Journal of the American Medical Informatics Association, 28(3), 2021, 444–452
doi: 10.1093/jamia/ocaa280
Advance Access Publication Date: 30 October 2020
Research and Applications

Research and Applications

FLANNEL (Focal Loss bAsed Neural Network EnsembLe) for COVID-19 detection

Zhi Qiao,[1] Austin Bae,[2] Lucas M. Glass,[2] Cao Xiao,[1] and Jimeng Sun[3]

[1]Analytics Center of Excellence, IQVIA, Beijing, China, [2]Analytics Center of Excellence, IQVIA, Cambridge, Massachusetts, USA and [3]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Corresponding Author: Cao Xiao, IQVIA, 201 Broadway Floor 5, Cambridge, MA 02139, USA; cao.xiao@iqvia.com

Received 3 June 2020; Revised 7 October 2020; Editorial Decision 25 October 2020; Accepted 26 October 2020

ABSTRACT

Objective: The study sought to test the possibility of differentiating chest x-ray images of coronavirus disease 2019 (COVID-19) against other pneumonia and healthy patients using deep neural networks.

Materials and Methods: We construct the radiography (x-ray) imaging data from 2 publicly available sources, which include 5508 chest x-ray images across 2874 patients with 4 classes: normal, bacterial pneumonia, non–COVID-19 viral pneumonia, and COVID-19. To identify COVID-19, we propose a FLANNEL (Focal Loss bAsed Neural Network EnsembLe) model, a flexible module to ensemble several convolutional neural network models and fuse with a focal loss for accurate COVID-19 detection on class imbalance data.

Results: FLANNEL consistently outperforms baseline models on COVID-19 identification task in all metrics. Compared with the best baseline, FLANNEL shows a higher macro-F1 score, with 6% relative increase on the COVID-19 identification task, in which it achieves precision of 0.7833 $\pm$ 0.07, recall of 0.8609 $\pm$ 0.03, and F1 score of 0.8168 $\pm$ 0.03.

Discussion: Ensemble learning that combines multiple independent basis classifiers can increase the robustness and accuracy. We propose a neural weighing module to learn the importance weight for each base model and combine them via weighted ensemble to get the final classification results. In order to handle the class imbalance challenge, we adapt focal loss to our multiple classification task as the loss function.

Conclusion: FLANNEL effectively combines state-of-the-art convolutional neural network classification models and tackles class imbalance with focal loss to achieve better performance on COVID-19 detection from x-rays.

Key words: COVID-19 detection, neural network ensemble, class imbalance, computer-assisted radiographic image interpretation

INTRODUCTION

Objective

The novel coronavirus disease 2019 (COVID-19) is a pandemic disease that has been spreading rapidly across the world and poses a serious threat to global public health. Up to July 2020, COVID-19 had caused more than 13 million infections and 600 000 deaths across the globe. Radiology plays a fundamental role in diagnosing the disease. Although less sensitive than chest computed tomography (CT), chest radiography (x-ray) is typically the first-line imaging modality used for patients with suspected COVID-19.[1] Owing to the international attention of the disease, there has been an increase in publicly available x-ray images of patients with COVID-19–related pneumonia. This enables us to identify patterns and construct models that could automatically detect COVID-19 from chest x-ray images.

Because COVID-19 is still relatively new, naturally there are significantly less x-ray images of COVID-19 compared with non–COVID-19 diseases. This inevitably means that these datasets will be imbalanced, with few COVID-19 samples, which can hinder classification accuracy and limit model generalization.

To address these challenges, we propose a FLANNEL (Focal Loss bAsed Neural Network EnsembLe) model. FLANNEL utilizes an ensemble structure, using 5 state-of-the-art convolutional neural network (CNN) classifiers as based models. The neural weight module combines each base model using a learnt weighted ensemble to get the final classification results. To handle the class-imbalanced dataset issue, the model adapts focal loss traditionally proposed for binary classification task to our multiclass classification task. This loss function back-propagates and updates the parameters of the neural weight module.

We train and evaluate the proposed FLANNEL architecture on a combination of publicly available datasets. The combined data consist of 5508 chest x-ray images across 2874 patients. There are 4 types of x-ray images: normal (n = 1118), bacterial pneumonia (n = 2787), COVID-19 pneumonia (n = 100), and non–COVID-19 viral pneumonia (n = 1503). These 4 types of x-rays naturally form a 4-way classification problem, in which our focus is to make accurate prediction on COVID-19 cases without hindering the performance of other classes.

Background and significance

Convolutional neural networks

CNNs have established themselves as the gold standard for most computer vision tasks such as image classification and object detection.[2–6] For the specific COVID-19 challenge, some works based on CNNs have been created for various imaging data. Most of them focus on region segmentation and COVID-19 diagnosis.[7–17] For example, Karanam et al[7] designed and developed a contactless patient positioning system for scanning patients in a completely remote and contactless fashion. Chest CT-based COVID-19 case identification or localization segmentation tasks have been studied by other authors.[8–17] Roy et al[17] presented a deep learning method for COVID-19 classification based on lung ultrasound data. Several of them made prediction on 3-dimensional images (sequential CT slice/ultrasound slice).[8,10–13,15–17] In contrast to those existing works, we focus on an identification task in x-ray images, which are 2-dimensional (2D) images, as they are more commonly used in routine clinical visits. As the first-line imaging modality for assessing patients with suspected COVID-19, 2D chest x-ray is less sensitive than chest CT and hence posts a bigger technical challenge. The most related work is COVID-Net, a neural network architecture for x-ray–based COVID-19 cases detection.[18] We added COVID-Net[18] and AI-COVID[8] as baselines in the experiments.

