Weighted Cross-Entropy for Unbalanced Data with Application on COVID X-ray images

Özgür Özdemir
Dept. Computer Engineering
Istanbul Bilgi University
Istanbul, Turkey
ozgur.ozdemir@bilgi.edu.tr

Elena Battini Sönmez
Dept. Computer Engineering
Istanbul Bilgi University
Istanbul, Turkey
elena.sonmez@bilgi.edu.tr

***Abstract*—Since December 2019 the world is infected by COVID-19 or Coronavirus disease, which spreads very quickly, out of control. The high number of precautions for laboratory access, which need to be taken to contain the virus, together with the difficulties in running the gold standard test for COVID-19, result in a practical incapability to make early diagnosis. Recent advances in deep learning algorithms allow efficient implementation of computer-aided diagnosis. This paper investigates on the performance of a very well known residual network, ResNet50, and a lightweight Atrous CNN (ACNN) network using a Weighted Cross-entropy (WCE) loss function, to alleviate imbalance on COVID datasets. As a result, ResNet50 model initialized with pre-trained weights fine-tuned by ImageNet dataset and exploiting WCE achieved the state-of-the-art performance on COVIDX-Ray-5K test set, with a top balanced accuracy of 99.87%.**

***Keywords*—Coronavirus; COVID-19; Deep Learning; Automatic Diagnosis; Weighted Cross-Entropy; Loss functions; Classification.**

I. Introduction

Since December 2019, COVID-19 spread rapidly from Wuhan, China, to the rest of the word. The Web page of the World Health Organization (WHO) [1] gives numbers at a glance: the total number of confirmed cases was 14,007,791, the confirmed fatalities were 597,105 for a total of 216 countries affected by the COVID disease, at the 19 of July 2020. The Web page is regularly updated.

Currently, Real Time Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) seems to be the most possible accurate test for diagnosis of COVID-19; however, RT-PCR is a time-consuming test and it requires a large quantity of RNA for detection [2].

Considering those shortcomings together with all restrictions recently imposed to access laboratories due to the risk to get infected by Coronavirus, the test RT-PCR it practically impossible in many realities.

That is, the current emergency together with the risk of further spread highlight the need of an automatic system capable to distinguish between COVID and non-COVID patients. This paper proposes a new algorithm for Computer-based Automatic Diagnosis (CAD) system, capable to support frontline doctors for early diagnosis.

Chest X-ray images, together with Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), are among the most used images in the medical field for therapeutic diagnosis. Every type of image has its own pros and cons; chest X-ray is the simple and cheap one, it can be acquire also with a portable device at the patients home; overall, since X-ray crosses the human body, it allows to analyse the internal structure of the body, without surgery.

The recent success of machine learning permits to model completely automatic systems for medical image analysis. However, automatic classification of medical images with deep learning algorithms requires a high number of images, per class, which may be not available. This paper focuses on COVID-19, with a case study on X-ray images. The subject is challenging due to the little number of available COVID images, together with a large intra-class and a small inter-class variance in the images.

The main contribution of this study can be summarizing as: (1) it introduces an extended version of COVIDX-Ray-5K training set by combining it with another publicly available dataset, i.e. COVID-ChestXray, (2) it proposes to use a lightweight network architecture Atrous CNN (ACNN) for medical image classification, (3) it underlines the imbalance on COVID datasets and provides a solution by exploiting Weighted Cross-entropy (WCE) loss function, (4) it reaches the state-of-the-art for the experiment on COVIDX-Ray-5K database introduced by [3].

The remainder of the paper is organized as follows. Section 2 introduced the previous works. Section 3 overviews the publicly available COVID datasets. Section 4 describes the methodology used in this study. Section 5 provides details about experiments and discusses the results. Lastly, conclusions are drawn in Section 6.

II. Related Work

The Coronavirus disease does not have specific clinical exhibition [4], however several researchers all over the word are struggling to contribute for vanquishing the virus. In the research areas resulting from the intersection between computer vision and machine learning, medical images of COVID and non-COVID patients are challenged to create computer-based automatic diagnosis systems.

Zu et al. [4] published one of the first study on Computed Tomography (CT) findings of COVID patients, which proves

that thin-slide chest CT allows early detection and tracking of the disease. Overall, CT is recommended because it is sensitive to the detection of ground-glass opacity, however, chest X-ray shows multifocal patchy opacity in both lungs, which may also be present in COVID-19 patients.

