Data Analytics for Business

DAB 402 Capstone Final Report
August 2022

Project: To provide an alternative testing method for prediction of COVID-19 and other lung diseases using lung ultrasound imagery with Deep-learning and Machine Learning based detection models.

Professor: Umair Durrani

Group: II

Abhinav Reddy Moddu – 0787883
Abhiroop Juttu – 0781192
Noorjahan Shaik – 0775162
Srinath Toparapu - 0786927
Twinkle Pemmasani – 0782001

GitHub Repository Link: https://github.com/NSHAIK2/DAB402---Capstone-
Dataset Link: https://github.com/nrc-cnrc/COVID-US

Content

1. Abstract ----------03
2. Introduction ----------03
3. Motivation of LUS utilization for COVID-19 screening ---------03
4. Solving the Problem ----------04
5. Research Question ----------04
6. Dataset Description ----------04
7. Metadata Study ----------05
8. Data Assessment and Processing ----------06
9. Limitations ----------07
10. Method of Analysis ----------07
11. Modelling ----------07
12. Results ----------08
13. Discussion ----------13
14. Conclusions ----------13
15. Recommendations ----------14
16. References ----------14

1. Abstract:

The COVID-19 pandemic has exposed the vulnerability of healthcare services worldwide, especially in underdeveloped countries. Clinical progression of COVID-19 is variable and can range from patients remaining asymptomatic and able to recover at home to patients developing severe respiratory failure requiring prolonged hospitalization and intensive care [1]. The main purpose of this work is to investigate and compare both deep learning and supervised machine learning techniques applied to Lung Ultrasound images for the detection of COVID-19. The opensource dataset from GitHub was the COVIDx-US dataset compiled from several sources like GrepMed, Radiopaedia, Clarius and other sources. We utilized the most recent version which includes 242 videos. We have retrieved 18628 processed images of patients with COVID-19 infection, Pneumonia, other fungal infections with lung disorders, normal healthy persons lung ultrasound recordings. We have built & trained different Machine Learning Models which includes Random Forest, KNN, Naïve Bayes, VGG16. The key outcome for this project is to identify given image has any type of category variant such as Covid, Pneumonia, other lung disease or if the case is normal. We have developed a web application that accepts random images as input and processes them to determine if they are covid, pneumonia, or normal using the prediction interface[2].

2. Introduction:

In December 2019, a novel coronavirus, named SARS-CoV-2, emerged in Wuhan, China, which caused the COVID-19 disease when infecting humans. COVID-19 is a serious illness that can lead to the death of the infected host. The threat posed by COVID-19 led the World Health Organization (WHO) to declare the COVID-19 pandemic by March 2020.

Coronaviruses are a group of highly diverse, enveloped, positive-sense, single-stranded ribonucleic acid (RNA) viruses and are widely spread in birds and mammals. Sometimes these viruses infect humans, causing mild to moderate respiratory diseases. Before SARS-CoV-2, two coronaviruses were known to cause severe human disease. SARS-CoV, which causes severe acute respiratory syndrome (SARS); and MERS-CoV, which causes Middle East Respiratory Syndrome (MERS). However, in contrast to SARS and MERS, the symptom onset for COVID-19 is significantly larger, or it may appear in a mild form, allowing infection spread by asymptomatic patients, which in turn has led to the current pandemic. Although the WHO has emphasized the need for massive testing and contact tracing to better tackle the pandemic, not all countries have the required laboratory infrastructure and reagents to effectively address this task. Additionally, getting results from some of these tests may take a couple of days, leading to non-confirmed COVID-19 patients with mild or no symptoms to further spread the disease while waiting for the test results.

3. Motivation of LUS utilization for COVID-19 screening [3]:

In the last few years, Lung Ultrasound (LUS) imaging has been proposed as an alternative to the use of CT or X-ray for screening and follow-up of lung diseases. For instance, it has been suggested that lung visualization through ultrasound imaging effectively replaces physical auscultation with stethoscopes. Moreover, when used correctly, LUS imaging could even help to reduce infections between patients and medical staff.

Recent medical correspondence has pointed out the advantages of using LUS imaging as a tool for early diagnosis and follow-up of COVID-19 patients . Some works highlight the benefits of using LUS in

the context of the COVID-19 pandemic, especially considering its portability, accessibility, no radiation, ease of disinfection (e.g., using disposable caps of the ultrasound probes), and low cost.

