COVID X-Ray Image Classification using Deep Learning

https://mediaspace.illinois.edu/media/t/1_dbkpe1io

Robert Wilcox, Sr.
U. of Illinois - Champaign
Monee, IL
rwilcox3@illinois.edu

Emil Varghese
U. of Illinois - Champaign
Burbank, CA
emilv2@illinois.edu

Tyler Bybee
U. of Illinois - Champaign
Des Moines, IA
tbybee2@illinois.edu

Charles Beyer
U. of Illinois - Champaign
Glenview, IL
ccbeyer2@illinois.edu

1. ABSTRACT

Background and Objective: Novel Coronavirus also known as COVID-19 originated in Wuhan, China around December 2019 and has now spread across the world. The study tries to classify COVID-19 chest X-Ray images against pneumonia and healthy patients using convolutional neural networks (CNNs).

Materials and Methods: In this study, we try to develop a prediction model as suggested in the paper, FLANNEL (Focal Loss bAsed Neural Network EnsembLe) for COVID-19 detection [2], which uses several standard convolutional neural network models and fuse them together for better accuracy using Ensemble methods. These publicly available datasets have been used for preliminary analysis, Kaggle Chest X-Ray images, IEEE COVID-19 data set and ChestX-ray database. The images belong to one of the four classes: "*Normal"* i.e. healthy, *"Bacterial Pneumonia"*, *"Viral Pneumonia"* and *"COVID-19".* The data set comprises of images that fall into the above categories. Some cleanup was done to the datasets to remove images with obvious leakage. This resulted in 293 COVID-19 examples, 1,576 normal examples, 2,773 bacterial pneumonia examples and 1,494 viral pneumonia examples for training and evaluation (compared to 100 COVID-19 examples, 1,118 normal examples 2,787 bacterial pneumonia examples and 1,503 viral pneumonia in the original FLANNEL paper [2]). Additionally, resizing and scaling was done as part of preprocessing, so the images are consistent across all sources. Step 1 of the process uses 9 standard CNN models trained using the standard image transforms and enhanced image transforms (random rotation, random brightness change, and random contrast change). The ensemble model is used to better address the class imbalance challenge and the quality of images. This model fuses the learnings from the best 5 base models and combines them using an importance weight and uses a focal loss to back propagate the learnings.

Results: The base training is done using models InceptionV3, VGG19-bn, ResNeXt101, Resnet152, Densenet161, Alexnet, ShuffleNet, MNASNet and SqueezeNet. Training was done using standard image transformations and image augmentation transformation. The best 5 models are selected based on the macro F1 score. Based on this criterion, the best performing models are Resnet152, Densenet161, Alexnet, VGG19-bn and ResNeXt101. The results are fed to the Ensemble model with focal loss and it achieved a F1 score of 0.81. For COVID-19 identification, the model achieves a macro-F1 score of 0.97, both of which are an improvement over the original FLANNEL [2] paper.

Discussion: Based on the results, most base models are not performing better with the enhanced image transforms (Table 1). As a result, enhanced image transformation was not used in developing the final models. The final model in this paper has improved performance over the base model evaluation in the original paper. The number of images in the dataset has increased from the time of publishing the original paper and this is one of the reasons for the better performance of base models. The number of publicly available images for COVID-19 is still very limited. Also, for this paper we have used data from 2 different data sets and the quality of X-Rays and additional markings train the model to predict COVID-19 images with much higher accuracy rate. The Ensemble step is designed to use the weights from multiple base learners to increase robustness of the classification and to tackle class imbalance issue [2]

On average we have seen around 5 mins per epoch, and this significantly slows down the process. Based on the limited resources available to students, it is time consuming to efficiently run and compare the various models. As a result, epochs were limited from the original 200.

Conclusion: The model proposed in this paper has improved performance over the original FLANNEL [2] paper as measured by the COVID-19 F1 and Macro-F1. The authors of this paper believe this is due to increased numbers of COVID-19 images in the dataset since the time of the original paper publication. Also, the models chosen in this paper performed better than the models chosen in the original paper.

Keywords
COVID-19 detection, Convolutional neural networks (CNNs), computer-assisted radiographic image interpretation, Class imbalance

2. INTRODUCTION

2.1 Objective

As of March, 2021, there have been over 100 million cases of the novel coronavirus disease 2019 (COVID-19) around the world, and almost 3 million deaths [1]. Due to the recency of the disease, there are relatively fewer X-Ray images related to the disease compared to non-COVID-19 related X-Ray images, which makes developing effective deep learning models difficult [2].

