COVID-19 X-ray Image Classification

Maneesh Kumar Singh
University of Illinois at Urbana-Champaign
mksingh4@illinois.edu

Raman Walwyn-Venugopal
University of Illinois at Urbana-Champaign
rsw2@illinois.edu

Satish Reddy Asi
University of Illinois at Urbana-Champaign
sasi2@illinois.edu

Srikanth Bharadwaz Samudrala
University of Illinois at Urbana-Champaign
sbs7@illinois.edu

ABSTRACT

Objective: Detecting COVID-19 using Chest X-Ray (CXR) images is becoming increasingly popular in deep learning research. When training deep neural networks, large and balanced datasets are preferred. However, since COVID-19 is new, there are a limited number of CXR images available which results in a challenge for training deep neural networks. Existing research has shown different approaches to address this imbalanced data issue. Two notable studies are FLANNEL (Focal Loss bAsed Neural Network EnsembLe) model and a patch-based classifier that works on segmented versions of the lung contours. We propose merging these two concepts together to improve performance of detecting COVID-19 in CXR images.

Materials and Methods: Using segmentation networks to create masks for the lungs as a pre-processing step. Replace base models in FLANNEL with patch-based classifiers that take the image and respective mask as their input. The patch-based classifiers will be used as the ensemble.

Results: We are able to reproduce FLANNEL and base models results using the updated datasets. We were also able to reproduce patch-based classification using X-ray images. We also created a segmentation network that can produce masks of the lung contours for CXR images. We successfully used this segmentation network to produce masks of the CXR images in the updated FLANNEL datasets.

Discussion: We saw improvement in metrics when training the base models and FLANNEL ensemble in detecting COVID-19 images. Since no parameters were changed, we suspect that this is due to the increase of COVID-19 images for the dataset in comparison to when the FLANNEL paper was written. We are in the process of training the patch-based classifiers to use as the base models for the ensemble.

Conclusion: With the success of performing segmentation on the dataset and the increased performance of the original base models due to the increased dataset size, we are hoping that this will lead to a further improvement once we finalize the patch-based classifiers. Our optimism is due to the patch-based classifiers outperforming their "global" counterparts that processed the whole non-segmented image.

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). It has spread worldwide leading to an ongoing pandemic. This pandemic has ravaged the world on an unprecedented scale. By April 2021, 141 million people have been infected and there are over 3 million deaths [10]. Chest X-Ray (CXR) is one of the important, non-invasive clinical diagnosis tools that helps to detect COVID-19 and other pneumonia for affected patients.

Using deep learning for X-ray classification is an ongoing research area and recently there have been promising models proposed for COVID-19 classification. The problem that all of these models face is an imbalanced dataset due to the limited number of COVID CXR images available.

FLANNEL is a COVID-19 CXR classification model proposed by Zhi Qiao et al. [7] that has been shown to accurately detect COVID-19 even when trained with only 100 available COVID-19 x-ray images. There are two core components for the FLANNEL architecture, the first is that it uses an ensemble [3] of five independent base models that predict the classification of the CXR. Each of the predictions are then passed through another neural weight network to determine the final prediction classification. The goal of the ensemble is to increase the robustness and accuracy of the network since each base model should capture patterns in the images independently [8]. The second core component for the FLANNEL is its use of the special Focal Loss [5] function, a modification of the standard cross-entropy loss that places a focus on the imbalance negatives by applying down-weights to well-classified examples. Focal Loss has been known to improve performance for imbalanced datasets.