A short review of LUS findings in COVID-19 patients is presented in [36]. Some of these findings are consistent with CT results, including multiple fused bilateral B lines, subpleural pulmonary consolidations, irregular pleural line, and poor blood flow. An important finding reported by Fiala [36] is that subpleural lesions in COVID-19 patients differ significantly from pulmonary diseases, including bacterial pneumonia, tuberculosis, and cardiogenic pulmonary edema, among others. Based on these observations, the author suggests adopting lung ultrasound for early detection of pulmonary alterations as a triage tool, particularly in environments with limited resources. Additionally, other authors reported that LUS imaging findings of COVID-19 progression in diagnosed patients are related to the observed patterns on CT images. Particularly, ground-glass opacity, consolidation shadow, and thickened pleura observed in CT images had special manifestations in LUS images as B-lines, consolidations, irregular or fragmented pleural line, pleural effusion, and absence of lung sliding. Furthermore, Soldati et al. Suggested that LUS findings in superficial pulmonary tissue are correlated with histopathological findings revealed in CT scanning.

Based on the previous discussion, it can be asserted that LUS imaging is a promising option for the screening, diagnosis, and follow-up of pulmonary diseases. Thus, over the past decade, interest in developing computational tools for computer-assisted analysis of LUS imaging has increased [24–30]. A brief description of the results reported in these works will be provided in the next section. We will first review works related to the identification of general lung conditions by means of computer-assisted analysis of LUS imaging. We will then describe previous work on computer-assisted screening of COVID-19

4. Solving the Problem:

An alternative proposal for diagnosing covid-19 patient with high in accuracy, precision, recall, F1-scores, and ROC can be developed using machine learning for classification of patient medical data. Detecting particles of virus patterns from images derived from the video scripts with the help of Convolutional neural network and pre-trained models to attain high accuracy and low error with least latency is the primary task.

5. Research Question:

i. To evaluate and compare the performance of deep-learning techniques vs Supervised Learning Methods for detecting COVID-19, pneumonia, Other fungal infections from lung ultrasound imagery.
ii. Which machine learning classifier can identify COVID-19 and other categories accurately using lung ultrasound images with the highest accuracy, F1-score, and Confusion Matrix?

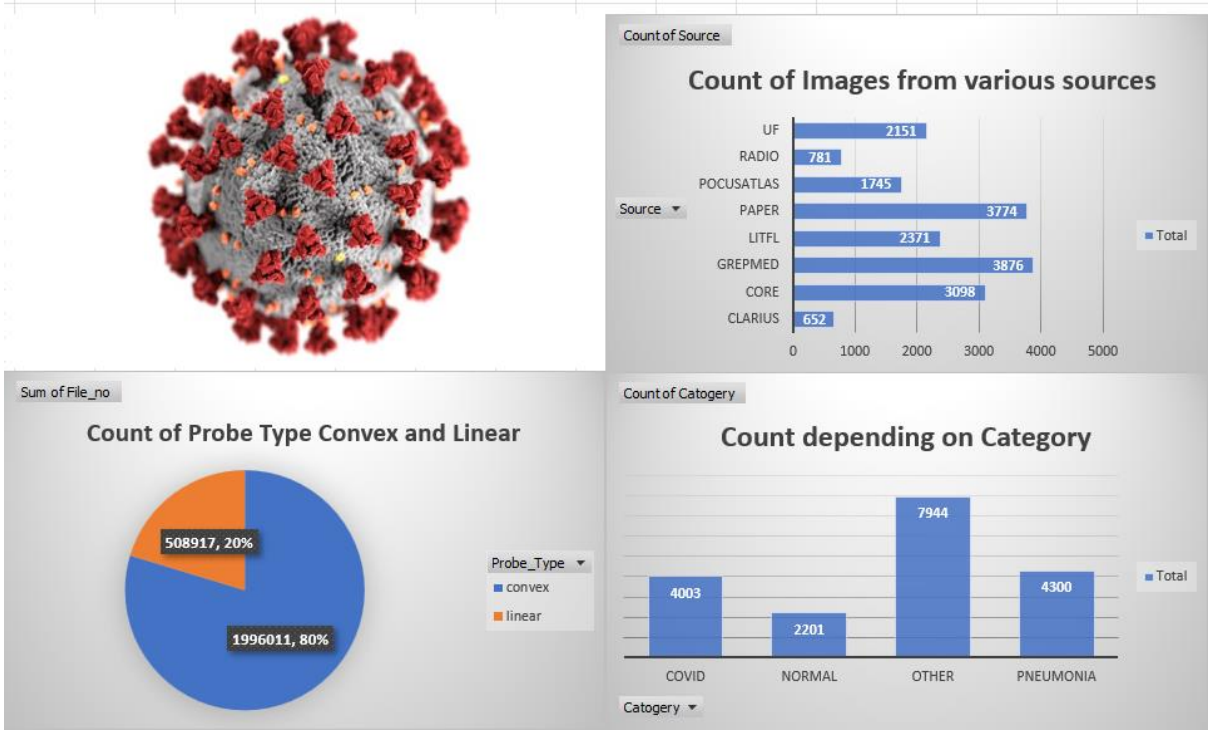
6. Dataset Description:

The dataset used for the project is the LUS dataset available in GitHub. (https://github.com/nrc-cnrc/COVID-US). This is an open-access dataset that was consolidated by Digital Technologies Research Centre of University of Waterloo and McGill University. The dataset was posted in LinkedIn through the National Research Council Canada Data. We downloaded the data following the scripts provided by NRC-CNRC (National Research Council Canada, 2021) [4]. The 188 videos (18448 images) recorded are with either convex or linear probes from a total of 188 patients. All the images will have labels in the particular manner following information the source, file_no, category,

probe type, Frame number. The images are categorised into linear and convex type. Here category means the images related to covid, Pneumonia, Normal or others.

7. Metadata Study:

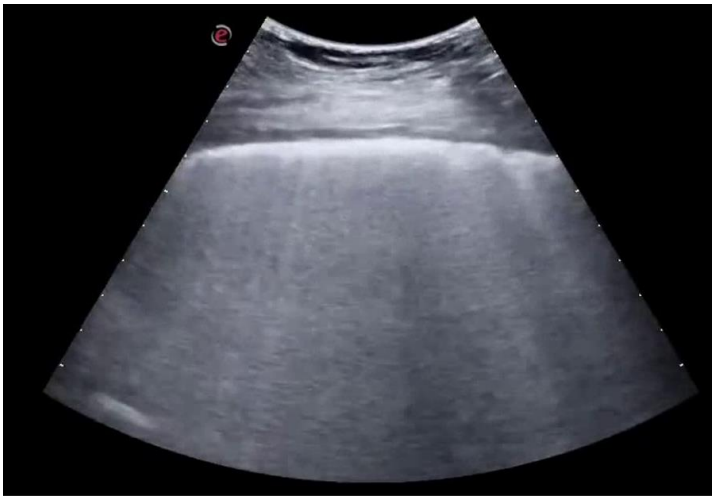
The number of lung ultrasound videos, video's data source, image classification ,ultrasound probe used, frames generated per video were analysed. LUS images were taken from the GrepMed, LITFL, The PocusAtlas, Radiopaedia, CoreUltrasound, University of Florida, Clarius and Paper sources. Probes used to capture lung ultrasound videos were either convex or linear, with images saved in JPEG format and in RGB mode. Variation in dimensions were addressed by resizing images to uniform dimensions. Finally, there are four (4) image classes available (i.e., COVID-19 (4,003 images), pneumonia (4,300 images), normal (2,201 images), and other (7,944 images)). A total of 18,448 LUS frames were used.



By analysis of metadata found on GitHub repository we can understand that GrepMed is major source of contribution and majority of images lie in other category while 80% of images received are convex.

The Probe_Type Variable has two types of images as Convex and Linear.

1. Convex have a curved array that allows for a wider field of view at a lower frequency



2. Linear uses high frequency ultrasound to create high resolution images of structures near the body surface

8. Data Assessment and Processing:

The Labels of images are used to split into different categories. Removing the noise from images is removing unwanted pixels. In the original data, the average image size is about 500,800.However after cropping it is reduced to 500,500. Therefore approximately 300*500 pixels are cropped out to reduce the time taken and resources needed for processing. We loaded the images into NumPy to optimize the memory allocation.

The following steps are followed for processing:
- Converting labels to ordinal - Initially data labels are in categorical datatypes, while training the model for better performance labels are converted to ordinal.
- Data Normalization - When taking the raw data, there can be lot of bias and variance. So, images are normalized.

- Splitting the data into train and test sets with 80% and 20% respectively. As we have imbalanced dataset while splitting the data, we have used stratify method for best performance.
- Training the model with processed images and respective labels to classify the class labels appropriately.
- After training the model, observing the performance of each model with accuracy, f1-score, and confusion matrix. And, we have checked if the models are overfitting.