Moreover we adopted several state-of-the-art CNN architectures as base models in FLANNEL for x-ray–based COVID-19 detection. Owing to the limited number of COVID-19 x-ray images that exist in training data, each of these models alone could suffer from high error and variance. In order to solve this issue and improve reliability, we utilize the power of ensemble methods.

Ensemble models

Ensemble learning is a process of combining different independent classifiers, called base learners, to increase the robustness and accuracy of the final classification. As long as these base learners are diverse and able to capture the patterns in the data independently, the ensemble model is able to generalize significantly better by reducing individual errors.[19] Instead of training a complex single neural network with a large number of layers and parameters, decomposing the architecture into smaller and simpler individual base models has been shown to be more accurate and require less time and memory for training.[20,21] The individual base learners in neural network ensembles capture patterns over different regions or granularities of the input space,[22] which contributes to its higher overall performance. Instead of relying on traditional methods of ensemble voting such as bagging, our proposed model introduces a neural ensemble layer to adopt a heterogeneous ensemble strategy that uses different state-of-the-art CNNs as base learners to get better accuracy compared with the individual classifiers.

Class imbalance challenge

Class imbalance in datasets is a very common but significant problem that often hinders model performance and generalization. Resampling, via undersampling or oversampling, is a common tactic to solve the imbalance issue.[23–25] However, oversampling tends to be error prone due to overfitting or added noise, and undersampling reduces the amount of training data that the model can learn from,[26] causing both methods to be ineffective in our setting. Focal loss is a special loss function proposed for an imbalance binary classification task, in which the standard cross-entropy loss of the model is reshaped, such that the well-classified examples are down-weighted, and it can focus on learning the hard imbalanced negatives.[27] In this study, we adapt focal loss to our multiple classification task.

MATERIALS AND METHODS

Study design

Our main task is to differentiate between chest x-ray images of COVID-19 patients and chest x-ray images from other pneumonia and healthy patients, which can be considered as a multiple classification task.

We combined 2 publicly available datasets to generate the experimental data: (1) COVID Chest X-ray dataset (https://github.com/ieee8023/covid-chestxray-dataset) and (2) Kaggle Chest X-ray images (pneumonia) dataset (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia). The choice of these 2 datasets for our experiment was guided by the fact that both are open source and fully accessible to the research community and the general public. As these public datasets grow, the model can continue to improve and scale continuously. The detailed statistics are shown in Table 1.

All Kaggle Chest X-ray images and most COVID Chest X-ray images are based on anteroposterior (AP) or posteroanterior (PA) views. For consistency purposes, we only included images with both AP and PA views in our final experimental dataset. The dataset comprises a total of 5508 chest x-ray images across 2874 independent patient cases. Because AP and PA are 2 different types of x-ray views, we introduced horizontal flips and random noise to convert PA view into AP view. This method of data preprocessing allows us to train the same model for both views and allows the model to be view-independent. We used a train-test split ratio of 4:1 to randomly generate the training and testing set to be used in the model. We furthermore used 5-fold cross-validation on training data to acquire 5 resulting models, and then recorded ensemble (average) performance of these 5 models on testing data. The reason we did this instead of 1-time standard train, validation, and test splits is that we wanted to maximize the sample usage because of the limited samples in this study. The detailed statistics of experimental data are shown in Table 1.

Table 1. Experimental data description

Source		Total	COVID-19	Viral	Bacterial	Normal
Original data	CCX data	119	100	11	7	1
	KCX data	5389	0	1492	2780	1117
View distribution	AP view	5413	24	1492	2780	1117
	PA view	95	76	11	7	1
Training/test splits	Training	4406	77	1194	2237	898
	Testing	1102	23	309	550	220
	Total	5508	100	1503	2787	1118