Park et. al [6] has also created a deep learning model that has been proven to be effective on detecting COVID-19 when trained with limited datasets. The approach taken was to first detect lung contours of the CXR and perform segmentation. The motivation for performing segmentation first is that the patch based model focuses on the lung area since it's the primary region of interest used to perform analysis. In general, standard biomarkers [6] from CXR images analyzed are the following

1. Lung Morphology
2. Mean Lung Intensity
3. Standard Deviation of Lung Intensity
4. Cardiothoracic Ratio (CTR)

Thus it could be observed that most of the initial diagnosis is carried out from CXR images by concentrating on the

lung area. We also find this strategy also makes the model less susceptible to noise happening outside the lung region. After the lungs have been segmented, patch-based classification is performed. Patch-based classification involves selecting random crops or patches across the image for a set number of times and then performing classification on each patch. Afterwards, the final prediction of the image is made by majority voting based on the prediction of each patch. Conclusion provided in the patch based paper [6] by Park and Ye, it is clear that the patch-based classification outperformed the models that used the whole image for a limited train set data. As we have an imbalanced dataset with limited COVID 19 CXR images, we are optimistic that utilizing patch-based classification models for the FLANNEL ensemble with the combination of focal loss optimization would result in a performance improvement.

Our goal is to take the novel ideas of each approach listed above with the goal of improving performance. To accomplish this we will make modifications to the existing FLANNEL architecture by first pre-processing the CXR images by performing segmentation of the lung contours. Afterwards, we will then update the independent base models in the ensemble to be patch-based classifiers. We call this new architecture "Patched FLANNEL". The network structure is shown in Figure 1 and similar to FLANNEL consists of two stage approach.

2. METHOD

The primary objective was to improve the detection of COVID-19 in CXR images with a multi-classifier model that can detect four categories: Normal, Pneumonia Viral, Pneumonia Bacteria and COVID-19. The baseline we will be comparing against is the original FLANNEL architecture. We used the same datasets that were used in the FLANNEL paper, the COVID Chest X-ray Dataset [2] from GitHub and the Kaggle Chest X-ray images dataset. Similar to the FLANNEL paper, we also restricted the types of images used to anteroposterior (AP) or posteroanterior (PA). The restricted images were then labelled appropriately into one of the four categories.

2.1 Segmentation Training

The first major data pre-processing step that we performed on our dataset was segmentation. In order to accomplish this, we recreated the same segmentation network that Park et al. used for their patch-based classification; FC-DenseNet103 [4]. We trained the FC-DenseNet103 model using PyTorch to produce a mask of the lung contours of a CXR image. The datasets that were used to train the segmentation network were the Japanese Society of Radiological Technology (JSRT) dataset which contained 247 PA CXR images and the Segmentation in Chest Radiographs (SCR) database which contains segmentation masks for the CXR images from the JSRT dataset. The JSRT/SCR dataset were randomly split where 80% of images were used for training and 20% were used for validation; this resulted in 197 images being used for training and 50 images being used for validation for the JSRT dataset as shown in Table 1. Since CXR images from different data sources will come in a wide variety of formats, the JSRT dataset was pre-processed by performing data type casting to float32, histogram equalization to adjust the contrast, gamma correction to adjust brightness and standardizing the image size by resizing it to 256x256. During training, the network parameters were initialized with a random distribution and the Adam optimizer was used with an initial learning rate of 0.0001. The learning rate was decreased by a factor of 10 when there was no improvement in the loss. The Jaccard Index (JI) was used to evaluate the model during training since we were comparing the similarity of the mask produced by the network to the mask provided in the SCR dataset. An early stopping strategy was used based on the validation performance to prevent the model from overfitting.

We then applied the trained FC-DenseNet103 segmentation model on the AP and PA CXR images from the Covid Chest X-ray and Kaggle X-ray datasets, this resulted in producing a mask for the lung contours for each of the images. We then split the segmented dataset using a train-test ratio of 4:1 to randomly generate train test splits. To ensure reporting accurate performance on the base models, we used five fold cross validation while training. The detailed statistics are shown in Table 2.

