Analysis and Detection of SARS-CoV-2

A Project Report submitted in partial fulfillment of the
requirements for the degree of
Bachelor of Technology

By

Shreyas Yelegaonkar	**EL-B-75**
Rachita Rao	**EL-B-41**
Revati Wankhede	**EL-B-71**
Dhruv Nagar	**EL-A-31**

Under the guidance of
Prof. Dr. Medha Wyawahare

DEPARTMENT OF ELECTRONICS ENGINEERING
VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE

Academic Year 2020-2021

CERTIFICATE

Bansilal Ramnath Agrawal Charitable Trust's
VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE-37
(An Autonomous Institute, Affiliated to Savitribai Phule Pune University) **Department of Electronics Engineering**

This is to certify that the **project report** entitled **"Analysis and Detection of SARS-CoV-2"** has been submitted in the academic year 2020-21 by

Shreyas Yelegaonkar	**1710157**	**EL-B-75**
Rachita Rao	**1710871**	**EL-B-41**
Revati Wankhede	**1710760**	**EL-B-71**
Dhruv Nagar	**1710158**	**EL-A-31**

Under the supervision of **Prof. Dr. Medha Wyawahare** in the partial fulfillment of the requirements for the degree of Bachelor of Technology in **Electronics Engineering.**

Guide/Supervisor **Head of the Department**

Prof. Dr. Medha Wyawahare Prof. Dr. Shripad Bhatlawande

External Examiner

Place: Pune

LIST OF FIGURE

ABSTRACT

SARS-CoV-2 or Covid-19 has been the worst pandemic in human history for over 100 years. Due to its widespread nature it has affected the entire human race in some way or the other. Most of the supposedly third world countries are finding it difficult to cope with this pandemic.

Hence we have put forth a webpage which will effectively work as a radiologist in the areas where it is difficult to find one. Our webpage will effectively detect Covid-19 virus through a patient CT-scan report. CT-scan is an effective image for detecting whether the patient has the virus along with the location where it is present.

Along with that we will also display the sentiment study for various social media platforms so that people visiting the webpage get a look at how people around the world are doing health wise which will act as a coping mechanism mentally. Our webpage will also display the current tracker of covid-19 across the world with the help of tableau platform.

Our motive is to provide a one stop go for all covid-19 related issues.

Keywords : Covid-19, Radiologist, CT-scan, Sentiment analysis, Tableau

CHAPTER 1

INTRODUCTION

With the development in research of covid-19, scientists have found variants across the globe which makes it more difficult in terms of finding data suitable for the task which we intend to do. The data collection for covid-19 detection is taken from Negin medical center which is located at Sari in Iran. This dataset is combined using three different data collections identified by the medical center. The images are HRCT radiology scans which are in the format of 16-bit DICOM format with 512x512 pixel resolution. If it were to be converted to a lesser resolution there might be data loss which can be risky in cases of small area of infection or patients with CT score less than 2.8. But reducing the size is key for getting faster results so for this reason we have converted the images into TIFF format as grayscale images.

One down side of doing this is that visualization becomes very difficult, in order to get rid of this problem we had to convert them into float by dividing the largest pixel value across the image. The second data collection which is used for sentiment study has been taken from twitter data which has been manually tagged for ease in analysis. The data contains location, time and date when the tweet was tweeted and the original tweet in the form of text. The labels we will be using for negative, positive, neutral, extremely positive and extremely negative. The data collection contains 41157 rows of data. The data will be first checked on the basis of hashtags used by the user.

The tweets comprises the United States, United Kingdom, India, Canada and Australia as its major location centers. Other countries across the world add to about 10 percent of tweets. This has to do with the number of users on twitter and not a bias from the data. The other part of the study comprises data taken from data world website which provides daily incremental data which is used for our visualization part. The visualization will consist of a map consisting of bubble structure for density of covid cases in that country. Making the visualization interactive was important for individual states from the countries to be grouped so that the user will get a better understanding of the covid case density in their area compared to the world.