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

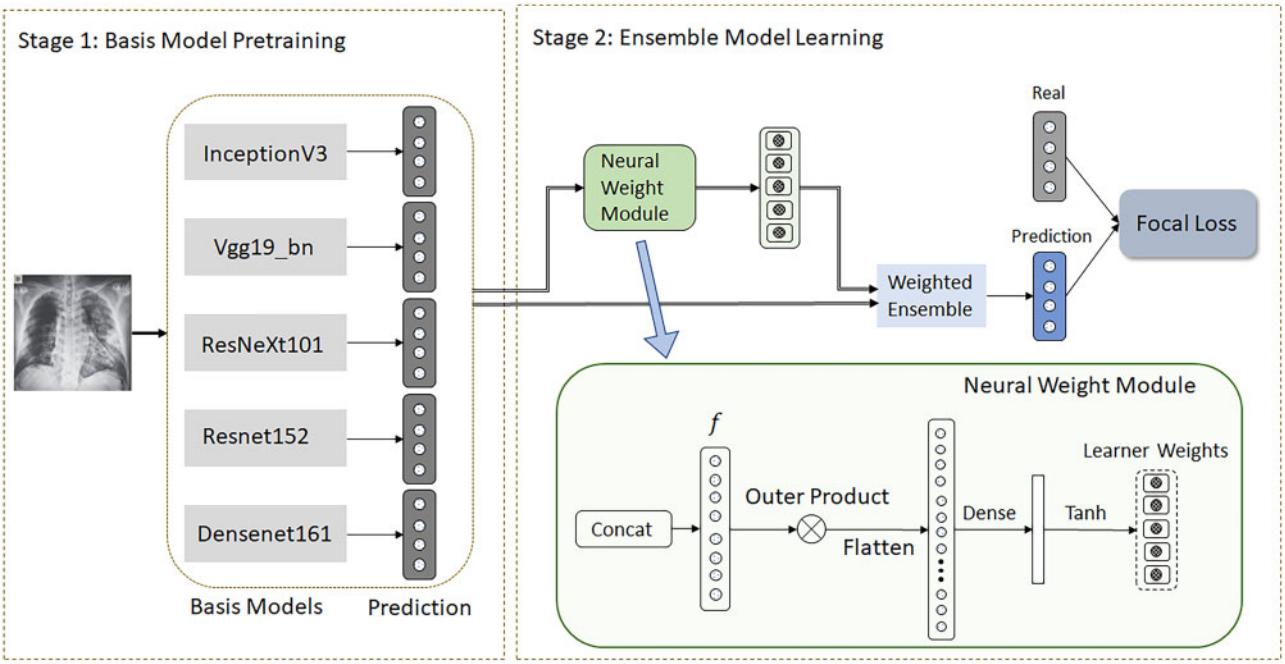

Figure 1. Framework of FLANNEL (Focal Loss bAsed Neural Network EnsembLe).

For image preprocessing, we resized the original input images first to size 256 × 256 and then randomly cropped them in the center to size 224 × 224. The original x-ray images had some symbols in the corners correlated with our labels that could lead to feature leakage, so we introduced a crop around the center to mask that. The resolution size of 224 × 224 was chosen, as it is the default input size for all of our base models. Although this downsampling in resolution may lead to loss of image details, it allows a simpler model that is easier to train and less prone to overfitting. Most state-of-the-art CNNs adopt this resolution for the same reason. The data are then augmented using random flip and noise and normalized before entering the model.

From Table 1, we can find that COVID-19 x-ray images are very rare compared with that of non–COVID-19 pneumonia because it is such a new disease with limited data.

Methods

As shown in Figure 1, our proposed FLANNEL model is composed of 2 stages. First, several base learners are independently trained for COVID-19 classification. Second, the resulting weights from all base learners are used to train the overall ensemble model. We also provide pseudocode in Table 2 to elaborate on the algorithm in detail.

Stage 1: Base learner training

CNNs have been widely used in image classification and get huge successes. Here, we choose 5 popular and state-of-the-art CNN classification models as base learners to model the COVID-19 identification task. These following models were chosen due to their flexibility and high performance with general image classification.

1. Inception v3: It is the third edition of Google's Inception CNN.[3,28]
2. VGG19-bn: The model architecture is from VGG group with batch normalization and consists of 19 layers.
3. ResNeXt101: This 101-layer architecture is designed by the ResNeXt group.
4. Resnet152: This is a 152-layer Deep Residual Neural Network that learns the residual representation functions.
5. Densenet161: This is a Densely Connected Convolutional Network with 161 layers.[29]

Table 2. 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
Base learners					
InceptionV3	0.5904 (0.27)	0.5864 (0.05)	0.8056 (0.01)	0.8771 (0.04)	0.7149 (0.09)
Vgg19_bn	0.6160 (0.06)	0.5349 (0.04)	0.7967 (0.02)	0.8691 (0.03)	0.7042 (0.02)
ResNeXt101	0.6378 (0.12)	0.5649 (0.03)	0.7959 (0.01)	0.8537 (0.02)	0.7140 (0.03)
Resnet152	0.6277 (0.11)	0.5506 (0.02)	0.7988 (0.01)	0.8700 (0.01)	0.7110 (0.03)
Densenet161	0.6880 (0.07)	0.5930 (0.02)	0.8017 (0.01)	0.8953 (0.01)	0.7445 (0.02)
Additional baselines					
COVID-Net [20]	0.7179 (0.13)	0.5592 (0.04)	0.8095 (0.02)	0.8787 (0.03)	0.7413 (0.03)
AI-Covid [22]	0.6391 (0.16)	0.5238 (0.07)	0.7504 (0.02)	0.7223 (0.03)	0.6589 (0.07)
Ensemble learning					
Voting	0.7684 (0.04)	0.6005 (0.03)	0.8214 (0.03)	0.9079 (0.01)	0.7745 (0.01)
Ensemble_MLP_l1	0.6247 (0.07)	0.6042 (0.03)	0.8185 (0.01)	0.9161 (0.01)	0.7409 (0.02)
Ensemble_MLP_l2	0.3735 (0.23)	0.6030 (0.02)	0.8206 (0.00)	0.9128 (0.01)	0.6775 (0.05)
Variant FLANNEL					
FLANNEL_w/o_Focal	0.7671 (0.06)	0.6001 (0.03)	0.8238 (0.01)	0.9135 (0.01)	0.7761 (0.01)
FLANNEL_w/o_Focal + Sampling	0.7837 (0.04)	0.5953 (0.04)	0.8245 (0.01)	0.9131 (0.01)	0.7791 (0.02)
FLANNEL	0.8168 (0.03)	0.6063 (0.02)	0.8267 (0.00)	0.9144 (0.01)	0.7910 (0.01)

The values in parentheses are the standard deviations.

Owing to the limited amount of training data of x-ray images, we use the listed pretrained models from the ImageNet Large Scale Visual Recognition Challenge (http://www.image-net.org/challenges/LSVRC/) and fine-tune each model with respect to the COVID-19 identification task. For fine-tuning, all parameters of the models were retrained with no layer being frozen. This method of fine-tuning models pretrained on a general image set speeds up the training process and also helps with generalization. We modify the last classification layer for each base learner such that it produces a 4-length vector for 4-way classification.