Zheng et al. [5] presented an accurate and rapid diagnosis system for COVID-19 suspected cases using CT images with weak labelling. For each patient, the lung region was segmented in a semi-automatic way; after that, the segmented areas were fed into a 3D deep neural network to predict for coronavirus. Results on their private database are promising.

The work of Tartaglione et al. [6] made several experiments by merging existing databases of COVID-19. It investigated on the possibility to use common and cheap Chest X-ray for COVID-19 early prediction. The proposed deep learning model uses (1) histogram equalization to make Chest X-ray pre-processing, (2) a U-net [7] for lung segmentation, and (3) a pre-trained deep Convolutional Neural Network (CNN) for classification.

Sethy et al. [8] used deep feature and Support Vector Machine (SVM) [9] for classification of COVID-19, on X-ray images available at Kaggle and GitHub. The performance of SVM using deep features produced by several deep learning models is compared, and the best accuracy is achieved by ResNet50 [10] plus SVM. The classification issue is posed as 3 class problem, it considers COVID-19 and common pneumonia patients, and healthy people; all X-ray images are collected from the Internet.

Apostolopoulos et al. [11] compared the performance of several deep learning models using a different database of X-ray images collected from GitHub and public repositories. They run both the 3 class problem, which considers (COVID-19, pneumonia, normal) X-ray images and the 2-class problem, considering only Covid and Non-Covid images. Best classification accuracy was reached by the VGG19 model [12].

The work of Narin et al. [13] tackled the 2 class problem using a little, homemade, database of 50 COVID and 50 NON-COVID X-ray images. They pre-trained ResNet50 model gave the best accuracy over others deep CNN based architectures.

Wang et al. [14] introduced to the research community COVID-Net, a deep CNN modelled for the detection of COVID-19 in X-ray images. The models reached good performance of the open access benchmark dataset COVIDx, which is presented by the same paper.

Minaee et al. [3] tackled the problem of Coronavirus disease early diagnosis by creating COVIDX-Ray-5K, which is one of the few publicly available datasets of COVID-19 X-ray images. The benchmark algorithm proposed by this paper uses pre-trained models and data augmentation to fine-tune them. The best performance was reached by SqueezeNet [15], with a sensitivity of 97.5%, specificity of 97.7%, balanced accuracy of 97.6% and Predicted Positive Prediction Rate of 97.7%.

This paper challenged the work of [3] by proposing a new CAD system, which improves the current performance on the COVIDX-Ray-5K dataset.

III. Databases

COVIDX-Ray-5K is one of the few datasets of COVID images, introduced by [3]. The database stores a total of 5,000 chest X-ray images, divided into (COVID, Non-COVID) classes; COVID images have been labelled by board certified radiologists[1]. Table I describes the database, which is highly unbalance with a little number of COVID images. The non-COVID images are a selection of images collected from the CheXpert dataset [16], which is described in the following paragraph and reported in Table I. The creators of the dataset paid attention to have all images of the same patient either in the training or in the test set.

COVID-ChestXray dataset [17] stores hundreds of frontal view x-ray images with metadata, such as time since first symptom, survival and intubation status and hospital location[2]. It has a total of 542 frontal chest images from 262 people from 26 countries. Because some of the provided samples from publicly available dataset are CT images, we had to dropout these samples. The details are given in Table I. Furthermore, the database comes with several benchmark algorithms to predict pneumonia severity, survival probability, need for intensive care estimation and the Leave-One-Country/Continent-Out algorithm.

Another publicly available database of chest images of Coronavirus disease is COVIDx [14]. The database stores a total of 13,975 chest X-ray images from 13,870 patients. The database has been collected by Wang et al.; it is the result of the merging of five different publicly available repositories[3]. X-ray images belongs to the 3 classes labelled as COVID-19, pneumonia and normal.

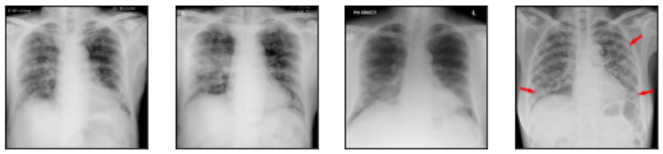
CheXpert is a large dataset storing 224,316 chest radiographs of 65,240 patients [16]. The database focuses on 14 most common pathologies. The validation set of 200 chest radiography from 200 patients was manually annotated by 3 board-certified radiologists, and the ground-truth of the 500 images from 500 patients of the test set has been fixed by the consensus of 5 board-certified radiologists. The database comes with a benchmark paper, which proposes an algorithm for automatic labelling of 5 pathologies. The results of the model were compared against the judgements of 3 additional board-certified radiologists. In 4 out of 5 pathologies the automatic system performed better than human. Table I gives the distribution of images. All images of this database may be used as Non-COVID images, samples of the negative class.