9. Limitations:

Low number of training images per class, hence we have stratified the data to avoid the bias.
Low training epochs for vgg-16 model.
Limited hyper parameter search space for all the models as the processing time is high as we have low resources.

10. Method of Analysis:

Deep learning (DL) has proved successful in medical imaging, and, in the wake of the recent COVID-19 pandemic, some works have started to investigate DL-based solutions for the assisted diagnosis of lung diseases[**Error! Reference source not found.**]. Leveraging these data, we introduce deep models and supervised learning models that address relevant tasks for the automatic analysis of LUS images. We have trained the model with 80% of images for VGG16, Random Forest, KNN and Naïve Bayes models. Testing the model with 20% of images and validating the output with classification report for precision, accuracy, f1 score the best performing model will be suggested for all the four categories. The confusion matrix will be used to understand the for which class the model is confused.

11. Modelling:

VGG16 Modelling: We have used a pretrained model VGG16 imagenet weights which are not trainable. So, using this layer we extracted the most important features from the input images. Thereafter we included a flatten layer to convert a 2-dimensional convolutional image to 1 dimensional vector.

This in turn is sent a dense multi-layer perceptron (mlp) as it is very good classifier. For this classifier we used 3 layers
Layer 1 - dense layer with 32 neurons with Relu as an activation function.
Layer 2 - dense layer with 20 neurons with Relu as an activation function.
And finally, the prediction layer - with 4 neurons as there are 4 class labels and activation function as SoftMax to give the probabilistic prediction of each class[5].
Then this model is compiled using Adam optimizer and we used sparse categorical cross entropy to reduce loss at each epoch.

For training the images we used 2000 data points at each instance with batch size of 10 for 5 epochs . As we have low computational resources, we needed to execute the code without internal error. Then after training for all the images in the similar way we used batches of 10 images to compute training and testing error and performance of the model.

Random Forest Modelling: Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes.

A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by most of the decision trees becomes the final output of the rain forest system. The random forest classifier definition is used for training the 80% of data and tested the 20% of data for all the four classes covid, normal, other and pneumonia.

KNN Modelling: The abbreviation KNN stands for "K-Nearest Neighbour". It is a supervised machine learning algorithm. The number of nearest neighbours to a new unknown variable that must be predicted or classified is denoted by the symbol 'K'.

The KNN Algorithm Load the data and Initialize K to your chosen number of neighbours. For each example in the data, it Calculate the distance between the query example and the current example from the data. Then adding the distance and the index of the example to an ordered collection. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances. Pick the first K entries from the sorted collection and get the labels of the selected K entries. Return the mode of the K labels

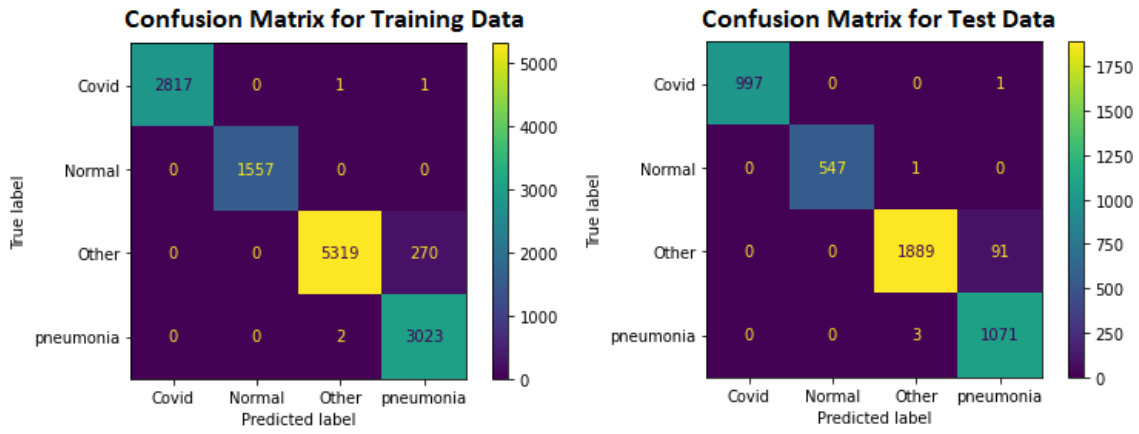
To select the K that's right for data, we run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm's ability to accurately make predictions when it's given data it hasn't seen before.