2.2 Base model training

The next improvement that we produced was creating patch-based classifiers. Similar to the original global base models in FLANNEL, the patch base models used were pre-trained from ImageNet[1] to account for the small size of the dataset. The pre-processed images were first resized to 1024x1024 to be as close to the original pixel distribution. The masks generated from the FC-DenseNet103 segmentation model were also upsampled to 1024x1024 to match the new CXR image size.The resized images were then masked with the lung-contours and passed as input to the patch-based classifier. The patch-based classifier then produced k number crops/patches of size 224x224 from the CXR. To limit patches outside the lung area, the random points were forced to be within the lungs and the random point was used as the center of the patch. During inference, the k should be large enough to ensure that the lung pixels are covered multiple times. Each patch is then fed into a network to produce a prediction. The confidence score was calculated for each category by calculating the percentage of predictions for each class based on the k patches. The optimization algorithm used during training was the Adam optimizer with a learning rate of 0.00001. An early stopping strategy based on validation performance was applied and a weight decay and L1 regularization were used to prevent overfitting. The best model is selected among 200 epochs training.

2.3 Ensemble model learning

Ensemble model learning step is similar to baseline FLANNEL paper. We take N base models predictions and concatenate them as f and feed them in neural weight module to learn base model weight.

We calculate outer product ff^t which is flattened and fed into dense neural network with TanH layer to map features into base models weights. Then we train the ensemble model to learn optimal weight combination by feeding linear combination of predictions and weights of base models. The neural weight module uses a modified Focal loss function to handle multiclass classification.

The neural weight module uses a modified Focal loss function to handle multiclass classification. It downweighs the

[1]http://www.image-net.org/challenges/LSVRC/index

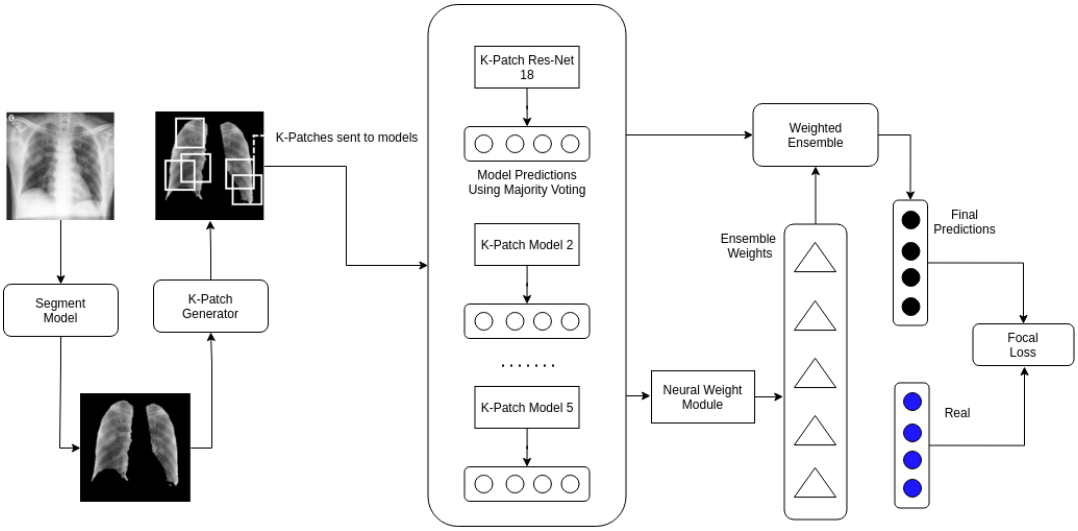

Figure 1: FLANNEL Improvement

well classified classes, in-favor of poorly classified classes so the model can focus on learning imbalanced examples.

$$LossFunc = FocalLoss(\hat{y}, y) \tag{1}$$

$$= \sum_{m=1}^{M} -\alpha_m y_m (1 - \hat{y}_m)^{\gamma} \log \hat{y}_m \tag{2}$$

Where $(1 - \hat{y}_m)^{\gamma}$ is a modulating factor with tunable focusing parameters γ and α_m [7]. α_m is set to be inverse class frequency of each class.
The overall algorithm is shown in Algorithm 1.

Table 1: FCDenseNet103 Segmentation Training & Validation Dataset

Dataset	Number of images
Training	197
Validation	50

2.4 Considerations

Here are our considerations discussed below -

- The pretrained model for segmentation (FC-Densenet103) will perform well on the datasets used in FLANNEL [7] paper.
- The number of patches(K) to be 100 to start with, will be fine tuned after evaluating the results.
- The patch image size will be chosen as 224X224 initially and will be fine tuned after evaluating the results.