This work uses the COVIDX-Ray-5K, COVID-ChestXray and CheXpert datasets; the use of the COVIDx database is part of our future work. Fig. 1 shows diseased image samples taken from the COVIDX-Ray-5K, ChestXray and CheXpert databases. Table I shows detailed distribution of datasets used in this study. Since the combination of COVIDX-Ray-5K and COVID-ChestXray contains adequate samples for both classes, we did not apply any augmentation. The total number of used train and test samples is detailed in Table I.

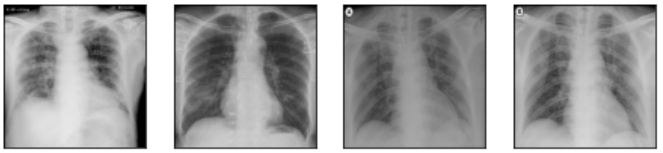
[1]The database is available at https://github.com/shervinmin/DeepCovid.

[2]The database is publicly available at: https://github.com/ieee8023/covid-chestxray-dataset.

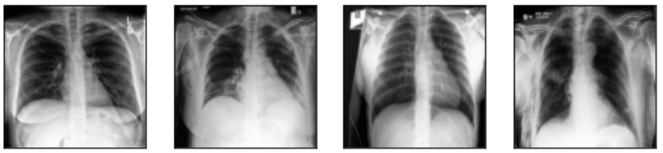
[3]Dataset generation scripts for constructing the COVIDx dataset is available publicly for open access at https://github.com/lindawangg/COVID-Net.



(a) COVIDX-Ray-5K



(b) COVID-ChestXray



(c) CheXpert

Fig. 1. Examples of images with COVID disease from the three datasets used in this study.

Table I. Distribution of images in the datasets. *Due to computational limitations, we had to take a subset of original dataset

Dataset	Training Set		Test Set	
	COVID	Non-COVID	COVID	Non-COVID
COVIDX-Ray-5K (w/out augmentation)	31	2000	40	3000
COVIDX-Ray-5K (with augmentation)	992	2000	40	3000
COVID-ChestXray	512	17	-	-
COVIDX-Ray-5K + COVID-ChestXray	543	2017	40	3000
CheXpert*	1500	1500	-	-

IV. Models

Recent advances in deep learning models overtook the classical machine learning approach of future extraction and classification in favour of models made up of several layers, which extract features at different level of details and make classification. Models are deep networks that transform the input images into a set of features, at different level of details, and classify them. The computational element of a network is a neuron with its activation function. Neurons are connected to each other's and weights are assigned to the connecting links. Neurons in hidden layers receive, as input, a weighted sum, which is filtered thought their activation functions before been outputted. The initial weights assigned to the links are fine-tuned via training, where input images are passed thought the network, classified and errors are back propagated. In Convolutional Neural Network (CNN) the weights are the coefficients of the filters, which convolve the input features. CNNs have become more and more deep to handle complex tasks. The problem of deep CNNs is the training time and the overfitting issue.

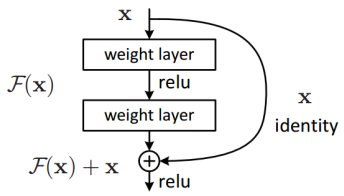


Fig. 2. Single residual block [10]

This issue was first solved by a model called AlexNet [18], which is a Deep Convolutional Neural Networks (DCNN), proposed by Krizhevsky et al., that won ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) competition, in 2012. Following AlexNet several other models have been proposed; this paper focuses on Residual Networks (ResNet) [10] and SqueezeNet [15]; it compares their performance against the Atrous CNN (ACNN) introduced by [19]. The following paragraphs give a general overview of the interesting architectures.

A. ResNet

Residual Networks [10] were introduced with the aim to tackle the vanishing gradient problem. That is, the more a network becomes deep the more it is difficult to train it because the error signal used to change the weights is too little, since it comes from the bottom layers, which are too far away. Residual Networks introduced a new type of building block, called Residual Block, which uses an identity arc to bypass one or more convolutional layers. Fig. 2 show the picture of a residual block.

As result, the Residual Block calculates the difference between the true output of the block, $H(x)$, and its input, x. In formula, $H(x) = F(x) + x$, which can be re-written as $F(x) = H(x) - x$, i.e. $F(x)$ is the residual. The good performance of ResNet confirmed the hypothesis that it is easy to optimize a residual instead of the original mapping.