There have been developments in ensemble neural networks, specifically FLANNEL (Focal Loss bAsed Neural Network EnsembLe) to classify X-Rays of COVID-related pneumonia, and we hope to increase the accuracy of a FLANNEL model by introducing preprocessing in the form of image augmentation and by attempting to improve the accuracy of the worst performing basis model of the ensemble by replacing it with a better performing model. Traditional data augmentation methods have demonstrated promise for increasing the accuracy of Convolutional Neural Networks and we hope that by increasing the accuracy of a basis model of the ensemble, the overall accuracy would increase [4]. This would allow for faster and more accurate diagnosis of patients suspected of having COVID-19, to assess appropriate treatment options faster

2.2 Background and Significance

2.2.1 FLANNEL Approach (CNNs, Ensemble, Class Imbalance

This paper leverages the work in the FLANNEL [2] paper and attempts to improve upon that work. Within that document is given an excellent background on the elementary Deep learning techniques that were employed to develop the FLANNEL approach. The authors of this paper defer to the expertise of the FLANNEL paper for their well-thought out discussion on these items. The main sections referenced from FLANNEL are on Convolutional Neural Networks, Ensemble Methods, and the Class imbalance challenge, which are covered well in these papers. As a brief recap, Convolutional Neural Networks are the choice solution for Image processing and this papers objective is image classification. Ensemble methods are a technique to combine multiple approaches into a single result with the goal of achieving a better overall performance than any one of the individual techniques can achieve. Class imbalance challenge refers to the limited amount of COVID-19 imagery as compared to the other classes. Having a class imbalance in a dataset is a problem in many ways. First, it is difficult to train a model with limited information as the model may learn features incorrectly with a limited dataset as opposed to a larger dataset. Second, when there is a class imbalance it means that are far fewer examples of a given type of class as compared to other classes. This is particularly a problem when the class is the main class of interest, as it is in the datasets being referenced in this paper. The problem occurs when the overall model performance does not reflect the performance of the imbalanced class type. Thus, someone reviewing the model could draw in an invalid conclusion on the model overall when the classes are imbalanced.

2.2.2 Preprocessing

Preprocessing of images is a technique to improve the overall performance of the model. During the review of a dataset, a key strategy is to look for data leakage. An example of this would be to have an arrow on an X-Ray image that points to a particular disease. That arrow "leaks" the fact that the image does not contain a normal condition and the models will learn that the presence of an arrow indicates an unhealthy image. Strategies for this type of leakage would be to attempt to remove either the leakage or to remove the image. The existing FLANNEL paper did crop images to remove data leakage on the borders of the image. In our work, we are adding in more datasets which has other data leakage considerations, such as the arrows.

Other pre-processing techniques include resizing, cropping, random horizontal flipping, and noise introduction, which were employed within the existing FLANNEL work. Random image rotation and contrast variation are additional techniques added by this paper in order to improve model performance.

2.2.3 Alternate Models

In ensemble, the model selection is crucial to the overall performance. There are several CNNs to choose from. The existing FLANNEL work used vgg19-bn, inceptionv3, ResNext101, Densenet161, and Resnet152. This paper investigates the use of Alexnet, Squeezenet, Shufflenet, and MNASNet as alternative models. As part of preprocessing, the paper also investigates additional image enhancements like random rotation, random brightness change, and random contrast change to verify the effect on model training.

3. RELATED WORK

FLANNEL paper produced by Prof. Sun, et al and is the basis of this paper. FLANNEL uses an ensemble technique combining 5 state-of-the-art Convolutional Neural Network (CNN) classifiers as based models [2]. Data used in the paper is available publicly as well as the basis of the code for the project. The model compared with the baseline shows a higher macro-F1 score with 6% relative increase on the COVID-19 identification task [2].

"COVID-19 detection from scarce chest x-ray image…"[5] uses a few-shot approach which is well-suited to scarcity in data, which still seems to be the case for our problem domain, to the author's best knowledge. Like the FLANNEL paper, the approach employs transfer learning, but the use of unsupervised learning and few-shot seem to be a potentially novel aspect of their approach. The work produced a 96.4% accuracy over the 83% baseline. The work also classifies normal, COVID-19, and

pneumonia in the models. The few-shot approach employs a Siamese network, which there are reference implementations for, though not included in pytorch. Given the capabilities of few-shot learning that are exhibited by GPT-3, the authors of this paper feel that this could be a potentially useful avenue to integrate into the FLANNEL approach.

