Computed Tomography Diagnosis of COVID-19 using Supervised Learning

Amantej Sran, Ruthvik Raja M.V, ENGG 6090(Image Analysis), University of Guelph

***Abstract*—During the Corona Virus outbreak, CT(Computed Tomography) is widely used for diagnosing COVID-19 patients. Due to many privacy concerns the CT images are not publicly available to implement Machine Learning and Deep Learning techniques for research and development of the AI-enabled algorithms to classify the CT images. To address this problem some researchers has created an open source dataset COVID-CT, which consists of 349 COVID-19 diagnosed CT images from 216 patients and 397 Non COVID-19 CT images. The usage of this dataset is confirmed by the senior Radiologist who has been treating and diagnosing COVID-19 patients since the outbreak of the novel Corona Virus. We also performed Machine Learning models like K-Nearest Neighbours, Support Vector Machine, Logistic regression and Deep Learning techniques like Convolutional Neural Network on the dataset to diagnose COVID-19.**

***Index Terms*—COVID-19, Classification, Supervised Learning, Convolutional Neural Network, DenseNet-169, ResNet-50.**

ABBREVIATIONS

CNN	Convolutional Neural Network
CoV-2	Corona virus 2
Cssl	Contrastive self Supervised Learning
CT	Computed Tomography
DA	Data Augmentation
DT	Decision Tree
KNN	K-Nearest Neighbours
LR	Logistic Regression
ML	Machine Learning
RT-PCR	Reverse Transcription Polymerase chain reaction
SARS	Severe acute respiratory syndrome
SVM	Support Vector Machine
TL	Transfer Learning

I. INTRODUCTION

Corona virus is a type of virus that is linked to worldwide pandemic illness. The novel corona virus that has caused this outbreak was first found in December of 2019 [1] and with the rapid spread of this virus throughout the world, an increased need for COVID-19 tests has been demanded [1].

Evidence suggests that transmission of the virus happens primarily through close contact with infected people, who infect others most commonly by saliva and respiratory droplets [2] and even with the early stages of vaccine doses being distributed, there are still approximately 5 million new cases worldwide per two-week period [3]. There have been over 120 million total cases worldwide as shown in Figure 1, Figure 2 shows the number of new cases till March 27th 2021 and almost 2.8 million deaths [3]. These numbers prove that more efficient and faster COVID-19 tests are necessary. One of the major problem with the Corona virus is controlling the spread of the virus from one person to another person. The most common current testing methods include a saliva test or a mucus sample from the nose or throat [4]. Both of these tests require additional lab testing to receive results which can take up to a week [4]. These tests are done with the help of RT-PCR test kits and there was a great shortage of these kits during the outbreak of COVID-19. As a result, many suspected cases were not tested in time and the spread of the virus continued to increase rapidly. Hence, more testing techniques are necessary to fasten the testing process so, hospitals have been utilizing alternative methods like CT scans of the patient is used to diagnose the disease. For example, in China, many hospitals have used CT scans of a patient to diagnose COVID-19 finally, this method has shown a positive response in identifying the virus in a patient.

Fig. 1: Worldwide total cases of COVID-19 (As of March/27/2021)

The CT scan of a patient can be used to predict whether the patient is diagnosed with Corona virus which is caused by the SARS-CoV-2 virus but the CT scan cannot be used to judge which viral pneumonia is causing COVID-19. However during the pandemic most of the viral pneumonia is caused by the SARS CoV-2.

During the outbreak of COVID-19, the radiologists in the rural regions found it very difficult in identifying the COVID-19 virus by seeing the CT scan of a patient because this virus is relatively new to the radiologists. To overcome these problems several works [14,15] has been done to develop Deep Learning models to predict whether a patient

is diagnosed with COVID-19 or not based on the CT scan of the patient but due to some privacy concerns the CT scans that are used in this work is not publicly available so some researchers [13] has created an open-source dataset COVID-CT, which consists of 349 COVID-19 diagnosed CT images from 216 patients and 397 Non-COVID-19 CT images. This dataset is publicly available and we have manually extracted the images from the dataset for training the model by reading the captions of the images which contains clinical findings of the Corona virus.