Along with that we intended to include cases count and death count on the region the user has selected on the map. We have further combined all these results to display them on a single web page which will be a go to for any user for all kinds of information regarding the disease. Our goal is to reach the worst hit areas which cannot afford radiologists to use this as concrete evidence based analysis. Many people who cannot afford a radiologist can use our webpage to get a fair idea of how infected they are. We do not promote using the web page as the sole source of deciding the medication and suggest people to visit a doctor for a correct diagnosis.

CHAPTER 2
LITERATURE SURVEY

[1]
"Willa Yu, Nora Luo, Xingxuan Zhang, Aijie Li"
Twitter Sentiment Analysis Based on News topics during Covid-19

Abstract: This paper deals with the analysis process of Covid-19 data using tools such as TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet. The analysis provides a potential approach to reveal the public's sentiment status and help institutions respond timely to it.

[2]
"Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, Liming Xia"
Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases

Abstract: This study included 1014 patients in Wuhan, China, who underwent both chest CT and RT-PCR tests between January 6 and February 6, 2020. With use of RT-PCR as the reference standard, the performance of chest CT in the diagnosis of COVID-19 was assessed. In addition, for patients with multiple RT-PCR assays, the dynamic conversion of RT-PCR results (negative to positive, positive to negative) was analyzed as compared with serial chest CT scans for those with a time interval between RT-PCR tests of 4 days or more.

[3]
“Sejal Dua”
Sentiment Analysis of COVID-19 Vaccine Tweets

Abstract**:** This paper focuses on twitter data collection using Tweepy. The tweets regarding different Covid-19 vaccines such as Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. For each vaccine, a relevant search term was used to query for recent tweets. TextBlob was the model used to detect polarity and subjectivity. This model is best suited to recognize only syntactic and semantic features from textual data rather than understanding the entire data.

[4]
“Ming-Yen Ng, Elaine Y. P. Lee, Jin Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, Christopher Kim-Ming Hui, Kwok-yung Yuen, Michael D. Kuo”
Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review

Abstract: Pulmonary manifestation of COVID-19 infection is predominantly characterized by ground-glass opacification with occasional consolidation on CT. Radiographic findings in patients presenting in Shenzhen and Hong Kong are in keeping with four previous publications from other sites.This was a retrospective study in Shenzhen and Hong Kong. Patients with COVID-19 infection were included. A systematic review of the published literature on radiologic features of COVID-19 infection was conducted.

CHAPTER 3

METHODOLOGY

The study is divided across various sections so we will delve through various factors working towards the study.

A. SARS-CoV-2 Detection

The Data preprocessing for the Medical center dataset can be categorized into two parts one for getting the data into model worthy format and the second part is data generator which is used to create more random data from the training part of our data. In the case of images, data generators can be useful to create mock images which increase the size of data. In turn it helps the model robust and save a lot of overhead memory. The images will be converted to 512x512x1 in TIFF format. The image data generator for training set will have the following attributes,

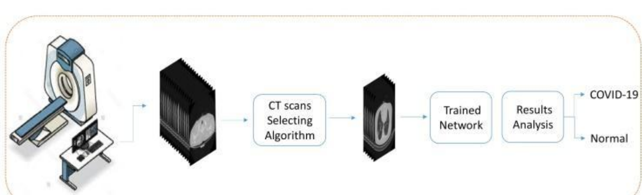
- Data = training data
- Batch_size = 14
- Class_mode = categorical
- Color_mode = grayscale
- Shuffle = true

Shuffle will ensure the augmented data is not segregated at a certain part of the set which will lead to bias in the data. Color mode as discussed earlier will be grayscale to ensure no loss and faster execution.