The COVD-Net paper [6] which was used in the FLANNEL paper to compare performance and is also referenced by a number of other papers in this area of research. The models achieved a 93.3% accuracy and employs CNNs in what the authors titled PEPIX to represent their Projection-Expansion-Projection-Expansion architecture [6]. In order to add explainability as well as audit the model, the authors used GSInquire. This tool highlights areas on the image that the model believes are features. This was then evaluated by radiologists to verify that the model was using valid features for classification.

The AICOVID paper[6] is another related work that was also used for comparison within the FLANNEL paper. In this approach, deep learning is used in the detection of COVID-19, normal, and pneumonia data. What is unique about this paper is that CT images were used instead of X-Rays. The paper is a good though CT images are not as readily available to the authors of this paper best knowledge and so would not be a source for a potential solution. This work had very good performance, reporting a 96% accuracy. However, in comparison of the F1 scores done in the FLANNEL paper [2] and considering the higher quality of CT imagery, the authors of this paper find the results surprising.

In Deep-COVID[8], the authors use ResNet18, ResNet50, SqueezeNet, and DenseNet-161 models to do transfer learning on COVID-19 classification. The authors used a fine-tuning technique on these existing models that is similar to the approach in FLANNEL. For a sensitivity rate of 98%, these models achieved a specificity rate of around 90% on average [8]. The authors also applied a heatmap solution similar to the one used in FLANNEL. In reviewing the heatmap against the radiologist's analysis, it appears that although the classification was correct, the features used in making that decision were off at times.

"Classification of COVID-19 chest X-rays…" [9] is the final related work reviewed by the authors of this paper. It is very similar in its approach to the Deep-COVID paper. This time the author compared the performance of AlexNet, GoogleNet, and SqueezeNet. The approach utilized CNN in their models. The author's approach was to gather 6 separate datasets, 2 of which used 3 class classification (normal, pneumonia, COVID-19) and focused again on adjusting the parameters of the pre-trained models. Accuracy measured on the range of 95.9 – 99.2%. The related works section in the paper is a very good resource for additional research on this topic.

4. DATA

The following data sets will be utilized for this project:

COVID-19 Chest X-Ray Dataset. This is a public dataset containing the chest X-Rays of patients who are positive or suspected of COVID-19 or other viral and bacterial pneumonias. This data is collected from public sources. This data set has 542 images from 262 people. The data set is available at https://github.com/ieee8023/covid-chestxray-dataset [14][15]

COVID-19 X-Ray Dataset This dataset contains X-Rays of patients with COVID-19, pneumonia, and no disease. This dataset is a combination of the data from multiple sources. This contains 127 images collected from COVID-19 dataset and Normal and pneumonia images are from ChestX-ray8 database. This data set is available at https://github.com/muhammedtalo/COVID-19 [13]

Chest X-Ray Images (Pneumonia) This dataset has enough non-COVID-19 and "normal" images to compare to COVID-19 images. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia [16]

As part of this paper, the COVID-19 datasets were combined to increase the size of the overall dataset. Manual inspection of the COVID-19 images showed many possible signs of data leakage. Examples of this are markings on the edge of the images or wires (such as those connected to patient monitors) crossing the image. There were no arrows found in any of the non-COVID-19 images, but there were wires visible in some of the non-COVID-19 images.
Because most images had markings on the edges, a center-crop of the image was used to remove those markings, in the same method as the original FLANNEL[2] paper. Any images with arrows that would not be removed by the crop were manually removed from the dataset. Images with visible wires in the X-Ray image were not removed as that would have significantly reduced the dataset. Also, it is likely that future X-ray would also have wires, i.e. removing images with wires in them could reduce real-world performance of the model.
The data sets used in the original FLANNEL[2] paper had fewer COVID-19 images and fewer types of "other" diagnoses. Images were sorted into "Normal" i.e. healthy, "Bacterial Pneumonia", "Viral Pneumonia" and "COVID-19" classes. Images were selected that fell into the previously mentioned categories from the datasets. Images with leakage were removed from the dataset. This resulted in 293 COVID-19 examples, 1,576 normal examples, 2,773 bacterial pneumonia examples and 1,494 viral pneumonia examples for training and evaluating from (compared to 100 COVID-19 examples, 1,118 normal examples 2,787 bacterial pneumonia examples and 1,503 viral pneumonia in the original FLANNEL [2] paper). Additionally, resizing and scaling was done as part of preprocessing, so the images are consistent across all sources.