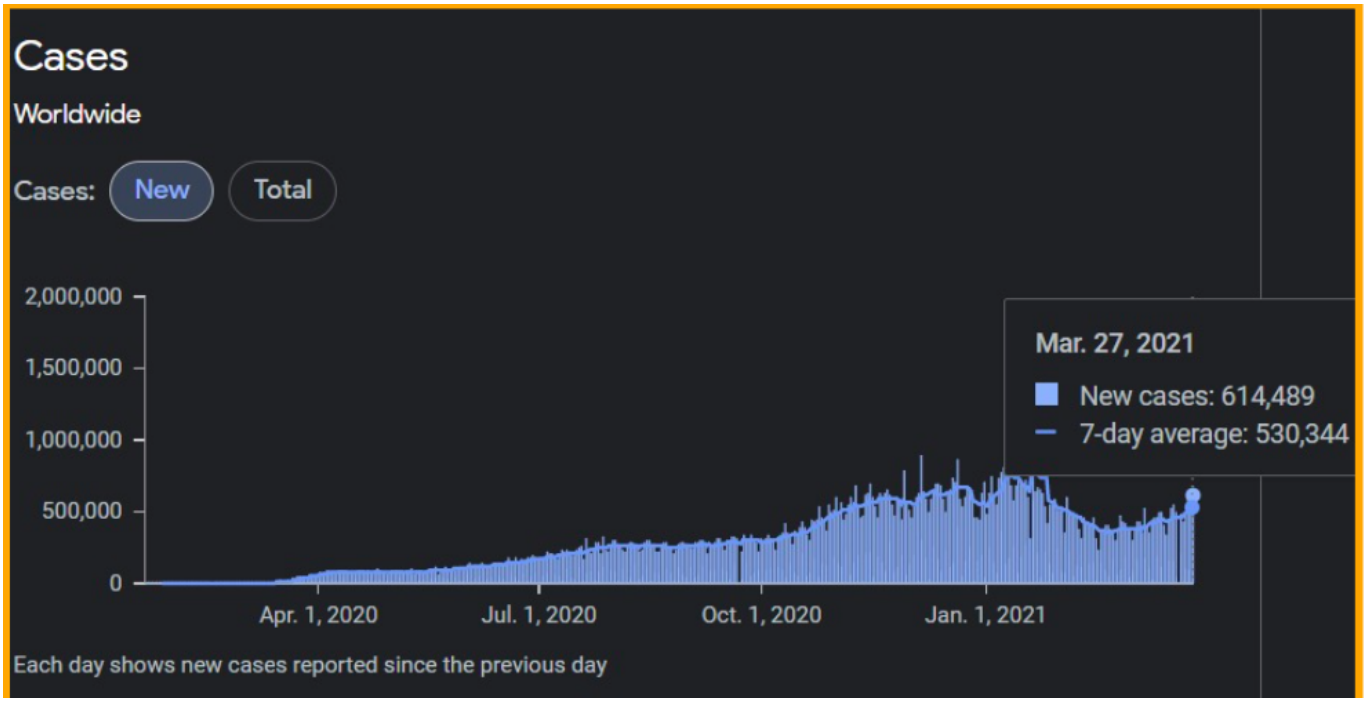

Fig. 2: Worldwide new cases of COVID-19 (As of March 27/2021)

There are few issues with the dataset like firstly when the original CT scans are put into papers the quality of the image is getting degraded due to which the diagnosis decisions are less accurate, secondly the number of bits is getting reduced per pixel, thirdly the CT scan consists of the number of slices and when they are put on the paper the number of slices is getting reduced due to which some important information is getting lost.

Experienced radiologists can make accurate predictions by seeing the images with poor quality but the Machine Learning or Deep Learning models may not find accurately by sending images with poor quality as test images so, to address this issue we used the original CT images for testing the model whereas we used the CT scans that are produced on the paper or clicked by the patient with the phone for training the model. This produces images with different sizes so we have reshaped all the training and testing images to a particular image shape. Researchers [13] has conducted experiments by training the model on the images that are extracted on to the paper and on the original CT images, and the former outperforms the latter so we also used paper extracted images for training the model. For testing the model we explicitly read the captions of the images which are original CT images that are obtained from the hospital for testing the accuracy of the model. To classify whether the input image is diagnosed with COVID-19 or not initially all the Non-Covid images are labeled with 0 and all the COVID-19 images are labeled with 1 before the images are sent for training the model and to test the goodness of the model accuracy score metric is implemented on the model. Since, the output is known i.e whether an image(CT Scan) is diagnosed with COVID-19 virus or not Supervised Learning techniques are implemented and the flow of the process is shown in Figure 3. The accuracy score is directly proportional to the goodness of the model.

Fig. 3: Supervised Learning

II. The COVID-CT Dataset

The dataset consists of 349 chest CT images of patients diagnosed with COVID-19 and 397 chest CT images of patients without COVID-19 [16]. Each image consists of only one CT image suppose if it contains multiple sub figures then it is manually split by previous researchers [13] into individual CT images. The final individual CT images are used for training and evaluating the model. The size of each CT image ranges from 153 to 1853 by height and from 124 to 1485 by width and the images are from 216 patient cases and the male patients are more than the female patients. Thereby, all the images are reshaped to a particular shape for training and evaluating the model. The above images are collected from different datasets and journals like Med Pix, LUNA, Radiopaedia website, PubMed Central (PMC) website etc. Various other datasets that are available in comparision with COVID-CT dataset is given in the Figure 4.

As the number of images for training the model is less Data Augmentation is implemented to generate more images by mentioning various parameters (or) all possibilities for a CT image having COVID-19 virus or not. The process is as follows:
Initially, all the input images are reshaped to a particular shape then the Data Augmentation function is implemented on the input images by mentioning different values to the hyper parameters and also the path where the newly generated images have to be stored and the number of images to be created. Finally the images are created and stored in the mentioned path.

