UC Berkeley
SCHOOL OF INFORMATION

Final Report
Date April, 12 2022

Evaluating the Efficacy of Detecting COVID through ML-Based Analysis of Throat Images

Joe Villasenor, Ryan Mitchell, Manohar Madhira, and Napoleon Paxton

MIDS W251 Deep Learning at the Edge and in the Cloud

ABSTRACT

Early detection of highly contagious diseases, such as COVID-19 is key to controlling their spread and minimizing the effects on those that are infected. In-person tests put the tester and others in the vicinity at risk of infection, and over-the-counter tests are not always readily available due to supply chain issues or perhaps lack of state or federal funding. In this study, we examine the utility of COVID-19 detection through throat image analysis. Our approach utilizes a data set of throat images procured through a network of collaborators in the medical field. These images are labeled as COVID, Bacterial, and Normal and were fed into a Deep Neural Network (DNN) composed of a Convolutional Neural Network (CNN) and a fully connected Artificial Neural Network (ANN) for training. Our architecture yielded classification accuracies of 87.5% (CNN using just patient throat image data), 92.3% (ANN using just patient metadata), and 100% when combining the CNN outputs on the image data with the patient metadata in an ANN. Additionally, we utilized a Generative Adversarial Network (GAN) to generate additional throat images for training the CNN. However, the model performance did not change on the validation and test data. Our results suggest that with continued procurement of training data and additional architecture tuning, DNN can be a useful tool for early COVID-19 detection.

Keywords: Neural Network,GAN, CNN, ANN, COVID-19

1 INTRODUCTION

When it comes to infectious disease, early detection is key to minimizing transmission. Unfortunately, the current methods to detect COVID-19 require either over-the-counter testing kits or in-person testing. Recent supply chain disruptions have shown us that over-the-counter testing is not always available, and in-person testing can place potentially "negative" patients in an environment where they may be exposed to COVID at the testing site itself.

However, additional means of testing for COVID may be possible, as preliminary research has shown that images of the pharynx can reveal evidence of the COVID-19 infection [1–5]. To that end, the goal of our group project is to utilize deep learning techniques at the edge to assist doctors and patients in the detection of COVID-19 and help differentiate between viral and bacterial infections in the throat. In order to acquire the necessary data, our team partnered with a network of doctors to collect and share images and additional HIPAA-compliant patient metadata.

In our approach, we utilized a Convolutional Neural Network (CNN) and transfer learning techniques to analyze images of a patient's throat for discriminating features of three designated classes (normal, COVID-19, bacterial pharyngitis). We also experimented with Generative Adversarial Networks (GAN) and other data augmentation techniques to generate new images to complement our existing dataset and improve the classification performance of our CNN.

The outputs from the CNN model were then layered in with additional patient metadata and passed through an Artificial Neural Network (ANN) to arrive at a final patient classification, as described by Ningrum, Dina Nur Anggraini et al [6]. A diagram of the end to end pipeline is visualized in the figure below, Figure 1. The balanced accuracy for the CNN + ANN model was higher (100%) than the CNN model using patient image data (87.5%) or the ANN model using only patient metadata (92.3%) in isolation. Combining the patient's image and metadata using both networks seems to have helped prevent overfitting/misclassifications from the respective networks.

Finally, our team created a throat detection model to extract only the relevant throat structures from patient/doctor submitted images. If our end to end implementation were to be used in a production setting, this detection model would greatly increase image processing throughput for further training and would be necessary for running inference on patient or doctor-submitted images at scale by automating the bounding box / region of interest identification in the image to pass to the CNN.

Preliminary research in the area of classification using deep learning networks has been used for throat images which were not acquired using a smartphone. The model is

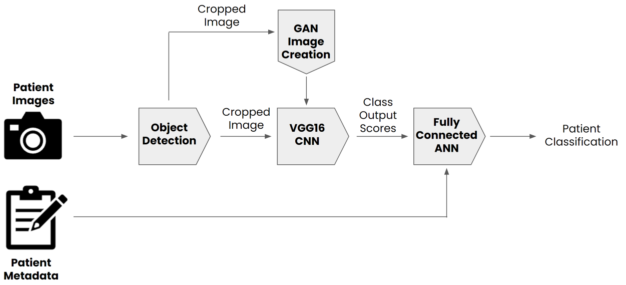


Figure 1. Architecture of DNN Covid Detection

a binary one classifying in pharyngitis or no-pharyngitis.