Stage 2: Ensemble model learning

We then take all the N (number of pretrained base learners and here is 5) M-length vectors (denoted as $P_i \in \mathbb{R}^M,\ i = 1, \ldots, N,$ where $\mathbb{R}$ represents a real number) and simply concatenate them (denoted as f) and feed it into the neural weight module to learn base learner weights, as shown in Figure 1.

For the neural weight module, we first construct feature interaction via outer production ff^T to capture more latent information. This is then flattened and fed through a Dense and Tanh layer to map features into base learner weights. Because the outputs of the Tanh function can output negative values, such weights allow the model to discount inaccurate predictions of the classifier.

We then take the linear combination of the base learner predictions using the resulting base learner weights (denoted as w) to get final prediction $\hat{y} = Softmax(\sum_{i=1}^{N} w_i P_i)$. Rather than relying on traditional methods of ensemble voting, we allow FLANNEL to self-learn the optimal combination between the output of the base learners. The neural weight module could be easily extended to accommodate more complex output formats of the base learners.

Then we define loss function for model training. The standard loss function used for training multiclass neural networks is cross-entropy loss as following,

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

where M represents the number of classes. Both $\hat{y} \in R^M$ and $y \in \{0,\ 1\}^M$ are M-length vectors and represent the predicted value and ground truth of class distribution, respectively.

However, heavy class imbalances in the data during training will overwhelm and dominate the gradient, making optimal parameter updates difficult. Focal loss is a loss function proposed for binary classification tasks, in which the well-classified examples are down-weighted and can focus on learning the hard imbalanced examples.[27] Here, we extend focal loss to multiclass classification in our model to address these imbalance issues. For each image, we define the focal loss as:

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

where $(1 - \hat{y}_m)^{\gamma}$ is a modulating factor with a tunable focusing parameter γ and α_m represents a weight factor vector that balances the importance of the different classes. When an example is misclassified and $\hat{y}_m$ is small, the modulating factor is close to 1 and the loss is unaffected. As $\hat{y}_m \to 1$, the factor goes to 0 and the loss for well-classified examples is downweighted. γ can be adjusted to tune the rate of this downweighting. As it increases, the effect of the modulating factor is likewise increased (we found $\gamma = 3$ to work best in our experiments). For α_m, we can set the bigger value for the minority class and the smaller value for the majority class to make all classes contribute equally to our loss. When $\{\alpha_m = 1,\ m = 1 \ldots M\}$ and $\gamma = 0$, focal loss is equivalent to cross-entropy loss. In our experiments, α_m is set to be inverse class frequency of each class. The resulting loss is then back-propagated to update the weights for the neural weighing module in the ensemble. During this stage, the parameters of the 5 base learners are frozen and not updated.

RESULTS

Baseline models for performance comparison

First, we compare FLANNEL with the chosen 5 base learners of FLANNEL framework (Resnet152, Densenet161, InceptionV3, VGG19-bn, and ResNeXt101). All of these models were fine-tuned using their default parameter settings and by using the Adam optimizer.[30]

Furthermore, we also compare FLANNEL with 2 recent COVID-19 deep learning models, COVID-Net[18] and AI-COVID.[8] COVID-Net is a tailored deep convolutional neural network design

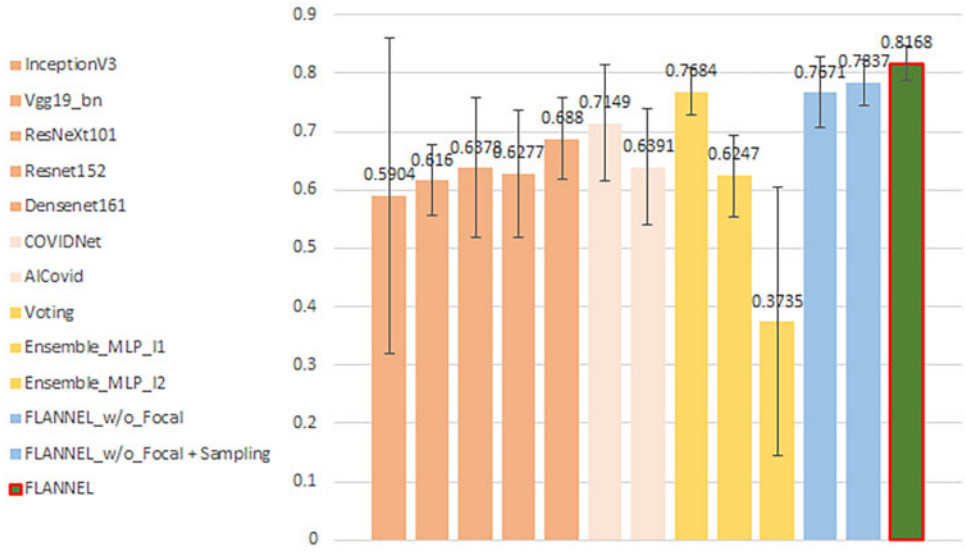

Figure 2. Illustrates the COVID-19 (coronavirus disease 2019) F1 score vs rest comparing different models. The error bars are from 5-fold cross-validation.

for COVID-19 cases detection. AI-COVID presented a deep network framework for sequential CT slice imaging, and hence we just use slice-level classification networks for our identification task because we only have 1 x-ray image per sample in our dataset.

