|  |
| --- |
| COMP702 |
| PROJECT PROPOSAL |
| Federated Learning for healthcare: analysing the recovery of covid-19 patients with long-term effect |

|  |
| --- |
| Shoukat, Usman  9-14-2021 |

Contents

[Problem Description 1](#_Toc83132438)

[Aims and Objectives 1](#_Toc83132439)

[Methodology 1](#_Toc83132440)

[Datasets 1](#_Toc83132441)

[Plan 1](#_Toc83132442)

[References 1](#_Toc83132443)

# Problem Description

Human life, health and productivity has largely been affected by current covid-19 pandemic. During the start of the covid, the focus was on acutely ill people to save their lives and minimise the risks of the elderly and vulnerable. But now, a troubling phenomenon is coming to light with a staggering number of covid infected people saying that they are not recovering completely from covid symptoms over a longer period of time. This defines the long covid, which is not recovering from the symptoms which are being suggested as covid-19 symptoms for several weeks or months. [1] According to one study five most common symptoms found in long covid positive patients are fatigue, headache, dyspnoea, hoarse voice and myalgia. And the long covid was more common with increasing age and body mass index, and female gender. [2]

As this long covid is affecting the lives of so many people, it is important to predict the long covid through the initial symptoms of covid so that it can be addressed earlier and in a better way by health practitioners.

Many different AI algorithms and models are recently being used to diagnose and predict the possibility of many high-risk diseases like breast cancer [3] and SARS-CoV-2 [4]. Thus, they can help us in the identification of COVID-19 as well [5]. Different AI algorithms and models can be trained to predict the possibility of long-term covid given the patient’s health conditions and initial covid symptoms. But to train these traditional models, data collection is very critical. Medical data is usually not shared publicly to protect the patient’s privacy concerns, which leads to insufficient datasets to train these models. [6] The non-availability of the datasets is limiting the number of research projects and solution in the healthcare sector.

Here comes the federated learning, a great tool to train a shared model by aggregating the locally computed updates, to rescue us. Google first introduced it in 2016. [7]

Diagram

Description automatically generatedEvery client (server, organisation, mobile phones) has a local dataset and local machine learning model in federated learning. The global central server is in a federated environment that has a centralised learning model which aggregates the model parameters of all the local clients. Each client trains the local model by using its local data and then share the parameters of that locally trained model with the global model. The global model goes through many iterations to take the update from distributed clients. [8] As data never leaves the local centres in this type of learning thus, it can be an excellent solution to the data privacy issues.

Figure 1 Federated Learning [8]

# Aims and Objectives

The aim of the project is to study the application of federated AI for the diagnosis of long-term covid. We will check feasibility the federated AI by comparing it with different classical machine learning algorithms. The main objectives of this project will be:

* To research about the long covid, its symptoms and health condition that might lead to long covid
* To research and compare classical machine learning algorithms with federated learning methods. The comparison will be done on same algorithms but between classic and federated learning architectures
* To learn about the API’s and tools like TensorFlow and Scikit learn needed to implement ML algorithms and federated learning
* To identify suitable datasets that contain the information about covid patients affected by long covid
* To train different machine learning models on the obtained datasets and compare their results with the federated learning architecture.

# Methodology

The diagnosis process will work in the following way. First the doctor will take the chest X-ray of the patient and feed it into the ML algorithm to check whether the patient have covid or not. If the patient has the covid, then the doctor will ask the patient to fill in a form with demographic details and details of the symptom. Then the doctor will feed this information along with the health record of the patient to the model which will then give the prediction on the possibility of long-term covid by using the association rules.

In this research, first datasets of chest X-rays images of the covid patient, pneumonia patient and normal person [9] will be obtained, and image pre-processing will be done. I am using this dataset as it is widely being used for the classification of covid patients. Moreover, this dataset can also help us in the determination of long covid as pneumonia is one of the health conditions that impacts the severity of covid-19 cases. I will be using the convolutional neural network algorithm for the classification of above-mentioned categories as it is the most used algorithm to diagnose covid based on Xray image. [10] I shall be training three different variations of convolutional neural network both in classic as well as federated learning architecture on above mentioned data to classify among covid, pneumonia and normal patient. And then I shall compare the results of both the architectures by comparing their accuracies, precision, recall and f-score.

The Implementation flow of this model is given below

Figure 2.1 Implementation Flow Diagram

Pre-Processing

 Comparison of Both the Architectures

Classification as Covid, Pneumonia and Normal Patient

Classic CNN Architecture

Federated Learning CNN Architecture

Data containing Chest X-ray of Pneumonia, Covid and Normal Patient

Classification as Covid, Pneumonia and Normal Patient

Then a numerical dataset containing the information about the age, gender, health conditions and symptoms of the covid that person is experiencing like fatigue, breathlessness, cough, muscle, and body aches., chest heaviness and pressure, diarrhoea, pins and needles will be collected. [11] [12] [13] I am using these datasets as these datasets gives the demographic information and the information about the symptoms of covid patients, which are the important factors for the development of long-term covid. Once the data is collected, then pre-processing techniques will be applied to the data for data cleaning, including handling the missing data and normalising the data. Logistic regression [14] [15], Support Vector Machines [15], KNN [16] and Multi-Layer Perceptions [17] are some of the supervised learning algorithms that are recently being used for the diagnosis of COVID-19. I shall be using and training these models both in classic as well as federated learning architecture to classify between long-term and short-term covid. And then I shall compare the performance of these models in both architectures (classic and federated) by using precision, accuracy, recall and f-score. The Implementation flow of this model is given below