2 METHODOLOGY AND RESULTS

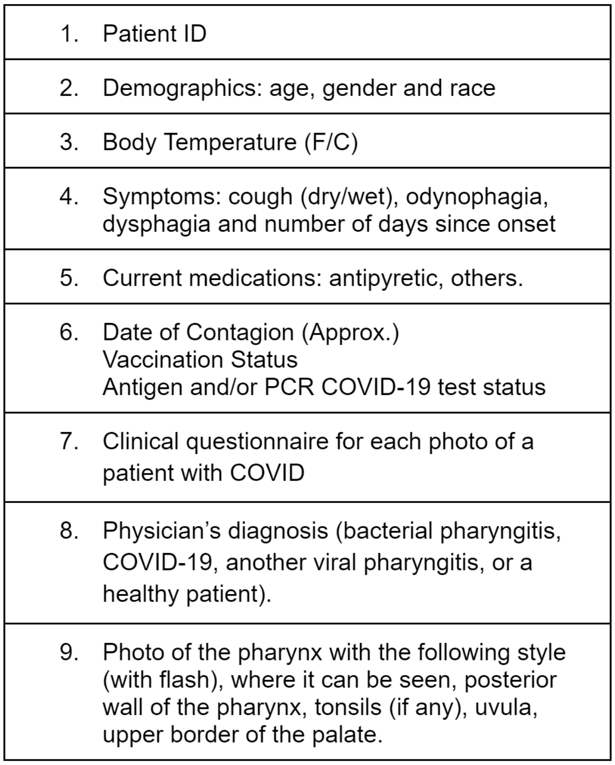
2.1 Data Collection & Preparation

We collaborated with different clinicians to get our pictures and surveys. Our clinical collaborators included specialists in the following fields: ENT, general medicine, family doctors, infectious disease, and internal medicine doctors. Some of the challenges were standardizing pictures, getting the clinical information, and complying with the research institution's standards of patient data protection. A one-day training at the hospital's COVID ward was also conducted to explain the procedure and get pictures from the patients.

For the data collection and preparation we asked the clinicians to collect the following data points for each patient and submit via a sharable Qualtrics survey shown in Figure 2.

In total We received 142 total patient records with 186 images across all classes. Qualtrics survey ensured that data within fields was streamlined. However, we faced the following issues:

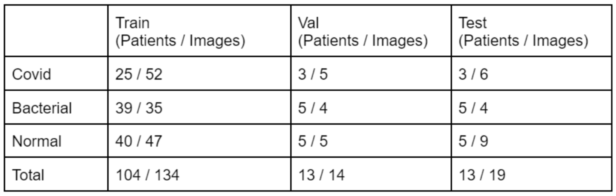
1. 7 records from survey testing were excluded from the analysis.
2. 3 records were classified as other (other than bacterial, covid, normal or viral). These three were dropped from the analysis.
3. 7 patients were classified as Viral to indicate Covid Phase 1 and Phase 2. For the analysis we combined all covid phases under Covid.
4. Poor Image quality - due to the wide range of image taking methods, and patient conditions, many images were not directly usable. To minimize impact we used Object (Throat) Detection and Image augmentation techniques to help improve image quality.
5. While we had collected Race, 78.5% of the respondents were reported as Other. Since images of pharynx are not impacted by race, and the data was skewed, we excluded race from the analysis.
6. Number of days since patients was not filled in many cases, hence these fields were not included in final analysis



1. Patient ID
2. Demographics: age, gender and race
3. Body Temperature (F/C)
4. Symptoms: cough (dry/wet), odynophagia, dysphagia and number of days since onset
5. Current medications: antipyretic, others.
6. Date of Contagion (Approx.)
   Vaccination Status
   Antigen and/or PCR COVID-19 test status
7. Clinical questionnaire for each photo of a patient with COVID
8. Physician's diagnosis (bacterial pharyngitis, COVID-19, another viral pharyngitis, or a healthy patient).
9. Photo of the pharynx with the following style (with flash), where it can be seen, posterior wall of the pharynx, tonsils (if any), uvula, upper border of the palate.