B. SqueezeNet

The new architecture SqueezeNet [15] was proposed with the aim to reduce the total number of parameters in the network while keeping the same accuracy. This objective was achieved by using three strategies: (1) replacing 3×3 filters with 1×1, (2) using squeeze layers to decrease the number of input channels, and (3) postponing the down sample operation by using strides greater than one only in convolutional layer toward the end of the network.

C. ACNN

Atrous CNN (ACNN) was designed by Zhou et al. [19] as a CNN ad-hoc for medical image segmentation. Atrous or dilated convolution is an alternative to the down-sampling layer. Atrous convolution inserts zeros between non-zeros filters' coefficients to sample the feature map, e.g. a 3×3 filter with dilation rate of 2 will have the same input space of a 5×5 filter while using only 9 coefficients. As the training

parameters reduces, the network becomes more light weight compared to regular convolutional architectures.

In this study, a similar implementation of Zhou et al. is applied, that is atrous blocks that contain two atrous convolution layer with a kernel size of 3×3 and rate of 1 and 3, respectively. The composition of atrous blocks constitutes the ACNN network. This paper tests the performance of ACNN for medical image classification.

V. Algorithms for Loss Functions

A loss function is an algorithm to evaluate how well a network models the input data. In classification the challenge is to assign every inputs signal to the correct class; in case of binary classification there are only two possible output classes, i.e. COVID or Non-COVID.

Possible loss functions for regression are Mean Square Error (MSE), also known as Quadratic Loss or L2 Loss, described in Equation (1) and the Mean Absolute Error (MAE), also known as L1 Loss, detailed in Equation (2). In all formulas, y is the true class, $\hat{y}$ the predicted class, m is the total number of training samples. In case of formula (1), on the left side the is the detailed equation, on the right part the simplified one, to increase the readability; that is, variable y must be read as y_i the same for $\hat{y}$, and the sum is over all training samples i, with $i = 1, ..., m$.

$$MSE(y, \hat{y}) = \frac{1}{2m} \sum\nolimits_i (\hat{y}_i - y_i)^2 \rightarrow MSE = \frac{1}{2m} \sum\nolimits_i (\hat{y} - y)^2 \tag{1}$$

In the following, we will use the simplified notation to increase readability.

$$MAE = \frac{1}{m} \sum \mid \hat{y} - y \mid \tag{2}$$

For classification the above formulas are not convex and, therefore, not usable. That is, since cost functions must be differentiable, it is necessary to insert the logarithm function inside the formula, which is now divided into two part, one for the positive and one for negative samples. The resulting loss functions is called Cross-Entropy (CE) and has the following formula:

$$CE = -(y \log \hat{y} + (1 - y) log(1 - \hat{y})) \tag{3}$$

This work implements a custom loss function, named Weighted Cross-Entropy (WCE), in favour of positive samples. Cross-entropy is among the most widely used types of loss function to train networks. However, training with cross-entropy might not be ideal for unbalanced datasets, because the error on minor class is more likely to vanish since the major class samples dominate the dataset. In this study, we analysed a utilized version of cross-entropy function, namely Weighted Cross-Entropy (WCE), on unbalanced COVID datasets. WCE can be expressed as

$$WCE = -(\beta y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{4}$$

where β is adjustment weights. The quantity of β is either scalar or vector, depends on utilization of cross-entropy function. The quantity becomes scalar when the function is binary; vector if the function is categorical.

Exploiting of adjustment weight allows handicapping either false negative or false positive predictions. The value of $\beta < 1$ penalizes the error on false positive samples, while the value of $\beta > 1$ penalizes the error on false negatives.

VI. Experimental Setup and Results

A. Data Preprocessing

COVIDX-Ray-5K was the first database that we challenged. Given the fact that it is highly dominated with negative samples, data augmentation has been required despite the utilized loss function, WCE. That is, to alleviate the imbalance between positive and negative classes, we employed horizontal flipping and we rotated the images from -5 to +5 degree while skipping 3 degrees for each rotation. Consequently, the training set of COVIDX-Ray-5K comprises 992 COVID and 2,000 Non-COVID X-ray images. The details about the used datasets are given in Table I.

