in the program. By Profession Mr. Ravindran is Sr. Software Engineer working in a Multinational IT Solutions company with 10+ years of experience in software engineering.

Faizan Khan is a graduate student from University of Illinois – Urbana Champaign pursuing MCS-DS and CS598 is his 6th course in the program. By Profession Mr. Khan is Sr. Software Engineer working for a Fortune 500 company with 10+ years of experience in software engineering