Figure 2. Qualtrics Survey

After correcting for the above issues, we used a randomized 80/10/10 split to divide patients into train/val/test categories. Given the small number of patients in the validation and test sets, we decided to aggregate validation and test performance for our model evaluations. All images belonging to a patient were included in the corresponding categories.



	Train (Patients / Images)	Val (Patients / Images)	Test (Patients / Images)
Covid	25 / 52	3 / 5	3 / 6
Bacterial	39 / 35	5 / 4	5 / 4
Normal	40 / 47	5 / 5	5 / 9
Total	104 / 134	13 / 14	13 / 19

Figure 3. Data Splits

Note: Patients with no usable images, all in the bacterial group, were dropped further from the analysis.

2.2 Object (Throat) Detection

A key component for a scalable end-to-end implementation of this project is standardized patient image data. The images submitted by our network of doctors spanned a wide range in terms of camera distance from the patient's mouth, wideness of mouth opening, resolution, brightness, and overall quality. Some patients also had objects inside their mouths in the images, including tongue depressors and ventilator tubes. To counteract some of this variability, and to prevent our image classification model from detecting extraneous features, our team manually cropped the entire set of images to focus just on the structures visible in the throat. The class labeling and bounding were done in Roboflow, and the bounding box parameters were passed to a script to crop the images before they were passed to the GAN or to the CNN networks.

Our team also used this cropped data to train an object detection model to extract images of the structure of the throat. We leveraged the ‘Yet Another EfficientDet Pytorch’ implementation (https://github.com/zylo117/Yet-Another-EfficientDet-Pytorch) for object detection and used transfer learning from pre-trained model weights to fine-tune the model for our dataset. This capability would greatly increase image processing throughput for further training and would be necessary for running inference on patient or doctor-submitted images at scale, should this project be continued and/or expanded in the future. Please refer to the figure below for a demonstration of the object detection model performance on an image outside of the model training set.

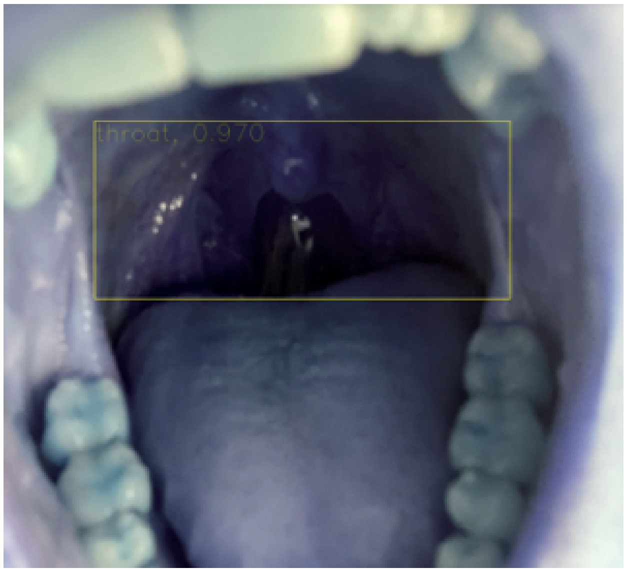


Figure 4. Throat Detection

2.3 Generative Adversarial Network

Due to the limited amount of data available to us through our network of collaborating doctors, our team also experimented with generating synthetic images to augment the training set for our CNN model. We settled on a GauGAN (Gaussian GAN) implementation due to the simplicity of the overall network and the fact that it only requires a random noise vector as an input. We first generated a Spatially Adaptive Denormalization (SPADE), in which the denormalization parameters are spatially invariant (the same values are applied to every position in the input activation). The implementation of residual blocks with SPADE normalization were performed before the encoder, generator, and discriminator.