The attributes for test set of the image data augmentation will be,

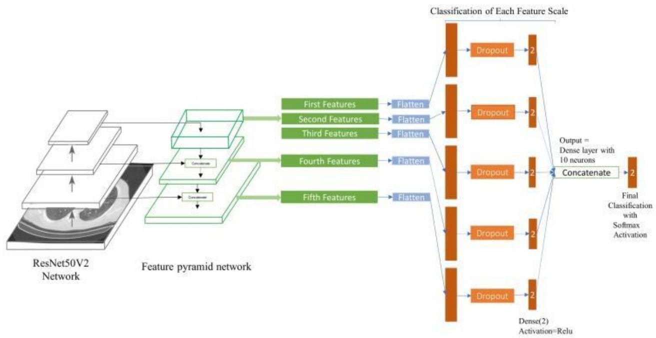
- Data = validation data
- Batch_size = 10
- Class_mode = categorical
- Color_mode = grayscale
- Shuffle = true

The batch size will be lesser compared to training image data generators as we don't want the augmented set in large numbers in the validation set. The test data will also have the same attributes as the validation set but only the source for the data generation will be the test set. Our data collection contains 48260 CT scans from 282 normal patients and 15589 CT scans from covid positive patients. Most of the previous models have been based on CNN. We have proposed a RESNET50V2 with a modified feature of FPN(Feature Pyramid Network).



Flowchart

We will delve into the architecture of FPN which is used with the 50V2 variant of ResNet.



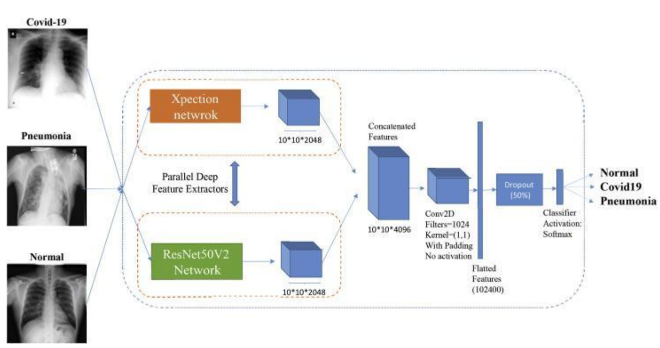
Architecture of ResNet50V2 with FPN

The **Feature Pyramid Network** combines the High resolution and weak semantic features with the low resolution and strong semantic features through a top-down path with lateral connections. The bottom up pathway uses feedforward computation of convolution networks. The output of the last layer of this method is used as an enhancing factor for top down approach.

Finally the convolution is appended into a full feature map consisting of each merged map for reducing the effect of upsampling during the top down approach. Using this network on the ResNet will enhance the predictions for our dataset.

Now we will delve into the structure of ResNet50V2,

Resnet50V2 uses the bottleneck approach. It takes input as multiple of 32, then uses the convolution and max pooling on the input then 3 residual blocks are present after that. In the next version that is V2 the pre activation of weighted layers instead of post activation. So there is a fundamental change in the architecture. There is an identity connection in the V2 as the second non linearity is removed. Also the batch normalization and ReLU activation is present before the multiplication of the weighted vectors.



