Detection of COVID-19 & its severity using Chest X-Rays and EHR

Dr. Mamatha H.R
PES University
Bangalore, India
mamathahr@pes.edu

Ria Kalia
PES University
Bangalore, India
riakalia@gmail.com

Amrutha S
PES University
Bangalore,India
amruthashetty299@gmail.com

Krithika P Suwarna
PES University
Bangalore,India
krithikap019@gmail.com

Abstract

COVID-19 first appeared in December 2019 in the city of Wuhan in China and has since then spread throughout the world. One year later and the virus continues to mutate and infect individuals at an increasingly alarming rate. Providing the right treatment to patients during the early stages of the infection is extremely vital for their survival. There is also a need for quicker testing since the fastest one takes 24 – 48 hours to provide results.

Such a situation demands for an automated COVID detection toolkit. Recent research using computer vision techniques suggest that Chest X-Rays contain important features about the effects of the virus in the chest region. Application of advanced deep learning techniques along with clinical imaging can help overcome the shortcomings of the current diagnostic tests. We propose an automated tool to solve this problem, for detecting the presence of the virus and its severity using deep learning models. The proposed model takes into consideration Chest X-Rays as well as health records of the patient to predict whether the patient has COVID or not, and to predict the severity for positive cases.

I. INTRODUCTION

Coronavirus or COVID-19 was declared a pandemic by the World Health Organisation more than a year ago, on 11th March 2020. Countries all over the world have been in a state of lockdown since months together, with people restricting themselves to their homes and staying out of public places as much as possible.

The first case of COVID-19 can be traced back to the city of Wuhan in China. A number of cases of pneumonia of unknown causes originating in Wuhan was reported to the WHO. Samples of the virus were then analysed, and the virus was referred to as SARS-CoV-2. As of 17 March 2020, over 121 million cases have been identified globally.

The incubation period of this virus is between 3 to 10 days, and it's transmitted through saliva or discharge from the nose/mouth. The symptoms of an infected person range include cold, breathing problems such as shortness of breath, and chest pain. Coronaviruses typically present with respiratory symptoms that can be between mild and moderate. However, in severe cases it can also lead to death. Old people and people with comorbidities are typically more prone to the severe effects of the disease and might require hospitalization. Patients with a severe case of the disease also develop COVID induced pneumonia.

Currently, there are two primary methods for the detection of the virus: the antibody test and the RT-PCR. The reverse transcription polymerase chain reaction test works by detecting nucleic acid in the upper and lower respiratory specimens. It typically takes 3 – 4 hours for the test results to come out, but there usually is a delay of 24 – 48 hours to receive the results, due to the large number of cases (and hence the amount of samples to be tested). In certain places, there is also the problem of shortage in the number of testing kits. Moreover, the test is expensive, and not all people can afford to spend huge amounts of money to get tested.

The other method for testing is the antibody test which looks for antibodies in the blood. However, this test is not very reliable and cannot detect COVID-19 cases in its initial phases. Hence, there is a need for a quicker way to detect COVID-19. Having such a tool can help prevent the spread of the virus and also allow patients to recover quickly.

There is a lot of ongoing research to look for such methods. A key domain in this area is clinical imaging. Clinical imaging basically refers to techniques that analyses various parts of the human body from images and tries to diagnose health conditions based on that. The images can include X-Rays, CT scans etc. Deep learning methods are employed to extract features from these images and classify them. Similarly, electronic health records can also be used to analyse the symptoms of patients and predict diseases. Again, this can be done using deep learning models.

However, a major limitation in models that use machine learning algorithms to detect diseases is the lack of a clinical context—the diagnosis is given by just using images. They do not consider other health parameters or symptoms (like fever, cough, loss of smell, etc) for the detection of the virus and are solely dependent on the chest X-rays which might result in errors. Therefore, it's critical to consider the complete health records of the patient along with the chest X-rays to increase the accuracy and the usage of such a system.

