A Comparative Study for Efficient Covid-19 Detecting Machine Learning Models on CT Images

Ankur Chaturvedi¹, Divyansh Mishra¹, Dr. Vikram Rajpoot¹, Janvi Gupta¹, and Aditi Sharma¹

¹Department of Computer Engineering & Applications, GLA University Mathura, NH-2 Delhi-Agra Highway, 281406, India

E-mail: ankur.chaturvedi@glau.ac.in

Abstract. Covid19 has impacted the whole world drastically, Lakhs of people are getting infected every day and thus it’s getting difficult to detect Covid infection in huge population. Main cause of death due to Covid19 is Covid infection in lungs. Infection in lungs is detected by CT-Scans of Lungs, but since doctors need time analyze CT-Scan and also, they can’t do it 24 into 7. Hence, many people who need extra medical care for Covid infection in lungs, need to wait for analysis of CT-Scan by doctors that whether their lungs are infected or not. Here in the paper, it is tried to make an efficient Machine Learning Model which could detect Covid19 infection in Lungs, in order to reduce human resource time (time taken by doctors to analyze lung CT-Scan) so that needy patients would be getting started with corresponding medical treatments more quickly. It may save many lives. Here in the paper, it’s tried to detect Covid19 infection in lungs analyzing CT-scans using SVM classifier, Logistic regressions classifier and Machine learning perceptron with two approaches of preprocessing of pixel values and then results have been compared. Dataset used for the purpose, is kaggle dataset (Covid-19 Lung CT Scan).

1. Introduction
It’s been seen from some time that Covid19 has impacted the whole world and many people are dying everyday due to the disease. The main cause of death due to Covid19 is Corona virus infection in lungs. A way to detect Corona infection in lungs is CT-Scan analysis. Detection of Corona virus infection in lungs by Machine Learning approaches can reduce time for detection and thus facilitate quick starting of treatment for the needy, which may save lives. Here, it’s tried to analyze and train the Machine Learning model on lung CT-Scans of Covid19 patients and healthy persons. Models used for the purpose are SVM and Logistic regression. These two (SVM and Logistic regression) models have been tried on three differently preprocessed pixel valued (NPV, MPV and APV) images and the accuracies of all these combinations have been compared.

Defining the abbreviations used in this paper, NPV, MPV and APV:

NPV, MPV and APV are three different approaches to preprocess pixel values of an image.

I. NPV: stands for Normal pixel value.

II. MPV: MPV stands for Multiplied pixel value.
III. APV: APV stands for Advanced pixel value.

IV. Training set
Training_set is a 2D matrix of size (m, b). Where ‘m’ is the number of images in training data set. And ‘b’ is the size of vector V, value of which depends on the approach (NPV, MPV or APV) used for preprocessing the image. Each row in the training data set represents an image.

V. Vector V
A vector V represents an image. Each row of training_set matrix is formed by vector V that represents an image as a numerical valued vector. Size of vector V is (1, b). Where ‘b’ depends on the approach (NPV, MPV or APV) used for preprocessing the image.

MPV, NPV and APV approaches to preprocess an image are explained in detailed way in section 3 of this paper. Python programming language is used to apply these approaches and to make and train models (Logistic regressions and SVM) in this paper. Python libraries used for the purpose are numpy, sklearn, opencv and matplotlib. And the dataset used for the purpose, is kaggle dataset (Covid-19 Lung CT Scans).

2. Literature Review
I. Ozsahin, B. Sekeroglu, M. S. Musa, M. T. Mustapha, and D. U. Ozsahin [1]. Here, they reviewed all techniques based on Machine Learning that are being used to diagnose Covid19 using lungs CT-Scans. They searched and read articles and papers related to deep learning, neural networks, Covid19, and CT-Scans of lungs. Their basic classifications of these studies are as Covid vs Non-Covid, Covid vs normal, Covid vs Non-Covid-Pneumonia and Covid severity. And results emphasize on benefits of masks and lesions masks in Covid19 diagnosis.