Algorithm 1: FLANNEL with patch-by-patch Training

Input :
X-ray Images, Class Labels
Segmentation Model
Base Models $\{Learner_1, Learner_2, \ldots, Learner_n\}$
(Define B as batch size)
(Define K patch count)
Stage 1:
Run segmentation network on the dataset to generate masks for each CXR image.
Stage 2:
For each *batch of images from input images and labels* **do**
1. Fetch the segmented mask and image.
2. Resize the CXR image to 1024x1024.
3. Upscale the mask to 1024x1024.
4. Separate image into random k patches.
5. Pass random k-patches to base models.
6. Perform majority voting to get get confidence scores/prediction values from each base model.
7. Get prediction values from all Base Models.
$P_i = Learner_i(X) \in R^{B \times 4}$, where i = 1,...n
8. Get Learner weights.
$W = NeuralWeightModule([P_i, i = 1, .., n]) \in R^{B \times 5}$
9. Linear Combination for Prediction
$\hat{Y} = Softmax(\sum_{i=1}^{n} W_i P_i) \in R^{B \times 4}$ (where W_i represents i-th column of W)
10. Loss = $FocalLoss(\hat{Y}, Y)$
11. Back-propagate on Loss and update parameters
End

3. RESULTS

We chose 5 base models for FLANNEL framework, Densenet161, InceptionV3, Resnet152, ResneXt101 and Vgg19_bn. These

Table 2: Experimental data description

Source		Total	COVID-19	Viral	Bacterial	Normal
Original data	CCX data	554	478	16	42	18
	KCX data	5856	0	1493	2780	1583
View Distribution	AP view	6163	282	1501	2789	1591
	PA view	247	196	8	33	10
Training/test splits	Training	5127	378	1509	2291	1288
	Testing	1283	100	339	531	313
	Total	6410	478	1509	2822	1601

AP: anteroposterior; CCX: COVID Chest X-ray; COVID-19: coronavirus disease 2019; KCX: Kaggle Chest X-ray; PA: posteroanterior.

Table 3: Performance comparison on F1 score: Class-specific F1 score is calculated using 1 class vs the rest strategy

	COVID-19	Pneumonia virus	Pneumonia bacteria	Normal	Macro-F1
Original Base Learners					
Densenet161	0.7694 (0.03)	0.5901 (0.05)	0.8030 (0.01)	0.8875 (0.02)	0.7625 (0.02)
InceptionV3	0.8938 (0.01)	0.6413 (0.03)	0.8112 (0.02)	0.9015 (0.03)	0.8120 (0.02)
Resnet152	0.8302 (0.02)	0.6218 (0.02)	0.8046 (0.01)	0.9080 (0.00)	0.7911 (0.01)
ResNeXt101	0.8197 (0.03)	0.6151 (0.04)	0.8016 (0.01)	0.9046 (0.01)	0.7852 (0.02)
Vgg19_bn	0.8753 (0.02)	0.6023 (0.01)	0.8016 (0.01)	0.8950 (0.00)	0.7936 (0.00)
FLANNEL_OldData	0.8168 (0.03)	0.6063 (0.02)	0.8267 (0.00)	0.9144 (0.01)	0.7910 (0.01)
Original FLANNEL	0.9239 (0.01)	0.6675 (0.02)	0.8306 (0.01)	0.9322 (0.00)	0.8385 (0.01)

The values in parentheses are the standard deviations.

models were fine-tuned using default parameter values, settings and by using the Adam optimizer. We compared FLANNEL with these 5 base models of the framework.

We are planning to create and train patch-based versions of the base models that will use the masked version of the same images from the dataset. We will then re-run the FLANNEL with the patch-based models and compare performance. In addition to the improvements, we are also planning to compare 2 recent COVID-19 deep learning models, COVID-Net [9] and AI-COVID [1].

