COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Dominique Duncan (University of Southern California), September 16, 2020

Title: COVID-ARC (COVID-19 Data Archive)

Dominique Duncan CIC Database Profile

NSF Award #: 2027456

YouTube Recording with Slides

September 2020 CIC Webinar Information

Transcript Editor: Macy Moujabber

Katie Naum:

We will be hearing from Dominique Duncan at the University of Southern California. Dominique we're ready whenever you are.

Dominique Duncan:

Slide 1:

Thank you. So, I'm Dominique Duncan from USC, from the Nerve Imaging and Informatics Institute and I'll be talking about our COVID-19 data archive. For short it's called COVID ARC.

Slide 2:

So, we have a lot of experience at our institute with large scale multimodal data repositories, so we decided to apply our experience and the tools that we've developed in these other projects and extend them to this COVID-19 data archive. So, you can see here on the bottom left, that's a screenshot of our home page. If you want to go to the website, I encourage everyone to go to covid-arc.loni.usc.edu to find out more about the project, the various data sets that we have, our analytic tools and to find out more about either uploading data or downloading existing data that we have. So, we have some publicly available data sets that we've been curating and organizing. We want to encourage researchers to be able to not only perform analysis on individual data sets but across data sets from various sites and so we want to make that process easier for them. And so we're also providing various tools like quality control tools that people can use to evaluate images and evaluate the quality using various metrics. We have various visualization tools for imaging data and other types of data and then a wide variety of analytic tools that people can use. And then for data sets that are not publicly available, data providers can decide if they want to store their data on our server or if they want to keep their data stored locally at their site and just provide us with the metadata so that we can let users know what data are available and then we just facilitate the process of requesting access to that data. So, if data providers give us their data, it's anonymized and the data providers maintain full control of access and they get to decide who gets access to it. If they want to wait to publish some findings, and then make their data sets publicly available, they're able to do that as well. And then we use ASPERA which is IBM's HIPAA compliant encrypted high-speed file transfer system for either uploading data to our server or for people who want access to the data. They can download it that way as well.

Slide 3:

I know you can't see any of this text but this just gives you an overview of what we have on our server. So, if you request access, this is the tree structure of the COVID ARC project and so it's separated into the data and then our analysis that we've been working on because we're also very involved in the analytics for this project. So, the data will be separated by the different sites and then there's more information on each of those.

Slide 4:

So, here is a table of the data sets that we currently have on the server so you can see the location, where the data were acquired for each of those data sets, and then the data format of those images. Right now, we're focusing mainly on chest CT but we're also interested in other types of data and soon we'll have some brain data, so brain MRI of COVID-19 patients as well as EEG. So here you can just see how many COVID images and non-COVID images and masks are there from each of those sites.

Slide 5:

Here you can see some class activation maps. So, there are many features, as we know, CT features of COVID-19 patients like ground glass opacity, consolidation, crazy paving pattern, reticular pattern, etc. So, this just highlights the regions that were most important for our classification that we're doing.

Slide 6:

One of the issues that we're discovering is that image quality tends to lead to misclassification and so we're assessing the quality of the images but then also considering various ways of improving the quality of the images so looking at different filtering methods and that's something that's currently in progress.

Slide 7:

We're also doing image thresholding on the lung masks so we're using image thresholding to find the best possible processed image type in the lung masks to improve the prediction rate of our neural networks.

Slide 8:

And this is a table on one of the data sets. That first one from Brazil that had about 1,200 COVID patients and 1,200 non-COVID images and we did a comparison looking at various methods to compare convolutional neural networks and their accuracy. So, we found that ResNet-18 performed the best for that.

Slide 9:

And just to summarize, because I think I'm at the end of the time these are the people that are working on the project including the REU students. And I'd like to thank the NSF for funding this project as well as Katie and Florence for organizing this and inviting me to talk. And please visit the website or email me if you have any questions and if you would like access to any of the data or if you have some data that you would like to contribute to the website. Thanks.