Dataset	COVID images	Patients	Open-Source
COVID-CT	349	216	Y
COVID-19 Image data collection	84	45	Y
SIRM COVID-19 Database	100	60	Y
COVID-CS	68626	400	N
COVID-19 CT Segmentation	379	20	Y

Fig. 4: Comparison of COVID-CT with other datasets

III. Literature Review

One possibility for faster COVID-19 testing is the use of CT computed tomography images [5]. A study pool done by Cochrane noted findings of COVID-19 through chest CT images [6]. This pool found that 90% of the time, the algorithms used were able to correctly find COVID-19. However, 38% of tests that did not have COVID-19 were found to have COVID-19 incorrectly. Perhaps this data set was too small and lack of distinguishable evidence of COVID-19 from other respiratory diseases could be a factor. The concepts of machine vision and deep learning could prove effective in accurately diagnosing real COVID-19 present CT images [7].

Machine learning is computer algorithms associated with improving an application through experience and data input [17]. The flow chart for machine learning typically is an algorithm that takes data (training samples), learns from them, and then makes a prediction on how to act [17]. There are different types of machine learning algorithms that are classified by the learning style (supervised learning, unsupervised learning, or semi-supervised learning) or similarity in function [17]. Each learning algorithm has three components including the representation, the evaluation, and the optimization (Figure 5). Representation is the set of classifiers that the computer is to recognize [17]. Evaluation is the accuracy or error scoring part of the function [17]. Optimization is a search method that is most commonly looking for the highest-scoring classifier [17]. Common representation classifiers include K-nearest neighbor (KNN), support vector machines (SVM), logistic regression (LR), and decision trees (DT).

KNN algorithm is a supervised machine learning algorithm used to solve classification [18]. Supervised algorithms use labeled input data to learn functions that provide an output [18].

The KNN algorithm is therefore a binary classification. It will assume that similar objects are in close proximity (neighbors) to each other (Figure 6) [18].

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Fig. 5: The three main components of machine learning algorithms [17]

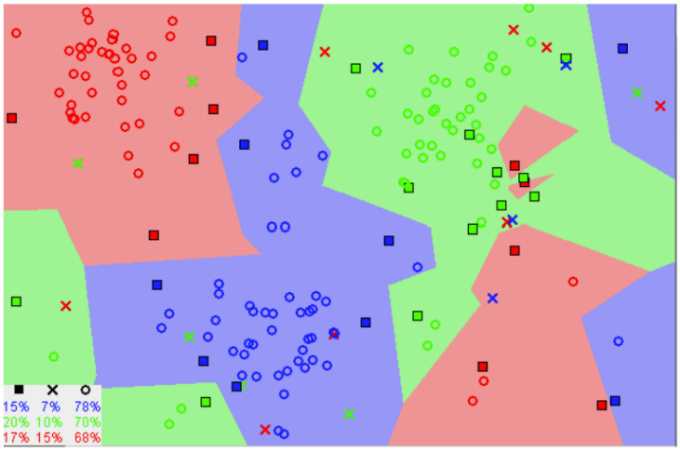

Fig. 6: Depiction of how similar data points will exist in close proximity [18]

With this method data is loaded in, K is initiated to the chosen number of neighbors, and then the distances between the comparison examples are found [18]. The ordered collection of distances are sorted in ascending order, and the labels of the selected K entries are found. The mode of the K labels is found [18].

Support vector machine algorithm takes a multi- dimensional space and has an objective to find a hyperplane that classifies the data points [19].

Although there are many possible hyperplanes, the aim is to find the maximum distance between data points of both classes [19]. Visual representation of hyperplanes in 2D and 3D space is shown in Figure 8.

Logistic Regression is another classification algorithm used to define datasets in a cost-based approach. The logistic approach decision boundary aims to use a cost function and limit it between 0 and 1, where 1 represents the highest probability of likeness [20]. LR is represented by a sigmoid function because it is not linear. A threshold probability of likeness is chosen between 0 and 1 to define the classification (Figure 7) [20].

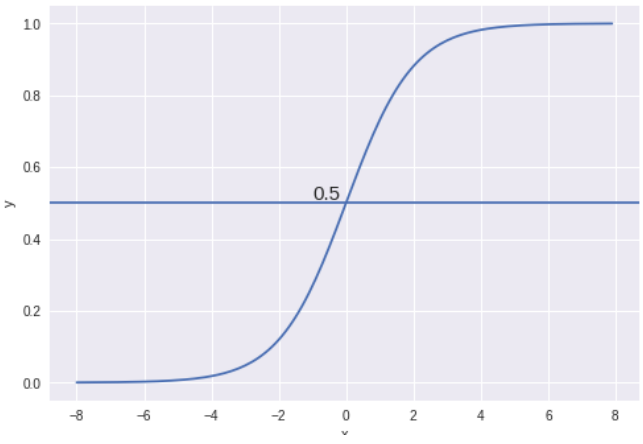

Fig. 7: Example decision boundary of cost function for LR, where the threshold is 0.5 [20]

Decision trees are another supervised learning algorithm.