3.1 Evaluation strategy

Our main goal is to study the detection of COVID-19 among different respiratory x-ray images. We first measured the overall accuracy and precision of all 4 classes of x-ray images (COVID-19 viral pneumonia, non-COVID-19 viral pneumonia, bacterial pneumonia and normal images).

For each class of image, we record precision and recall values for each fold. We calculate F1-score for each fold and then average them to calculate the mean F1 score.

We are evaluating the global base models independently, the global ensemble, the patch-based models independently and the patch-based ensemble.

3.2 Implementation Details

The FC-DenseNet103 segmentation model was implemented in PyTorch and trained on a NVIDIA 1080 GPU.

FLANNEL with patch-based classification are implemented in PyTorch and are trained on 4 different Amazon Web Services Elastic Compute Cloud virtual machines each featuring a single NVIDIA Tesla V100 GPU. The base models are fine tuned using pre-trained models. The data are augmented with random flips, crops and scaling during the fine tuning process.

After the base models are trained, FLANNEL is trained by passing in the concatenated output layers of the base models as the input features.

3.3 Experimental Results

3.3.1 Segmentation Training

Training the Segmentation Network on the JSRT/SCR dataset had a Jaccard Index (JI) score of 93.39% for creating the lung contour masks. The figure 2 shows the mask creation and applying the mask on the image.

Figure 2: Segmentation Training

3.3.2 FLANNEL

In this section, We compared FLANNEL ensemble performance with the independent performance for each of the base models. Due to imbalance dataset, overall accuracy is not the appropriate measure for evaluation of the model. With this imbalance, even a significant increase in COVID-19 detection performance will not affect overall accuracy much. So, we will use F1 score for COVID-19 vs the rest comparing different models. As shown in the figure 3, clearly FLANNEL outperformed the other state-of-the-art models in detecting COVID-19 cases.

In Table 3, we also show F1-score for each classification and macro F1-score for all classes. In Table 3, we can see FLAN-

NEL performed better than base models. We also see with improved data FLANNEL scores are better than when run with old dataset containing only 100 COVID-19 images [7].

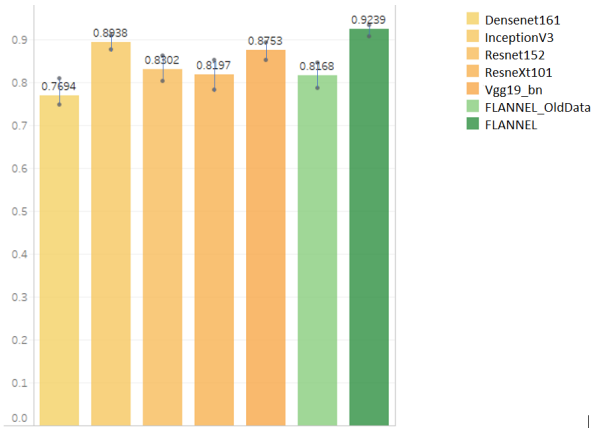

Figure 3: FLANNEL Improvement

We also provided the visual description of FLANNEL performance, via confusion matrix as shown in figure **??**. As the matrix depicts, For COVID-19 identification, FLANNEL has higher precision and recall than the other 2 types of pneumonia. This proves FLANNEL can distinguish pneumonia images from normal images and differentiate chest x-ray images of COVID-19 against other pneumonia images.

4. DISCUSSION

4.1 Slow Model Training

- Training the original FLANNEL base model is slow already where training a single model takes on average X hours.
- Training the patch-based models is even slower due to two reasons. The first is that the patch-base model dataloader was originally responsible for applying the mask to the cxr image, this was addressed by transitioning the responsibility to the segmentation task to immediately apply the mask to the CXR. The second reason for the patch-based models being slow is that a classification is run for each patch sequentially. This increases the complexity from N to kN where k is the number of patches. We were unable to address this due to time constraints but we believe parallelizing the patch predictions could lead to major improvements.