Since the images in the datasets are collected from different sources, they vary in shapes. Therefore, all images are resized to a fixed shape, i.e. 256x256. Moreover, the intensity of the pixel values is fluctuating from image to image. To alleviate the inconsistency of pixel values between images, normalization is applied for each image in the datasets.
All models queried on a test set, which is made up of 40 COVID and 3,000 Non-COVID images provided by Minaee et al. The results of the used algorithms, with configuration details, are given in Table II.

B. Network details and hyper-parameters

Besides the fact that the network weights are randomly initialized for training, we conducted some experiments on using weights from previously trained networks. That is, weights from ResNet50 model [10], which was pre-trained on a dataset of 1000 object, i.e. ImageNet, were exploited to initialize the training. Moreover, the network architectures were fine-tuned on CheXpert dataset to classify healthy or diseased lungs from X-Rays. Although original CheXpert dataset [16] is very big, we used only a subset of it, due to computational limitations.

At training time, the data is split into batches with fixed size of 64, and 20% of the training set is used as validation set, to detect over-fitting. Adam optimizer with learning rate of 0.0002 and exponential decay rate of 0.5 is used for optimizing the loss function.

A fully connected network (FCN) with 7000 trainable parameters is employed for producing likelihoods of predictions. Despite memory overhead, the true labels are constructed of the sparse vector representations to avoid fine-tuning a threshold value for regression. Therefore, the last layer of FCN, that is softmax, constitutes of 2 neurons indicating the binary classes.

Since ResNet50 [10] contains 16 residual blocks, we set the block size to 16 also for ACNN model, so as to compare the architecture sizes. Consequently, ResNet50 consists of 23M trainable parameter, whereas ACNN model has only 8M parameters.

Table II. Results of used algorithms with their configurations

Training set	Configurations			Results				
	Model	**Pre-training**	**β (Eq. (4))**	**Sensitivity**	**Specificity**	**bACC**	**PPCR**	**Gain(%)**
COVIDX-Ray-5K (with augmentation)	ResNet50	-	-	0.9000	0.9993	0.9497	0.9980	-0.30%
	ResNet50	-	0.75	0.8000	1.000	0.9000	0.9973	-2.88%
	ResNet50	ImageNet	-	0.7500	1.000	0.8750	0.9967	-4.19%
	ResNet50	ImageNet	0.75	0.7500	1.000	0.8750	0.9967	-4.19%
	ResNet50	ChestXpert	-	0.825	0.9993	0.9121	0.9970	-2.27%
	ResNet50	ChestXpert	0.75	0.6667	0.9987	0.8327	0.9908	-6.65%
	ACNN	-	-	0.7500	0.9983	0.8742	0.9951	-4.31%
	ACNN	-	0.75	0.7000	0.9997	0.8498	0.9957	-5.53 %
	ACNN	ChestXpert	-	0.9250	0.9690	0.9470	0.9684	-1.95%
	ACNN	ChestXpert	0.75	0.8000	0.9993	0.8997	0.9967	-2.92%
COVIDX-Ray-5K (w/out augmentation) + COVID-ChestXray	ResNet50	-	-	0.9500	0.9397	0.9448	0.9398	-3.53%
	ResNet50	-	0.75	0.9500	0.9977	0.9738	0.9970	+0.89%
	ResNet50	ImageNet	-	1.000	0.9963	0.9982	0.9964	+2.10%
	ResNet50	ImageNet	0.75	**1.000**	**0.9973**	**0.9987**	**0.9974**	**+2.18%**
	ResNet50	ChestXpert	-	0.9750	0.9680	0.9715	0.9681	-0.74%
	ResNet50	ChestXpert	0.75	0.9500	0.9967	0.9733	0.9961	+0.81%
	ACNN	-	-	1.000	0.8653	0.9327	0.8671	-7.87%
	ACNN	-	0.75	1.000	0.9776	0.9645	0.9299	-3.03%
	ACNN	ChestXpert	-	1.000	0.7730	0.8865	0.7760	-14.90%
	ACNN	ChestXpert	0.75	0.9750	0.9783	0.9767	0.9783	+0.08%
COVIDX-Ray-5K	ResNet50 [3]			0.9750	0.9010	0.9380	0.9020	-
	Squeezenet [3]			0.9750	0.9773	0.9762	0.9773	-

