Covid19 X-rays Image Classification Using Machine Learning and Deep Learning

[1]Nikhil Fande, [2]Roshan Fande, [3]Rupesh Wadibhasme, [4]Bhavesh Wadibhasme

[1,2,3,4]M.Tech in Mathematical Modeling and Simulation
Department of Modeling and simulation,
Savitribai Phule Pune University,Pune(MH), India

***Abstract*: The novel Coronavirus (Covid19) has firstly started in China and widely spread in various countries and approximately 4.1 Million cases have been found worldwide. There are a limited number of COVID-19 testing kits available in hospitals due to gradually increasing in cases on a daily basis. Therefore, in order to increase the testing, it is necessary to implement an auto detection system which will prevent the spread of Covid-19. In this work, we build the binary classifiers based on the machine learning and deep learning models on real image data in predicting positive case probability and provide comparison study of each model. The top features of images have been used for further modeling processes to test stability of binary classifiers by comparing their performance on separate data. We observe that support vector machines (SVM) and Logistic Regression is more stable than convolutional neural networks.**

***Index Terms*: COVID-19, Coronavirus, Machine Learning, Convolutional Neural Network, Chest X-ray, Support Vector Machine, Logistic Regression**

1.INTRODUCTION

The WHO has declared a coronavirus outbreak as a pandemic. The corona virus is a family of viruses that can cause a range of illness in humans like common cold and more severe forms like SARS (Severe Acute Respiratory Syndrome) and MERS (Middle East Respiratory Syndrome) which are life threatening. The name of the virus took the form of a crown with protrusion around it and hence its known as coronavirus.

Coronavirus cases total number is approximately 4,015,107 and 276,268 of them died and 1,387,478 were recovered. Currently infected patients' number is 2,351,361 [1]. While 95% of the number of infected patients survive the disease slightly, 5% the rest have a serious or critical illness. The recent outbreak of coronavirus is believed to have occurred in a market for illegal wildlife in the central Chinese city of Wuhan. Chinese health authorities and the WHO are investigating the outbreak of the recent coronavirus. Worries are mounting that prolonged quarantines, supply chain disruptions and a sharp reduction in tourism and business travel could weaken the global economy or even cause a recession due this pandemic. The distribution of COVID-19 cases seen worldwide till 8 May 2020. is shown in Figure 1.[1]

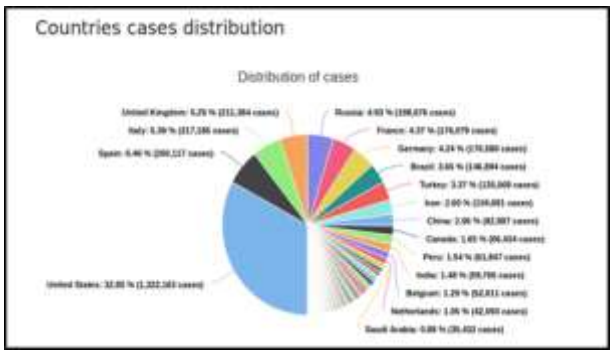

Figure 1. The distribution of COVID-19 cases seen worldwide till 8 May 2020

Even in many developed countries, the health system has come to the point of collapse due to the increasing demand for intensive care units simultaneously. Vaccine would normally take years, if not decades to develop. Most experts think vaccines are likely to become available by mid-2021.

2.PRESENTATION OF MODELS

In this section we present the structure of models with different (i) parameters, (ii) stopping criteria and (iii) activation functions. References are provided for more details.

2.1 Support Vector Machine (SVM)

SVM is the well-known supervised learning approach used for both classification and regression. It is based on statistical learning theory and tries to find out the optimal hyperplane for separation of two classes if it is the binary classification problem. Optimal hyperplane is one which maximizes the margin between examples of different classes. Following Figure 2. Shows the best optimal

hyperplane classifying the two classes i.e. class A and class B.[2]

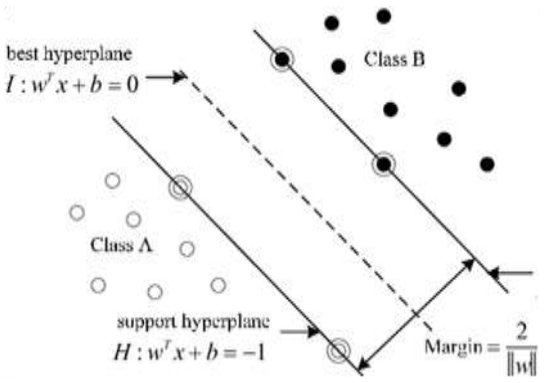

Figure 2. Support Vector Machine

The main objective of SVM is to classify the new example into their correct class. There are many hyperplanes which may classify the data but the best hyperplane is one which maximize the margin between two classes.[7]