The encoder for GauGANs encodes the original image into a mean and standard deviation from which to sample noise, which is then passed to the generator. The generator uses the information from the semantic label map and adds that to each batch normalization layer, at which point the noise is reshaped and upsampled to the target image size. The discriminator’s architecture follows a multi-scale design with InstanceNorm applying spectral normalization to all convolutional layers. It also takes as input the image concatenated with the semantic label map. The Loss function imposed a soft 0.05 weight Kullbach-Leiber divergence (KLD) loss term on the Gaussian statistics generated by the encoder.

Some of the pictures generated using the GauGAN had impressive results, and others were consistently dark, since our initial set for COVID patients were dark.

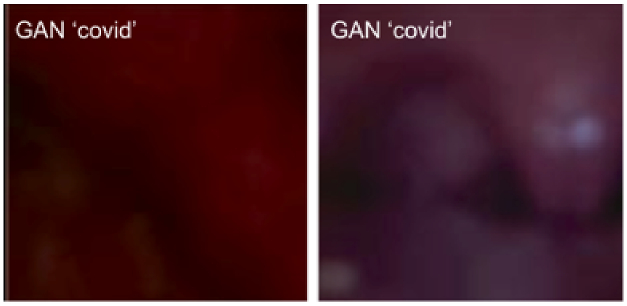


Figure 5. Initial GAN Images

After removing the dark images from the original set we were able to improve the quality of the images. Only data from the training set was used to train the GAN to prevent any data ‘leakage’ from the validation and test sets.

2.4 Convolutional Neural Network

For the convolutional neural network, we opted to use a slightly modified version of a well documented architecture (VGG-16) due to its strong performance in published research on similar medical classification tasks. During training, we determined that the pre-trained weights generalized better on the test image set compared to a training run with randomized initial weights. We ultimately explored several different variations of this CNN model - with varying degrees of data augmentation - and compared image and patient level classification accuracies after 100 training epochs. Class predictions were averaged across images for patients that submitted multiple images. Patient level accuracy scores tended to be higher than image level accuracy scores because, for patients that submitted 2+ images, the models tended to consistently misclassify the images (as opposed to classifying one image correctly and the remaining images incorrectly).

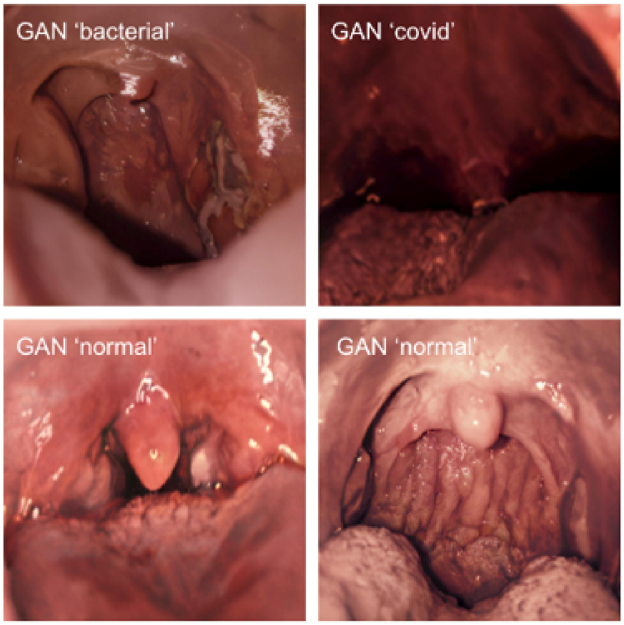
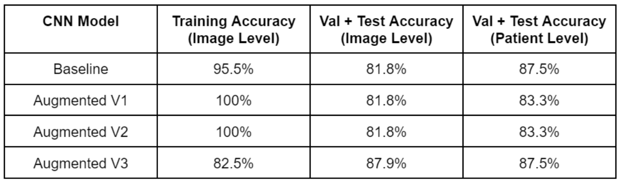


Figure 6. Improved GAN Images

The baseline model was trained on the cropped throat images from the original mouth image dataset. The Augmented V1 model trained on the original image set plus the final output image from a set of random permutations of scale/shift/rotate, RGB color shifts, multiplicative noise, hue contrast shifts, and brightness/contrast changes. The Augmented V2 model was trained on the original image set plus every image iteration from the augmentations mentioned above. The Augmented V3 model was trained on the original image set (as the baseline model had the best performance) plus a series of synthetic images generated from our Gaussian GAN model.