C. Evaluation Metrics

Due to the imbalance of the data, sensitivity and specificity metrics are more informative to evaluate the performance of the applied algorithms rather than the metrics widely used for classification like accuracy, precision[4], and F1-score. In other words, we considered also Balanced Accuracy (bACC) and Predicted Positive Condition Rate (PPCR), as they are two informative measurements in case of classification of imbalance data [20]. Sensitivity, specificity, bACC, and PPCR are, respectively defined as:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$bACC = \frac{Sensitivity + Specificity}{2}$$

$$PPCR = \frac{TP + FP}{TP + FP + TN + FN}$$

where TP stands for true positive samples, TN stands for true negative samples, FP stands for false positive samples, and FN stands for false negative samples. The test results of trained networks summarized in Table II, in addition to showing the results of the networks used by Minaee et al. Their best performance was reached using SqueezeNet; the authors gave Specificity and Sensitivity; we added Balanced Accuracy (bACC) and Predicted Positive Condition Rate (PPCR) with respect to their confusion matrices. Moreover, the Gain column indicates the contribution of the used configurations and evaluates the success of per network compared to SqueezeNet, the best model from Minaee et al., in terms of $\frac{bACC+PPCR}{2}$.

[4]Besides precision, recall is another widely used metric. However, it is also referred to sensitivity.

D. Discussion

Table II summarizes the performance evaluation of the trained networks, with configuration details. As the table shows merging COVIDX-Ray-5K and COVID-ChestXray training sets improved the results substantially, up to 2.18%. The network trained on the combined data with the weights initialized from fine-tuned values from ImageNet and penalized with β (in Equation (4)) value of 0.75 achieved the best performance. To the best of our knowledge, this configurations becomes the state-of-the-art among all models queried on COVIDX-Ray-5K test set [3].

Careful attention must be exercised in networks that exploit WCE for error calculation. The results on the combined databases show that penalizing the false positive samples improves the performance for each network. However, the network trained on purely COVIDX-Ray-5K training set were contradictory to this achievement. In our detailed analysis on distribution of the predictions reported that the models trained on this dataset are already good at predicting the positive samples. Therefore, the false positive samples were not the dominating problem. We thus deduced that $\beta < 1$ degrades the performance, when false negative samples dominate the false positives. Since $\beta < 1$ penalizes the false positives, this proves the effect of WCE on distribution of false samples. It is an intriguing optimization problem for future works, that is finding the optimal adjustment value to alleviate imbalance of dataset in favour of both false negative and false positive samples.

Furthermore, the performance comparison between ACNN and ResNet50 models is reported in Table II. Despite the fact that there is some inconsistency, ACNN achieved significantly close performance to ResNet50. Highlighting the lightweight of the network, ACNN results are promising for medical image classification along with segmentation.

VII. Conclusion and Future Work

COVID-19 created a situation of emergency, which is still out of control. This paper challenges [3] and reaches the state-of-the-art on the COVIDX-Ray-5K test set with the top balanced accuracy of 99.87%. Furthermore, our study analyses the performance of the ACNN network architecture for medical image classification task and exploiting weighted cross entropy (WCE) loss function to alleviate the imbalance on COVID datasets.

A successful CAD system may be helpful to support frontline doctors, or, in some realities, can, actually, provide the only available diagnosis. However, the current number of available databases of COVID-19 images is still too limited, and there is not a common database used as benchmark to compare the proposed models. In summary, it is desirable to have a couple of big and balanced databases with X-ray and CT images to be used, as benchmark, by the entire research community.

Further work includes (1) the use of the proposed algorithm with other databases of COVID-19 X-ray images to test its robustness, (2) the adaptation of the suggested method for tomography images, and (3) the investigation of other deep learning algorithms.

References

[1] "Coronavirus disease (covid-19)," Last access: 19.07.2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] "Reverse transcription polymerase chain reaction," Last access: 12.07.2020. [Online]. Available: https://en.wikipedia.org/wiki/Reverse_transcription_polymerase_chain_reaction

[3] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *arXiv preprint arXiv:2004.09363*, 2020.

[4] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang, "Coronavirus disease 2019 (covid-19): a perspective from china," *Radiology*, p. 200490, 2020.

[5] C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang, "Deep learning-based detection for covid-19 from chest ct using weak label," *medRxiv*, 2020.

[6] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto, "Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data," *arXiv preprint arXiv:2004.05405*, 2020.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[8] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," *Preprints*, vol. 2020030300, p. 2020, 2020.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.

[14] 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," *arXiv preprint arXiv:2003.09871*, 2020.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[16] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[17] 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 preprint arXiv:2006.11988*, 2020.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] X.-Y. Zhou, J.-Q. Zheng, P. Li, and G.-Z. Yang, "Acnn: a full resolution dcnn for medical image segmentation," *arXiv preprint arXiv:1901.09203*, 2019.

[20] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.