Considering this, we propose a system for the detection of COVID-19 and its severity using Chest X-Rays as well as Electronic Health Records. We aim to construct a system that takes Chest X-Rays (Posterior Anterior view) and the health records of a patient as input. A deep learning model will then process these inputs to predict whether the patient has COVID or not and in the case that the patient tests positive, the model will also try to predict the severity of the case by looking for adverse effects of the virus on the lungs. The proposed model takes into consideration the problems with other diagnostic tools such as time and cost and tries to overcome them.

This paper is organized as follows: Section II describes the design and the modules in the system. Section III provides the implementation of all the models. Section IV includes details about demonstration and results. Conclusion and future work is detailed in Section V.

II. DESIGN AND METHODOLOGY

1. Detailed Design

Pre-existing models evaluate Chest X-Rays for COVID but they do not consider the clinical context of patients. Therefore, we aim to make use of electronic health records to evaluate the same and thereby provide a clinical context while diagnosing COVID.

The proposed product can be divided into primarily three modules:

1. Detection of COVID-19
2. Prediction of recovery/death using EHR
3. Severity Prediction

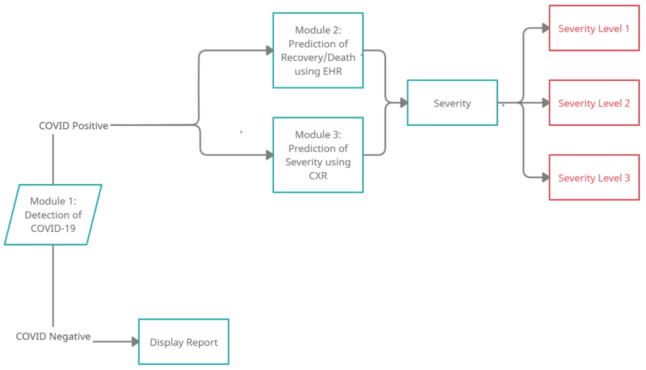
Module 2: Prediction of Recovery/Death using EHR
Severity Level 1
Severity
Severity Level 2
COVID Positive
Module 3: Prediction of Severity using CXR
Severity Level 3
Module 1: Detection of COVID-19
COVID Negative
Display Report

Figure 1 : Design of the proposed system

Figure 1 summarizes the sequence of execution of the modules - the X-Rays are first processed to determine if the patient is COVID positive or not. If the patient is negative, no further processing takes place and the result is displayed to the user. If the patient is COVID positive, two things happen in parallel: Prediction of Severity, and the Prediction of recovery/death. The results of the modules are aggregated into the report, which is made available to the user.

1.1 Module 1: Detection of COVID-19

This module makes use of CXR. The input image provided by the patient is first pre-processed by performing image resizing (to 224 x 224 pixels) and RGB reordering. This is a crucial step since the images being sent to the model must be uniform. Next, the image is sent to the deep learning model which provides a prediction of 1 if the CXR indicates that the patient is COVID positive, and 0 for negative.

The deep learning model has been implemented by using transfer learning with VGG16 as the base model. We made a few tweaks to the topmost layers of the VGG16 model, added a few layers of our own, to obtain a 19-layer model that makes predictions with an accuracy of 98%. The model was trained using 1200 images (600 healthy lungs CXR and 600 COVID infected lungs CXR) and around 350 images were used for testing (80/20 split).

1.2 Module 2: Prediction of Recovery/Death using EHR

This module makes use of the EHR provided by the patient, which makes use of the following data:

- Gender
- Age
- Symptoms
- From Wuhan/not
- Has visited Wuhan/not

The symptoms considered are Fever, Cough, and Difficulty in breathing, Pain, Fatigue, Diarrhea, Cold, Pneumonia, Vomiting and Malaise. This data is passed to a supervised machine learning model which makes a prediction of 0 if the patient is likely to recover, and 1 if the patient is unlikely to recover without medical aid.