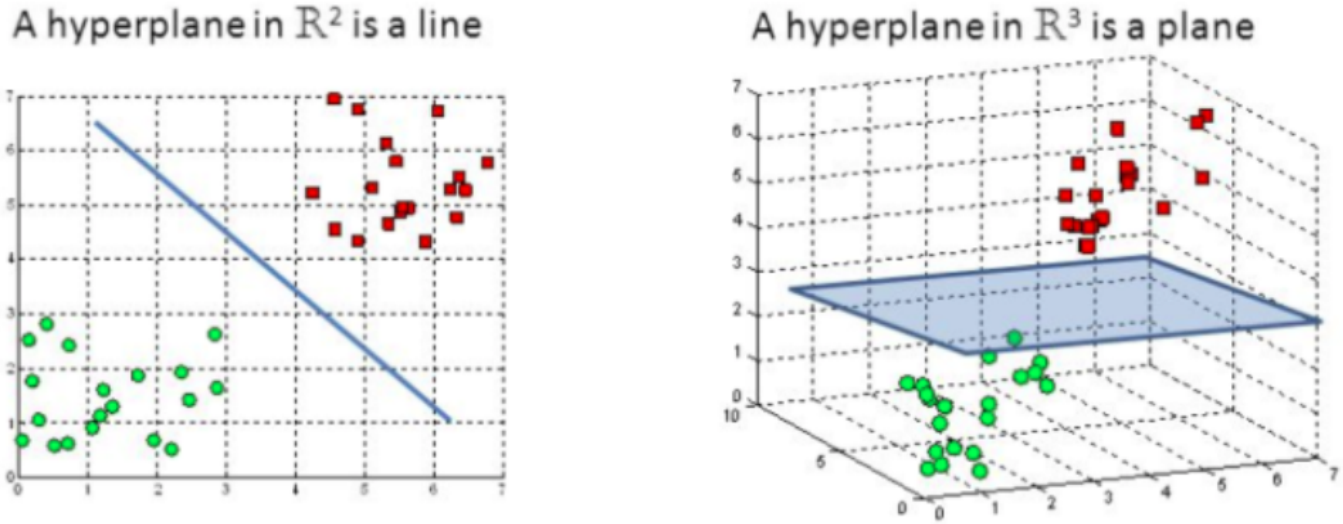

Fig. 8: Hyperplanes in 2D and 3D feature space [19]

Input data are given, and a training output parameter is given for the input to correspond to [21]. The "decision tree" comes into place by decision nodes where the data is split and then the "leaves" which are the final outcomes of each decision. An example decision tree is shown in Figure 9.

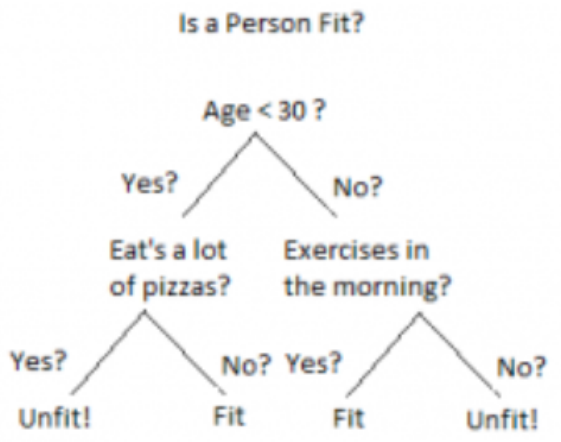

Fig. 9: Example decision tree [21]

Deep learning is an area of machine learning that deals with algorithms related to how the brains in humans works and perceives things (artificial neural networks) [8]. This method allows the computer to learn by example. The model will learn to perform a classification task from an image and train using a large set, enabling a "learn as it goes" approach [8].

One study which considered deep learning on CT images for COVID-19 detection used a voting-based scheme and cross-dataset analysis [7].

They noticed that a problem with regular 4 detection of COVID-19 in CT images is the independent segmentation that is compared with other test images. The data sets become generalized and can create false diagnoses. Their solution was to take patient images and classify it to a certain group in a voting-based system. The results of this attempt however showed a lower accuracy in real COVID-19 test sample CT images. The accuracy when using cross-dataset analysis went from 87.7% to 56%.

Another study using deep learning applied a two- step method in the use of CT images for diagnosis of COVID 19 to improve the clinical performance [9]. They used 467 positive COVID-19 training images [9]. To train deep learning models given the small information they had, they had two methods. One was to have additional segmentation masks of lung regions and segmentation masks of lesion regions (Figure 10) which would pay more attention to lung regions with manifestation of COVID-19 [9]. The second method was to deep learn through better visual representations with contrastive self-supervised learning (CSSL) [9]. This CSSL method creates augmented CT examples and then learns the visual representation [9]. Out of the two methods the best was a combination transfer learning plus CSSL method which used lung masks and lesion masks through multi-task learning and then CSSL to improve learned depictions [9]. The overall accuracy of this method was 89% [9].

The attempts made to accurately detect COVID-19 in CT images have been promising but require tweaking to increase accuracy and viability of the method. It has been proposed that the deep learning concept of convolutional neural networks could increase the accuracy [10]. Convolutional neural networks (CNN) are a deep learning algorithm that analyzes images and assigns importance to aspects/objects in the image to be able to differentiate them [10].