ResNet50V2 architecture

In our study we will be using 7 pyramid structures for the feature pyramid network which will eventually be used in the Resnet. Then a 5 layer network will be used with a dropout of 0.5 and then the prediction of each layer will be concatenated to form our final prediction model. All the layers will have relu as the activation function, only the concatenated network will have a softmax activation function. The optimizer we will be using is NADAM optimizer. It is called Nesterov-accelerated Adaptive Moment Estimation. In cases of curvature gradient nadam works better than adam optimizer. It uses the moving average of gradient as an exponential decay for acceleration of learning rate. Learning rate is initially taken as 0.0001.

```
Epoch 1/20
267/267 [==============================] - 532s 2s/step - loss: 0.4647 - accuracy: 0.7512 - val_loss: 0.6750 - val_accuracy: 0.8399
/usr/local/lib/python3.7/dist-packages/keras/utils/generic_utils.py:497: CustomMaskWarning: Custom mask layers require a config and m
  category=CustomMaskWarning)
Epoch 2/20
267/267 [==============================] - 432s 2s/step - loss: 0.1543 - accuracy: 0.9549 - val_loss: 0.6750 - val_accuracy: 0.7785
Epoch 3/20
267/267 [==============================] - 431s 2s/step - loss: 0.0907 - accuracy: 0.9710 - val_loss: 0.2411 - val_accuracy: 0.9287
Epoch 4/20
267/267 [==============================] - 431s 2s/step - loss: 0.0932 - accuracy: 0.9687 - val_loss: 0.2683 - val_accuracy: 0.9627
Epoch 5/20
267/267 [==============================] - 431s 2s/step - loss: 0.0648 - accuracy: 0.9796 - val_loss: 0.2477 - val_accuracy: 0.9539
Epoch 6/20
267/267 [==============================] - 431s 2s/step - loss: 0.0542 - accuracy: 0.9853 - val_loss: 0.2178 - val_accuracy: 0.9704
Epoch 7/20
267/267 [==============================] - 431s 2s/step - loss: 0.0717 - accuracy: 0.9780 - val_loss: 0.1619 - val_accuracy: 0.9759
Epoch 8/20
267/267 [==============================] - 431s 2s/step - loss: 0.0556 - accuracy: 0.9825 - val_loss: 0.2622 - val_accuracy: 0.9660
Epoch 9/20
267/267 [==============================] - 431s 2s/step - loss: 0.0571 - accuracy: 0.9834 - val_loss: 0.8667 - val_accuracy: 0.7193
Epoch 10/20
267/267 [==============================] - 431s 2s/step - loss: 0.0609 - accuracy: 0.9829 - val_loss: 0.3112 - val_accuracy: 0.9298
```

Training Epoch(1-10)

```
Epoch 11/20
267/267 [==============================] - 431s 2s/step - loss: 0.0418 - accuracy: 0.9857 - val_loss: 0.5673 - val_accuracy: 0.8849
Epoch 12/20
267/267 [==============================] - 431s 2s/step - loss: 0.0478 - accuracy: 0.9851 - val_loss: 0.2159 - val_accuracy: 0.9529
Epoch 13/20
267/267 [==============================] - 430s 2s/step - loss: 0.0431 - accuracy: 0.9840 - val_loss: 0.2638 - val_accuracy: 0.9857
Epoch 14/20
267/267 [==============================] - 430s 2s/step - loss: 0.0248 - accuracy: 0.9924 - val_loss: 0.4459 - val_accuracy: 0.9529
Epoch 15/20
267/267 [==============================] - 430s 2s/step - loss: 0.0585 - accuracy: 0.9833 - val_loss: 0.2214 - val_accuracy: 0.9430
Epoch 16/20
267/267 [==============================] - 431s 2s/step - loss: 0.0352 - accuracy: 0.9868 - val_loss: 0.3491 - val_accuracy: 0.9770
Epoch 17/20
267/267 [==============================] - 431s 2s/step - loss: 0.0759 - accuracy: 0.9792 - val_loss: 0.2281 - val_accuracy: 0.9803
Epoch 18/20
267/267 [==============================] - 431s 2s/step - loss: 0.0255 - accuracy: 0.9908 - val_loss: 0.2033 - val_accuracy: 0.9704
Epoch 19/20
267/267 [==============================] - 431s 2s/step - loss: 0.0380 - accuracy: 0.9880 - val_loss: 0.2315 - val_accuracy: 0.9693
Epoch 20/20
267/267 [==============================] - 431s 2s/step - loss: 0.0310 - accuracy: 0.9909 - val_loss: 0.8498 - val_accuracy: 0.7127
<keras.callbacks.History at 0x7f72b48e7450>
```

Training Epoch(10-20)

B. Sentiment Analysis of SARS-CoV-2

Data preprocessing -

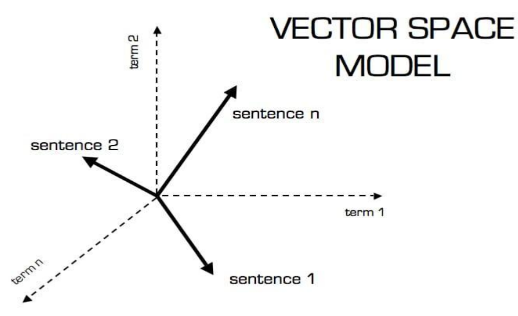
Data preprocessing is an essential step in building a Machine Learning model and depending on how well the data has been preprocessed; the results are seen. In NLP, text preprocessing is the first step in the process of building a model.The various text preprocessing steps are:

- Tokenization

- Lower casing
- Stop words removal
- Stemming
- Lemmatization

These various text preprocessing steps are widely used for dimensionality reduction. In the vector space model, each word/term is an axis/dimension. The text/document is represented as a vector in the multi-dimensional space. The number of unique words means the number of dimensions. Image for postImage for post Installation: The python library I'll be using to implement the text preprocessing tasks is nltk pip install nltk==3.4.5
Tokenization: Splitting the sentence into words.

Stop words removal: Stop words are very commonly used words (a, an, the, etc.) in the documents. These words do not really signify any importance as they do not help in distinguishing two documents.



Vector model space

Lemmatization: Unlike stemming, lemmatization reduces the words to a word existing in the language.Either Stemming or Lemmatization can be used. Libraries such as nltk, and spaCy have stemmers and lemmatizers implemented. These are built based on a rule-based approach.Stemmer is easy to build than a lemmatizer as the latter requires deep linguistics knowledge in constructing dictionaries to look up the lemma of the word.For lemmatization to resolve a word to its lemma, part of speech of the word is required. This helps in transforming the word into a proper root form. However, for doing so, it requires extra computational linguistics power such as a part of speech tagger.

countvectorizer and tfidfvectorizer

Vectorization is the general process of turning a collection of text documents into numerical feature vectors. This specific strategy (tokenization, counting and normalization) is called the Bag of Words or “Bag of n-grams” representation. Documents are described by word occurrences while completely ignoring the relative position information of the words in the document.

CountVectorizer converts a collection of text documents to a matrix of token counts: the occurrences of tokens in each document. This implementation produces a sparse representation of the counts.

Note that for each sentence in the corpus, the position of the tokens (words in our case) is completely ignored. When constructing this bag-of-words representation, the default configuration tokenizes the string by extracting words of at least 2 alphanumeric characters (punctuation is completely ignored and always treated as a token separator).

We can further transform a count matrix to a normalized tf: term-frequency or tf-idf: term-frequency times inverse document-frequency representation using Tfidf Transformer. The formula that is used to compute the tf-idf for a term t of a document d in a document set is:

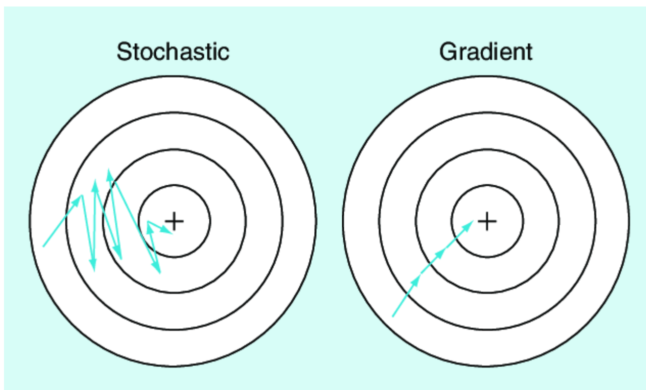
$$\text{tf-idf}(t, d) = \text{tf}(t, d) * log(\frac{n}{df(t) + 1})$$

The effect of adding 1 to the denominator in the equation above is that terms with zero idf, i.e., terms that occur in all documents in a training set, will not be entirely ignored. At the end, each row is normalized to have unit Euclidean norm (by dividing l2 norm of itself).

The goal of using TF-IDF instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus..

Stochastic Gradient Descent

It is an optimization algorithm which updates the coefficients after each training instance instead directly at the end of a training cycle. It is a discriminative algorithm which is used as a classifier. The coef_ parameter used in SGD classifier is used for updating weight assigned to each feature. This classifier was proven to be the best model for our twitter data collection compared to the other evolved models such as CatBoost,XGboost classifiers.