X. Li, W. Zeng, X. Li, H. Chen, L. Shi, X. Li, H. Xiang, Y. Cao, H. Chen, C. Liu and J. Wang [2]. Here, they collected data about 131 confirmed Covid19 positive patients with help of three Chinese hospitals and analyzed the symptoms and manifestations that were found common among these patients. They also analyzed the evolution features of lungs CT-Scans as well as characteristics.

T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, U. R. Acharya [3]. Here, in this paper, they introduce a model to detect Covid19 using raw lungs CT-Scans. This model performs over Covid vs Non-Covid (binary classification) data and Covid vs Non-Covid vs Pneumonia data (multi-class classification).

X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang and P. Xie [4]. They try to do some experiments on data (images of lungs CT-Scans) in this paper study and these experiments further emphasize about benefit of this data in developing Machine Learning Based softwares to detect Covid19. And they further try to develop methods and programs based on multi-task learning and self-supervised learning to recognize Covid infection in lungs.

J. Matos, F. Paparo, I. Mussetto, L. Bacigalupo, A. Veneziano, S. P. Bernardi, E. Biscaldi, E. Melani, G. Antonucci, P. Cremonesi, M. Lattuada, A. Pilotto, E. Pontali and G. A. Rollandi [5]. Here, the technology, Computed tomography (CT) is told to be very useful in finding severe acute coronavirus 2 infection as well as coronavirus (2019) disease too. It's also told in this study that CT helps in predicting the disease precisely. It further shows that SARS-CoV-2 burden can be estimated by measuring the VoD in CT post processing tool.

S. Sharma [6]. It a survey of a large number of research papers related to Machine Learning techniques used for Covid19 diagnosis. It says that papers and programs made for the purpose, have achieved an average accuracy of 91% and most of these models work with typical hazy patches on the outer edges of lungs caused by pneumonia of coronavirus. It also says that accuracy of Covid19 infection detecting models can be improved by training and testing on more images.

J. Shuja, E. Alazani, W. Alasmary and A. Alashaikh [7]. It is a survey in which they surveyed four open source efforts that are: how can be Machine learning models can be made to detect Covid19 using lungs CT-Scans, X-rays and cough sounds; how to estimate Covid19 transmission, further positive cases of
Covid and prognosis with the help of demographic, epidemiological and mobility data; analysis of sentiments and emotions emerged in people due to Covid19 outburst, with the help of social media; Survey and comparison of scholarly articles and papers related to technology developing to diagnose Covid19.

S. Kadry, V. Rajinikanth, S. Rho, N. S. M. Raja, V. S. Rao and K. P. Thanaraj [8]. They try to develop a Machine Learning based program or model to detect Covid19 infection. They use lungs Ct-Scans for their purpose. They try sequence of methods to make the model such as multi-thresholding, image multi-thresholding, separating images with threshold filter, finding features to train on, and then classification. They used Naive bays, SVM with linear kernel, K-Nearest neighbors, random forest and decision tree classifiers for their purpose.

S. A. Harmon, T. H. Sanford, B. Turkbey [9]. Here, they tried multiple deep learning techniques on data which was collected from multiple nations. And got accuracy around 98%.

H. Y. F. Wong, H. Y. S. Lam, and M. Y. Ng [10]. They performed a clinical survey on some number of Covid19 patients and studied the severity of the Covid19 infection in lungs using RT-PCR and chest radiographic examinations. Their findings reveal that the severity of Covid19 infection in lungs peaked at 10-12 days after the onset of symptoms of the infection.

3. Image Pre-processing

Since, Image consists of pixels and each pixel consists of three values R, B, G.

In this section, three different approaches (NPV, MPV and APV) are used to preprocess an image using R, B, G values of each pixel.

3.1. NPV

In NPV approach, each set of pixel values is taken as it is and a vector is formed using these values. If there’s a dataset which have CT-Scans images with p × q pixels.