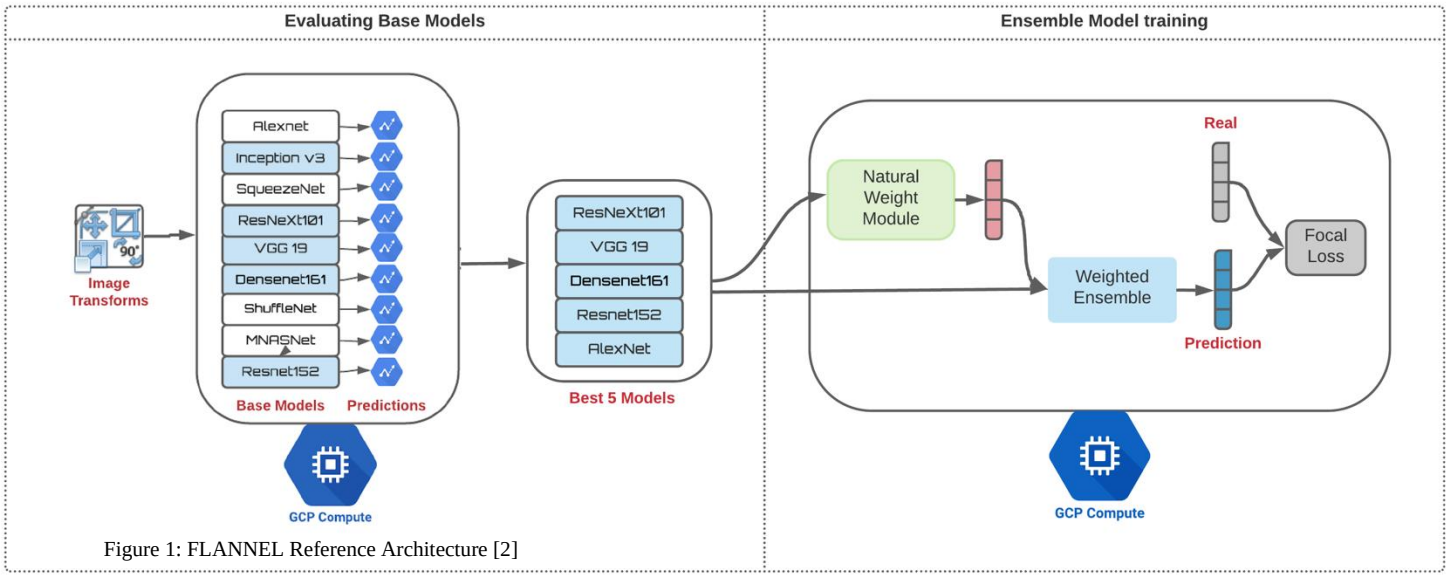

Figure 1: FLANNEL Reference Architecture [2]

5. MATERIALS AND METHODS

5.1.1 Approach

In this paper, two approaches are taken to try to improve the original FLANNEL basis models. The first was to attempt to improve the performance of the pretrained models by performing traditional image augmentation of the training data. The original preprocessing consisted of resizing all images to the same size, doing a center crop of that to get a 224x224 image (295x295 for inceptionV3), randomly horizontally flipping the image, and introducing noise to the images. The goal is to increase the generalizability of the model by also adding a random rotation, random brightness change, and random contrast change [3][4]. We added rotation up to +/-30 degrees as not all of the images were perfectly vertical. Images also had varying amounts of brightness and contrast, so those were also added at +/- 20% of the original values as well. Performance of the models trained on the original image modifications was compared to models trained on the more-modified images using F1-Score. Additionally, the train/evaluation split was performed prior to image modifications, so the data splits for each type of model are on the same base images.