Each neuron's behavior is defined by a weight. Given the pixel values from the image, the "artificial" neurons of the convolutional neural network will discern and recognize certain characteristics [11]. Figure 11 shows a visual of how the transfer function for CNN works including the initial inputs (neurons), the weights, the transfer function algorithm of weighted sums, and the outcome "activation map". Convolutional neural networks have been a promising algorithm for detection of COVID-19 in CT scans of lungs [9].

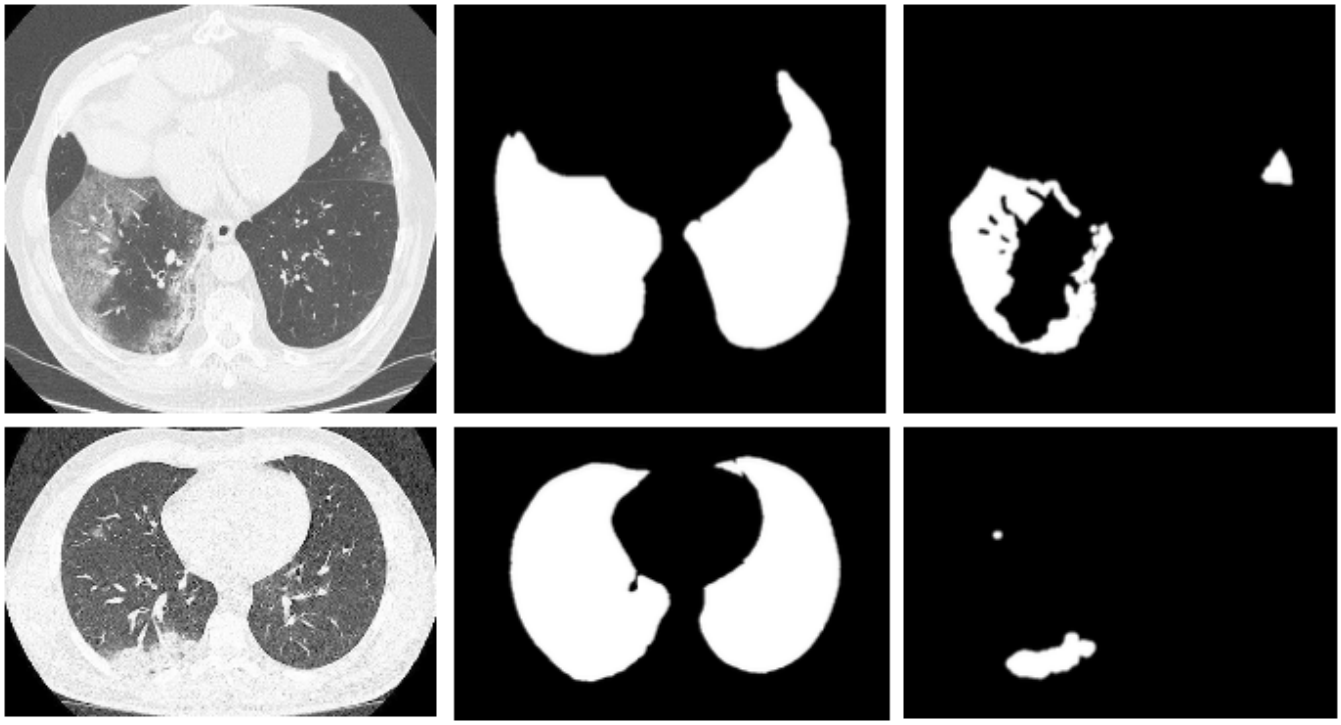

Fig. 10: (a) First column: CT images of positive COVID training set, (b) Second column: segmentation masks of lung regions, (c) Third column: segmentation masks of lesion regions

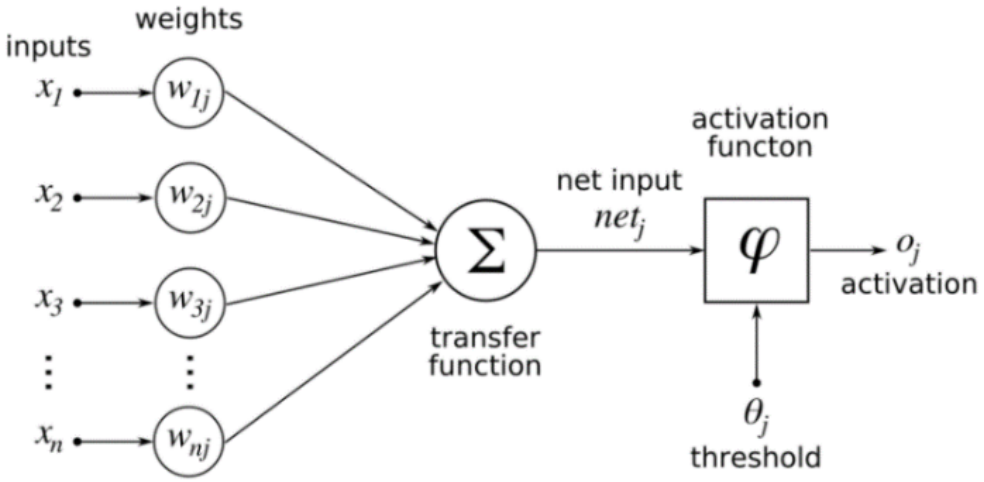

Fig. 11: Convolutional neural network transfer function [12]