So, for each image, total pixels in the image = p × q

Since, each pixel has 3 corresponding R, B, G values ranging from 0 to 255.

So, total numerical pixel values in an image = p × q × 3

Then it’s needed to define the vector V for each image such that vector V consists of all p × q × 3 pixel values of the corresponding image.

So, size of vector V would be (1, p × q × 3).

Where vector V represent an image.

If there are ‘m’ number of training examples, then our training_set would be a 2D matrix of size (m, p × q × 3).

And 2D matrix training_set will be given to model for training.

For example,

There’s a dataset which have 500 CT-Scan images with 256×256 pixels.

Now, total numerical pixel values in an image = 256 × 256 × 3 = 65536

thus, size of vector V would be (1, 65536).

And size of matrix training_set would be (500, 65536).

3.2. MPV

Each pixel has 3 numerical values (R, B, G). In MPV approach, combinations of multiplied R, B, G values have been used to form the vector V. Total 4 combinations of R, B, G values have been used here,

first combination is R × G/255,
second combination is R × B/255,
third combination is B × G/255,
And fourth combination is R × G × B/255 × 255.

To have the values of combinations within limit 0 to 255, it’s needed to divided the values of first combination, second combination and third combination by 255 and values fourth combination by (255)^2.
Now, the vector $V$ would be formed by horizontally stacking the values of first combination, second combination, third combination and forth combination.

If there’s a dataset which have CT-Scans images with $p \times q$ pixels.

So, for each image, total pixels in the image = $p \times q$

Now, each combination (first combination, second combination, third combination and forth combination) would be calculated for every pixel of the image.

So, total number of calculated values of each combination would be equal to the total number of pixels in the image, since each combination is calculated for each pixel. And since, each pixel has 4 combinations (first combination, second combination, third combination and forth combination), so

Total number of calculated values from all combinations = $p \times q \times 4$

And thus, size of vector $V$ would be (1, $p \times q \times 4$).

And if there are ‘m’ number of images in training data set, then size of our 2D matrix training_set would be (m, $p \times q \times 4$).

For example,

There's a dataset which have 500 CT-Scan images with 256×256 pixels.

Now,

Total number of calculated values from all combinations = $256 \times 256 \times 4$

= 262144

thus, size of vector $V$ would be (1, 262144).

And size of matrix training_set would be (500, 262144).

3.3. APV

In APV approach, $V$ vectors from NPV and MPV approaches have been taken and stacked horizontally to form the vector $V$ for APV approach.

Let vector $V$ for NPV approach = $V_{npv}$

Let vector $V$ for MPV approach = $V_{mpv}$

Let vector $V$ for APV approach = $V_{apv}$

So, $V_{apv} =$ horizontally stacking vectors $V_{npv}$ and $V_{mpv}$

From section 3.1 and 3.2, it’s known that size of $V_{npv}$ is (1, $p \times q \times 3$) and size of $V_{mpv}$ is (1, $p \times q \times 4$). Therefore, size of $V_{apv}$ would be (1, $(p \times q \times 3) + (p \times q \times 4))$ that is equal to (1, $p \times q \times 7$).

And if there are ‘m’ number of images in training data set, then size of the 2D matrix training_set would be (m, $p \times q \times 7$).

For example,

There's a dataset which have 500 CT-Scan images with 256×256 pixels.

Now, Total number of values of $V_{npv}$ and $V_{mpv}$ = $(256 \times 256 \times 4) + (256 \times 256 \times 3)$

= 458752

thus, size of vector $V$ would be (1, 458752).

And size of matrix training_set would be (50, 458752).

4. Modelling the Data

2D matrix training_set evaluated from each approach (NPV, MPV and APV) is passed as a numpy array to Logistic regression classifier with lbfgs solver and 500 iterations.

2D matrix training_set is also passed to SVM classifier as a numpy array.

And the accuracies and F1-scores given by each model have been analyzed.