Comparison between SGD and GD

Cat Boost Classifier

CatBoost converts categorical values into numbers using various statistics on combinations of categorical features and combinations of categorical and numerical features.
CatBoost documentation says that-

"CatBoost is a high-performance open source library for gradient boosting on decision trees."

So, CatBoost is an algorithm for gradient boosting on decision trees.It is a readymade classifier in scikit-learn's conventions terms that would deal with categorical features automatically.It can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML. It can work with diverse data types to help solve a wide range of problems (described later) that businesses face today. It is developed by Yandex researchers and engineers, and is used for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other tasks.

Also, it provides best-in-class accuracy. It is especially powerful in two ways:

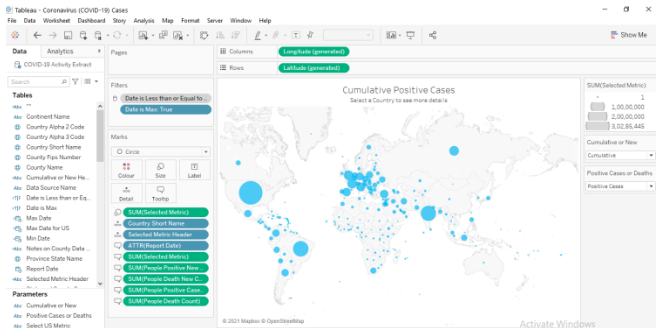
1. It yields state-of-the-art results without extensive data training typically required by other machine learning methods, and
2. Provides powerful out-of-the-box support for the more descriptive data formats that accompany many business problems.

“CatBoost” name comes from two words - **“Category”** and **“Boosting”**. It works well with multiple categories of data, such as audio, text, image including historical data. **“Boost”** comes from gradient boosting machine learning algorithm as this library is based on gradient boosting library. Gradient boosting is a powerful machine learning algorithm that is widely applied to multiple types of business challenges like fraud detection, recommendation items, forecasting and it performs well also. It can also return very good results with relatively less data, unlike DL models that need to learn from a massive amount of data.

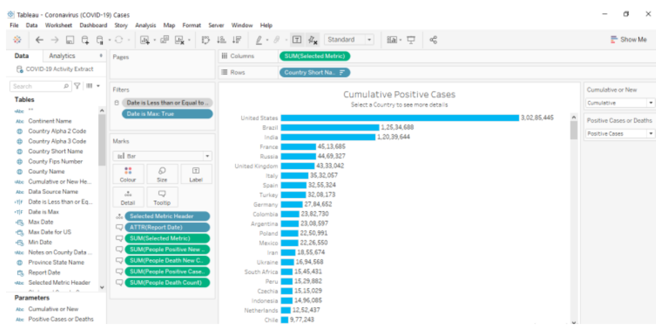
It is open-source and can be used by anyone. Both the models gave approximately similar accuracy, with SGD classifier edging the CatBoost algorithm. Due to the random selection of the training and test set we cannot differentiate between the two.

C. Tableau Visualization

Tableau is an easy to use visualization tool that provides with data merging, categorization etc. The data world dataset is a daily updating dataset that uses government data across the world to accumulate it into csv files that are reusable. The visualization are as follows,



World Map of Covid cases



Live tracker for each country

CHAPTER 4
RESULTS

```
models/ResNet50V2-FPN-fold1-03-0.9287.hdf5
tp:  8021 fp:  301
models/ResNet50V2-FPN-fold1-04-0.9627.hdf5
tp:  8150 fp:  172
models/ResNet50V2-FPN-fold1-06-0.9704.hdf5
tp:  8064 fp:  258
models/ResNet50V2-FPN-fold1-01-0.8399.hdf5
tp:  6023 fp:  2299
models/ResNet50V2-FPN-fold1-07-0.9759.hdf5
tp:  8034 fp:  288
models/ResNet50V2-FPN-fold1-13-0.9857.hdf5
tp:  8150 fp:  172
```