To verify the advantages of FLANNEL on ensemble learning, we also selected 2 traditional ensemble strategies voting and stacking to ensemble the 5 chosen base learners. These will be denoted as Voting and Ensemble_MLP. For Ensemble_MLP, the concatenated prediction values from base models are fed into the multilayer perceptron for final prediction. Here, we use separately Ensemble_MLP_l1 for 1-layer MLP and Ensemble_MLP_l2 for 2-layer MLP.

To verify the advantages of FLANNEL for the imbalanced datasets, we compare it to a version of FLANNEL that replaces the Focal Loss with multiclass standard cross-entropy loss (denoted as FLANNEL_w/o_focal). Another comparison model adds on top of FLANNEL_w/o_focal and utilizes resampling strategies commonly used for class balancing (denoted as FLANNEL_w/o_focal_sampling).

Implementation details

All the base models and FLANNEL are implemented in PyTorch and trained on 3 NVIDIA Tesla P100 GPUs (NVIDIA, Santa Clara, CA) over 200 epochs. The 5 base models (InceptionV3, Densenet161, Resnet152, ResNeXt101, and Vgg19_bn) are fine-tuned using the respective pretrained models with the default model architecture. The data are augmented with random flips, crops ,and scaling during the fine-tuning process.

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

Evaluation strategy

In order to overall verify the prediction accuracy, we first measure the overall accuracy of the model in distinguishing the 4 classes (COVID-19 viral pneumonia, non–COVID-19 viral pneumonia, bacterial pneumonia, and normal images). The main intention of the study is the detection of COVID-19 among kinds of respiratory-related x-ray images. For each class of images, the classification metric F1 score, which conveys the balance between the precision and the recall, is recorded.

Experimental results

In this section, we present the experimental results to show the performance of our proposed FLANNEL and all baseline methods.

First, we note that overall accuracy is not a great metric for evaluation the model. Because the classes in our dataset are heavily unbalanced in favor of non–COVID-19 pneumonia images, even a significant increase in COVID-19 detection performance will not affect overall accuracy very much. Therefore, we present the F1 score for COVID-19 vs the rest, comparing different models in Figure 2. Obviously, it shows that FLANNEL outperforms other state-of-the-art models in detecting COVID-19 cases.

Moreover, we also present F1 score for each disease classification and macro-F1 score for all classes of the label set, which is shown in Table 2. From Table 2, we can observe that COVID-Net achieves better performance than other baseline models, as it is

Box 1. The FLANNEL algorithm

Algorithm 1 FLANNEL Training

Input:
X-ray Images, Class Labels
Base Models $\{\text{Learner}_1, \text{Learner}_2, \ldots, \text{Learner}_n\}$
(Define B as batch size)
Stage 1:
Fine-tune all base models with respect to inputs images and labels.
Stage 2:
For each batch ($X \in \mathrm{R}^{B\times 1\times 224\times 224}$, $Y \in \mathrm{R}^{B\times 4}$) from inputs images and labels do
–> Step1: Get prediction values from all Base Models
$P_i = Learner_i(X) \in \mathrm{R}^{B\times 4}$, where $i = 1, \ldots, n$
–> Step2: Get learner weights
$W = NeuralWeightModule([P_i,\ i = 1, \ldots,\ n]) \in \mathrm{R}^{B\times 5}$
–> Step3: Linear Combination for Prediction
$\hat{Y} = Softmax(\sum_{i=1}^{n} W_i P_i) \in \mathrm{R}^{B\times 4}$
(where W_i represents i-th column of W)
–> Step4:
$\text{Loss} = FocalLoss(\hat{Y},\ Y)$
Back-propagate on Loss and update parameters
End For