Reasons for patch-based models for not improving F1-Score over original FLANNEL based models:

- Mask production of CXR images was inaccurate. This is difficult to judge as we would need experts to determine if this was an actual problem
- Currently predictions for patch-based classifiers are performed using majority voting. This can potentially work against us if detections of Covid-19/Pneumonia are localized in certain areas of a CXR
- Local-patterns within a lung areas of a CXR are less important than the global patterns within the lung

We realized very early in the project that running FLANNEL model using five base models and 200 epochs for five foldes is going to take at least 3-4 days on a Tesla V100 GPU or equivalent. We had to find a solution on both cost and runtime in order to have results early. We tried multiple optimization methods such as increasing the number of workers, mounting training images in memory to bring the runtime down. We also used spot instances in parallel to bring the cost down. We were able to complete training and evaluation of the FLANNEL model in 36 hours and considerably less cost around 100 dollars.

5. CONCLUSION

As we are making progress, we have run the base models and FLANNEL on the new dataset. With the improved distribution of COVID-19 data we see FLANNEL outperforms the metrics as seen in the base FLANNEL paper by 13%. We created the Segmentation model that can produce mask of lung contours from Chest x-ray images. We are working on integrating the patch-based model in the FLANNEL ensemble model that would accept masked images as input. We will compare the performance of patched FLANNEL against the original FLANNEL. In addition to measuring classification performance, we will also measure timing performance to note how much of an impact the patch-based classification technique has on runtime.

6. CONTRIBUTION

All authors were actively involved in patched-FLANNEL development and implementations. Major contribution from authors are Maneesh: AWS Setup, batch run of FLANNEL and patch-by-patch models on AWS, debugging and fixing base FLANNEL and patched-FLANNEL issues, project report, Github documentation Satish: Sync checkpoint, output graphs and reports development, Github documentation Raman: patch-by-patch segmentation, classification development, project report, GitHub documentation Srikanth: patch-by-patch segmentation, classification development, project report,GitHub documentation

7. ACKNOWLEDGEMENT

GitHub source code for both FLANNEL and Patch-by-Patch classification [6] paper are used as baseline for this improvement.