A study published by Xu et al. used the deep learning multiple models of convolutional neural networks to test CT scans of lungs for COVID-19. A total of 618 CT samples were collected and 528 samples were used in training and validation sets [12]. Only 189 of the samples were COVID-19 patients while others consisted of pneumonia or healthy people's lung CT scans. The remaining 90 sets were used as the test set (30 COVID-19, 30 pneumonia, and 30 healthy) [12]. Three different processes were used on the images. The first step was a pre-processing to extract effected pulmonary regions [12]. The second step was a 3D convolutional neural network model to segment image cubes [12]. The last step was to classify the images into one of three categories [12]. The three categories were COVID-19, pneumonia, and no infection. The analysis report for each sample was calculated using a function called Noisy-or Bayesian [12]. The process used in this study is shown in additional details in the flow chart in Figure 12. The overall accuracy of the methods used to distinguish COVID-19 positive CT scans from pneumonia scans was 86.7% which is very promising [12]. Overall, CT scans are proving to be a valuable method of diagnosing COVID-19 as CT scans can detect earlier and faster presence of COVID-19 than sputum samples [12].

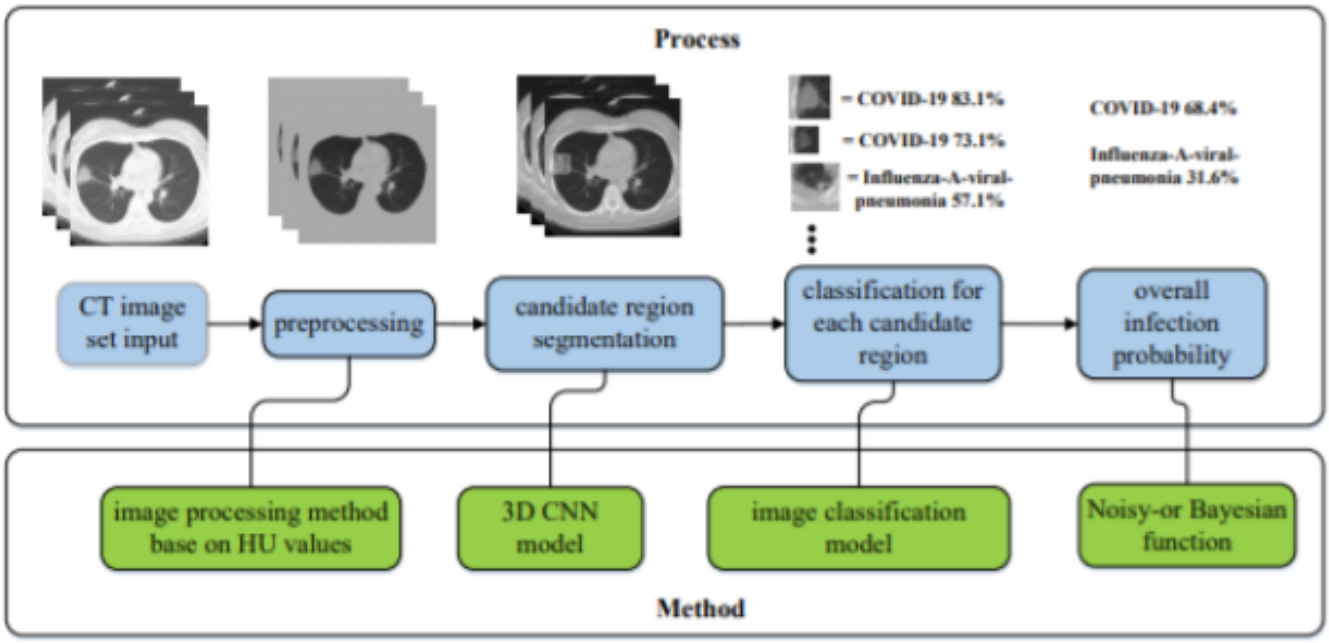

Fig. 12: Process flow chart of CNN process used in study by Xu et al. [12]

IV. Proposed Methods

The classification of a CT image can be done using various methodologies that are available in Deep Learning(Figure 13) and Machine Learning Algorithms (Figure 14). Some of the popular Machine Learning methods like K-Nearest Neighbors, Support Vector Machines, Logistic Regression and Decision Trees are implemented for doing binary classification to classify whether a CT scan is diagnosed with COVID-19 or not. Initially we implemented ML models to do binary classification but did not produce high accuracy score and also the Decision Tree failed to execute in the allowed execution time. so, later we implemented CNN model i.e a Deep Neural Network with 7 Layers consisting of 2 convolutional layers with 3x3 filters, 2 max pooling layers with pool size of 2x2 to extract the features and lesion that are present in the image, 1 flatten layer to flatten the size of the input image to a 1D array, 1 dense layer with 64 neurons and finally the output layer with 1 neuron. Initially intensity values of all the input images are normalized by dividing each intensity value with 255.0 since all the input images are of gray scale images before sending it to the model for faster computation. Since this is a classification algorithm we used sigmoid activation function at the output layer to classify the input image based on the probability. For the input layers we used relu activation function and to fit the model we used adam optimizer for optimizing the values, binary cross entropy loss function is used since the problem is of type binary classification and finally the accuracy metric is used to calculate the goodness of the model. The accuracy score for the training data is around 80% but for the test data it is just around 60-65% when the image size is 480x480 but when the image size is reduced to 224x224 the accuracy score has increased to 68%. This resembles that the model got overfitted because the difference between the accuracy score on training data and testing data is somewhat high so, Dropout Layer, different weight initialisation techniques and Hyper parameter tuning

can help to achieve a high accuracy score for the test data also.