The second approach to improve the accuracy of the FLANNEL model is to replace the lowest performing models of the original basis models, with better performing models according to F1-score. There are various CNNs evaluated for this, and pretrained weights (parameters) will be used in the same method as FLANNEL. Specifically, the pretrained model first has its final layer replaced with 4 outputs pertaining to the X-Ray labels, at which point it is trained with all of the weights frozen besides the output layer. The models that were evaluated in pursuit of a better performing base models are, InceptionV3, VGG19-bn, ResNeXt101, Resnet152, Densenet161, Alexnet, ShuffleNet, MNASNet and SqueezeNet. All these models were trained using the data set with traditional image augmentation as well as the enhanced image augmentation with random rotation, random brightness change, and random contrast change.

Once all the models are trained on both datasets, the best scoring 5 models are used as the basis of the FLANNEL architecture **[see figure above]**.
Depending on their performance, some models could be trained on the highly preprocessed data, and the rest of the models could be trained on the less-preprocessed data.

5.1.2 Methods

As shown in the architecture diagram, the model architecture consists of 2 stages. In the first stage the various CNN models are trained to find the best 5 models. The results of the 5 best models are used to train the ensemble model.

Stage 1: Base learners

CNN models InceptionV3, VGG19-bn, ResNeXt101, Resnet152, Densenet161, Alexnet, ShuffleNet, MNASNet and SqueezeNet are evaluated and the best 5 models were chosen. The chosen base learners are

- **Alexnet**: Alexnet is one of the best performing CNN models. It is similar to LeNet but with more layers per filter and stacked convolution layers.
- **VGG19-bn**: The model architecture is considered to be a successor of the AlexNet, but it was created by a different group named as Visual Geometry Group at Oxford.
- **ResNeXt101**: This 101-layer architecture is developed by the by Facebook's AI Research and UC San Diego.

- **Resnet152**: This 152-layer Deep Residual Neural Network uses skip connection between layers which make deeper networks possible.
- **Densenet161**: This is a 161 layer Deep Neural Network where every layer is directly connected to every other layer in a feed forward fashion [10]

Stage 2: Ensemble Model

In this step, the predictions output of the ***N*** base learners (N being 5 for the top models in this case) along with the 4 classes are saved data files. The data files are used as input to the ensemble model training. The base learner weights are calculated by taking the ***N*** prediction outputs and concatenating it to feed a neural weight module[2].

The neural weight module takes the concatenated output, ***f*** and uses it to create the feature interaction by doing the outer product ff^T. This captures more latent information. The result is then flattened and passed through the ***Tanh*** activation function. Outputs of ***Tanh*** could have negative values, and this helps to discount inaccurate predictions of the model. [2]

The linear combination of the base learner predictions using the base learner weights (***w***) are used to get the final predictions.

The model was tested with the standard cross entropy loss as defined by the following equation [2]

$$LossFunc = CELoss(\hat{y}, y) = \sum_{m=1}^{M} -y_m \log(\hat{y}_m)$$

where ***M*** represents the number of classes and ***y^*** represents the predicted values and ***y*** represents the actual values.

This approach does not lead to making optimal parameter updates since the data set is heavily imbalanced with the number of COVID-19 images being much less than the other classes. So, a focal loss approach was used to address this issue. The distribution of the image classes is used as weights to balance the unbalanced dataset. [2]

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

Here α_m is the inverse class frequency of each class. This will make sure that all the classes contribute equally to the loss. The resulting loss is back propagated, and the weights of the neural weighing module is updated. The parameters of the base learners are frozen and not updated at this stage.[2]

6. SETUP

The following hardware/software was used for this project:

Software: Python, PyTorch, Torchvision. Jupyter Notebooks

Hardware:
Google cloud VM (2 vCPUs, 13 GB memory), 1 x NVIDIA Tesla T4

Google cloud VM (2 vCPUs, 13 GB memory), 1 x NVIDIA Tesla K80

7. EXPERIMENTAL RESULTS

As our experiment has multiple experimental components, the experimental results are provided in sections covering each specific experiment.

Base Learner F1 Score Performance

As outlined in previous sections of the document, the first task is to identify the "Best 5" base learner models which will be subsequently pushed through the Ensemble Model Learning process. Each model was evaluated against the source data using basic preprocessing as well as the enhanced image pre-processing. The enhanced image pre-processing did not result in better performance based on F1 scores and the authors of this paper decided it would not be utilized for further model development.