- FLANNEL GitHub Repository
- Patch-by-Patch classification

TODO: Specify commit ID for COVID-xray data 78543292f8b01d5e0ed1d0

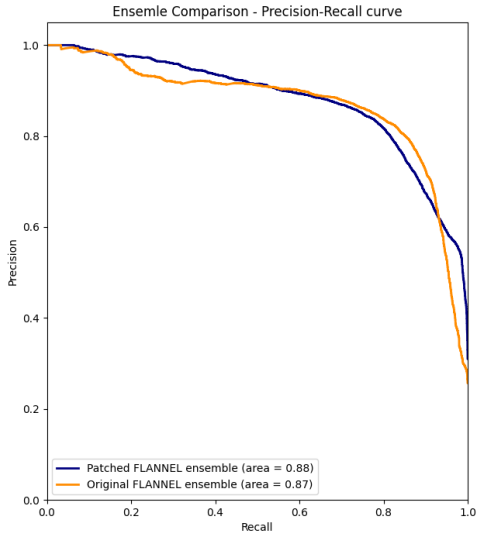

Figure 4: first figure

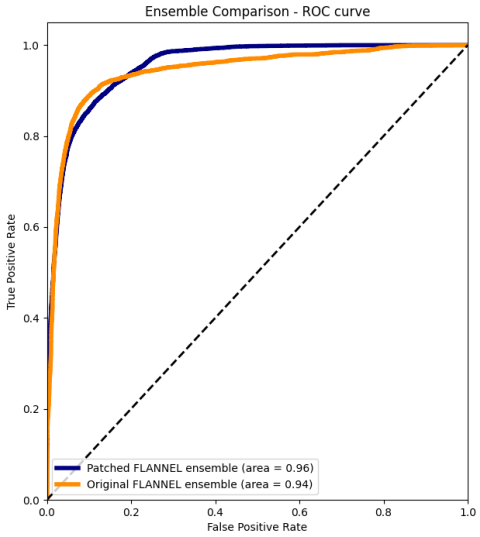

Figure 5: second figure

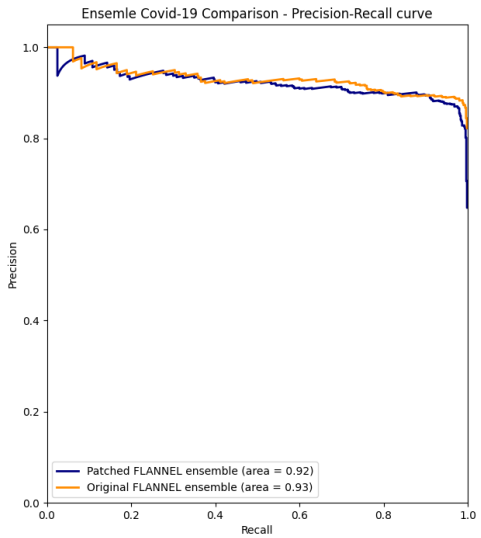

Figure 6: second figure

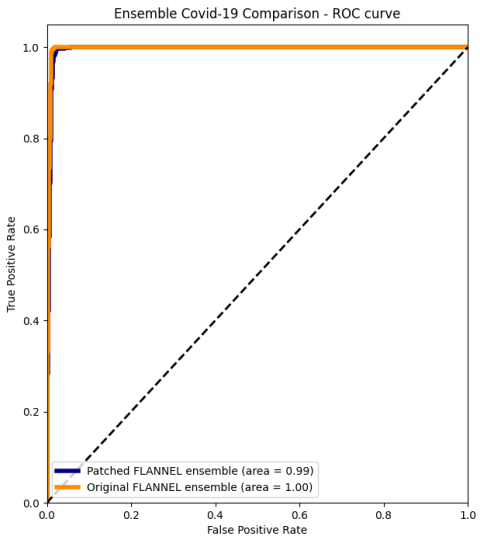

Figure 7: first figure

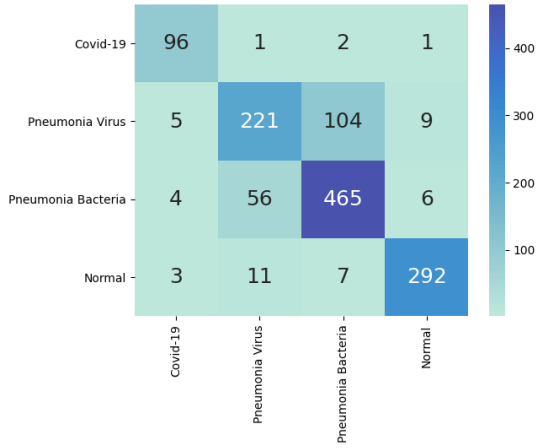

Figure 8: second figure

Figure 9: first figure

8. REFERENCES

[1] H. X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T. M. L. Tran, J. W. Choi, D. C. Wang, L. B. Shi, J. Mei, X. L. Jiang, I. Pan, Q. H. Zeng, P. F. Hu, Y. H. Li, F. X. Fu, R. Y. Huang, R. Sebro, Q. Z. Yu, M. K. Atalay, and W. H. Liao. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology*, 296(3):E156–E165, 09 2020.

[2] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*, 2020.

[3] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[4] S. Jégou, M. Drozdzal, D. Vázquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *CoRR*, abs/1611.09326, 2016.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018.

[6] Y. Oh, S. Park, and J. C. Ye. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Trans Med Imaging*, 39(8):2688–2700, 08 2020.

[7] Z. Qiao, A. Bae, L. M. Glass, C. Xiao, and J. Sun. FLANNEL (Focal Loss bAsed Neural Network EnsembLe) for COVID-19 detection. *Journal of the American Medical Informatics Association*, 28(3):444–452, 10 2020.

[8] A. Sharkey. On combining artificial neural nets. *Connect. Sci.*, 8:299–314, 12 1996.

[9] L. Wang and A. Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, 2020.

[10] WHO Emergency Response Team. Weekly epidemiological update on covid-19 - 20 april 2021. Technical report, WHO, April 2021.