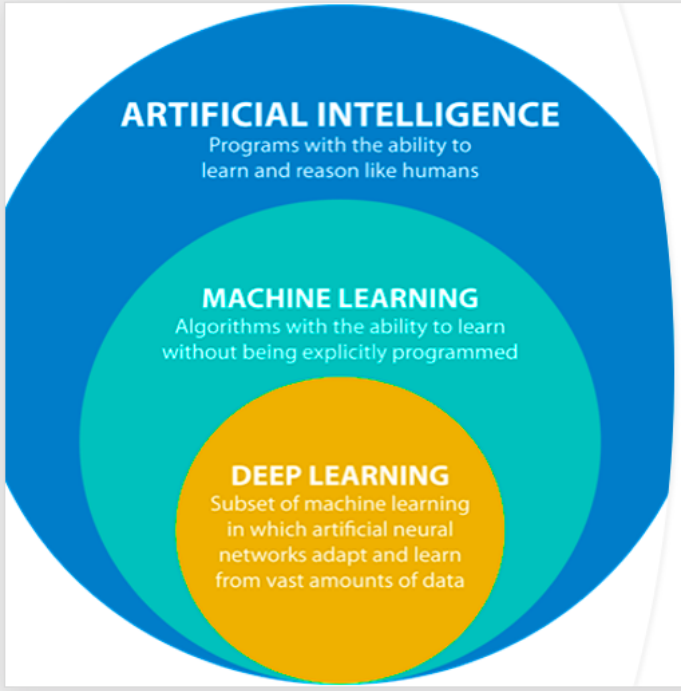

Fig. 13: Deep Learning vs Machine Learning

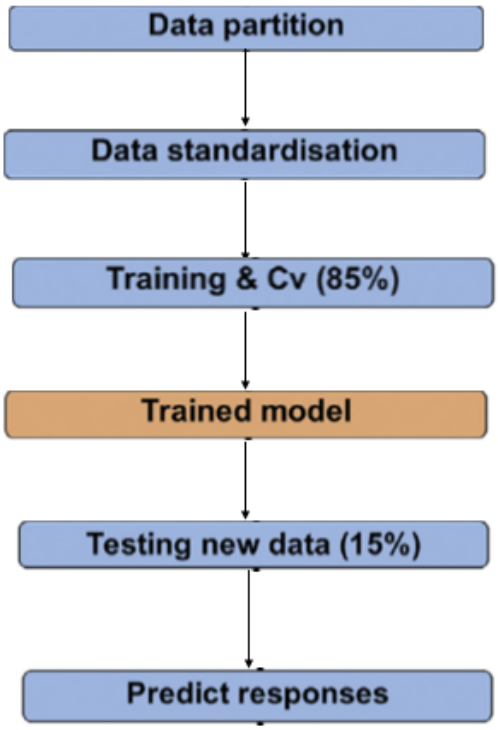

Fig. 14: Workflow of a Supervised Learning algorithm

The accuracy score of the model can be further increased by implementing the following method:
Initially, we have to segment the lung part alone from the image by defining a mask so that our model won't concentrate or train much on the background of an image then we have to segment the lesion regions that are present inside a lung by defining a lesion mask and these can be achieved by using algorithms that are available for segmentation and finally the cropped images can be used to train our models and do classification with the help of pre-trained models like DenseNet-169 or ResNet-50 algorithms [13]. Implementing Transfer learning can help in achieving better accuracy score as the weights are already optimised and the model has been trained for different images.

V. Discussion

The number of CT-Covid and CT-Non Covid images available publicly were only 349 and 397. There are so many images that are available online in various database repositories but due to so many restrictions imposed by the hospitals we were unable to retrieve the images. So, to train the CNN or Machine Learning model we used Data Augmentation to generate all kind of possible images from a pre-defined training set. This helps the model to not overfit and to produce high accuracy score by training under all possible scenarios. Initially we used Machine Learning models like K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Logistic Regression, Decision Trees etc but these models failed to produce high accuracy even after performing Hyper parameter tuning. The KNN model produced a highest accuracy score of 64% when k=7 and distance metric is Manhattan. The Logistic Regression achieved a highest accuracy score of 60% when we set the hyper-parameter number of iterations to 1000. The SVM algorithm achieved a accuracy score of 63% and finally the Decision Trees failed to execute in the allowed time because the size of each input image is 480x480, there are nearly 2000 images so the model failed to create decision trees in the allowed time complexity.

Finally we have implemented the CNN model by loading all the original and generated images, then all the images are re- scaled to same shape because each image has different size and then the images are appended with the labels(0 for Non-Covid and 1 for Covid). After labelling the images we have split and shuffled the images for training(80%) and testing(20%) the CNN model.