Naïve Bayes Modelling: Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. we have used the GaussianNB classifier to fit it to the training dataset. We can also use other classifiers as per our requirement. Convert the given dataset into frequency tables. Generate Likelihood table by finding the probabilities of given features.

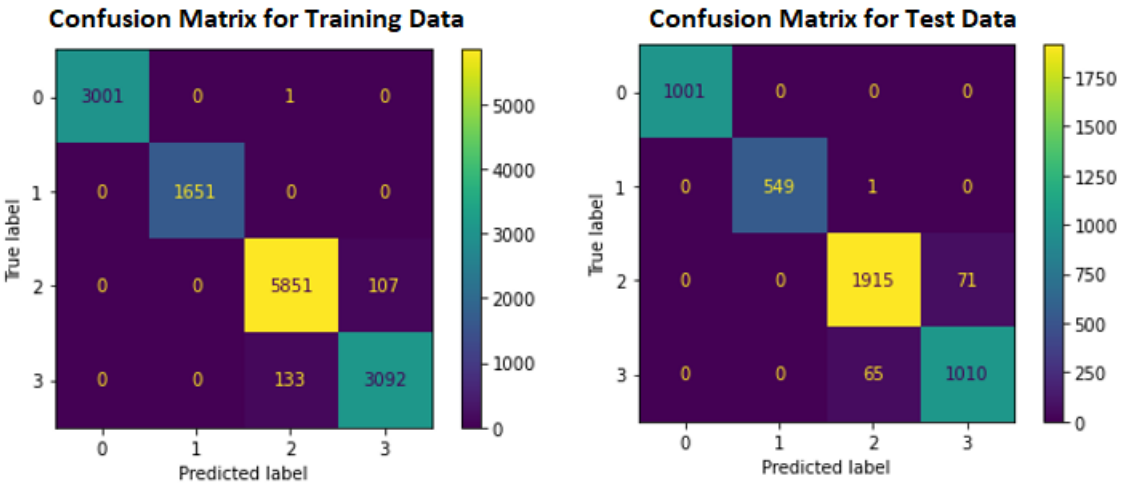
12.Results:

A. Confusion Matrix*:* A confusion matrix is a matrix (table) that can be used to measure the performance of a machine learning algorithm, usually a supervised learning one. Each row of the confusion matrix represents the instances of an actual class, and each column represents the instances of a predicted class.

Vgg16 Model Confusion Matrix: By seeing the below image we understand that the model for test data is confused for covid as pneumonia once and correctly predicted the covid for 997 images. For Normal only once it is confused as Other and correctly predicted 547 images. For the class Other the model is confused 91 images as Pneumonia and correctly predicted 1889 images. And for pneumonia model is confused 3 times as other and correctly predicted images are 1071.

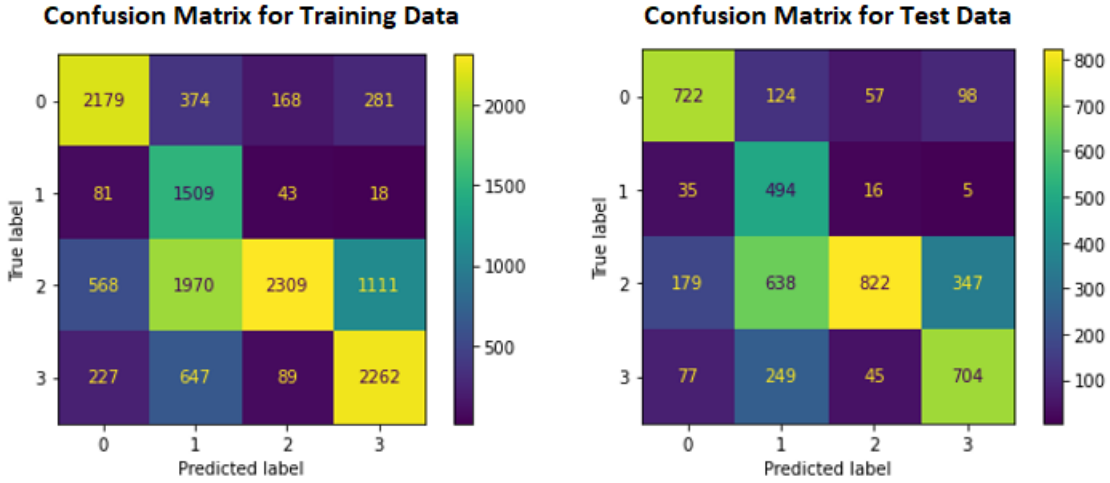


KNN Model Confusion Matrix: By seeing the below image we understand that the model for test data correctly predicted covid for 1001 images. For Normal images for once model is confused as Other and correctly predicted 547 images. For the class Other the model is confused 71 images as Pneumonia and correctly predicted 1915 images. And for pneumonia the model is confused 65 images as other and correctly predicted images are 1010. So, we understand that for covid and other class prediction KNN has best performance.