Table 1 contains the performance, based on F1 Score, for the

Table 1. Performance comparison on F1 score for standard image pre-processing (Shading Indicates chosen models)

	COVID-19	Viral	Bacterial	Normal	Macro-F1	Accuracy	Avg. Train Time (per Epoch)
Base Leaners							
Inception V3	0.84	0.37	0.76	0.83	0.71	0.71	194 seconds
VGG19_bn	0.96	0.67	0.83	0.91	0.82	0.82	198 seconds
ResNeXt101	0.98	0.65	0.84	0.95	0.85	0.83	676 seconds
Resnet152	0.99	0.67	0.83	0.94	0.86	0.83	389 seconds
Densenet161	0.97	0.65	0.83	0.95	0.85	0.83	372 seconds
Alexnet	0.96	0.62	0.81	0.95	0.83	0.81	129 seconds
ShuffleNet	0.95	0.53	0.82	0	0.57	0.76	128 seconds
MNASNet	0.51	0.31	0.7	0	0.25	0.49	150 seconds
SqueezeNet	0.91	0.63	0.79	0.89	0.81	0.78	142 seconds

Table 2. Performance comparison on F1 score for augmented image pre-processing

	COVID-19	Viral	Bacterial	Normal	Macro-F1	Accuracy	Avg. Train Time (per Epoch)
Updated Leaners							
Inception V3	0.28	0	0.77	0.87	0.69	0.69	210 seconds
VGG19_bn	0.96	0.65	0.84	0.96	0.83	0.83	202 seconds
ResNeXt101	0.97	0.61	0.81	0.95	0.83	0.81	682 seconds
Resnet152	0.96	0.69	0.83	0.95	0.86	0.83	400 seconds
Densenet161	0.97	0.66	0.84	0.95	0.86	0.84	384 seconds
Alexnet	0.94	0.57	0.81	0.94	0.81	0.79	135 seconds
ShuffleNet	0.91	0.53	0.82	0	0.56	0.73	125 seconds
MNASNet	0.49	0.31	0.7	0	0.24	0.48	159 seconds
SqueezeNet	0.91	0.63	0.7	0.8	0.72	0.72	141 seconds

base learner models utilizing the source data with only basic preprocessing.

Table 2 contains the performance, based on F1 Score, for the base models utilizing the enhanced image pre-processing.

The chosen models are highlighted in the in the tables. The ResNeXt101, Resnet152, Alexnet, VGG19-bn and Densenet161 models performed the best for the standard image pre-processing. All these models were fine-tuned using their default parameter settings and using the SGD optimizer [11]

The base learners were selected by running 25 epochs and by updating the last layer to match the 4 way classification.

Evaluation strategy

The classification metric, F1 score, which is the harmonic mean between precision and recall, has been used to rate the overall accuracy of the models to predict the 4 classes (COVID-19, viral pneumonia, bacterial pneumonia, and normal images). The main objective of the original paper and this study is the detection of COVID-19 among the various kinds of respiratory related X-Ray images.

In addition to the F1 table results, confusion matrices help to graphically represent the predictive capability of the model for COVID-19, Pneumonia, and Normal images. The authors of this paper generated matrices for all the base learner models, but only include one such matrix in the document for the sake of brevity.

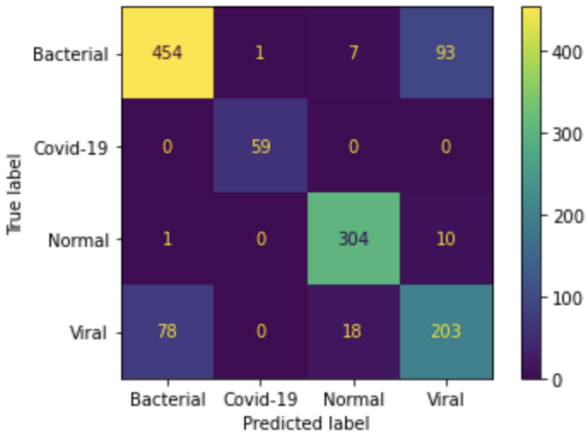

Figure 2 - ResNeXt101 Confusion Matrix

Ensemble Model Learning

The second experimental component consists of benchmarking the Best 5 base learners against the original FLANNEL base learners. While previous experiment's F1 scores give insight into which 5 models perform the best individually, the final result will compare the performance of the ensemble model learning to the original paper. [2]. The confusion matrix for the ensemble model shows the classification accuracy of the images and how much misclassification occurred. The model does a good job of identifying the COVID-19 images. The model appears to have an issue with detecting bacterial pneumonia and viral pneumonia.