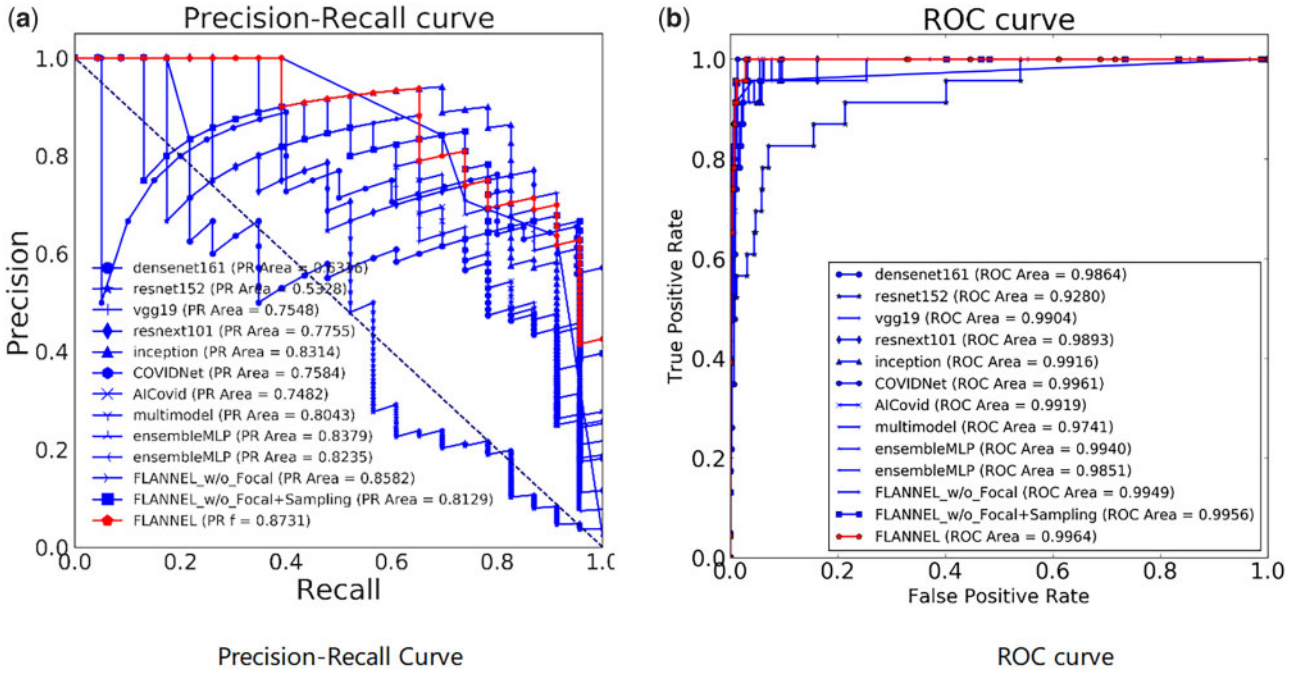

Figure 3. The diagnostic ability of the compared models for COVID-19 (coronavirus disease 2019) classification. Panel A shows the precision-recall curve, while Panel B shows the receiver-operating characteristic curve. FLANNEL: Focal Loss bAsed Neural Network EnsembLe; PR: precision-recall; ROC: receiver-operating characteristic.

specifically designed for COVID-19 detection. AI-COVID does not show apparent performance improvement on COVID-19 detection. It is probably because AI-COVID mainly focuses on fast features extraction followed by multislice ensemble for final 3-dimensional CT image prediction, not specifically for 2D image classification. A multilayer perceptron–based ensemble strategy can improve detection performance on other majority classes but has poor performance on minority class (COVID class). Further, COVID-19 detection accuracy degrades with employment of more layers. More fully connected layers cannot effectively model limited features (concatenated outputs from base learners) specifically for this class-imbalance challenge. From Table 2, FLANNEL shows a higher macro-F1 score, with a 2% increase, especially a 6% relative increase for COVID-19 cases over the best-performing baseline model.

Compared with FLANNEL_w/o_Focal just using cross-entropy loss instead of focal loss, FLANNEL shows an increase of almost 6% in F1 score for COVID-19 cases. It is also clear that resampling strategies help improve performance in case of class imbalance. Traditional resampling strategies (FLANNEL_w/o_Focal + Sampling) increased F1 score for COVID-19 cases by almost 2%, while still almost 4% behind compared with FLANNEL. Also most importantly, FLANNEL was able to increase the performance of COVID-19 classification without negatively impacting the performance for other classes.

In order to understand the trade-off in COVID-19 classification performance for different threshold values, we also introduce the plot curves. A popular metric for classification systems is the receiving-operating characteristic (ROC), which can be summarized by its area under the curve. Because ROC curves can be misleading when the class distribution is imbalanced, we also show the precision-recall (PR) curve. The focus of the PR curve on the minority class deems it an effective diagnostic for imbalanced binary classification models. The experimental results are shown in Figure 3, in which the PR curve and ROC curve are separately used to show the diagnostic ability of compared models.

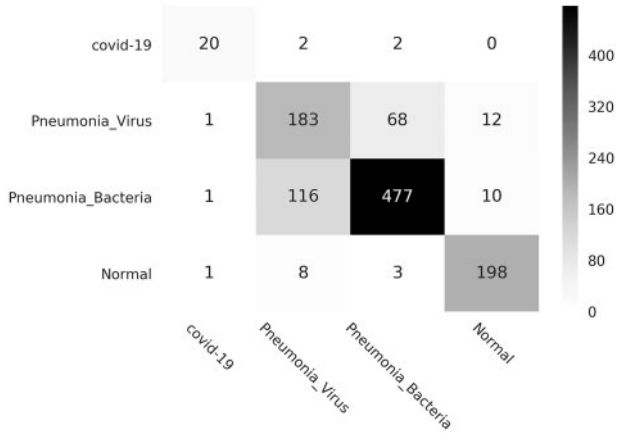

Figure 4. Confusion matrix (instance counts). COVID-19: coronavirus disease 2019.

Finally, we also provide the visual description of FLANNEL performance, via confusion matrix, shown in Figure 4. It shows that (1) the cases predicted as pneumonia are mainly from the GroundTruth pneumonia-related cases and (2) for COVID-19 identification, FLANNEL has higher precision and recall than the other 2 types of pneumonia. It means FLANNEL can distinguish pneumonia images from normal images and differentiate chest x-ray images of COVID-19 against other pneumonia images.

In order to verify model performance under single-view condition, we show performance evaluation (specifically for F1 score for COVID-19 vs the rest) on AP-view x-ray images in Figure 5. From Figure 5, we find that all models have performance degradation because of heavier imbalance, with fewer samples with the COVID-19 label. Nevertheless, our proposed FLANNEL still has the best F1 score compared with all the baselines.