Equation of Hyperplane in 2 dimensions

$$\beta 0 + \beta 1X1 + \beta 2X2 = 0 \qquad \text{Equation...} \quad (1)$$

Equation of Hyperplane in p dimensions

$$\beta 0 + \beta 1X1 + .. + \beta 2pX2p = 0 \qquad \text{Equation...} \quad (2)$$

Where, X is the feature vector

As we know, many times real time data comes with nonlinear structure. Thus, in such cases it is very difficult to separate the data with linear-hyperplane. In order to separate the nonlinear data SVM uses the kernel trick. Kernel tricks transform the data from original space to higher dimensional space. The major drawback of SVM is the computing cost.

2.2 Gradient Boosting Machine

Gradient Boosting Machine or simply GBM is the tree based supervised learning algorithm used for both classification and regression.[5] In GBM learning procedure consecutively fits a new model to provide the accurate estimate of response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic squared-error loss, the learning procedure would result in consecutive error-fitting. In general, the choice of the loss function is up to the researcher, with both a rich variety of loss functions derived so far and with the possibility of implementing one's own task-specific loss.[6]. The high flexibility of GBM makes it more customizable for any particular data-driven task. Particular data-driven task.

Input : Data set $\mathcal{D}$.
A loss function L.
A base learner $\mathcal{L}_\Phi$.
The number of iterations M.
The learning rate η.

1 Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg\min_\theta \sum_{i=1}^{n} L(y_i, \theta)$;
2 **for** $m = 1,2,..,M$ **do**
3 $\quad \hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;
4 $\quad \hat{\phi}_m = \arg\min_{\phi \in \Phi, \beta} \sum_{i=1}^{n} \left[(-\hat{g}_m(x_i)) - \beta\phi(x_i)\right]^2$;
5 $\quad \hat{\rho}_m = \arg\min_\rho \sum_{i=1}^{n} L(y_i, \hat{f}^{(m-1)}(x_i) + \rho\hat{\phi}_m(x_i))$;
6 $\quad \hat{f}_m(x) = \eta\hat{\rho}_m\hat{\phi}_m(x)$;
7 $\quad \hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$;
8 **end**

Output: $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$

Figure 3. Algorithm for GBM

2.3 Logistic regression

The underlying concept of logistic regression is the logit function (natural logarithm of odd ratio). It is most interpretable algorithm followed by equation,[4]

$$Y = \alpha + \beta x \qquad \text{Equation (3)}$$

Where,
Y is the response variable to be predicted
X is the independent variables
β are the variable coefficient

Logistic regression is well suited for describing and testing the hypothesis about the relationship between categorical outcomes or one or more continuous variables. Logistic regression predicts the logit of Y from X. As stated earlier, logit is the natural log of odd of Y and odd ratio is the probability of Y happening to probability of y not happening.
The simple logistic regression has form

$$\text{logit}(Y) = natural\log(odds) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

The value of coefficient β determines the X and logit Y. If β is greater then there is a larger change in Y with respective to X keeping rest of X constant. Similarly, as β is less then there is less change in Y with respective to X keeping rest of X constant. As β equal to zero then there is no any relationship exit between X and Y.

2.4 Convolutional Neural Network (CNN)

Convolutional Neural Network is similar to the ordinary neural network. It is made of firing neurons, learnable weights and biases. Each neuron receives some input and pass to intermediates layers and finally provides the output. The structure of CNN is given below.[4]

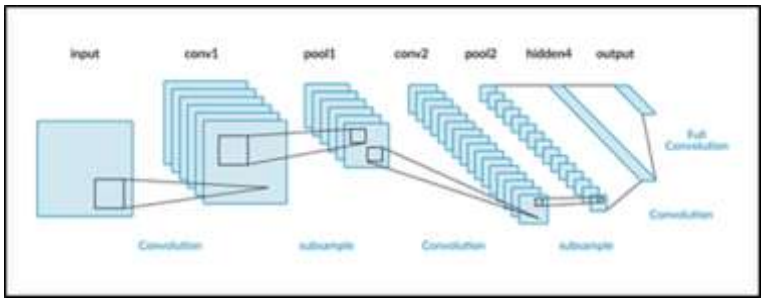

Figure 4. Architecture of CNN

Architecture of CNN consist of following layers as,

1. **Input:** It holds the raw image value. In this case image may be in any size (example: 32 × 32) with 3 channels i.e. RGB.
2. **Convolutional layer:** Output of input is fed to the convolutional layer and it extracts the feature using convolution operation.
3. **Pooling layer:** Output of convolutional layers given to the pooling layer. Objective of this layer is to reduce the spatial size of the convolved features which reduce the computation power for processing the data.
4. **Hidden layer:** Objective of the hidden layer is to add the complexity in the model for better learning. Each hidden layer consists of activation functions
5. **Flatten:** Once final convolved features have been obtained then it is converted into 1-D array and fed to the fully connected layer for learning.
6. **Fully connected layer:** It is a simple kind of feedforward network which accept the input and pass through the different hidden layer.[9]