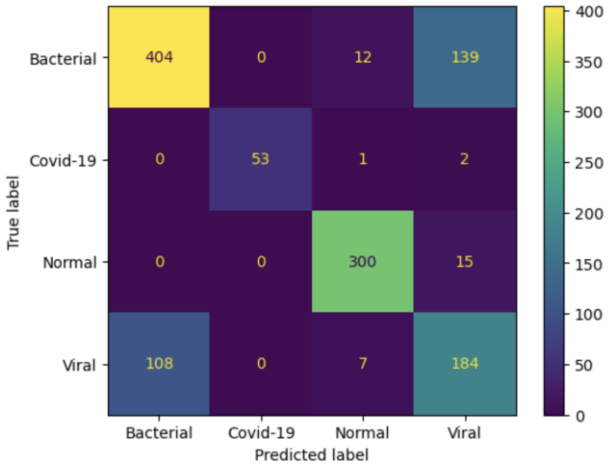

Figure 3 – Ensemble model Confusion Matrix

The ensemble model addresses the class imbalance challenge since it fuses the learnings from the best 5 base models and combines them using an importance weight and uses a focal loss to back propagate the learnings. The ensemble model learning resulted in a better overall performance compared to the original FLANNEL paper. See Table 3 for details. Notably, this paper had a COVID-19 F1 of .97 vs .81689 and a Macro-F1 of .81 vs .791 in the original paper.

Table 3. F1 score for Augmented FLANNEL vs original FLANNEL

	FLANNEL	AUGMENTED
COVID-19	0.8168	0.97
Viral	0.6063	0.58
Bacterial	0.8267	0.76
Normal	0.9144	0.94
Macro-F1	0.791	0.81

The ensemble model does not result in better performance as compared to the individual base ensemble models. However, since the F1 values are very high in the individual models, it is the author's assertion that the ensemble model, with it's reduced macro-F1, will result in better general model performance due to the likelihood of overfitting of the individual models.

8. DISCUSSION

The most notable observations/challenges from our results are

- Class Imbalance
- FLANNEL Baseline Challenges
- High COVID-19 Classification Rate
- Performance issues for augmented image preprocessing (*for certain base learners)*

Class imbalance

The primary takeaway from the classification task is the class imbalance issue. With a total of 4,901 images, COVID-19 accounts for 5%, Bacterial Pneumonia 45%, Normal 26% and Viral Pneumonia 24%. The models also have high misclassification rate for Viral Pneumonia and Bacterial pneumonia. The ensemble model in this paper tries to overcome this by using focal loss along with the weighed base models using the same strategy in the FLANNEL[2] paper.

FLANNEL Baseline Challenges

Given that the FLANNEL research paper included the original source code and data sources, little effort was expected in recreating the original baseline results. Unfortunately, that was not the case. The codebase is incomplete and needed considerable rework.

The first issue was that the provided source code was not complete and was lacking sections of code. (E.g. Model Training) Additionally, code defects were encountered as well as sections that required updates due to changes in the data.

A second issue was the data itself. Even though the original FLANNEL[2] data sources were used for this paper, the data has changed since the original FLANNEL implementation. Variances were encountered with image quality, quantity, and even classification labeling. While the original labels simply consisted of Bacterial Pneumonia / Viral Pneumonia / COVID-19 / Normal, far more specific labels were assigned to images in the newer dataset requiring updates to the original code to recreate the original labelling.

The final issue we have encountered is in the equipment for processing the models. Colab, Azure, AWS, and Google Cloud Platform(GCP) were all explored by the authors of this paper. Issues were encountered in some of the platforms. Colab terminated our training mid-run several times and was abandoned. In GCP, access to GPUs was a problem as our free accounts required a manual enabling of GPUs and then the limit was only allowed to be one without escalated approval. Once the GPU limit was removed, finding a zone with an available GPU was also a challenge and several hours were spent finding zones. Lastly, github has a maximum file size of 100MB. All models in this paper were larger, so the repository had to be changed to gitlab which was a 5GB limitation.

High COVID-19 Classification Rate

Compared to Pneumonia and Normal X-Ray images, the positive classification rate for COVID-19 was consistently higher across our testing. While the authors of this paper would like to think that these models are simply exceptional, there may be another explanation.

COVID-19 patients, who are sick enough to be in a hospital setting and receiving an X-Ray, are likely to be experiencing significant medical issues. Because of this, they may have a higher incidence of other items appearing in the X-Ray images.(e.g. wires for monitors, tubes, etc.) that may be resulting in unintended data leakage.