1.2 Module 3: Prediction of Severity

This module makes use of the CXR to predict the severity of the patient. This is done by looking for lung opacity in the CXR. The severity is classified into 1 (lung opacity present) or 0 (lung opacity absent). This is done by processing the CXR using a deep learning model to predict the presence of lung opacity.

2. Final Result.

Patients who want to test for COVID-19 first upload their Chest X-Ray image along with their symptoms, age, gender, etc. Once the form is submitted the image goes through Module 1 which predicts if the patient tested positive or negative.

If the patient tests negative, then there is no further step to be performed and the result is displayed on the website.

If the patient tests positive, then the data of the patient is passed onto Modules 2 and 3. Module 2 provides a result of

1 (severe) or 0 (not severe), while Module 3 predicts 1 ((opacity present) or 1 (opacity absent). The result of the modules is merged to give the final result of severity by classifying the severity into:

- Mild (Level 1)
- Moderate (Level 2)
- Severe (Level 3)

. Table 1 depicts the criteria for classification into different severity levels.

EHR Result	Severity Result	Final Result
0	0	1
0	1	2
1	0	2
1	1	3

Table 1 : Criteria for classification

Once the final severity has been calculated, the result along with the severity level is displayed on the website which the user can download in PDF format. The website also has an option to look at a patient's previous test results.

III. IMPLEMENTATION

MODULE 1 : Detection of COVID-19

1. Dataset

The dataset of Chest X-Rays was available on Kaggle and is maintained by the National Institute of Health Chest X-Rays. We took 800 healthy Chest X-ray images and an equal number of COVID affected Chest X-rays.

2. Preprocessing

The images in the dataset had a variety of characteristics, and we had to pre-process them to ensure accurate results. We first resized the images to 224 x 224 pixels and applied RGB reordering on it to ensure uniformity.

3. Deep Learning Model

After applying all the pre-processing techniques, the final output images are fed to the training model. We applied transfer learning, where the VGG-16 model acts as the base, to train our dataset. On top of the base model, we applied additional layers to obtain a training accuracy of 98%.

MODULE 2 : Prediction of Recovery/Death using HER

1. Dataset

For this module, we made use of a dataset from Kaggle containing health records of patients diagnosed with COVID. The dataset had 1085 rows and 21 columns. The columns include id, country, reporting date, summary, gender, age, symptoms, hospital date, etc. The final column 'death' indicates whether that individual recovered from COVID-19 or not.

2. Preprocessing

Preprocessing involved dropping unnecessary columns (such as country name, id, summary, etc.), converting dates to date-time format, consolidating multiple columns to one, filling missing values etc.

It was observed that the 'symptoms' column contained unstructured data, where the symptoms were written in text format, separated by commas. Further analysis of the dataset showed that there were ten recurring symptoms in most patients. Based on this observation, we created separate columns for each of the following symptoms: Fever, Cough (including sore throat), Difficulty in breathing, Pain (including headache), Fatigue, Diarrhea, Cold (including a runny nose), Pneumonia, Vomiting, and Malaise. These columns contain a value of either 0 (for symptom absent), or 1 (for symptom present).

As the final step we performed one hot encoding on all the columns containing values that could be classified, such as Gender, and all the columns representing symptoms.

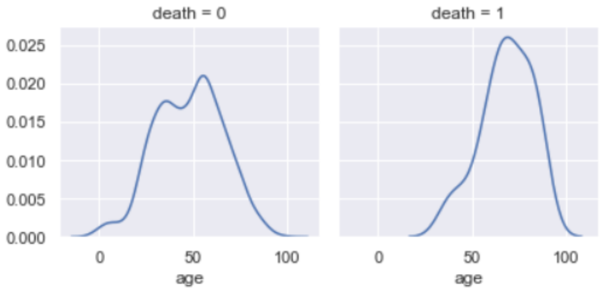

Figure 2 : Graph of Age vs Death

We applied various data visualization techniques on the final columns that were obtained. Figure 2 indicates one such analysis. From the bar graph, it is clear that most patients showed symptoms of fever when they got COVID.