3.THE CRITERIA

There are several performance measures to compare the models including Matthews Correlation Coefficient,Accuracy, F1-score, precision and recall. In this paper we will mainly present the result on F1 Score criteria. For each model we will Calculate Cross Validation Performance Measures. Finally We Will Take an Average of Performance Measures for each Model and Check Which Model Gives a higher F1-score and MCC (Matthews Correlation Coefficient).

4. DATA AND MODELS

In this section we provide an overview of data structure and models.

4.1 The Data

In this study chest X-ray images of 125 Covid-19 have been obtained from an open source GitHub repository shared by Dr. Joseph Cohen and some open sources.[11] This repository consists of chest X-ray images of Covid-19, Middle East respiratory syndrome (MERS), Severe Acute respiratory syndrome and pneumonia. We are only interested in Covid-19 X-ray images. Additionally, 500 normal chest X-ray images taken into account from Kaggle ("Chest X-ray Images (Pneumonia)") for modeling purposes.

In our experiments, we have created the dataset as, considering 105 covid-19 chest X-ray images as positive cases and 480 normal chest X-ray images as negative cases. Further we split this into a training and validation set as 80:20 ratio. We have created a test dataset consisting of 20 Covid Images and 20 Normal Images. Following Fig 5 and 6 shows the chest X-ray images of Covid-19 and normal patients respectively.

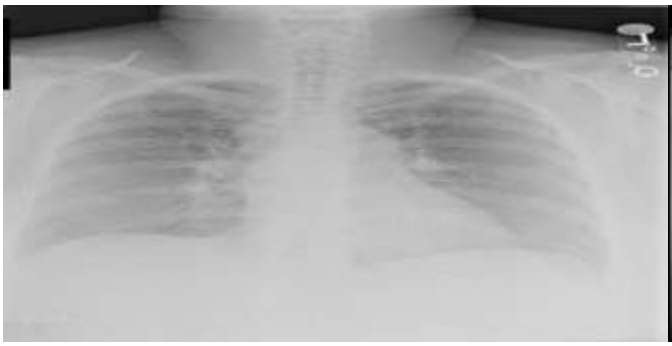

Figure 6. Chest X-ray image of normal patient

Figure 5. Chest X-ray image of Covid-19 patient

4.2 The models

The models we use has been discussed in the previous section. We mostly focus on four models: Support Vector Machine (SVM), Gradient Boosting Machine, Logistic Regression and Convolutional Neural Network. To rank the model with respective quality MCC and F1-score criteria are used.

In this study, input images of size 128 × 128 × 3 have been used for model building. Basic preprocessing has been done on top of each image such as segmentation and sharpening. Here features have been extracted using CNN and used to build convolutional neural networks. Similarly removing the last layer of CNN, the remaining features have been used for building SVM, GBM and Logistic Regression. The dataset is randomly split into two independent datasets with 80% and 20% for training and validation respectively.

4.2.1 Parameters and hyperparameters used for Models

Following table 1. Shows the models with their hyperparameters.

Models	Hyperparameters
SVM	C=1.0, kernel='rbf', degree=3, tol=0.001
GBM	Leraning_rate=0.1, estimators=100, max_depth=3
Logistic Regression	tol=0.0001, C=1.0
CNN	Activation function=LeakyReLU, Dropout=0.1, optimizer=adam, loss=categorical_crossentropy epoch=40, batch size=30

5. RESULTS AND DISCUSSION

We convert Images into arrays and resize the picture to reduce computation. For Feature Extraction from Image we use 32 and 64 Filter with 5*5 Convolution Filter size with 2*2 Stride Followed by Leaky Relu activation Function .To Increase the Example we use Data Augmentation using slight rotation and zooming of Image. But we are not flipping Image horizontally and vertically. Performance Measurement for such a Model, We use Matthews correlation coefficient, F1 Score, Precision, Recall.. We verify our model Performance Based on these Measures. We train our Model on 460 Images and Validate on 117 Images, and use 40 Images (20 COVID + 20 Normal) For testing Purpose, Testing Images are not included in Training and Validation Set. To avoid Biasing in train_test_split we used Stratified Sampling with Five-Fold Cross Validation.