A review of the COVID-19 images was completed to remove images where there were obvious markers. Different data sets for COVID-19 and other image classes, the quality of images are different. As a result of this, the models could be learning some unknown markers within COVID-19 data set giving it higher accuracy.

Performance Issues for augmented image preprocessing

The augmented image prepreprocessing did not improve the COVID-19 F1 Scores of any of the models and in a number of cases the performance decreased (See Tables 1 and 2). The variances are significant and unexpected. The authors of this paper did research into augmentation and believe that the amount of data augmentation, which has a regularizing effect,[2] is causing the models to now underfit as is seen in the results.

9. CONCLUSION

Based on the results, the base models used in this paper are performing better than the original FLANNEL [2] paper. The authors of this paper attribute this to 1) The number of COVID-19 images in the dataset has increased from the time of the original paper publication and 2) the models are running into unbalanced data problem with a bias towards COVID-19 images. The images are also sourced from different data sets and they might not be adhering to the same quality standards.

The ensemble model proposed in this paper improves upon the original FLANNEL[2] paper. The proposed model combines the outputs of the best performing CNN models and uses focal loss to handle the unbalanced dataset and create a more accurate model. The ensemble model has the ability to diagnose COVID-19 from X-Ray images with high accuracy rate even though the number of available images are limited.

As COVID-19 does not appear to be eradicated in the near future, models such as this could prove valuable in the detection of COVID-19 in patients as well as demonstrating the efficacy of ensemble methods applying effectively to learning and prediction of imbalanced classes. Additionally, noticeable improvements in the pretrained basis models, due only increasing the number of COVID-19 samples to about 300 images, from 100 originally, shows the usefulness and versatility of leveraging pretrained models on relatively small amounts of data, but still obtaining good results.

AUTHOR CONTRIBUTIONS

TB created and gave the presentation. All authors were involved in implementing the methods, conducting experiments, and writing the paper. The experiments took a while since training each model takes up time and resources. The authors were all using individual GCP virtual machines to train, validate and test the various base models and ensemble performance.

10. REFERENCES

[1] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. Access: 2021-03-29.

[2] Z. Qiao, A. Bae, L. Glass, C. Xiao, J. Sun. FLANNEL (Focal Loss bAsed Neural Network EnsembLe) for COVID-19 detection *Journal of the American Medical Informatics Association*, 28(3), 2021, 444–452. (https://github.com/qxiaobu/FLANNEL)

[3] L.F. Rodrigues, M.C. Naldi, J.F. Mari. Comparing convolutional neural networks and preprocessing techniques for HEp-2 cell classification in immunofluorescence images. *Computers in Biology and Medicine*, 116, 2020.

[4] J. Wang, L. Perez. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv*:1712.04621, 2017.

[5] S. Jadon. COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach. Medical

Imaging 2021: *Imaging Informatics for Healthcare, Research, and Applications*, Proceedings Vol. 116010, 2021.

[6] L. Wang, Z.Q. Lin, & A.Wong. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10, 2020.

[7] X. Bai, R. Wang, Z. Xiong, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology*, 296, 2020.

[8] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, G.J. Soufi. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning, *Medical Image Analysis*, Volume 65, 2020.

[9] T.D. Pham. Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning? *Health Information Science and Systems*, 9, 2021.

[10] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017.

[11] Kingma D, Ba J, Adam A. Method for stochastic optimization. In: ICLR 2015: International Conference on Learning Representations; 2015.

[12] Ivanov, Slav, "37 Reasons why your Neural Network is not working", https://blog.slavv.com/37-reasons-why-your-neural-network-is-not-working-4020854bd607, 7/25/2017.

[13] T. Ozturk, M. Talo, A. Yildirim, U. Baloglu, O. yıldırım, U.R. Acharya. Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images. Computers in Biology and Medicine. 121. 10.1016/j.compbiomed.2020.103792, 2020

[14] J.P. Cohen, P. Morrison, L. Dao. COVID-19 image data collection, arXiv:2003.11597, 2020. https://github.com/ieee8023/covid-chestxray-dataset

[15] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi. COVID-19 Image Data Collection: Prospective Predictions Are the Future, arXiv:2006.11988, 2020. https://github.com/ieee8023/covid-chestxray-dataset

[16] D.S. Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172.5 1122-1131, 2018.