VI. Results

The results for the above algorithms are as follows:

Algorithm	Accuracy Score(%)
KNN(K=3, distance="euclidean")	60%
KNN(K=5, distance="euclidean")	60%
KNN(K=5, distance="manhattan")	61%
KNN(K=7, distance="manhattan")	64%
KNN(K=7, distance="cityblock")	64%
KNN(K=9, distance="manhattan")	62%
LR(iterations=100)	57%
LR(iterations=1000)	60%
SVM	63%
CNN(Image size: 480x480)	60-65%
CNN(Image size: 224x224)	68%

Fig. 15: Results

VII. Conclusion

This article has presented different methods that are available in Supervised Learning for doing binary classification whether a patient is diagnosed with COVID-19 or not based on the CT scan of the patient and also discussed in detail the concepts that are available in classification. In the near future, the goal is to increase the accuracy score of the model by implementing Transfer Learning, Segmentation algorithms and different methods as discussed above on the input images.

References

[1] L.M.Sauer, "What Is Coronavirus?", 2021.[Online]. Available:*https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus*. [Accessed: 25-Mar- 2021].

[2] WHO.int, 'Transmission of SARS-CoV-2: implications for infection prevention precautions', 2020. [Online]. Available: *https://www.who.int/news-room/commentaries/detail/transmissionof- sars-cov-2-implications-for-infection-prevention-precautions*. [Accessed: 25-Mar-2021].

[3] Google News, 'Coronavirus (COVID-19)' 2021. [Online]. Available: *https://news.google.com/covid19/map?hl=en-CA&gl=CA&ceid=CA%3Aen*. [Accessed: 25- Mar-2021].

[4] FDA, 'Coronavirus Disease 2019 Testing Basics', 2020. [Online]. Available: *https://www.fda.gov/consumers/consumer-updates/coronavirus-disease-2019-testing-basics*. [Accessed: 25-Mar-2021].

[5] Grand Challenge, 'CT diagnosis of COVID-19', 2020. [Online]. Available: *https://covidct.grand-challenge.org/*. [Accessed: 25-Mar-2021].

[6] N. Islam, J.P. Salameh, M. M. G. Leeflang, L. Hooft, T. A. McGrath, C. B. van der Pol, R. A. Frank, S. Kazi, R. Prager, S.S. Hare, C. Dennie, R. Spijker, J.J. Deeks, J. Dinnes, K. Jenniskens, D.A. Korevaar, J.F. Cohen, A. Van den Bruel, Y. Takwoingi, J. van de Wijgert, J. Wang, and M.D.F. McInnes, "How accurate is chest imaging for diagnosing COVID-19," Cochrane, Nov. 2020.

[7] P. Silva, E. Luz, G. Silva, G. Moreira, R. Silva, D. Lucio, and D. Menotti, 'COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis,' Inform Med, Sep. 2020.

[8] Mathworks.com, 'Deep-Learning', n.d. [Online]. Available: *https://www.mathworks.com/discovery/deep-learning.html*. [Accessed: 25-Mar-2021].

[9] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-Dataset: A CT Image Dataset about COVID-19," Medrxiv, Mar. 2020.

[10] S. Saha, 'A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way', 2021. [Online]. Available: *https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53*

[11] B. Dickson, 'What are convolutional neural networks (CNN)?', 2020. [Online]. Available: *https://bdtechtalks.com/2020/01/06/convolutional-neural-networks-cnn-convnets/*

[12] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Y. Chen, J. Su, G. Lang, Y. Li, H. Zhao, K. Xu, L. Ruan, W. Wu, 'Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia,' arXiv, Feb. 2020.

[13] He, Xuehai and Yang, Xingyi and Zhang, Shanghang, and Zhao, Jinyu and Zhang, Yichen and Xing, Eric, and Xie, Pengtao, Sample-Efficient Deep Learning for COVID19 Diagnosis Based on CT Scans, medrxiv, 2020.

[14] Lu Huang, Rui Han, Tao Ai, Pengxin Yu, Han Kang, Qian Tao, and Liming Xia. Serial quantitative chest ct assessment of covid-19: deep-learning approach. Radiology: Cardio- thoracic Imaging, 2(2):e200075, 2020.

[15] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zheng- han Fang, Qi Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology, page 200905, 2020.

[16] Zhao, Jinyu and Zhang, Yichen and He, Xuehai and Xie, Pengtao, a CT scan dataset about COVID-19, arXiv preprint arXiv:2003.13865, 2020.

[17] Retrieved from *https://emerj.com/ai-glossary-terms/what-is-machine-learning/*

[18] Retrieved from *https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761*

[19] Retrieved from *https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47*

[20] Retrieved from *https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148#: :text=Logistic%20Regression%20is %20a%20Machine,on%20the%20concept%20of%20probability.&text=The%20hypothesis%20of%20logistic%20regression,function%20between%200%20and%201%20.*

[21] Retrieved from *https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html#: :text=Introduction%20Decision%20Trees%20are%20a,namely%20decision%20nodes%20and%20leaves*.