In Modeling Precision is defined as the fraction of the examples which are actually Positive among all the examples which we predicted positive. And Recall is among all the examples that are actually positive, what fraction did we detect as Positive. For Skewed Dataset We did not prefer Accuracy Score, Such time we Prefer F1 Score (harmonic mean of Precision and Recall) and Matthews Correlation Coefficient.

We Judge our Model on this Four Performance Measurement,

As mentioned earlier we have built the four different model (i.e. SVM, GBM, Logistic regression and CNN) for Covid-19 X-ray

images classification. In this study we have mostly focused on cross validation performance measures (Fivefold Performance Measures has been calculated. Each fold consists of 20% validation data and 67% train data for modeling purposes.

Following table shows the five-fold performance measure on Validation data. We calculated all performance measures on minority class (COVID Positive).

Performance Measures	CV1	CV2	CV3	CV4	CV5	Mean
Precision	0.94	0.8	0.87	0.857	0.92	0.87
Recall	0.76	0.76	0.66	0.57	0.61	0.67
F1-Score	0.84	0.78	0.75	0.68	0.74	0.75
MCC	0.81	0.73	0.72	0.65	0.71	0.72

Fig.1 Results of Convolutional Neural Network Classifier

Performance Measures	CV1	CV2	CV3	CV4	CV5	Mean
Precision	0.9	0.75	0.9	0.77	0.84	0.83
Recall	0.85	0.85	0.85	0.80	0.76	0.82
F1-Score	0.87	0.8	0.87	0.79	0.8	0.82
MCC	0.85	0.75	0.74	0.74	0.76	0.79

Fig.2 Results of Support Vector Machine Classifier

Performance Measures	CV1	CV2	CV3	CV4	CV5	Mean
Precision	0.93	0.77	0.86	0.90	0.77	0.84
Recall	0.66	0.80	0.61	0.90	0.80	0.75
F1-Score	0.77	0.79	072	0.90	0.79	0.79
MCC	0.75	0.74	0.68	0.88	0.74	0.75

Fig 3 Results of Gradient Classfier

Performance Measures	CV1	CV2	CV3	CV4	CV5	Mean
Precision	0.94	0.69	0.85	0.76	0.77	0.80
Recall	0.85	0.85	0.85	0.95	0.80	0.86
F1-Score	0.90	0.76	0.85	0.85	0.79	0.83
MCC	0.88	0.71	0.82	0.82	0.74	0.79

Fig 4. Results of Logistic Regression Classifier

As Per our Analysis average f1-score of SVM is Higher than CNN, Logistic regression and Gradient Boosting classifier.

We chose the SVM Model to predict test examples. In test example we have total 40 Images (20 COVID + 20 Normal)

Following Table shows the Performance of SVM on test data.

Precision	Recall	F1-Score	MCC
1.0	0.95	0.97	0.95

Out of 20 COVID Images Model correctly detect 19 Images with Zero False Positive. & for 20 Normal Images Model detect all (20) Normal Images with One False Negative. For such a type of Problem statement both Precision and Recall have to be high. Precision shows how many examples it detects correctly as COVID Positive out of total COVID Positive Detection. Recall Shows how many examples it detect correctly as COVID Positive out of Total COVID Example.

6. CONCLUSION

In this study we have shown the results of different algorithms on Covid19 dataset.In future doctors may use these models for testing the patients in hospital against this fight.Also we have shown different performance measures for each algorithm.In this we found SVM is beating the other algorithms,but it may change as dataset increases.

ACKNOWLEDGMENT

I wish to thank Roshan G. Fande,Rupesh Wadibhasme and Bhavesh Wadibhasme for his extra effort and guidance which add the more lights on work.

REFERENCES

[1] https://www.worldometers.info/coronavirus/
[2] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," submitted to Data Mining and Knowledge Discovery, 1998.
[3] Yushi Chen, Deep feature extraction and classification of hyperspectral images based on convolutional Neural Network
[4] Jonne Peng , an Introduction of Logistics regression analysis and reporting
[5] Alexey Natekin, Alois knoll,Gradient Boosting Machines, a tutorial
[6] Halefom Tekle Weldegebriel A New Hybrid Convolutional Neural Network and eXtreme Gradient Boosting Classifier for Recognizing Handwritten Ethiopian Characters.
[7] X. X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," Pattern Recognition, pp. 1318– 1325, Sep. 2011.
[8] D. S. Maitra, U. Bhattacharya, and S. K. Parui, "CNN Based Common Approach to Handwritten Character Recognition of Multiple Scripts," presented at the International Conference on Document Analysis and Recognition (ICDAR), 2015.
[9] Mohamed Medhat Gaber, Classification of COCID-19 in chest X-ray using DaTraC Deep Convolutional Neural Network