Generally speaking, augmenting the training data did not materially impact the classification performance, and there is evidence that the model began to overfit on the training data with the augmented images.



CNN Model	Training Accuracy (Image Level)	Val + Test Accuracy (Image Level)	Val + Test Accuracy (Patient Level)
Baseline	95.5%	81.8%	87.5%
Augmented V1	100%	81.8%	83.3%
Augmented V2	100%	81.8%	83.3%
Augmented V3	82.5%	87.9%	87.5%

Figure 7. CNN Accuracy

The Confusion Matrix for the validation and test sets show fairly consistent classification accuracy for each class, with bacterial images classified with 100% accuracy, COVID-19 with 82% accuracy, and normal with 71% accuracy.

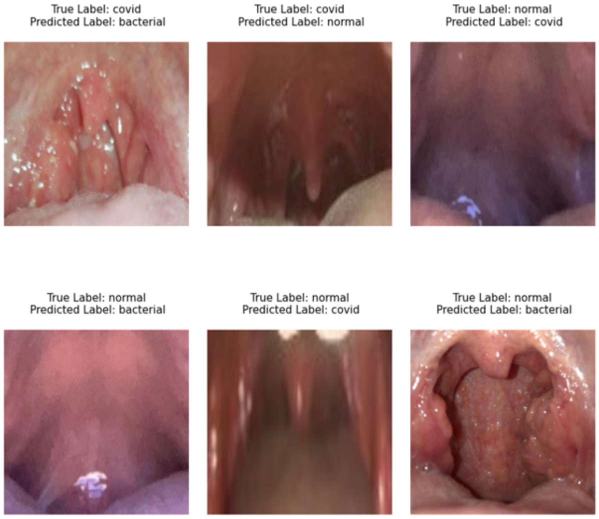


Figure 8. CNN Misclassifications

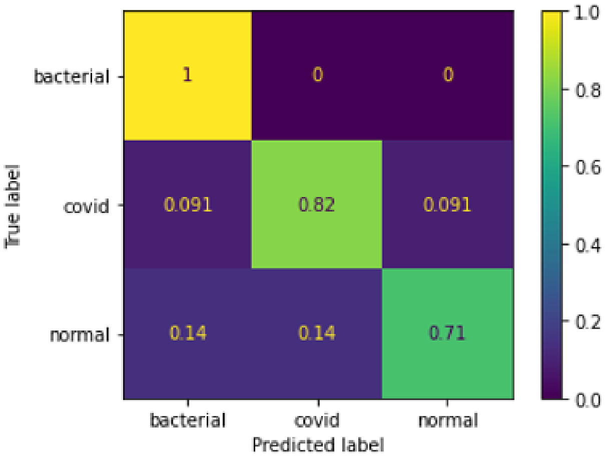


Figure 9. CNN Confusion Matrix

2.5 Artificial Neural Network

Ningrum et al. [6] showed that the balanced approach of a CNN+ANN model performed better than their CNN model in isolation. Combining patients metadata with ANN also prevents the overfitting that may occur with CNN.

Our team followed this same approach and combined the patients' metadata with the CNN predicted probabilities for each class. We used a simple fully connected layer with 1 input layer, 1 hidden layer and 1 output layer. Each layer had a bias term and used ReLU activation function.

Combining the results with the CNN improved the validation accuracy to 100% from 87.5% obtained by CNN alone. ANN alone with CNN had an accuracy of 92.3% on the validation dataset.