3. Deep Learning Model

The next step was training the models on the pre-processed data. We made use of SKLearn to do this, where we first divided our dataset into training and testing data, with an 80:20 split.

The following models were then applied on the data:

- Multiple Linear Regression, which yielded an accuracy of 10.1 %
- Logistic Regression, which yielded an accuracy of 95.3 %
- Decision Tree, which yielded an accuracy of 93 %. Figure 6.12 depicts the decision tree produced.

- Pruned decision tree, with a depth of 3 gave an accuracy of 95.85 % (Figure 6.13)
- Random Forest, which also gave an accuracy of 95.85 %
- Boosted random forest, which gave an accuracy of 96%.
- Neural networks, which yielded an accuracy of 95.4 %
- A 3-layer CNN model which gave an accuracy of 94 %

4. *Performance Comparison*

We then plotted graphs of the performance of the various models based on their Accuracy, Precision, Recall and F1-Score. Figures 3, 4, 5, 6 depict the same. It can be observed from these that Boosted Random Forest was the best performer for all the aforementioned metrics.

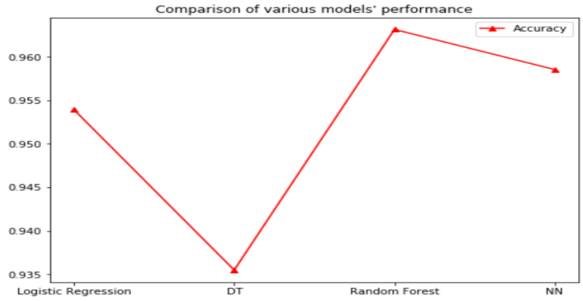

Figure 3: Comparison of model's performance based on Accuracy

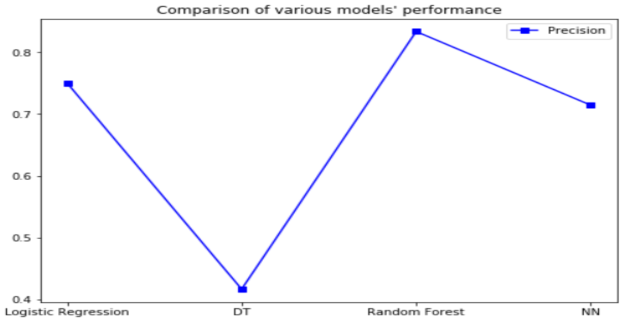

Figure 4: Comparison of model's performance based on Precision

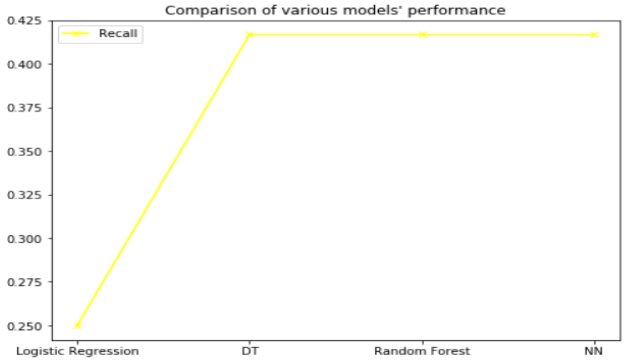

Figure 5: Comparison of model's performance based on Recall

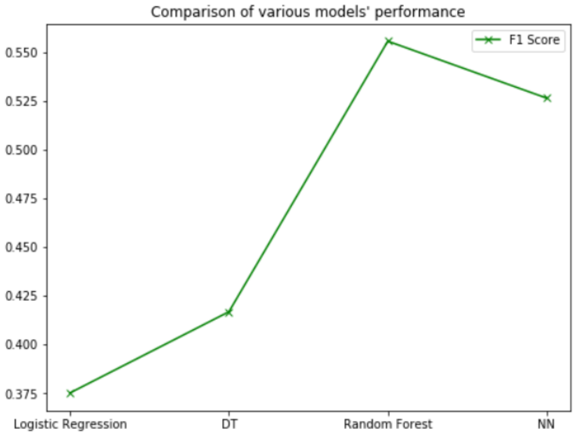

Figure 6: Comparison of model's performance based on F1 Score

MODULE 3 : PREDICTION OF SEVERITY

1. *Dataset*

The dataset used for this module was obtained from Kaggle and contains CXR along with a classification of 'Lung Opacity' or 'No Lung Opacity'.

2. *Preprocessing*

Analysis of the dataset suggested that there was an imbalance in the data with respect to the number of images in the classifications. To rectify this, we made use of all the images in the smaller class, and an equal number of randomly chosen images from the dominant class. We also performed RGB reordering and image resizing where the images were resized to 224 x 224 pixels.

3. *Deep Learning Model*

Two models were constructed using this data – a CNN model and a model using MobileNet (Transfer Learning).

The CNN model contained 12 layers and had an accuracy of 80%.

The second model makes use of MobileNet, a computer vision model for Tensorflow. The MobileNet model has been trained on more than a million images from the ImageNet database. We made use of transfer learning where the top layers of the model were replaced with layers created by us. The total number of layers of the model is 35. The model gave an accuracy of 85%.

IV RESULTS AND DEMONSTRATION