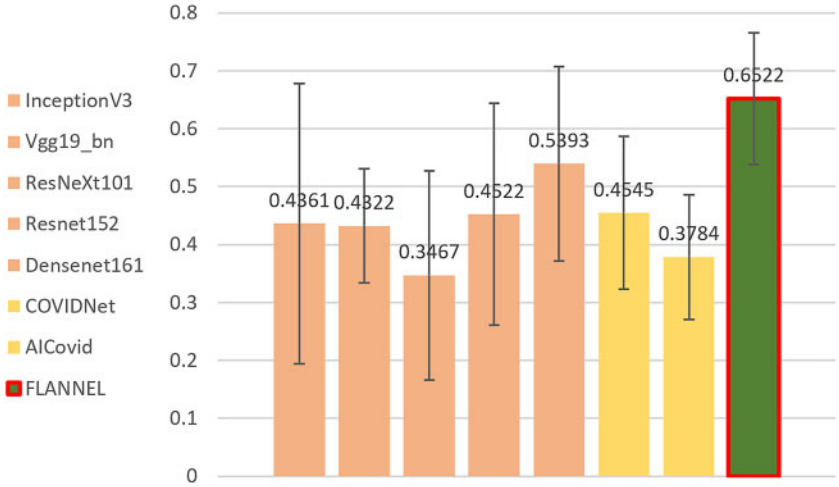

Figure 5. Performance verification (F1 score for COVID-19 [coronavirus disease 2019] vs rest) on anteroposterior view images comparing different models. FLANNEL: Focal Loss bAsed Neural Network EnsembLe.

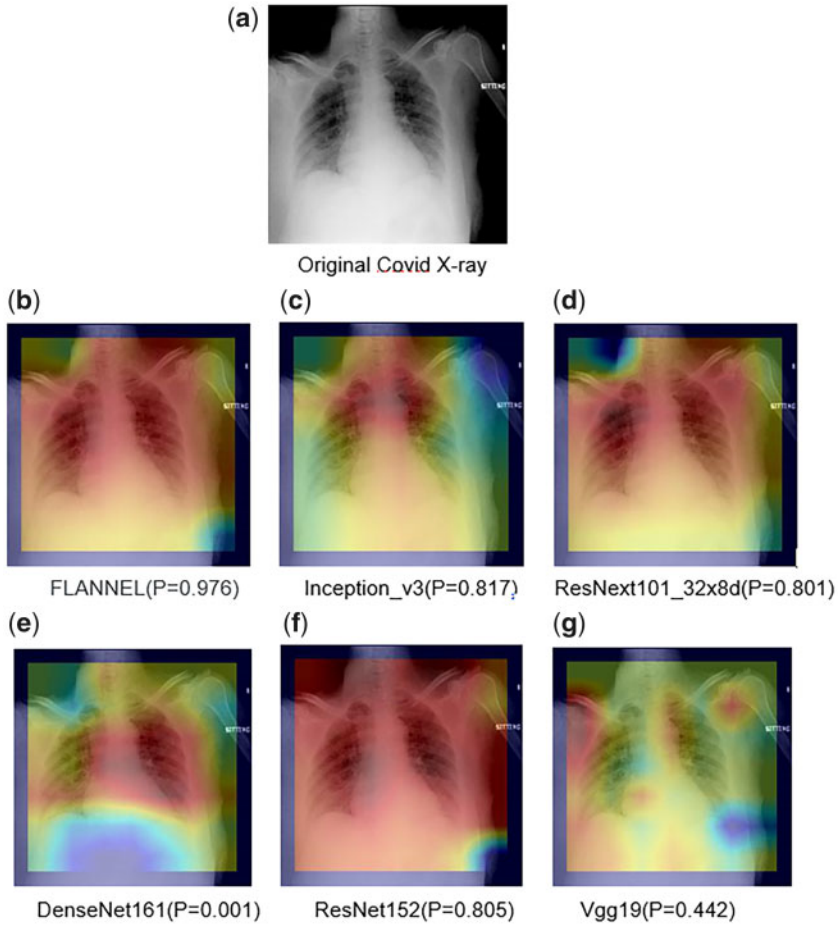

Figure 6. Visual explanations for the classification results. FLANNEL: Focal Loss bAsed Neural Network EnsembLe;

Then, Figure 6 presents the visual explanations for the classification results. We use class activation mapping to visualize the models' attention over inputs. The visual results are shown subsequently. Here, we show each of the 5 base models' visual attention heatmaps. We observe that FLANNEL can pay more attention to specific regions in base model parameters, which allows it to make the correct overall prediction.

DISCUSSION

The most notable observation from the base models is that most of them perform poorly at detecting the pneumonia caused by viruses (both COVID-19 and non–COVID-19). This poor performance is likely due to the fact that (1) there are more bacterial and normal

than viral images in our dataset and (2) the addition of the COVID-19 x-ray images makes it more difficult to differentiate between the 2 different classes of viral pneumonia because they are more similar to each other than they are to others. Our proposed FLANNEL model has the possibility of differentiating chest x-ray images of COVID-19 against those of other pneumonia and healthy patients. Moreover, the significant feature of the proposed FLANNEL model is that it manages to sharply enhance COVID-19 identification performance without decreasing that of other classes. In fact, FLANNEL had the highest F1 score of every single category except the normal class.

However, owing to the limited specific COVID-19 pneumonia–related images, it is inevitable to combine multiple source data from different datasets to construct experimental data as described in detail in the Materials and Methods. Even standard x-ray images look slightly different depending on the data source. However, it is still a challenge to conduct source agnostic image classification that utilizes big-scale data from multiple data sources to learn class features and mapping rules to infer the classification task without considering the different data source impacts. Source-free classification for COVID-19 is future work that might significantly improve the performance of the model.