Accuracy = \(\frac{(TP + FN) \times 100}{(TP + TN + FP + FN)}\) \hspace{1cm} (1)

F1- Score = \(\frac{TP}{[TP + 0.5 (FP + FN)]}\) \hspace{1cm} (2)

Where,

TP = true positives
TN = true negatives
FP = false positives
FN = false negatives
5. Result
Table 1. is Accuracy table, comparing accuracies from each classifier (SVM and Logistic regression) with respect to each approach (NPV, MPV or APV).

| Approaches | SVM       | Logistics Regression |
|------------|-----------|----------------------|
| NPV        | 99.59%    | 99.19%               |
| MPV        | 68.82%    | 70.85%               |
| APV        | 100%      | 99.59%               |

Table 2. is F1-score table, comparing F1-scores from each classifier (SVM and Logistic regression) with respect to each approach (NPV, MPV or APV).

| Approaches | SVM       | Logistics Regression |
|------------|-----------|----------------------|
| NPV        | 0.9953    | 0.9908               |
| MPV        | 0.5599    | 0.6363               |
| APV        | 1         | 0.9954               |

6. Observations and Conclusion
With this data set (mentioned in section 1), it’s seen that NPV approach gives very good accuracy but accuracy of MPV approach wasn’t that good. But when these two approaches are combined together in APV approach, then accuracy enhances with both the classifiers (SVM and Logistic regression) and rather with SVM classifier it reaches to even 100%.

7. References
[1] Ozsahin I, Sekeroglu B, Musa Sani M, Mustapha M, and Ozsahin D U 2020 Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence
[2] Li X, Zeng W, Chen H, Shi L, Li X, Xiang H, Cao Y, Chen H, Liu C and Wang J 2020 CT imaging changes of corona virus disease 2019(COVID-19): a multi-center study in Southwest China
[3] Ozturk T, Talo M, Yildirim E, Baloglu U B, Yildirim O U, Acharya R 2020 Automated detection of COVID-19 by cases using deep neural networks with X-ray images
[4] Yang X, He X, Zhao J, Zhang Y, Zhang S and Xie P 2020 COVID-CT-Dataset: A CT Image Dataset about COVID-19
[5] Matos J, Paparo F, Mussetto I, Bacigalupo L, Veneziano A, Bernardi S P, Biscaldi E , Melani E, Antonucci G, Cremonesi P, Lattuada M, Pilotto A, Pontali E and Rollandi G A 2020 Evaluation of novel coronavirus disease (COVID-19) using quantitative lung CT and clinical data: prediction of short-term outcome
[6] Sharma S 2020 Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients
[7] Shuja J, Alanazi E, Alasmary W and Alashaikh A  2020 Covid-19 Open-Source Data Sets: A Comprehensive Survey
[8] Kadry S, Rajinikanth V, Rho S, Raja N S M, Rao V S and Thanaraj K P 2020 Development of a Machine-Learning System to Classify Lung CT Scan Images into Normal/COVID-19 Class
[9] Harmon S A, Sanford T H, Xu S, Turkbey E B, Roth H, Xu Z, Yang D, Myronenko A, Anderson V, Amalou A, Blain M, Kassin M, Long D, Walker N V S M, Bagci U, Ierardi A M, Stellato E, Plensich G G, Franceschelli G, Girlando C, Irmici G, Dominic Labella D, Dima Hammoud D,
Ashkan Malayeri A, Jones E, Summers R M, Choyke P L, Xu D, Flores M, Tamura K, Obinata H, Mori H, Patella F, Cariati M, Carrafiello G, P, An P, Wood B J and Turkbey B 2020 Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets

Wong H Y F, Lam H Y S, Fong A H T, Leung S T, Chin T W Y, Yen Lo C S Y, Lui M M S, Lee J C Y, Chiu K W H, Chung T W H, Lee E Y P, Wan E Y F, Hung I F N, Lam T P W, Kuo M D and Ng M Y 2020 Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19