This study makes use of three modules to predict if a person is affected by COVID-19 by considering both Chest X-Rays and Electronic Health Records. The presence of the virus is detected using just the CXR, while severity is measured using both EHR and CXR. The health records contain information regarding the symptoms shown by the patient.

Module 1 (Detection of COVID-19) which uses transfer learning with VGG16 as the base model predicts with an accuracy of 98%. Module 2 (Prediction of Recovery/Death using EHR) which uses the Random Forest algorithm predicts with an accuracy of 96%. Module 3 (Module 3: Prediction of Severity) which uses the Random Forest algorithm predicts with an accuracy of 96%.

V CONCLUSION AND FUTURE WORK

It's been more than one year since COVID-19 struck the world. Multiple vaccines have been developed to provide protection from the virus, but it could take years to inoculate everyone in the world. Meanwhile the virus continues to mutate into variants that are becoming increasingly lethal and can spread quickly from one individual to another.

There is a need for a tool that can quickly detect the virus in a non-invasive and quick way. This study aims to provide such a tool by making use of CXR and EHR to predict whether a patient has COVID or not, and in the case that they do, the tool also aims to predict the severity of progression of the virus.

There is a shortage of hospital rooms / ICUs all over, as the world tries to combat the virus, and analysis of severity could help indicate whether the patient might require hospitalization or not. Moreover, this tool could also help reduce the load of healthcare workers / frontline workers since patients could get tested for COVID from the comfort of their homes.

Future work on this study involves consulting radiologists, who could cross check the results generated by the model and ensure that it can be used in real time.

The model currently predicts the severity of COVID-19, but if the patient is affected by other lung diseases, then the model fails to provide accurate predictions. Further analysis can be performed on this facet to ensure accurate predictions.

Lastly, the data generated by the model can be analysed on a day to day basis to keep a tally of the total active cases, new cases per day, recoveries per day etc. Present methods to keep track of these metrics are quite inaccurate, and a centralized system could act as a fool-proof way to measure these values. The data produced can also be used to identify containment zones or areas with a large number of cases. Data regarding the severity of patients can be leveraged to identify the number of ICUs required on a daily basis.

There are various use cases for the data generated by the tool, and using it could be extremely advantageous to the healthcare industry by taking off some of the pressure from doctors, nurses and frontline workers who have been working round the clock since more than a year now.