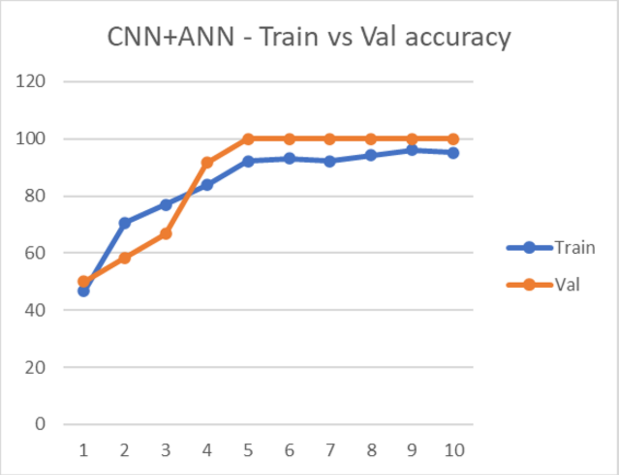


Figure 10. Combined CNN + ANN Accuracy by Epoch

3 DISCUSSION & LIMITATIONS

Of course our research, while promising, does have some limitations. One such limitation is that our models are trained on data that is representative of the patient population that visited our network of collaborators. The classification accuracy scores we measured should be consistent on newly submitted patient images/metadata, provided they come from the same network of doctors and that the population prevalence of the diseases we classify remain stable. However, if we were to sample or test randomly in the overall population, we would significantly overestimate the proportion of patients with bacterial pharyngitis and COVID-19 because the population prevalence of those diseases is far lower than what we observe in our training data. This issue of imbalanced training data on convolutional neural networks and the need to compensate for prior class probabilities is discussed by Buda et al. [7]. There is also the risk of patient misclassification, where false negative predictions for COVID-19 and bacterial pharyngitis could cause patients to not seek out preventative care and increase the probability of disease transmission if the patient does not self-isolate.

Another limitation of our research is the finite amount of data we were able to collect. One potential implication of this is that our classification model would not have specificity for COVID versus other viral throat infections because we did not have access to non-COVID viral infection images or patient metadata. Our data came from doctors that work in secondary and tertiary level of care hospitals (where more complicated cases are referred by primary care physicians), also most of our patients that volunteered for our experiment were in stage 2 (Pneumonia/respiratory symptoms) and stage 3 (Organ failure), whereas the prevalence of COVID-19 is higher of those in stage 1: Flu-like symptoms or asymptomatic. Furthermore, because most of the COVID patients we had access to were at stages 2 and 3, many had a difficult time opening their mouths. As a result, it is possible that the CNN was unintentionally detecting non-structural patterns or features (eg. reduced brightness) in the COVID training images. We did our best to prevent this by cropping out any irrelevant image data (such as tongue depressors or ventilator equipment in the mouth), but we cannot discount the possibility that some non-structural features were learned. It is also likely that our ANN model was able to better learn how to detect COVID-19 due to a higher prevalence of symptoms in our patient pool (which is skewed towards later stages) than would be seen in the overall population, where many COVID positive patients may be asymptomatic. It stands to reason that stage 1 COVID-19 patients would be more difficult to correctly identify with our current models. Of course, retraining the models with additional data on early stage COVID patients would help to mitigate this issue.

3.1 GitHub Repo

You can find the code for this project at the following location: https://github.com/rmadhira86/w251-finalproject

REFERENCES

[1] Zoabi, D.-R. & Shomron. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *Digit. Med.* **4** (2021).

[2] Mohammad-Rahimi & Hossein. Application of machine learning in diagnosis of covid-19 through x-ray and ct images: A scoping review. *Front. cardiovascular medicine* **8**.

[3] Noah, K. & Mosadeghi. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. *Digit. Med* **1** (2018).

[4] Walker, H., Tong & Palmer. Patient expectations and experiences of remote monitoring for chronic diseases: Systematic review and thematic synthesis of qualitative studies. *Int J. Medicine Inf.* **2** (2019).

[5] Yoo & Keun, T. Toward automated severe pharyngitis detection with smartphone camera using deep learning networks. *Comput. biology medicine* **125** (2020).

[6] Yoo & Keun, T. Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection. *J. Multidiscip. Healthc.* **14** (2021).

[7] Buda, M. M., Mateusz & Maciej. A systematic study of the class imbalance problem in convolutional neural networks. neural networks. *Neural Networks* **106** (2017).