Random Forest Model Confusion Matrix: By seeing the below image we understand that the model for test data correctly predicted covid for 1001 images. For Normal images for once model is confused as Other and correctly predicted 549 images. For the class Other the model is confused 85 images as Pneumonia and correctly predicted 1901 images. And for pneumonia the model is confused 75 images as other and correctly predicted images are 1000. So, we understand that for covid and other class prediction Random Forest has best performance.

Naïve Bayes Model Confusion Matrix: By seeing the below image we understand that the model the model is not performing well for any class of images.



B. Classification Report:

It is one of the performance evaluation metrics of a classification-based machine learning model. It displays the model's precision, recall, F1 score and support. It provides a better understanding of the overall performance of the trained and test model [7].

Metrics	Definition
Precision	Precision is defined as the ratio of true positives to the sum of true and false positives. Precision = True Positive / (True Positive + False Positive)
Recall	Recall is defined as the ratio of true positives to the sum of true positives and false negatives. Recall = True Positive / (True Positive + False Negative)
F1 Score	The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is. Recall = True Positive / (True Positive + False Negative)

Support	Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models; it just diagnoses the performance evaluation process.

The Classification of each category for all four models are displayed below. So as per the analysis we performed shows that VGG-16 model is best performing model.

VGG-16 Classification Report:

Category	Accuracy		Precision		Recall		F1 Score	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Covid	98%	98%	1.00	1.00	1.00	1.00	1.00	1.00
Normal			1.00	1.00	1.00	1.00	1.00	1.00
Other			1.00	1.00	0.95	0.95	0.97	0.98
Pneumonia			0.92	0.92	1.00	1.00	0.96	0.96

KNN Model Classification Report:

Category	Accuracy		Precision		Recall		F1 Score	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Covid	98%	97%	1.00	1.00	1.00	1.00	1.00	1.00
Normal			1.00	1.00	1.00	1.00	1.00	1.00
Other			0.98	0.97	0.98	0.96	0.98	0.97
Pneumonia			0.97	0.93	0.96	0.94	0.96	0.94

Random Forest Model Classification report:

Category	Accuracy		Precision		Recall		F1 Score	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Covid	98%	97%	1.00	1.00	1.00	1.00	1.00	1.00
Normal			1.00	1.00	1.00	1.00	1.00	1.00
Other			0.98	0.96	0.98	0.96	0.98	0.96
Pneumonia			0.96	0.92	0.97	0.93	0.97	0.93

Naïve Bayes Model Classification report:

Category	Accuracy		Precision		Recall		F1 Score	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Covid	60%	59%	0.71	0.71	0.73	0.72	0.72	0.72
Normal			0.34	0.33	0.91	0.90	0.49	0.48
Other			0.89	0.87	0.39	0.41	0.54	0.56
Pneumonia			0.62	0.61	0.7	0.65	0.66	0.63

C. ROC Curve Classification: An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives [8].

This curve plots two parameters:

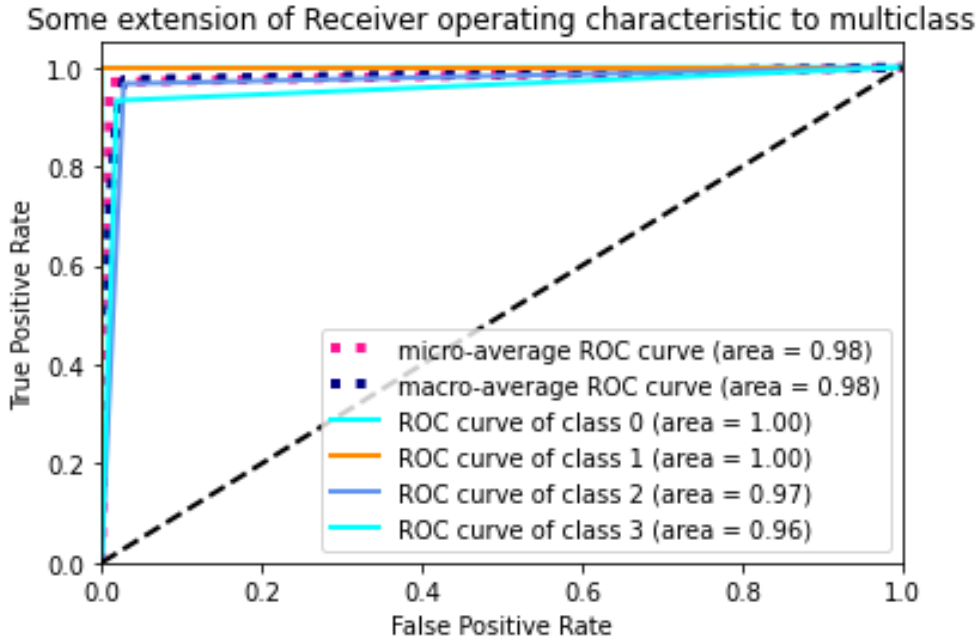
i. True Positive Rate

ii. False Positive Rate

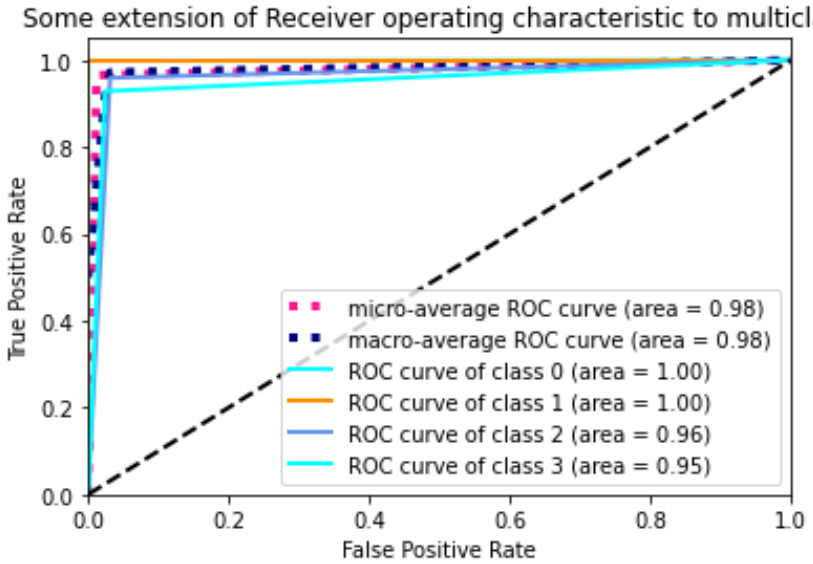
True Positive Rate (TPR) is a synonym for recall and is therefore defined as:
TPR = True Positive (TP) / [(True Positive (TP) + False Negative (FN)]

False Positive Rate (FPR) is defined as follows:
FPR = False Positive (FP) / [False Positive (FP) + True Negative (TN)]

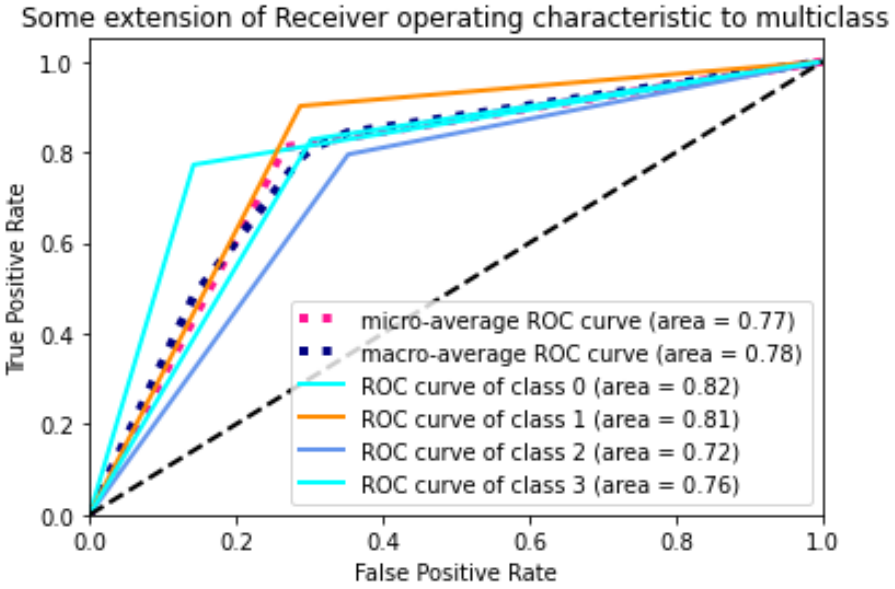
ROC for KNN Model: The ROC score for covid and normal class is at 100% , for other class at 97% and for pneumonia at 96%. As per supervised learning models the KNN is best predicting models for all the classes.