Another direction to consider is to incorporate hierarchical classification. In our case, we can consider first separating healthy from sick patients based on their x-rays, then further separating bacteria, viral pneumonia, and COVID-19 among the sick. The naïve approach requires training multiple models at different levels, which can create challenges at the granular level (eg, viral pneumonia vs COVID-19) due to the small sample size. Multiple models also introduce more computation and maintenance challenges in the long run. We consider this topic as an important future work.

CONCLUSION

COVID-19 is an acute resolved pneumonia disease with high fatality rate. As the most commonly ordered imaging study for patients with respiratory complaints, x-ray can help COVID-19 diagnosis. Hence, it is significant to study automatic diagnosis of the disease. Our objective is to explore effective deep learning methods to model x-ray images and improve prediction performance in identifying COVID-19 from other pneumonia images and healthy images.

With the power of ensemble learning, FLANNEL has the ability to detect and diagnose COVID-19 from pneumonia x-ray images with high accuracy, even when trained on just around 100 available COVID-19 x-ray images. We have shown that it is able to automatically combine and use the outputs of individual base learners as features to create a more accurate global model. Focal loss allows us to use all the training samples effectively without having to sample our already limited imbalanced dataset and solves the imbalance problem that hinders other traditional loss functions such as cross-entropy loss. FLANNEL vastly outperforms all other state-of-the art CNN architectures, especially on the COVID-19 detection, without much added model complexity or parameters.

This model could be used to supplement current COVID-19 diagnosis kits to improve testing availability and alleviate supply shortages through just examining x-ray images. With millions of confirmed cases only expected to grow in the future, neural networks could make a real impact in limiting the spread of the disease.

AUTHOR CONTRIBUTIONS

ZQ implemented the method and conducted the experiments. All authors were involved in developing the ideas and writing the paper.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

REFERENCES

1. Wong H, Lam H, Fong A, *et al.* Frequency and distribution of chest radiographic findings in COVID-19 Positive Patients. *Radiology* 2020; 296 (2): E72–8.
2. Krizhevsky A, Sutskever I, Hilton G. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Advances in Neural Information Processing Systems 25 (NIPS 2012). New York, NY: ACM Press; 2012.
3. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*: 1409.1556; 2015.
4. Lin M, Chen Q, Yan S. Network In Network. *arXiv*: 1312.4400; 2014.
5. He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition. In: proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016.
6. Xie S, Girshick R, Dollar P, Tu, Z, He, K. Aggregated residual transformations for deep neural networks. *arXiv*: 1611.05431; 2017.
7. Karanam S, Li R, Yang F, *et al.* Towards contactless patient positioning. *IEEE Trans Med Imaging* 2020; 39 (8): 2701–10.
8. Bai X, Wang R, Xiong Z, *et al.* Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology* 2020; 296 (3): E156–65.
9. Oh Y, Park S, Ye J. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans Med Imaging* 2020; 39 (8): 2688–700.
10. Wang X, Deng X, Fu Q, *et al.* A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Trans Med Imaging* 2020; 39 (8): 2612–25.
11. Mei X, Lee H, Diao K, *et al.* Artificial intelligence-enabled rapid diagnosis of COVID-19 patients. *Nat Med* 2020; 26 (8): 1224–8.
12. Ophir G, Maayan F, Hayit G, *et al.* Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. *arXiv:* 2003.05037; 2020.
13. Zhang K, Li X, Shen J, *et al.* Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* 2020; 182 (5): 1360.
14. Wang G, Liu X, Li C, *et al.* A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. *IEEE Trans Med Imaging* 2020; 39 (8): 2653–63.
15. Kang H, Xia L, Yan F, *et al.* Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. *IEEE Trans Med Imaging* 2020; 39 (8): 2606–14.
16. Ouyang X, Huo J, Xia L, *et al.* Dual-sampling attention network for diagnosis of COVID-19 from CAP. *IEEE Trans Med Imaging* 2020; 39 (8): 2595–605.
17. Roy S, Menapace W, Oei S, *et al.* Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Trans Med Imaging* 2020; 39 (8): 2676–87.
18. Wang L, Qiu Z, Wang A, *et al.* COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. *arXiv: 2003.09871;* 2020.
19. Zhang M, Zhou Z. Exploiting unlabeled data to enhance ensemble diversity. *Data Min Knowl Disc* 2013; 26 (1): 98–129.
20. Hansen L, Salamon P. Neural network ensembles. *IEEE Trans Pattern Anal Machine Intell* 1990; 12 (10): 993–1001.
21. Sharkey A. On combining artificial neural nets. *Connect Sci* 1996; 8 (3–4): 299–313.

22. Brown G, Wyatt J, Tino P. Managing diversity in regression ensembles. *J Machine Learn Res* 2005; 6: 1621–50.
23. Shen L, Lin Z, Huang Q. Relay backpropagation for effective learning of deep convolutional neural networks. *arXiv*: 1512.05830; 2016.
24. Geifman Y, El-Yaniv R. Deep active learning over the long tail. *arXiv: 1711.00941;* 2017.
25. Zou Y, Yu Z, Kumar B, *et al*. Domain adaptation for semantic segmentation via class-balanced self-training. *arXiv*: 1810.07911; 2018.
26. He H, Bai Y, Garcia E, *et al*. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: proceedings of 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008.
27. Lin T, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. *arXiv*: 1708.02002; 2017.
28. Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. In: proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016.
29. 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.
30. Kingma D, Ba J, Adam A. Method for stochastic optimization. In: ICLR 2015: International Conference on Learning Representations; 2015.