Figure - Output

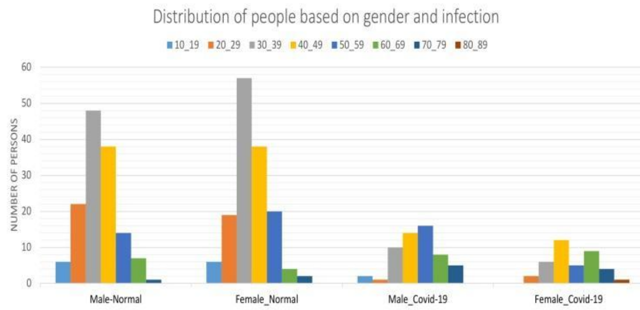


Figure – Distribution on gender

Average between five folds	Overall Accuracy	COVID sensitivity	Normal sensitivity
ResNet50V2 with FPN	98.49	94.96	98.7
Xception	96.55	98.02	96.47
ResNet50V2	97.52	97.99	97.49

Figure- ResNet50V2 accuracy

	Model	Test accuracy
4	Stochastic Gradient Decent	0.862488
1	Logistic Regression	0.859451
6	CatBoost	0.852162
0	Support Vector Machines	0.845603
2	Random Forest	0.827381
3	Naive Bayes	0.791667
5	XGBoost	0.739553

Figure – Sentiment Analysis

CHAPTER 5
FUTURE SCOPE

This Paper does not focus on visualization in detail. In future we can develop a SIRD model based system to enhance the visualization and can incorporate an allround system like Power BI. There is scope for development in terms of web development using frameworks like django which will help on the spot prediction of sentiment analysis and CT scan results.

CHAPTER 6
REFERENCES

[1] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. Di Napoli, ‘‘Features, evaluation and treatment coronavirus (COVID-19) [updated 2020 Apr 6],’’ in StatPearls [Internet]. Treasure Island, FL, USA: StatPearls Publishing, Jan. 2020. [Online]. Available: https:// www.ncbi.nlm.nih.gov/books/NBK554776/

[2] World Health Organization. Coronavirus disease (COVID-19) Pandemic. Accessed: Apr. 30, 2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel-coronavirus-2019

[3] SARS Basics Fact Sheet. (Dec. 2017). Centers for Disease Control Prevention (CDC) [Online]. Available: https://www.cdc.gov/sars/about/fssars.html

[4] World Health Organization. Middle East Respiratory Syndrome Coronavirus (MERS-CoV). Accessed: Apr. 20, 2020. [Online]. Available: https://www.who.int/emergencies/mers-cov/en/

[5] World Health Organization. (Apr. 2020). Middle East Respiratory Syndrome Coronavirus (MERS-CoV) The Kingdom of Saudi Arabia. [Online]. Available: https://www.who.int/csr/don/08-april-2020-merssaudi-arabia/en/

[6] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia – **Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in China: a report of 1014 case** Radiology (2020), p. 200642View Record in ScopusGoogle Scholar[2]

[7]
M.Y. Ng, E.Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M.M.-s. Lui, C.S.-Y. Lo, B. Leung, P.-L. Khong, *et al.*

Imaging profile of the covid-19 infection: radiologic findings and literature review
Radiology: Cardiothoracic Imaging, 2 (1) (2020), Article e200034
CrossRefGoogle Scholar

[8]
https://towardsdatascience.com/twitter-sentiment-analysis-based-on-news-topics-during-covid-19-c3d738005b55

[9]
https://www.analyticsvidhya.com/blog/2021/02/sentiment-analysis-predicting-sentiment-of-covid-19-tweets/

[10]
https://towardsdatascience.com/sentiment-analysis-of-covid-19-vaccine-tweets-dc6f41a5e1af

[11]
https://www.mdpi.com/2078-2489/11/6/314/htm

[12]
https://thesai.org/Downloads/Volume12No2/Paper_52-Public_Sentiment_Analysis_on_Twitter_Data.pdf

[13]
https://thesai.org/Downloads/Volume12No1/Paper_72-
Topic_based_Sentiment_Analysis_for_COVID_19_Tweets.pdf

[14]
https://ejmcm.com/pdf_3797_3d758af0de2b8ca74bc7cdfba9198fef.html