ACKNOWLEDGMENT

We would like to express our gratitude to Prof. Mamatha H.R., Department of Computer Science and Engineering, PES University, for her guidance and support throughout the project. We would also like to thank PES University for providing us with various opportunities and pushing us to do better. Finally, we'd like to thank our family and friends for their constant encouragement.

REFERENCES

[2] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J., 2019. A guide to deep learning in healthcare. Nature medicine, 25(1), pp.24-29.

[3] Panwar, H., Gupta, P.K., Siddiqui, M.K., Morales-Menendez, R. and Singh, V., 2020. Application of Deep Learning for Fast Detection of COVID-19 in X-Rays using nCOVnet. Chaos, Solitons & Fractals, p.109944

[4] Elaziz, M.A., Hosny, K.M., Salah, A., Darwish, M.M., Lu, S. and Sahlol, A.T., 2020. New machine learning method for image-based diagnosis of COVID-19. Plos one, 15(6), p.e0235187.

[5] Islam, M.Z., Islam, M.M. and Asraf, A., 2020. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Informatics in Medicine Unlocked, 20, p.100412.

[6] Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O. and Acharya, U.R., 2020. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine, p.103792.

[7] Fiszman, M., Chapman, W.W., Aronsky, D., Evans, R.S. and Haug, P.J., 2000. Automatic detection of acute bacterial pneumonia from chest X-ray reports. Journal of the American Medical Informatics Association, 7(6), pp.593-604.

[8] Cohen, J.P., Dao, L., Morrison, P., Roth, K., Bengio, Y., Shen, B., Abbasi, A., Hoshmand-Kochi, M., Ghassemi, M., Li, H. and Duong, T.Q., 2020. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. arXiv preprint arXiv:2005.11856.

[9] Mahmud, T., Rahman, M.A. and Fattah, S.A., 2020. CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. Computers in biology and medicine, 122, p.103869.

[10] Alakus, T.B. and Turkoglu, I., 2020. Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals, 140, p.110120..

[11] Bharati, S., Podder, P. and Mondal, M.R.H., 2020. Hybrid deep learning for detecting lung diseases from X-ray images. Informatics in Medicine Unlocked, 20, p.100391.

[12] Thanh, D. and Surya, P., 2019. A review on CT and X-ray images denoising methods. Informatica, 43(2).

[13] Ahmed, S., Yap, M.H., Tan, M. and Hasan, M.K., 2020. Reconet: Multi-level preprocessing of chest x-rays for covid-19 detection using convolutional neural networks. medRxiv.

[14] Khalifa, N.E.M., Taha, M.H.N., Hassanien, A.E. and Elghamrawy, S., 2020. Detection of coronavirus (COVID-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest X-ray dataset. arXiv preprint arXiv:2004.01184.

[15] Haghanifar, A., Majdabadi, M.M. and Ko, S., 2020. Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning. arXiv preprint arXiv:2006.13807.

[16] Punn, N.S. and Agarwal, S., 2020. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. arXiv preprint arXiv:2004.11676.

[17] Iwendi, C., Bashir, A.K., Peshkar, A., Sujatha, R., Chatterjee, J.M., Pasupuleti, S., Mishra, R., Pillai, S. and Jo, O., 2020. COVID-19 Patient health prediction using boosted random forest algorithm. Frontiers in public health, 8, p.357.

[18] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. and Cheng, Z., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), pp.497-506.

[19] Wagner, T., Shweta, F.N.U., Murugadoss, K., Awasthi, S., Venkatakrishnan, A.J., Bade, S., Puranik, A., Kang, M., Pickering, B.W., O'Horo, J.C. and Bauer, P.R., 2020. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. Elife, 9, p.e58227.

[20] Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P., 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), pp.1589-1604.

[21] Solares, J.R.A., Raimondi, F.E.D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Gomes, A.C.P., Payberah, A.H., Zottoli, M., Nazarzadeh, M. and Conrad, N., 2020. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. Journal of Biomedical Informatics, 101, p.103337.