ROC for Random Forest Model: The ROC score for covid and normal class is at 100% , for other class at 96% and for pneumonia at 95%.



ROC for Naive Bayes Model: The ROC score for covid class is 82%, for normal class is at 81% , for other class at 72% and for pneumonia at 76%. So, we understand that Naïve Bayes is least performing model.



ROC Curve Classification Table:

ROC Curves	KNN Model	Random Forest Model	Naïve Bayes Model
Micro-Average ROC Curve	0.98	0.98	0.77
Macro-Average ROC Curve	0.98	0.98	0.78
ROC Curve of Covid Class	1.00	1.00	0.82
ROC Curve of Normal Class	1.00	1.00	0.81
ROC Curve of Other Class	0.97	0.96	0.72
ROC Curve of Pneumonia Class	0.96	0.95	0.76

13. Discussion: As per our research with 18448 images with four different classifications as covid, normal, other and pneumonia we understand that VGG-16 is best performing in terms of overall accuracy, and for precision, f- score and recall the best performance is for covid, normal, other and pneumonia categories. Image data generation and transfer learning was also utilized to improve machine learning model performance.

14. Conclusions:

With strong performance in distinguishing LUS images of COVID-19 from mimicking pathologies, a trained neural network will exceed human interpretation ability and raises the possibility of disease-specific, subvisible features contained within LUS images. The use of these techniques in rapid diagnostic decision-making of COVID-19 can be a powerful tool for radiologists to reduce human error and can assist them to make decisions in critical conditions and at the peak of the disease. This research supports the idea that DL algorithms are a promising way for optimizing healthcare and improving the results of diagnostic and therapeutic procedures. We envision the proposed tool as a step towards decision making to aid diagnosis by providing a “second opinion” to increase reliability.

The promising results of our model are to be validated in a controlled clinical study that investigates the predictive power for automatic detection of COVID-19 [9].

15. Recommendations:

As future research lines, working on multi-criteria classification to distinguish images from datasets mixing patients with lung problems due to several possible diseases, such as tuberculosis, AIDS, COVID-19, etc . [10]. Moreover, we have not found datasets with metadata including stages of the disease to diagnostic the severity of the symptoms. It would be better to plan to work in this aspect in cooperation with doctors at hospitals. We should work closely with diagnostics with the real time scenario and understand the need for prediction and train the models for minimizing the human errors besides providing fast and accurate predictions.

16. References:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8560933/
2. Wong, A. (2022, January 26). *Alexander Wong on linkedin: #icu #ai #EdgeAI*. Alexander Wong on LinkedIn: #ICU #AI #EdgeAI | 10 comments. Retrieved May 24, 2022, from https://www.linkedin.com/posts/alexander-wong-90650216_icu-ai-edgeai-activity-6892104558341152768-2P2h/?utm_source=linkedin_share&utm_medium=android_app
3. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0255886#pone.0255886.ref001
4. Nrc-Cnrc. (n.d.). *NRC-CNRC/COVID-US: Open benchmark dataset of COVID-19 related ultrasound imaging data, curated and systematically validated - ensemble de données de référence ouvert d'imagerie échographique liées à la covid-19, organisé et systématiquement Validé*. GitHub. Retrieved May 24, 2022, from https://github.com/nrc-cnrc/COVID-US
5. Google. (n.d.). *ML Practicum: Image Classification | google developers*. Google. Retrieved May 24, 2022,from https://developers.google.com/machine-learning/practica/image-classification/convolutional-neural-networks
6. https://www.nature.com/articles/s41598-021-99015-3
7. https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/#:~:text=A%20classification%20report%20is%20a,this%20article%20is%20for%20you.
8. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
9. https://bmjopen.bmj.com/content/11/3/e045120
10. https://www.hindawi.com/journals/jhe/2021/6677314/