Figure 3. Implementation Flow Diagram

Pre-Processing

 Comparison of Both the Architectures

Classification as Long-Covid and Short-Covid

Classic ML Architecture

Federated Learning Architecture

Data containing Demographic Information and symptoms of Covid

Classification as Long-Covid and Short-Covid

At the end the results of both the above models will be combined with association rules to obtain the prediction on long covid. Which is shown below by a diagram

Figure 4. Implementation flow Diagram

Association Rules

Long-Covid and Short-Covid Model’s output

X-ray Images Model’s output

Health Record

Classification as long and Short Covid

The models will be coded in python 3. Tensorflow, Pytorch and Scikit learn will be used for the training of the models and, Seaborn and Matplotlib will be used to visualise the data and the results.

# Datasets

Until now, I have collected the following data sets:

1. Chest X-ray (Covid-19 and pneumonia)
   * This data set contains the chest X-ray images of Covid-19, Pneumonia and normal patients [9]
   * This dataset will help in distinguishing between covid, pneumonia and normal patient
2. Israeli Government database
   * This repository has two datasets that are of our interest. One contains the information about the covid symptoms experienced by the different patients, and the other dataset has information about the recovery period of different patients [13]
   * This dataset can be useful in distinguishing between long-term and short-term covid depending upon the demographic information and the recovery period of the patient
3. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK
   * This dataset contains the information about the self-reported long covid symptoms and duration in the UK [11]
   * This data can be useful in identifying the most important symptoms or health conditions for long-term covid
4. Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada
   * This dataset gives information about the onset of symptoms, asymptotic status, hospitalised status, death status, virus transmission status, age and gender of the patient [12]
   * This data can be useful in getting the demographic information of the covid patients

# Plan

The plan of the project is given by following Gant chart

Timeline

Description automatically generated

Figure 5. Gant Chart

# References

|  |  |
| --- | --- |
| [1] | N. Nabavi, “Long covid: How to define it and how to manage it.,” 2020. |
| [2] | C. Sudre, B. Murray, T. Varsavsky, M. Graham, R. Penfold, R. Bowyer, J. Pujol, K. Klaser, M. Antonelli, L. Canas and E. Molteni, “Attributes and predictors of long COVID,” pp. 626-631, 2021. |
| [3] | K. Malherbe, “Tumor microenvironment and the role of artificial intelligence in breast cancer detection and prognosis,” *The American journal of pathology,* 2021. |
| [4] | C. D'Ambrosia, C. Henrik and A.-S. Eliah, “Computing SARS-CoV-2 infection risk from symptoms, imaging, and test data: Diagnostic model development.,” *Journal of Medical Internet Research,* vol. 12, 2020. |
| [5] | B. Liu, B. Yan , Y. Zhou , Y. Yang and Y. Zhang, “Experiments of federated learning for covid-19 chest x-ray images,” *arXiv preprint arXiv:2007.05592,* 2020. |
| [6] | W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S. Lo and F. Wang, “Dynamic fusion-based federated learning for COVID-19 detection,” *IEEE Internet of Things Journal,* 2021. |
| [7] | McMaham, Brendan, M. Eider , R. Daniel , H. Seth and y. A. Blaise Aguera , “Communication-efficient learning of deep networks from decentralized data,” *Artificial intelligence and statistics,* no. PMLR, pp. 1273-1282, 2017. |
| [8] | S. Abdul , Mustafa, T. Sanaa and R. Mohamed, “COVID-19 detection using federated machine learning,” no. Plos one, 16(6), e0252573, 2021. |
| [9] | P. Patel, “Chest X-ray (Covid-19 & Pneumonia),” 2020. [Online]. Available: https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia. |
| [10] | N. El-Rashidy, A. Samir , A. Tamer , A. Eslam , A. Farman , H. Jong-Wan and E.-S. Shaker , “Comprehensive survey of using machine learning in the COVID-19 pandemic,” *Diagnostics,* 2021. |
| [11] | D. Ayoubkhani, P. Pawelek and M. Boswort, “Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK,” Office of National Statistics Uk, 2 September 2021. [Online]. |
| [12] | “Preliminary dataset on confirmed cases of COVID-19,” Public Health Agency of Canada, 13 August 2021. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/13-26-0002/132600022020001-eng.htm. |
| [13] | “COVID-19 reservoir,” Israeli Ministry of Health, 26 August 2021. [Online]. Available: https://data.gov.il/dataset/covid-19. |
| [14] | P. Podder, B. Subrato , . H. M. M. Rubaiyat and K. Utku , “Application of Machine Learning for the Diagnosis of COVID-19.,” *Data Science for COVID-19,* pp. 175-194, 2021. |
| [15] | A. M. U. D. Khanday, R. Syed Tanzeel, K. Qamar Rayees , R. Nusrat and M. M. U. Din, “Machine learning based approaches for detecting COVID-19 using clinical text data,” *International Journal of Information Technology,* 2020. |
| [16] | A. Hamed, S. Ahmed and N. Hamed , “Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a K NN Variant Algorithm,” *Arabian Journal for Science and Engineering,* 2021. |
| [17] | G. Guo , Z. Liu , S. Zhao , L. Guo and T. Liu, “Eliminating indefiniteness of clinical spectrum for better screening COVID-19,” *IEEE Journal of Biomedical and Health Informatics,* 2021. |